RGNZ/MGPB

Advanced Clustering Techniques for Tin Deposit Classification in Malaysia: A Machine Learning Approach

Rudarsko-geološko-naftni zbornik
(The Mining-Geology-Petroleum Engineering Bulletin)
DOI: 10.17794/rgn.2025.3.10

Original scientific paper

Meisam Saleki[1], Kamran Mostafaei[2]*, Zakaria Bin Endut[1], Mohammad Nabi Kianpour[1]

[1] *School of Materials and Mineral Resources Engineering, Engineering Campus, Universiti Sains Malaysia (USM), Malaysia.*
[2] *Department of Mining, Faculty of Engineering, University of Kurdistan, Sanandaj, Iran.*

Abstract

This study explores the application of advanced clustering techniques – Spectral Clustering, Gaussian Mixture Models (GMM), and a hybrid approach combining Autoencoders with K-Means – to classify tin deposits in Malaysia. Geochemical data from 28 tin samples across regions such as Pengkalan Hulu North, Menglembu, Klian Intan, and Sungai Lembing were analysed to identify distinct mineralization patterns. The results revealed that the integration of Autoencoders with K-Means yielded the highest clustering quality, with a Silhouette Score above 0.4 and a Calinski-Harabasz Index of 90 at four clusters, outperforming the other methods. The classification effectively distinguished between Pegmatite, Hydrothermal Veins, Polymetallic, and Disseminated deposits, aligning with the geological characteristics of the regions. These findings enhance the understanding of tin deposit distribution, offering significant potential for optimizing exploration strategies and mining operations, thereby contributing to the sustainable economic development of Malaysia's tin mining industry.

Keywords:
tin deposits, clustering techniques, geochemical analysis, Malaysia, mineral exploration

1. Introduction

Machine learning has seen rapid advancements, becoming a cornerstone in various scientific fields, including geoscience. It enables the processing and analysis of complex, large-scale datasets, which are increasingly common in modern geological research (**Mostafaei et al., 2024**). By applying machine learning techniques, geoscientists can uncover hidden patterns, make predictions, and automate decision-making processes that were previously difficult or impossible (**Smith et al., 2023; Mishra, 2022; Lary et al., 2016**). Among these techniques, clustering stands out as a particularly powerful tool for exploratory data analysis, allowing for the grouping of data points based on their inherent similarities without needing predefined labels (**Kaski et al., 2023; Konopka et al., 2018**). Clustering algorithms are essential in geoscience due to the complex, multidimensional nature of geological data. These algorithms enable the identification of natural groupings within data, which can be critical for tasks such as ore classification, mineral exploration, and geological mapping (**Meyrieux et al., 2024; Chen et al., 2023**). By leveraging clustering methods, geoscientists can better understand subsurface structures, categorize geological formations, and classify mineral deposits, thereby improving resource estimation and exploration strategies (**Reddy, 2018; Riquelme and Ortiz, 2024; Pedregosa et al., 2011**). Several advanced clustering techniques are commonly used in geoscience, each offering unique advantages depending on the dataset and the specific task. Among these, Spectral Clustering, Gaussian Mixture Models (GMM), and Autoencoder combined with K-Means have demonstrated particular efficacy.

Spectral Clustering leverages the eigenvalues of a similarity matrix to reduce dimensionality before clustering in a reduced space. This method is particularly effective for identifying non-convex clusters and complex data distributions, which are typical in geological datasets (**Fouedjio, 2017**). In geoscience, Spectral Clustering has been successfully applied to classify geological formations and mineral deposits, identifying distinct geochemical signatures associated with different types of mineralization (**Talebi et al., 2020**). For instance, it has been used to delineate mineral zones within polymetallic deposits, aiding in the identification of economically viable ore bodies (**Zuo et al., 2023**). Additionally, Spectral Clustering has been utilized to improve the spatial awareness of clustering algorithms, making them more effec-

* Corresponding author: Kamran Mostafaei
e-mail address: k.mostafaei@uok.ac.ir

Received: September 2024. Accepted: 6 February 2025.
Available online: 3 July 2025

tive for geostatistical data (**Ergüner et al., 2019**). This method also helps in defining climate zones and coastal environments, which are crucial for environmental studies (**Sinha et.al, 2023**). Furthermore, it has been applied to identify hazardous waste sites, enhancing environmental safety and management **(Im et al., 2012**).

GMM is a probabilistic clustering method that models data as a mixture of several Gaussian distributions. This method is more flexible than K-Means because it can accommodate clusters of varying shapes and sizes, making it suitable for geoscientific data (**Cheng et al.,2024; Chen et al., 2024; Reynolds, 2009**). In the context of geoscience, GMM has been applied to multi-element geochemical data for classifying ore types and understanding underlying geological processes (**Yu et al., 2024; Bourdeau et al., 2023**). GMM is particularly effective in distinguishing between various mineralization processes in geochemical surveys, providing insight into the spatial distribution of elements like gold and copper (**Zuo and Carranza, 2023**).

A combination of Autoencoder and K-Means has been applied to classify ores in polymetallic deposits, where the high dimensionality of the data poses significant challenges. By reducing the data's dimensionality integration, integration of Autoencoder and K-Means allows for more accurate clustering, leading to better resource estimation and more efficient mining operations (**Zhou et al., 2024; Li et al., 2020; Lou and Zuo, 2024**).

While significant advancements have been made in applying machine learning to geosciences, there remains a gap in assessing the comparative effectiveness of different clustering techniques for classifying complex geological data. This study specifically addresses this gap by evaluating Spectral Clustering, Gaussian Mixture Models (GMM), and a hybrid Autoencoder with K-Means approach. Highlighting these methods in the context of geoscientific data analysis responds to the existing limitations in accurately clustering non-linear and high-dimensional datasets, thus providing an enhanced understanding of their suitability for mineral deposit classification.

In our research, we aim to utilize the aforementioned clustering methods to classify tin samples collected from tin deposits in Malaysia. The selection of sampling locations – Pengkalan Hulu North, Menglembu, Klian Intan, and Sungai Lembing – was strategically made to represent the variation in tin deposit types found in Malaysia. Each location was chosen based on its geological significance: Pengkalan Hulu North and Menglembu are known for Pegmatite deposits, Klian Intan features Hydrothermal Vein deposits, and Sungai Lembing is associated with Disseminated and Polymetallic deposits. This selection ensures a comprehensive analysis that reflects the diversity of tin mineralization in the region.

The main scientific contribution of this research is to demonstrate the efficacy of advanced clustering techniques – specifically Spectral Clustering, Gaussian Mixture Models (GMM), and a hybrid approach combining Autoencoders with K-Means – in classifying tin deposits based on geochemical data. By evaluating these methods on a dataset of 28 tin samples from Malaysia, we provide a comparative analysis of their performance, highlighting the advantages of the hybrid Autoencoder with K-Means approach. Our findings contribute to the field of geosciences by offering a data-driven approach to mineral deposit classification, which can enhance exploration strategies and resource estimation. Furthermore, this study sets a precedent for the application of hybrid machine learning models in geological data analysis, paving the way for more nuanced and accurate interpretations of complex geological datasets.

2. Case study

Tin is a vital metal with numerous industrial uses, such as in electronics, soldering, plating, and alloys. Its unique characteristics, including corrosion resistance and low toxicity, make it essential in modern manufacturing. The demand for tin has been steadily rising, driven by technological progress and the need for sustainable materials (see **Figure 1**). Consequently, effective exploration and classification of tin deposits are crucial for meeting global supply demands and fostering economic growth (**International Tin Association, 2020**). Malaysia, historically a major tin source (see **Figure 2**), was once the world's leading producer. Although production has significantly decreased since the mid-20th century, tin remains a key mineral for the country's mining industry. Efforts are underway to revitalize tin mining activities, focusing on discovering new deposits and reassessing previously mined areas. Modern exploration techniques, including machine learning and advanced geochemical analysis, are being utilized to identify economically viable tin deposits (**Iglesias et al., 2020**).

