IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Preliminary Study for AI Application of Mineral Exploration Using Sediment Geochemical Data

Sunghyun Yoon[1], Hyoju Jun[2], Minwoo Chung[3], William Jae Hu Lee[4]

1. North London Collegiate School Jeju, 33 Globaleduro 145 beongil, Daejeongeup Seogwiposi, Jeju-do, 63644, South Korea
2. Saint Paul Preparatory Seoul, 14-8, Seochojungang-ro 31gil, Seocho-gu, Seoul, 06593, South Korea
3. The Masters School 49 Clinton Ave, Dobbs Ferry, New York, 10522, USA
4. Mercersburg Academy 100 Academy Dr, Pennsylvania, 17236, USA;
sunghyun.yoon1003@gmail.com

ABSTRACT: As mineral deposits near the surface are progressively depleted, there is an increasing need to invest more time and resources into developing new methods and protocols for exploration in deeper and covered conditions. This need is particularly critical for strategic resources such as essential minerals for electric vehicle batteries, which are becoming increasingly scarce. This study aims to aid geochemists in discovering mineral deposits by applying machine learning (ML) techniques on various geoscience datasets. Specifically, our preliminary study focuses on enhancing the efficacy of mineral exploration by modifying the model structure using ML methods based on sediment geochemical data from South Korea. To analyze the features of three key minerals (Ni, Co, and Li), three distinct engineering techniques were implemented to extract the top 10 features highly correlated with each target mineral. By incorporating features strongly associated with these minerals, our findings indicate an improvement in the accuracy of mineral exploration. Furthermore, we propose that future research should apply learning algorithms under geological constraints using various datasets to enhance exploration efficacy further.

KEYWORDS: Earth and Environmental Sciences, Geosciences, Machine Learning (ML), Sediment Geochemical Data, Feature Engineering.

Introduction

Mineral resources are a vital aspect that contributes to the sustenance of modern civilization. Mineral resource, commonly defined as "a concentration of naturally occurring material in or on the Earth's crust in such form and amount that economic extraction of a commodity from the concentration is currently or potentially feasible," is essential to modern technologies' industrial processes and manufacturing.[1-3] To enhance the success of mining projects, mineral resource estimation is conducted to determine the quantity of mineral deposits and establish confidence in geological interpretation. Therefore, a methodology that allows accurate reserve estimation is essential.

Conventional estimation methods for resource exploration are divided into geometric and geostatistical techniques. Despite their utility, these techniques possess inherent limitations, such as the tendency to underestimate or overestimate resources and display poor performance with highly heterogeneous datasets. Moreover, both methods involve expensive and time-consuming drilling operations.

Recently, Machine learning (ML) technologies have been actively employed in mineral resource estimation to mitigate these challenges. Machine learning (ML) techniques are "algorithms capable of learning and modeling complex nonlinear patterns in a large dataset"[4] With the integration of deep learning methods, there have been significant advancements in mineral resource exploration and estimation. At the core of deep learning are neural networks with multiple layers, allowing for the transformation of data representations from lower to higher levels of abstraction. As data passes through successive layers, more straightforward features are gradually combined to form more abstract and complex representations. This hierarchical approach enables deep learning models to capture and process the complex nature of geochemical patterns. In addition, AI analysis can be done without human intervention, reducing the risk of mineral exploration in hazardous or inaccessible areas.

Researchers have uncovered hidden geological and mineralogical information from multivariate geochemical datasets by leveraging deep learning techniques. Furthermore, integrating deep learning with big data analytics has enabled researchers to process massive amounts of geochemical data and extract meaningful insights. However, this abundance of data poses significant challenges, including the inherent difficulties of geoscience progress, data connection issues, and a need for ground truth and samples.[5] The inadequacy of conventional methods to address these challenges has spurred the application of machine learning for its ability to abstract high-level representations.[6] In scenarios with limited labeled data, Generative Adversarial Networks (GANs) have been utilized to train deep learning models like Convolutional Neural Networks (CNNs) for advanced image classification,[7,8] showing potential for exceeding the performance of contemporary methodologies in mineral prospectivity mapping and similar geoscience classification endeavors. Deep learning methods and remote sensing (RS) are widely applied in geological mapping and the exploration of mineralization processes,[9] and these are

DOI: 10.36838/v6i12.3

considered extreme events. The application of these methods demonstrates their broad applicability and potential[10,11] in enhancing geochemical and mineral exploration.

This study aims to 1) review the trend of resource exploration from conventional methods to ML techniques, 2) perform an exploratory analysis of 3 strategic minerals (Ni, Co, and Li) on geochemical data of stream sediment in South Korea, and 3) analyze and understand the whole process of extracting, interpreting and validating feature selection, which is applied to an essential process for ML algorithm models for three strategic minerals(Ni, Co, and Li).

Methods

1. Review:

It is necessary to evaluate the research trends in machine learning (ML) technologies over the past years (1993-2021). Figure 1. summarizes the various ML techniques applied in mineral resource estimation. Interestingly, artificial neural network (ANN) is the most used technique, followed by support vector machine (SVM). The remaining techniques are ensemble super learner (Ensemble), inverse distance weighted and artificial neural network (IDW-ANN), genetic algorithms (GA), k-nearest neighbor algorithm (kNN), support vector machines (SVM), support vector regression (SVR), random forest (RF), relevance vector machines (RVM), Gaussian process (GP), and new machine learning-based algorithm (GS-Pred). This is likely due to the robust capabilities of these techniques in pattern recognition and modeling complex systems. Unlike geostatistical methods, these ML techniques allow the modeling of physical features in complex systems without requiring explicit mathematical representations or exhaustive experiments. Therefore, ML can potentially provide a more accurate and efficient solution for estimating mineral grades.