Malaysia is home to various types of tin deposits, each with unique geological characteristics. Our study focuses on Pegmatite, Hydrothermal Veins, Polymetallic, and Disseminated deposits (**Lehmann, 2021; Basori, 2022; Cao et al., 2020; Adnan et al., 2024**):

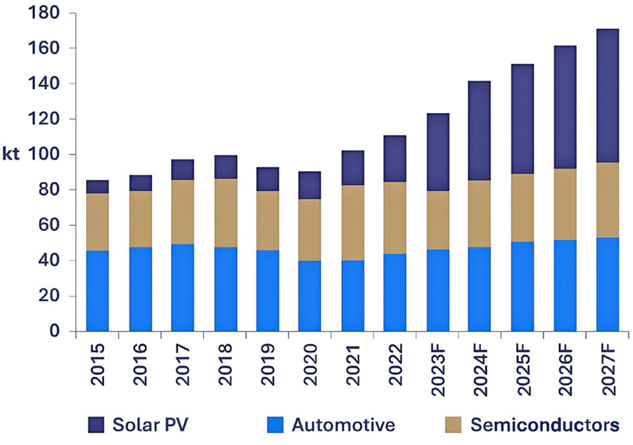

Figure 1. Demand of tin (F-forecast) (**Bloomberg, 2023**)

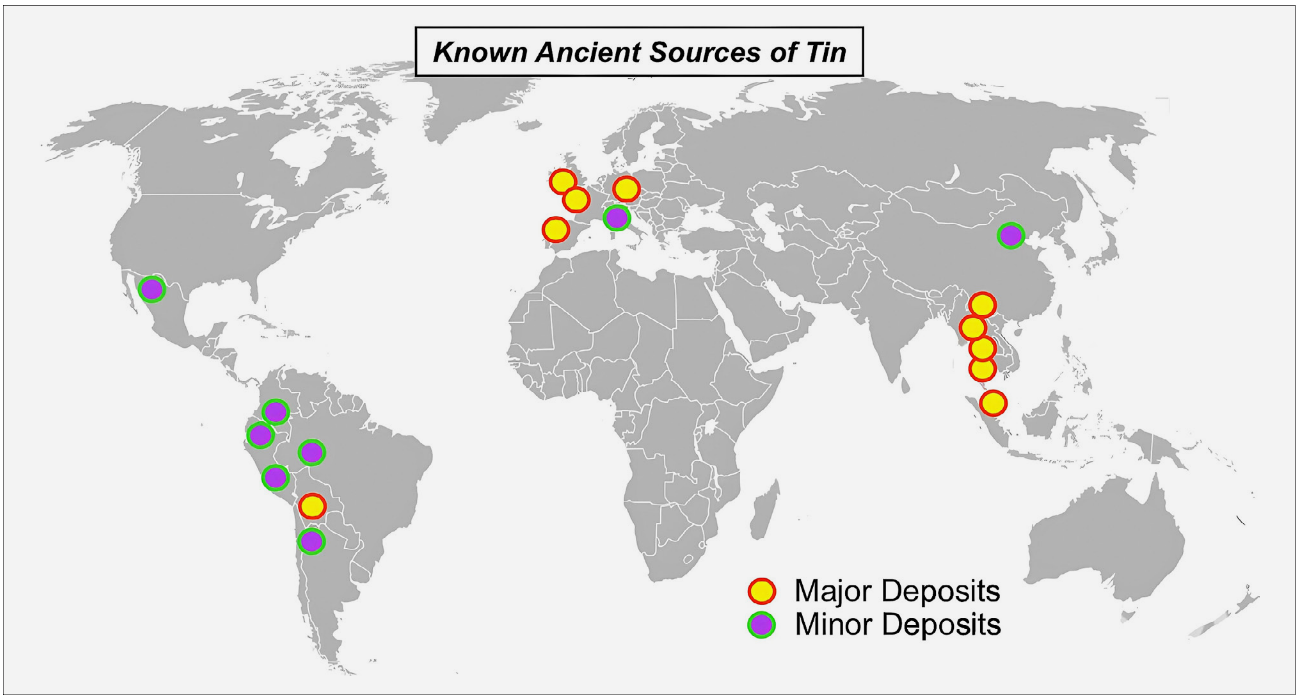

Figure 2. The distribution of tin ore deposit around the world (**Metallurgy in the Americas – Part VII, 2019**)

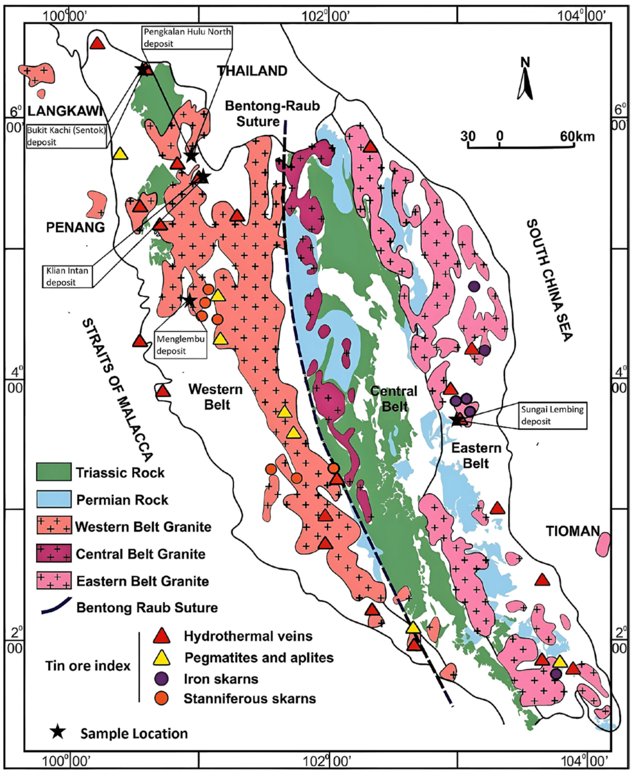

Figure 3. Simplified Malaysia geological map along with the location of studied deposits (**Yeap, 1993**)

A. Pegmatite Deposits: these deposits consist of coarse-grained igneous rocks formed during the final stages of magma crystallization. Pegmatites are important sources of tin and other rare elements, typically found in association with granite intrusions. In these deposits, tin is usually present as cassiterite (SnO_2) crystals.

B. Hydrothermal Veins: these deposits arise from hot, mineral-rich fluids that precipitate minerals within fractures and cavities of host rocks. Hydrothermal vein deposits of tin are typically associated with granitic intrusions and can contain high-grade ore bodies. Cassiterite is the primary tin mineral found in these veins.

C. Polymetallic Deposits: these deposits are rich in multiple metals, including tin, copper, zinc, and lead. They form through intricate geological processes such as hydrothermal activity and magmatic differentiation. Polymetallic deposits are particularly valuable due to their diverse metal content, which provides a range of mining opportunities and economic benefits.

D. Disseminated Deposits: these deposits feature a widespread distribution of tin minerals throughout the host rocks, rather than being concentrated in specific veins or discrete ore bodies. Extracting tin from disseminated deposits often requires large-scale mining techniques and sophisticated processing methods to be economically viable.

In our study, we utilized data from various regions across Malaysia (see **Figure 3**), Pengkalan Hulu North, Menglembu, Bukit Kachi (also known as Sintok), Klian Intan, and Sungai Lembing namely.

3. Methodology

3.1 Tin geochemistry and analysis

Cassiterite (SnO_2) is the main source of tin, and its geochemical analysis is vital for mineral exploration. Laser Ablation Inductively Coupled Plasma Mass Spectroscopy (LA-ICPMS) is used to determine the elemen-

Table 1. Chemical analysis of studied ore samples

Smple Location	Ti	V	Mn	Fe	Zr	Nb	Sb	W	Ta	Hf	U
Pengkalan Hulu	180.55	1.16	0.02	67.58	1.56	16.69	328.45	84.36	3.83	1.61	0.27
Pengkalan Hulu	182.41	0.44	0.01	63.59	1.96	17.15	68.09	8.71	20.60	2.80	0.15
Pengkalan Hulu	112.60	0.44	0.16	35.25	0.67	39.95	1550.66	5175.72	8.02	0.60	10.46
Pengkalan Hulu	7.55	0.02	0.04	158.28	0.10	0.21	168.39	503.05	0.05	0.12	0.41
Pengkalan Hulu	138.79	0.85	0.02	25.66	1.67	14.84	295.02	102.24	18.36	3.49	0.77
Pengkalan Hulu	28.45	0.01	0.14	31.19	0.39	3.80	1585.14	10671.17	0.32	0.29	13.74
Menglembu	429.46	0.25	0.09	165.40	0.07	551.77	855.00	54692.97	55.64	0.04	204.94
Menglembu	20.46	0.04	0.05	302.57	0.00	5.44	744.41	35970.97	3.09	1.48	88.86
Menglembu	22.38	0.03	0.33	157.65	0.01	23.27	1162.56	49658.54	26.60	0.01	110.54
Menglembu	328.59	0.41	0.11	153.85	0.29	337.04	155.29	62.20	17.16	0.19	4.90
Menglembu	4.04	0.00	0.11	218.76	0.08	20.08	190.96	2094.96	86.74	0.23	3.47
Menglembu	242.73	0.26	0.58	214.63	3.19	2092.60	601.69	10233.43	6860.79	6.86	259.65
Klian Intan	119.64	34.83	0.02	1142.95	0.00	0.00	2.61	2.79	0.00	0.00	0.28
Klian Intan	98.28	26.91	0.21	1032.17	0.00	0.21	2.49	4.95	0.00	0.00	0.29
Klian Intan	67.36	1.96	0.02	1915.50	0.07	0.06	4.56	5.33	0.00	0.01	0.64
Klian Intan	94.76	11.13	1.21	1007.32	0.41	6.61	6.94	605.06	0.00	0.00	1.26
Sungai Lembing	1542.54	122.45	5.13	2568.84	3.14	1.10	6.48	2554.98	0.00	0.08	0.81
Sungai Lembing	476.20	17.21	9.64	10688.29	2.54	3.79	50.93	811.08	0.05	0.05	4.05
Sungai Lembing	3274.30	424.46	8.92	1069.34	160.06	91.29	2.94	40.18	6.23 8.46	1.88	
Sungai Lembing	2510.39	19.50	1.06	417.54	19.05	336.59	0.84	3803.59	1.72	0.63	0.48
Bukit Kachi	5540.92	0.01	1214.54	2962.95	2639.72	18607.70	0.22	9813.59	4858.97	407.75	17.75
Bukit Kachi	2042.31	0.34	109.95	2879.91	893.05	3457.61	0.35	124.65	9120.17	303.26	9.47
Bukit Kachi	411.25	0.36	42.02	3105.02	1148.50	2623.31	3.24	3937.89	8201.65	239.28	22.57
Bukit Kachi	5087.27	0.00	400.48	2664.46	3481.08	7538.86	0.61	376.00	2549.31	500.72	83.32
Bukit Kachi	3826.08	2.84	2015.10	4638.62	1847.57	8389.18	6.41	370.99	29434.98	549.37	5.01
Bukit Kachi	425.91	0.09	99.28	3323.53	685.31	5598.07	11.10	1341.03	13880.27	153.64	10.07
Bukit Kachi	473.00	0.15	166.39	6961.73	1039.73	6419.50	6.11	3469.30	30088.60	256.91	18.39
Bukit Kachi	349.41	0.34	213.14	12523.20	1005.14	7079.85	15.39	2398.81	41647.92	257.76	16.35

tal composition of cassiterite grains. By examining the element concentrations in cassiterite, geologists can classify the ore into various deposit types, such as Pegmatite, Hydrothermal veins, polymetallic, or disseminated. This classification is crucial for understanding the geological environment and assessing the economic potential of the deposits (**Wang et al., 2022; Kumar et al., 2024; Guo et al., 2018**). The study analyzed 28 samples, a number that was dictated by the available data at the time of the research. While this number was not pre-determined through statistical methods, it represents the most comprehensive dataset accessible from the selected locations. This sample size enables the study to explore preliminary patterns and validate the clustering methods used, while recognizing that further research with an expanded sample pool would strengthen and build upon these initial findings.

Elements like Fe, Sb, Mn, Zr, Hf, Ta, Nb, V, Ti, U, and W are important in cassiterite studies due to their geochemical properties and their roles as indicators of different geological processes and environments. Iron (Fe) and manganese (Mn) are typically linked to hydrothermal processes. Zirconium (Zr) and hafnium (Hf) often replace tin (Sn) in the cassiterite structure, and their ratios can reveal the type of mineralization. Tantalum (Ta) and niobium (Nb) help distinguish between magmatic and hydrothermal deposits. Vanadium (V) and titanium (Ti) provide insight into the redox conditions during mineral formation. Uranium (U) and tungsten (W) concentrations indicate specific types of tin deposits and assist in determining the thermal history and origin of the deposits (**He et al., 2022; Liu et al., 2021**).

The trace element contents in selected cassiterite grains from various deposits were analyzed using Laser Ablation-Inductively Coupled Plasma-Mass Spectrometry (LA-ICP-MS) at the Centre for Ore Deposit and Earth Sciences (CODES), University of Tasmania. The instrument used combines a Resolution 193 nm excimer laser with an Agilent 7700x ICP-MS. To map trace element distribution, the signal was acquired in time-resolved mode with a 5 μm laser beam, a laser fluency of ~3.5 J/cm², and a repetition rate of 5 Hz. The signal from the carrier gas without ablation was regularly acquired to correct for instrumental background. To manage values that fell below the detection limit, we applied the rule, substituting them with 75% of the detection limit to provide a conservative yet realistic estimate. For values exceeding an upper threshold, we used the rule, substituting them with 133% of the detection limit to maintain data consistency while addressing potential overestimations. After substitution values according to the detection limit rules, we employed the Interquartile Range (IQR) method to identify outliers. To ensure data integrity and maintain sample size, we preferred to cap extreme outliers at the 95th percentile. This capping approach preserves the dataset's comprehensiveness, reduces the undue influence of outliers, and acknowledges the potential for rare but valid geochemical variations. **Table 1** presents the geochemical analysis of 28 samples from selected areas in Malaysia.

The elements were selected according to previous studies (**He et al., 2022; Liu et al., 2021**), that identified them as highly effective for distinguishing tin deposits. These elements – including Fe, Zr, Nb, Ta, Hf, and U – significantly influence statistical analyses by providing distinct geochemical signatures that enhance clustering accuracy and facilitate the classification of different mineralization types.

To validate the influence of these elements, we conducted t-tests and F-tests to examine differences in their concentrations across sample locations. Significant differences in means were found for Fe (p-value = 0.0012), Zr (p-value = 0.0096), Nb (p-value = 0.0080), Ta (p-value = 0.0153), Hf (p-value = 0.0053), and U (p-value = 0.0312). These results confirm that these elements contribute distinctively to the classification of tin deposit types. Although most elements did not show significant variance differences, Fe displayed a marginal result (p-value = 0.0713), indicating a potential difference in variance that warrants further investigation. These statistical analyses reinforce the importance of these elements in enhancing clustering outcomes and confirm their role in providing geochemical differentiation among tin deposits.

3.2. Dendrogram Clustering

Dendrogram clustering, a key technique in hierarchical clustering, involves creating a diagram to represent the arrangement of clusters formed by hierarchical algorithms (**Forina et al., 2002; Sangaiah et al., 2022**). The process starts with each data point as its own cluster and progressively merges the closest pairs of clusters based on a chosen distance metric, such as Euclidean or Manhattan distance. This merging continues until all data points are grouped into a single cluster. The resulting dendrogram visually illustrates the hierarchical relationships between clusters, with branch lengths indicating the distance or dissimilarity between them (**Ogasawara and Kon, 2021**).

In this study, we used Ward's linkage method, an agglomerative hierarchical clustering technique that minimizes the variance within clusters by merging clusters that result in the smallest increase in total within-cluster variance. This approach tends to create clusters of relatively equal size and is effective in identifying compact, spherical clusters. The distance measure used in this analysis is Euclidean distance, which is calculated as the straight-line distance between two points in a multidimensional space. It is a common distance metric used in clustering algorithms due to its simplicity and effectiveness in capturing the geometric distance between data points.

While this clustering approach has its strengths, it can be misleading, especially if the cut-off line falls near zero, as happened in our dataset. In such cases, considering alternative machine learning clustering methods can provide more robust and scalable solutions. These methods can handle larger datasets, noise, and varying cluster shapes more effectively.

3.3. Spectral Clustering

Spectral clustering is a powerful technique used to partition data points into clusters by leveraging the eigenvalues and eigenvectors of a similarity matrix derived from the data (**Von Luxburg, 2007**). The process begins with data preprocessing, where the input data, typically an n×m matrix (with n representing the number of data points and m the number of features), is normalized if necessary to ensure equal contribution from each feature to the similarity measurement (**Ng et al., 2001**).

Following this, a similarity matrix is constructed using a defined similarity measure, such as the Gaussian (RBF) kernel, which calculates the similarity between data points x_i and x_j based on their Euclidean distance:

$$S(i,j) = \exp(-\frac{x_i - x_j{}^2}{2\sigma^2}) \quad (1)$$

Where $\left\| x_i - x_j \right\|$ is the Euclidean distance between data points x_i and x_j, and σ is a scaling parameter that controls the width of the Gaussian (**Shi and Malik, 2000**).

Once the similarity matrix *S* is established, the degree matrix *D* is computed as a diagonal matrix where each diagonal element *D*(*i,i*) is the sum of the corresponding row of the similarity matrix:

$$D(i,i) = \sum_{j=1}^{n} S(i,j) \quad (2)$$

The Laplacian matrix, a key component in spectral clustering, is then derived by subtracting the similarity matrix from the degree matrix:

$$L = D - S \quad (3)$$

Alternatively, a normalized Laplacian matrix can be used:

$$L_{sym} = D^{-0.5} \, LD^{-0.5} \quad (4)$$

This matrix transformation ensures that the resulting clusters are balanced, especially in terms of the number of data points (**Chung, 1997**). Subsequently, eigen decomposition is performed on the Laplacian matrix to obtain its eigenvalues and eigenvectors. The first k eigenvectors corresponding to the smallest eigenvalues are selected to form a new feature matrix U, where each row represents a data point in the reduced-dimensional space. This feature matrix is then subjected to k-means clustering, which assigns each data point to a cluster based on its representation in this lower-dimensional space (**Von-Luxburg, 2007**).

Finally, the implementation of spectral clustering requires careful consideration of parameter tuning, such as the number of clusters k and the scaling parameter σ for the similarity measure, to optimize performance (**Ng et al., 2001**). The computational complexity, particularly during eigen decomposition, needs to be managed, especially for large datasets, where approximate methods or exploiting sparsity in the similarity matrix may reduce computational costs (**Zelnik-Manor and Perona, 2004**).

3.4. Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMM) clustering is a probabilistic approach that assumes data is generated from a mixture of several Gaussian distributions, each representing a different cluster. The process begins with data preprocessing, where the data matrix is normalized to ensure each feature contributes equally to the clustering process. Optionally, dimensionality reduction techniques like Principal Component Analysis (PCA) may be applied to reduce the number of features, improving clustering performance by mitigating the curse of dimensionality. The next step involves initializing the GMM by determining the number of Gaussian components (clusters), a critical hyperparameter that can be chosen using model selection criteria such as the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC) (**Fraley and Raftery, 2007; McLachlan, 2000**).

Initialization can be performed using random methods or by leveraging the results from k-means clustering to initialize the means of the Gaussian components. Once initialized, the Expectation-Maximization (EM) algorithm is employed to iteratively refine the parameters of the Gaussian components. In the E-step, the algorithm calculates the responsibility (posterior probability) that each data point belongs to each Gaussian component based on the current parameter estimates. The M-step then updates the parameters – mean, covariance, and mixing coefficients – by maximizing the expected log-likelihood function given these responsibilities. The algorithm alternates between these steps until convergence, typically determined by the change in the log-likelihood or after a fixed number of iterations. The GMM clustering process produces soft assignments, meaning each data point has a probability of belonging to each cluster, which is particularly useful for handling overlapping clusters (**Murphy, 2012**).

Additionally, the selection of the covariance matrix type – spherical, diagonal, tied, or full – can significantly impact the clustering outcome, with more complex covariance structures offering greater flexibility at the cost of increased computational complexity. This process is computationally intensive, particularly for large datasets, but can be optimized using techniques like parallelization or variational inference methods (**Blei et al., 2017**).

3.5. Integration of autoencoders and K-means

The combination of autoencoders and K-means clustering is a powerful hybrid approach for handling high-dimensional data, where traditional clustering methods may struggle. The process begins with data pre-processing, including normalization to ensure that each feature contributes equally to the clustering process. An autoencoder, which consists of an encoder and a decoder, is then trained in an unsupervised manner to compress input data into a lower-dimensional latent space while minimizing the reconstruction error between the input and output data (**Goodfellow, 2016**). Once trained, the encoder is used to transform the original high-dimensional data into this compact latent representation. K-means clustering is then applied to the data in this latent space, where the reduced dimensionality typically allows for better cluster separability and improved clustering performance (**Xie et al., 2016**). Hyperparameters, such as the number of clusters K and the structure of the autoencoder, including the number of layers and neurons, are carefully tuned to enhance clustering effectiveness (**Guo et al., 2017**). This combined approach effectively leverages deep learning to learn meaningful data representations, making clustering more robust and scalable, especially for complex and noisy datasets.

4. Results

Table 2 refers to the parameters and their values for three mentioned clustering methods, while **Figure 4** illustrates the applied clustering approaches.

The quality of the clustering is assessed using internal validation measures (**Arbelaitz et al., 2013; Xu and Tian, 2015; Van der Maaten and Hinton, 2008**). The performance of applied clustering methods was analysed using two metrics: the Silhouette Score and the Calinski-Harabasz Index. Each metric provides unique insight into the clustering quality, allowing compression of the effectiveness of these methods when choosing 4 clusters (4 primary mineralization have been proposed in areas). The applied methods used to assess the effectiveness of clustering algorithms without relying on external labels or ground truth data. These measures provide insight into how well the clusters are separated and how compact the data points are within each cluster. In this study, we used two widely recognized internal validation measures: the Silhouette Score and the Calinski-Harabasz Index. The Silhouette Score quantifies how similar each data point is to its own cluster compared to other clusters, with higher values indicating better-defined clusters. Specifically, the Silhouette Score ranges from -1 to 1, where values closer to 1 suggest that the data points are well-clustered, while values near 0 indicate overlapping clusters. The Calinski-Harabasz Index measures the ratio of between-cluster dispersion to within-cluster dispersion, with higher values indicating more distinct

Table 2. Parameters and their corresponding values used for clustering

Method	Parameter	Value
Spectral Clustering	[‘number of clusters’, ‘affinity’, ‘similarity matrix’]	[4, ‘precomputed’, ‘RBF kernel similarity matrix’]
Gaussian Mixture Models (GMM)	[‘number of components’, ‘random of state’, ‘covariance type’]	[4, 0, ‘full’]
Integration of Autoencoder and K-means	[‘input dim’, ‘encoding dim’, ‘epochs’, ‘batch size’, ‘optimizer’, ‘loss’, ‘number of clusters’, ‘random_state’, ‘init ‘, ‘number of init ‘]	[11, 2, 50, 256, ‘adam’, ‘mse’, 4, 0, ‘k-means++’, 10]

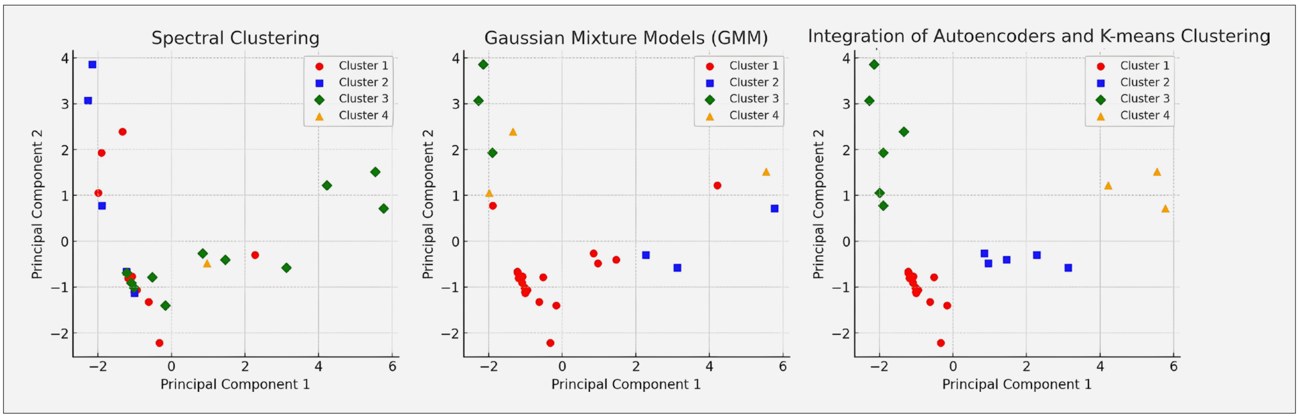

Figure 4. Clustering plots using

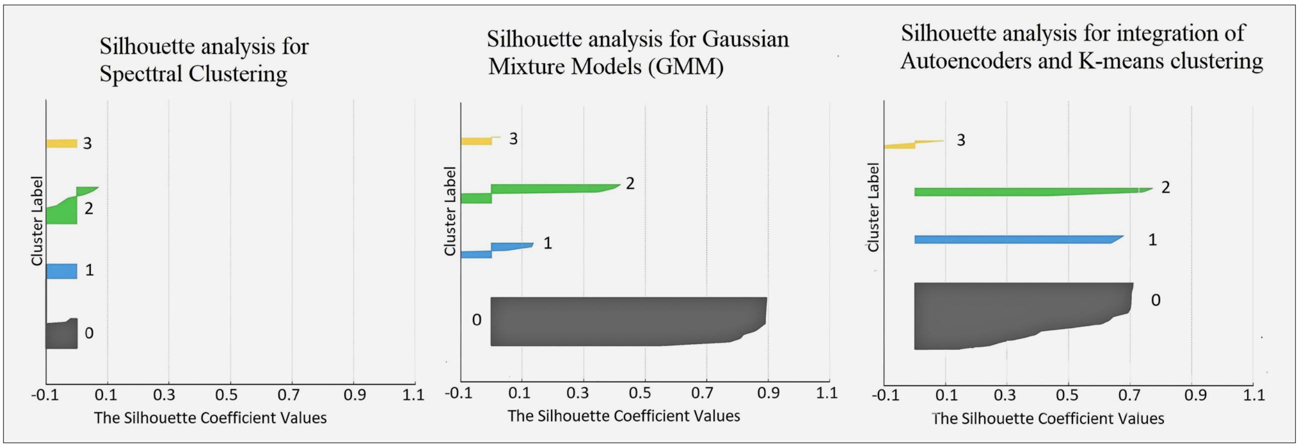

Figure 5. Silhouette Score plot

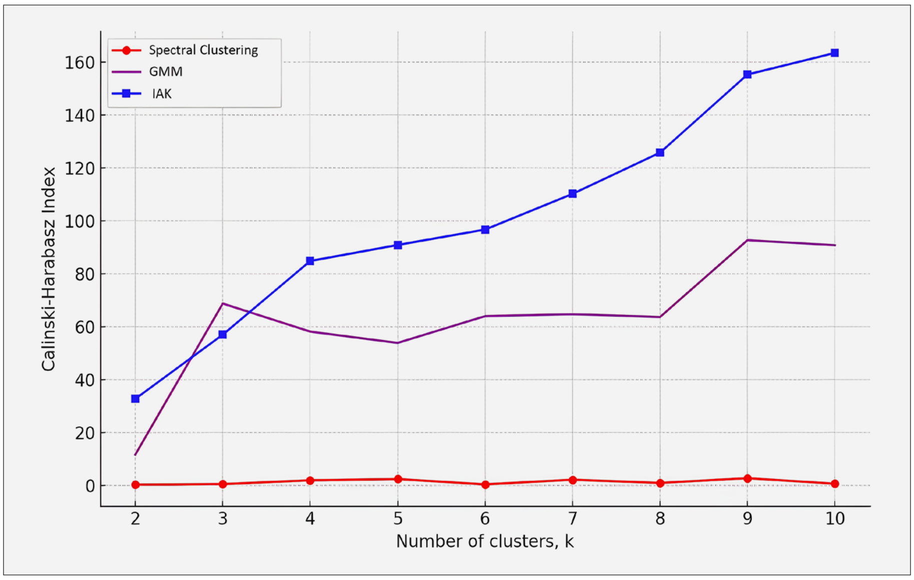

Figure 6. Calinski-Harabasz Index plot

and well-separated clusters. These metrics allow us to objectively compare the performance of different clustering methods and ensure that the clusters formed are meaningful and representative of the underlying geochemical patterns.

Figure 5 displays the Silhouette Score plot for applied clustering methods. By comparing the silhouette coefficient values for each cluster label across the three methods, it was observed that Spectral Clustering has a varied distribution of silhouette scores with most clusters having coefficients below 0.4, indicating that some points may be poorly clustered or lie between clusters. The Gaussian Mixture Models (GMM) method shows a similar trend with most silhouette scores around 0.3 to 0.4, but with slightly better-defined clusters than Spectral Clustering as fewer scores are negative. The integration of Autoencoders and K-means (IAK) Clustering appears to perform the best, with most silhouette scores above 0.4, suggesting more distinct and well-defined clusters. This method demonstrates a more balanced clustering performance, minimizing overlap between clusters and suggesting it is more effective in finding well-separated groups in the data.

Figure 6 compares the Calinski-Harabasz Index for different clustering methods – Spectral Clustering, Gaussian Mixture Models (GMM), and Autoencoders with K-means – as the number of clusters, k, increases from 2 to 10. The Calinski-Harabasz Index is a metric used to evaluate the quality of clustering, where higher values indicate better-defined clusters. At k=4, the integration of Autoencoders and K-means (IAK) exhibits the highest Calinski-Harabasz Index, around 90, indicating it achieves the most distinct and well-separated clusters among the three methods. Gaussian Mixture Models (GMM) has a moderate Calinski-Harabasz Index value around 70 at k=4, suggesting reasonable cluster quality but not as effective as IAK. Spectral Clustering consistently shows the lowest Calinski-Harabasz Index across all values of k, hovering around 5, indicating poor clustering quality and less distinct separation between clusters.

5. Discussion

5.1. Findings

To further elucidate the relationships between the geochemical elements and their contributions to the clustering of tin deposit types, we conducted hierarchical clustering analysis. The resulting diagram provides a visual representation of how the elements group together based on their geochemical similarities. This analysis complements the clustering results and offers deeper insight into the underlying patterns that differentiate the mineralization types. The dendrogram was generated using Ward's linkage method, which minimizes the variance within clusters, and the distances between clusters reflect the dissimilarity between the elements. This hierarchical approach allows us to identify which elements are statistically significant in differentiating the deposit types and how they contribute to the overall clustering structure.

The hierarchical cluster diagram (see **Figure 7**) reveals significant insight into the geochemical relationships among the elements analyzed in this study. The dendrogram shows the separation of W, Ta, and Sb from other elements in the hierarchical cluster diagram indicating their strong association with hydrothermal processes, which are common in Hydrothermal Vein and Polymetallic deposits. Similarly, elements like Mn, Hf, and V group together, suggesting their roles in indicating redox conditions and hydrothermal alteration, which are critical for distinguishing between different mineralization types (**Coyte and Vengosh, 2020; Schreiber, 1977; Scholtysik et al., 2020; Underwood et al., 2013; Nederbo, 2023; Gao et al., 2021**). Fe is indeed a redox-sensitive element, similar to Mn, Hf, and V (**Debret et al.,**

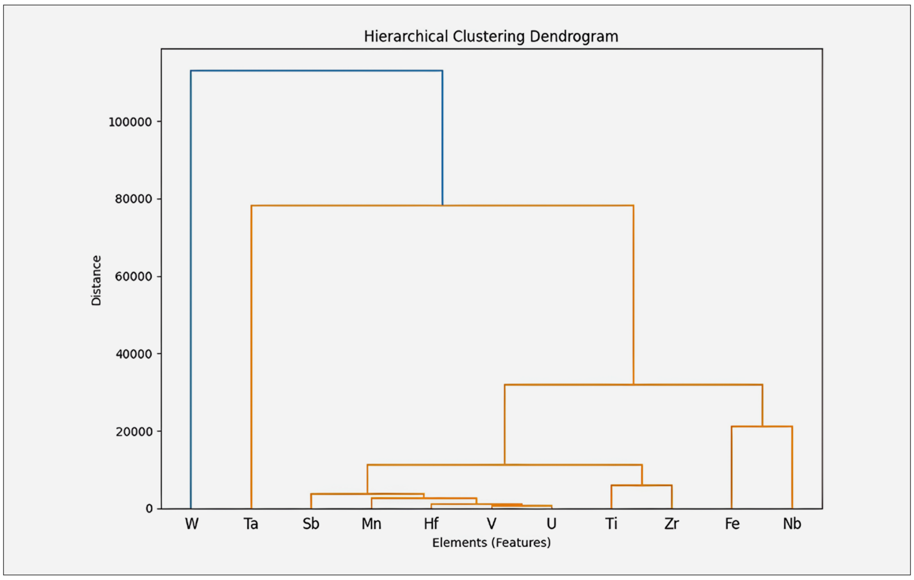

Figure 7. Hierarchical clustering dendrogram

2020), which indicates its involvement in redox conditions and hydrothermal alteration processes (**Studemeister, 1983**). However, the reason Fe might not be grouped with Mn, Hf, and V in the dendrogram could be due to several factors. This refers to distinct geochemical environments and processes. Fe may share stronger similarities with elements like Zr and Nb in other geochemical aspects, such as magmatic processes or the formation of Pegmatite deposits, where these elements are often found together (**Yan et al., 2018; Möller and Williams-Jones, 2017**).

Additionally, variations in the concentrations and interactions of elements within our samples could lead to Fe forming a separate cluster, influenced by factors like pH, salinity, and the presence of other elements. Fe's association with Zr and Nb might highlight its role in identifying Pegmatite deposits, reflecting its importance in magmatic processes (**Gysi et al., 2016**). This distinction illustrates the complexity and multi-faceted nature of geochemical clustering, providing a comprehensive understanding of the elemental relationships within our study. Notably, Fe, Zr, and Nb form another cluster, highlighting their importance in identifying Pegmatite deposits, where these elements are often associated with magmatic processes. Even in the absence of significant Fe-bearing minerals like tourmalines (e.g. schorl) or other Fe-rich minerals in pegmatite deposits, there are still several possible reasons why Fe might cluster with Zr and Nb. Fe might be present as trace element inclusions within the pegmatite matrix or in minor minerals, sufficient to influence clustering patterns (**Thomas et al., 2019**). Additionally, Fe could be associated with Zr and Nb due to similar geochemical processes during pegmatite formation, leading to their co-precipitation or involvement in similar geochemical environments (**Hadlich et al., 2024**). The grouping of U, Ti, and Zr further emphasizes their role in differentiating Disseminated deposits, where these elements are indicative of widespread mineralization within host rocks (**Mikysek et al., 2019**). These groupings align with the statistical significance of these elements, as highlighted in the t-tests and F-tests, which confirm their distinct contributions to the classification of tin deposit types.

The clustering analysis (see **Figure 8**) reveals insightful patterns regarding the mineralization types and geological settings of the samples. The overall cluster distribution pie chart indicates how the samples are grouped into four distinct clusters, each characterized by specific mineralization types. The zero cluster contains a mix of Disseminated, Hydrothermal Vein, and Polymetallic samples, suggesting overlapping geochemical signatures. The bar plot of mineralization type distribution by cluster further illustrates that zero cluster has the most diversity, comprising 6 Disseminated, 4 Hydrothermal Vein, and 4 Polymetallic samples. This mix suggests a geological environment where multiple mineralization processes might occur simultaneously or in succession, leading to transitional or hybrid samples.

The 1 and 3 clusters are exclusively composed of Pegmatite samples. Cluster 1 has 5 Pegmatite samples, while Cluster 3 has 3. The clear separation into two clusters, as shown in the bar plot, suggests possible subtypes within the Pegmatite samples or variations due to different geological conditions at their respective sample locations, such as Menglembu and Pengkalan Hulu North. Comprising 6 Disseminated samples, Cluster 2 reflects a homogeneous mineralization type. The consistent geochemical profile across these samples suggests a stable geological environment that favors Disseminated mineralization, primarily from the Pengkalan Hulu North location.

The sample location distribution by cluster bar plot, on the other hand, provides a geographical perspective on the clustering results. Samples in the zero cluster originate from diverse locations, including Pengkalan Hulu North and Menglembu. This geographical diversity aligns with the cluster's mixed mineralization types,

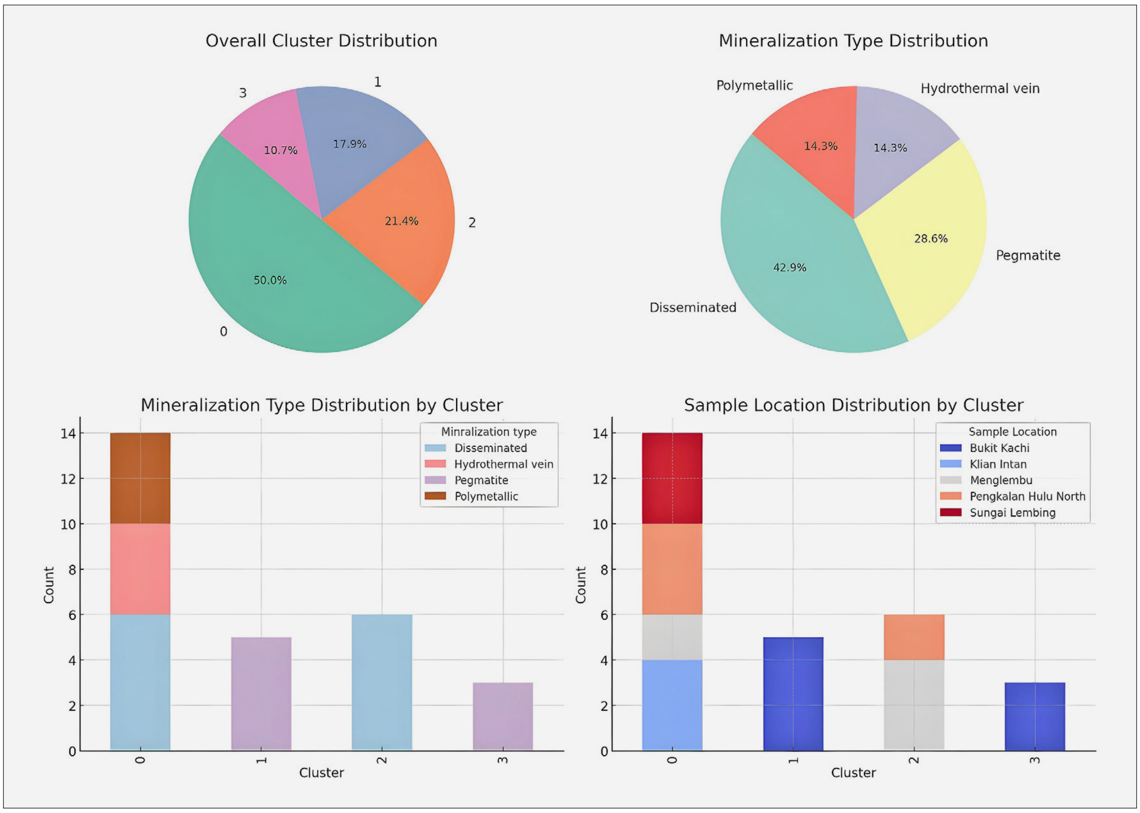

Figure 8. Cluster Distribution plots

suggesting that these locations are geologically complex. The overlap of Disseminated, Hydrothermal Vein, and Polymetallic samples within this cluster may point to transitional zones where multiple mineralization processes interact, such as areas of active hydrothermal activity coupled with disseminated mineralization.

The samples in the 1 and 3 clusters are mainly from distinct locations. Cluster 1 samples, primarily from Menglembu, and Cluster 3 samples, predominantly from Pengkalan Hulu North, highlight that different geological environments or zones within these areas may influence the formation of Pegmatite subtypes. The clear grouping suggests distinct geological conditions that are conducive to Pegmatite formation in these specific areas.

Cluster 2 mainly consists of samples from Pengkalan Hulu North, indicating a uniform geological setting that supports Disseminated mineralization. The consistency in mineralization type across this cluster suggests a stable geological environment with a single, dominant mineralization process. In other words, a stable geological environment refers to regions where consistent geochemical profiles are observed, supporting uniform mineralization processes. For instance, the samples in Cluster 2 demonstrate a homogenous geochemical composition that reflects a stable environment favoring Disseminated mineralization. Such stability is often a result of continuous, long-term geological conditions that allow for the uniform distribution of tin within the host rocks. The clusters containing Pegmatite samples (Clusters 1 and 3) suggest distinct geological zones within the respective sampling areas. The clear demarcation between these clusters indicates that the geochemical conditions at these sites favored Pegmatite formation, potentially influenced by localized magmatic activity or late-stage crystallization processes.

The hierarchical cluster diagram (see **Figure 7**) provides a complementary perspective to the pie chart in **Figure 8**, which summarizes the distribution of mineralization types across the clusters. The dendrogram reveals that the elements contributing to the mixed mineralization types in Cluster 0 (Disseminated, Hydrothermal Vein, and Polymetallic) are grouped together, suggesting that these deposits share overlapping geochemical signatures. For instance, the grouping of Fe, Zr, and Nb in the dendrogram corresponds to the Pegmatite-dominant clusters (Clusters 1 and 3), indicating that these elements are critical for distinguishing Pegmatite deposits from others. Similarly, the elements associated with Disseminated deposits (Cluster 2) form a distinct cluster, reinforcing the homogeneity of this mineralization type. The dendrogram thus provides a clearer understanding of the geochemical basis for the clustering results and validates the statistical significance of the elements in differentiating the deposit types.

5.2. Interpretation of results

The clustering analysis performed in this study reveals significant insight into the geochemical characteristics and geological settings of the tin deposits. The geochemical analysis involved a detailed examination of trace elements, including Fe, Sb, Mn, Zr, Hf, Ta, Nb, V, Ti, U, and W. Each of these elements plays a crucial role in identifying and distinguishing different mineralization types. Iron and Mn are typically associated with hydrothermal processes and provide insight into the nature of hydrothermal alteration. Zirconium and Hf often substitute Sn in the cassiterite structure, and their ratios can indicate the mineralization type (e.g. magmatic or hydrothermal origins) (**Breiter and Škoda, 2017; Claiborne et al., 2006**).

Tantalum and Nb are key indicators for distinguishing between magmatic and hydrothermal deposits. Vanadium (V) and Ti help reveal the redox conditions prevalent during the formation of the deposits. Uranium and W concentrations contribute to understanding the thermal history and genesis of the deposits (**Ahmed, 2022; Ballouard et al, 2016**).

The application of clustering methods – Spectral Clustering, Gaussian Mixture Models (GMM), and the integration of Autoencoders with K-Means – resulted in the identification of four distinct clusters. Each cluster represents unique geochemical and geological characteristics. Cluster 0 includes a combination of Disseminated, Hydrothermal Vein, and Polymetallic samples, suggesting overlapping geochemical signatures. The presence of multiple mineralization types implies transitional zones where diverse geological processes may coexist or interact. Clusters 1 and 3 (Pegmatite Dominance) exclusively contain Pegmatite samples. Their separation into distinct groups may indicate subtypes of Pegmatites or variations due to different geological environments at the sampling locations (e.g. Pengkalan Hulu North and Menglembu). Cluster 2 (Disseminated Samples), Composed primarily of Disseminated samples from Pengkalan Hulu North, highlights a stable geological environment conducive to Disseminated mineralization.

5.3. Clustering Significant

The trace elements analyzed not only differentiate mineralization types but also highlight the geological environments that favor specific deposit formations. For example, the higher concentrations of Ta and Nb in Pegmatite samples align with their typical formation in magmatic environments, while elevated Fe and Mn levels in the Hydrothermal Vein samples are consistent with hydrothermal activity. The clustering results, supported by these geochemical variables, provide an understanding of the underlying processes governing tin deposit formation. The observed groupings enhance the interpretation of the mineralization processes and point towards areas that may require further geological investigation to identify economically viable deposits. By expanding on the geochemical data and integrating it with clustering outcomes, this discussion clarifies the rationale behind the clustering and elucidates the influence of stable geological environments on mineralization patterns.

5.4. Method Comparison

The clustering results revealed distinct advantages and limitations associated with each method applied. Spectral Clustering demonstrated its strength in detecting non-convex clusters, which is beneficial for data with complex distributions; however, its sensitivity to parameter selection and computational intensity may limit its scalability for larger datasets (**Ng et al., 2001**). The Gaussian Mixture Model (GMM) provided flexible clustering that handled clusters of varying shapes and sizes well, although it struggled with high-dimensional data, sometimes leading to ambiguous assignments (**Reynolds, 2009**). The combination of Autoencoders and K-Means proved to be the most effective, as it reduced data dimensionality and improved cluster separation. Despite its higher computational requirements, this hybrid approach provided robust performance by effectively capturing non-linear relationships. These insights suggest that while each method has strengths, the Autoencoder with K-Means approach is particularly advantageous for high-dimensional geochemical data.

Comparing these findings with previous research in the field, studies such as those by **Zuo et al.** (**2023**) and **Chen et al.** (**2023**), our approach using a hybrid model aligns with trends seen in recent literature, where integrated methods are used to enhance clustering accuracy. Unlike traditional clustering efforts focused solely on GMM or K-Means, our research demonstrates that combining deep learning techniques like Autoencoders can yield superior results. This comparison underscores the potential for leveraging hybrid methods to push beyond conventional approaches in geological data analysis.

5.5. Limitations and uncertainties

Potential sources of error in this study include sample quality and measurement accuracy. The variability in the geochemical composition of samples, potentially stemming from environmental exposure or collection inconsistencies, could impact the reliability of the clustering outcomes. Additionally, while advanced techniques like Autoencoder with K-Means reduce noise through dimensionality reduction, uncertainties remain in model parameter tuning and data preprocessing. These factors may lead to subtle shifts in clustering assignments and interpretations. Acknowledging these uncertainties is crucial for accurate analysis and suggests that future studies should incorporate improved sampling protocols and cross-validation with larger datasets to mitigate such issues.

6. Conclusions

This research contributes to the field of geosciences and the mining industry by demonstrating the efficacy of advanced clustering methods, particularly the combination of Autoencoders with K-Means, in classifying tin deposits based on geochemical data. It highlights how modern machine learning techniques can improve the understanding of mineral distribution and inform exploration practices. The study sets a precedent for employing hybrid machine learning models in geological analysis, paving the way for more nuanced data interpretations and strategic mining operations. These contributions add to the existing body of knowledge by showcasing practical, data-driven solutions for mineral resource management.

The findings of this study offer practical implications for the mining industry by improving the efficiency of exploration strategies. For instance, the use of the Autoencoder with K-Means approach demonstrated superior cluster differentiation, which can be utilized to prioritize exploration sites and reduce costs. Mining companies can apply these methods to better categorize ore deposits, enhancing resource estimation and decision-making. By adopting these techniques, operations can more effectively allocate resources and refine extraction processes. An example application could be integrating these models into geographic information systems (GIS) to create more detailed mineral maps that inform fieldwork.

However, this study faced several limitations, including a relatively small sample size and potential biases introduced by the haphazard selection of samples based on availability. While the clustering methods applied provided meaningful results, there were challenges in fine-tuning parameters that might influence consistency. The computational demands of advanced techniques, particularly the integration of Autoencoders, were also a constraint. Future research could overcome these limitations by using larger and more balanced datasets, optimizing parameter selection through automated processes, and implementing cross-validation techniques to improve reliability.

Moreover, for future research, it is recommended to expand the dataset by including additional sample sites that capture more diverse geological settings to enhance the generalizability of the results. Incorporating more advanced machine learning techniques, such as deep clustering algorithms or ensemble methods, could provide further insight into the complexities of tin deposit classification. Additionally, performing longitudinal studies to track how changes in environmental and geological conditions impact clustering results would offer valuable perspectives. Collaborative studies that integrate field investigations with geochemical analysis using updated tools like high-resolution LA-ICP-MS are also encouraged to improve data accuracy.

7. References

Adnan, N., Endut, Z., Ariffin, K. S., Makoundi, C., and Jimoh, O. A. (2024). mineralogy and geochemistry of the tin-tungsten deposit in sintok, kedah, malaysia. Malaysian Journal of Microscopy, 20(1), 110-122.

Ahmed, A. H. (2022). Magmatic – Hydrothermal Deposits (Intrusion-Related Deposits). In Mineral Deposits and Occurrences in the Arabian–Nubian Shield (pp. 167-243).

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. Pattern recognition, 46(1), 243-256.

Ballouard, C., Poujol, M., Boulvais, P., Branquet, Y., Tartèse, R., & Vigneresse, J. L. (2016). Nb-Ta fractionation in peraluminous granites: A marker of the magmatic-hydrothermal transition. Geology, 44(3), 231-234.

Basori, M. B. I., Zaw, K., McNeill, A., and Large, R. R. (2022). A mineralogical and geochemical application for determining hydrothermal alteration zones associated with volcanic-hosted massive sulphide deposits at Tasik Chini area, Peninsular Malaysia. Applied Geochemistry, 145, 105404.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. Journal of the American statistical Association, 112(518), 859-877.

Bloomberg (2023). NEF, Rho Motion, Macquarie Strategy, September 2023.

Bourdeau, J. E., Zhang, S. E., Lawley, C. J., Parsa, M., Nwaila, G. T., & Ghorbani, Y. (2023). Predictive geochemical exploration: Inferential generation of modern geochemical data, anomaly detection and application to northern Manitoba. Natural Resources Research, 32(6), 2355-2386.

Breiter, K., & Škoda, R. (2017). Zircon and whole-rock Zr/Hf ratios as markers of the evolution of granitic magmas: Examples from the Teplice caldera (Czech Republic/Germany). Mineralogy and Petrology, 111, 435-457.

Claiborne, L. L., Miller, C. F., Walker, B. A., Wooden, J. L., Mazdab, F. K., & Bea, F. (2006). Tracking magmatic processes through Zr/Hf ratios in rocks and Hf and Ti zoning in zircons: an example from the Spirit Mountain batholith, Nevada. Mineralogical Magazine, 70(5), 517-543.

Cao, H. W., Li, G. M., Zhang, Z., Zhang, L. K., Dong, S. L., Xia, X. B., ... and Zhang, Y. H. (2020) Miocene Sn polymetallic mineralization in the Tethyan Himalaya, southeastern Tibet: A case study of the Cuonadong deposit. Ore Geology Reviews, 119, 103403.

Chen, G., Kusky, T., Luo, L., Li, Q., & Cheng, Q. (2023). Hadean tectonics: Insights from machine learning. Geology, 51(8), 718-722.

Chen, H., Wang, T., Zhang, Y., Bai, Y., & Chen, X. (2023). Dynamically weighted ensemble of geoscientific models via automated machine-learning-based classification. Geoscientific Model Development, 16(19), 5685-5701.

Cheng, L., Abraham, J., Trenberth, K. E., Boyer, T., Mann, M. E., Zhu, J., ... & Lu, Y. (2024). New record ocean temperatures and related climate indicators in 2023. Advances in Atmospheric Sciences, 41(6), 1068-1082.

Chung, F. R. (1997). Spectral graph theory (Vol. 92). American Mathematical Soc.

Coyte, R. M., & Vengosh, A. (2020). Factors controlling the risks of co-occurrence of the redox-sensitive elements of arsenic, chromium, vanadium, and uranium in groundwater from the Eastern United States. Environmental science & technology, 54(7), 4367-4375.

Debret, B., Reekie, C. D. J., Mattielli, N., Beunon, H., Ménez, B., Savov, I. P., & Williams, H. M. (2020). Redox transfer at subduction zones: insights from Fe isotopes in the Mariana forearc. Geochemical Perspectives Letters, 46-51.

Ergüner, Y., Kumar, J., Hoffman, F. M., Dalfes, H. N., & Hargrove, W. W. (2019). Mapping ecoregions under climate change: a case study from the biological 'crossroads' of three continents, Turkey. Landscape Ecology, 34, 35-50.

Fouedjio, F. (2017). A spectral clustering approach for multivariate geostatistical data. International Journal of Data Science and Analytics, 4(4), 301-312.

Forina, M., Armanino, C., & Raggio, V. (2002). Clustering with dendrograms on interpretation variables. Analytica Chimica Acta, 454(1), 13-19.

Fraley, C., & Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. Journal of classification, 24(2), 155-181.

Guo, X., Liu, X., Zhu, E., & Yin, J. (2017). Deep clustering with convolutional autoencoders. In Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II 24 (pp. 373-382). Springer International Publishing.

Gao, Z., Zhu, X., Wang, D., Pan, C., Yan, B., & Li, J. (2021). Insights into hydrothermal controls and processes leading to the formation of the Late Ediacaran Gaoyan stratiform manganese-carbonate deposit, Southwest China. Ore Geology Reviews, 139, 104524.

Goodfellow, I. (2016). Deep learning (Vol. 196). MIT press,767 p.

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.

Guo, J., Zhang, R., Sun, W., Ling, M., Hu, Y., Wu, K., Luo, M. and Zhang, L. (2018). Genesis of tin-dominant polymetallic deposits in the Dachang district, South China: Insights from cassiterite U–Pb ages and trace element compositions. Ore Geology Reviews, 95, pp.863-879.

Gysi, A. P., Williams-Jones, A. E., & Collins, P. (2016). Lithogeochemical vectors for hydrothermal processes in the Strange Lake peralkaline granitic REE-Zr-Nb deposit. Economic Geology, 111(5), 1241-1276.

Hadlich, I. W., Neto, A. C. B., Pereira, V. P., Botelho, N. F., Ronchi, L. H., & Dill, H. G. (2024). Mn–Fe-rich genthelvite from pegmatites associated with the Madeira Sn–Nb–Ta deposit, Pitinga, Brazil: new constraints on the magmatic-hydrothermal transition in the albite-enriched granite system. Mineralogical Magazine, 88(2), 111-126.

He, X., Bao, C., Lu, Y., Leonard, N., Liu, Z. and Tan, S. (2022). LA–ICP–MS U–Pb Dating, Elemental Mapping and In Situ Trace Element Analyses of Cassiterites from the Gejiu Tin Polymetallic Deposit, SW China: Constraints on the Timing of Mineralization and Precipitation Environment. Minerals, 12(3), p.313.

Iglesias, C., Antunes, I. M. H. R., Albuquerque, M. T. D., Martínez, J., and Taboada, J. (2020). Predicting ore content throughout a machine learning procedure–An Sn-W enrichment case study. Journal of Geochemical Exploration, 208, 106405.

Im, Jungho, John R. Jensen, Ryan R. Jensen, John Gladden, Jody Waugh, and Mike Serrato (2012). "Vegetation cover analysis of hazardous waste sites in Utah and Arizona using hyperspectral remote sensing." Remote Sensing 4, no. 2, 327-353.

Kumar, A. A., Sanislav, I., Huang, H., and Dirks, P. (2024). Cassiterite trace element discrimination diagrams to facilitate critical mineral exploration. Journal of Geochemical Exploration, 107530.

Liu, S., Liu, Y., Ye, L., Wei, C., Cai, Y. and Chen, W. (2021). Genesis of Dulong Sn-Zn-In polymetallic deposit in Yunnan Province, South China: Insights from cassiterite U-Pb ages and trace element compositions. Minerals, 11(2), p.199.

Kaski, S., Nikkilä, J., & Kohonen, T. (2003). Methods for exploratory cluster analysis. Intelligent Exploration of the Web, 136-151.

Konopka, B. M., Lwow, F., Owczarz, M., & Łaczmański, Ł. (2018). Exploratory data analysis of a clinical study group: Development of a procedure for exploring multidimensional data. PloS one, 13(8), e0201950.

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers, 7(1), 3-10.

Lehmann, B. (2021) Formation of tin ore deposits: A reassessment. Lithos, 402, 105756.

Li, Y., Luo, X., Chen, M., Zhu, Y., & Gao, Y. (2020). An Autoencoder-Based Dimensionality Reduction Algorithm for Intelligent Clustering of Mineral Deposit Data. In Proceedings of 2019 Chinese Intelligent Automation Conference (pp. 408-415). Springer Singapore.

Luo, Z., & Zuo, R. (2024). Causal Discovery and Deep Learning Algorithms for Detecting Geochemical Patterns Associated with Gold-Polymetallic Mineralization: A Case Study of the Edongnan Region. Mathematical Geosciences, 1-28.

McLachlan, G. J. (2000). Finite mixture models. A wiley-interscience publication.

Meyrieux, M., Hmoud, S., van Geffen, P., & Kaeter, D. (2024). CLUSTERDC: A New Density-Based Clustering Algorithm and its Application in a Geological Material Characterization Workflow. Natural Resources Research, 1-28.

Metallurgy in the Americas – Part VII (2019).

Mishra, S. (Ed.). (2022). Machine learning applications in subsurface energy resource management: state of the art and future prognosis.

Mikysek, P., Trojek, T., Mészárosová, N., Adamovič, J., & Slobodník, M. (2019). X-ray fluorescence mapping as a first-hand tool in disseminated ore assessment: sandstone-hosted U–Zr mineralization. Minerals Engineering, 141, 105840.

Möller, V., & Williams-Jones, A. E. (2017). Magmatic and hydrothermal controls on the mineralogy of the basal zone, Nechalacho REE-Nb-Zr deposit, Canada. Economic geology, 112(8), 1823-1856.

Mostafaei, K., Kianpour, M. N., & Yousefi, M. (2024). Delineation of Gold Exploration Targets based on Prospectivity Models through an Optimization Algorithm. Journal of Mining and Environment, 15(2), 597-611.

Murphy, K. P. (2012). Machine Learning, a probabilistic perspective. MIT Press,1104 p.

Nederbo, J. (2023). Vanadium Partitioning Between Magnetite, Hematite, and A Hydrothermal Fluid: Implications for Iocg and IOA Deposits (Master's thesis, Northern Illinois University).

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems, 14.

Ogasawara, Y., & Kon, M. (2021). Two clustering methods based on the Ward's method and dendrograms with interval-valued dissimilarities for interval-valued data. International Journal of Approximate Reasoning, 129, 103-121.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Reddy, C. K. (2018). Data clustering: algorithms and applications. Chapman and Hall/CRC.

Reynolds, D. A. (2009). Gaussian mixture models. Encyclopedia of biometrics, 741(659-663).

Riquelme, Á. I., & Ortiz, J. M. (2024). A Riemannian tool for clustering of geo-spatial multivariate data. Mathematical Geosciences, 56(1), 121-141.

Sangaiah, A. K., Javadpour, A., Ja'fari, F., Zhang, W., & Khaniabadi, S. M. (2022). Hierarchical clustering based on dendrogram in sustainable transportation systems. IEEE transactions on intelligent transportation systems, 24(12), 15724-15739.

Scholtysik, G., Dellwig, O., Roeser, P., Arz, H. W., Casper, P., Herzog, C., & Hupfer, M. (2020). Geochemical focusing and sequestration of manganese during eutrophication of Lake Stechlin (NE Germany). Biogeochemistry, 151(2), 313-334.

Schreiber, H. D. (1977). Redox states of Ti, Zr, Hf, Cr, and EU in basaltic magmas-an experimental study. In In: Lunar Science Conference, 8th, Houston, Tex., March 14-18, 1977, Proceedings. Volume 2. (A78-41551 18-91) New York, Pergamon Press, Inc., 1977, p. 1785-1807. (Vol. 8, pp. 1785-1807).

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence, 22(8), 888-905.

Sinha, S., Fasullo, J., Nerem, R. S., & Monteleoni, C. (2023). Multi-decadal Sea Level Prediction using Neural Networks and Spectral Clustering on Climate Model Large Ensembles and Satellite Altimeter Data. arXiv preprint arXiv:2310.04540.

Smith, C. M., Faulds, J. E., Brown, S., Coolbaugh, M., DeAngelo, J., Glen, J. M., ... & Ayling, B. F. (2023). Exploratory analysis of machine learning techniques in the Nevada geothermal play fairway analysis. Geothermics, 111, 102693.

Studemeister, P. A. (1983). The redox state of iron: a powerful indicator of hydrothermal alteration. Geoscience Canada.

Thomas, R., Davidson, P., & Appel, K. (2019). The enhanced element enrichment in the supercritical states of granite–pegmatite systems. Acta Geochimica, 38, 335-349.

Talebi, H., Peeters, L. J. M., Mueller, U., Tolosana-Delgado, R., & van den Boogaart, K. G. (2020). Towards geostatistical learning for the geosciences: A case study in improving the spatial awareness of spectral clustering. Mathematical Geosciences, 52(8), 1035-1048.

Underwood, C. C., McMillen, C. D., Chen, H., Anker, J. N., & Kolis, J. W. (2013). Hydrothermal chemistry, structures, and luminescence studies of alkali hafnium fluorides. Inorganic Chemistry, 52(1), 237-244.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).

Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and computing, 17, 395-416.

Wang, C., Zhao, K.-D., Chen, J., and Ma, X. (2022) Examining fingerprint trace elements in cassiterite: Implications for primary tin deposit exploration. Ore Geology Reviews, 149, 105082.

Xie, J., Girshick, R., & Farhadi, A. (2016, June). Unsupervised deep embedding for clustering analysis. In International conference on machine learning (pp. 478-487). PMLR.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. Annals of data science, 2, 165-193.

Yan, Q. H., Qiu, Z. W., Wang, H., Wang, M., Wei, X. P., Li, P., ... & Liu, J. P. (2018). Age of the Dahongliutan rare metal pegmatite deposit, West Kunlun, Xinjiang (NW China): Constraints from LA–ICP–MS U–Pb dating of columbite-(Fe) and cassiterite. Ore Geology Reviews, 100, 561-573.

Yeap, E. B. (1993) Tin and gold mineralizations in Peninsular Malaysia and their relationships to the tectonic development. Journal of Southeast Asian Earth Sciences, 8(1-4), 329-348.

Yu, S., Deng, H., Liu, Z., Chen, J., Xiao, K., & Mao, X. (2024). Identification of Geochemical Anomalies Using an End-to-End Transformer. Natural Resources Research, 1-22.

Zelnik-Manor, L., & Perona, P. (2004). Self-tuning spectral clustering. Advances in neural information processing systems, 17.

Zuo, L., Wang, G., Carranza, E. J. M., Pang, Z., Ren, H., Cao, K., ... & Gao, M. (2023). Deep Vector Exploration via Alteration Footprints and Thermal Infrared Scalars for the Weilasituo Magmatic–Hydrothermal Li–Sn Polymetallic Deposit, Inner Mongolia, NE China. Natural Resources Research, 32(5), 1871-1895.

Zuo, R., & Carranza, E. J. M. (2023). Machine learning-based mapping for mineral exploration. Mathematical Geosciences, 55(7), 891-895.

SAŽETAK

Napredne tehnike klasteriranja za klasifikaciju naslaga kositra u Maleziji: pristup strojnoga učenja

Ova studija istražuje primjenu naprednih tehnika klasteriranja – spektralno klasteriranje, Gaussove modele miješanja (GMM) i hibridni pristup koji kombinira autoenkodere s metodom *k-means* u svrhu klasifikacije naslaga kositra u Maleziji. Geokemijski podatci 28 uzoraka rude kositra u regijama Pengkalan Hulu North, Menglembu, Klian Intan i Sungai Lembing analizirani su kako bi se identificirali različiti obrasci mineralizacije. Rezultati su otkrili da je integracija autoenkodera s metodom *k-means* dala najvišu kvalitetu klasteriranja, sa Silhouette vrijednosti iznad 0,4 i Calinski-Harabasz indeksom od 90 na četiri klastera, nadmašujući ostale metode. Klasifikacija je učinkovito razlikovala pegmatitne, hidrotermalne, polimetalne i diseminirane tipove ležišta, što je u skladu s geološkim karakteristikama regija. Ovi nalazi poboljšavaju razumijevanje distribucije naslaga kositra nudeći znatan potencijal za optimiziranje strategija istraživanja i rudarskih operacija, čime se pridonosi održivomu gospodarskom razvoju malezijske rudarske industrije kositra.

Ključne riječi:
ležišta kositra, tehnike klasteriranja, geokemijska analiza, Malezija, istraživanje minerala

Author’s contribution

Meisam Saleki (Lecturer, Mining) Data interpretation and writing. **Kamran Mostafaei** (Assistant Professor, Mineral Exploration) Data interpretation, presentation of the results, and writing. **Zakaria Bin Endut** (Assistant Professor, Economic Geology) provided the data. **Mohammad Nabi Kianpour** (Researcher, Mineral Exploration) Coding and data analysis, presentation of the results, and writing.
All authors have read and agreed to the published version of the manuscript.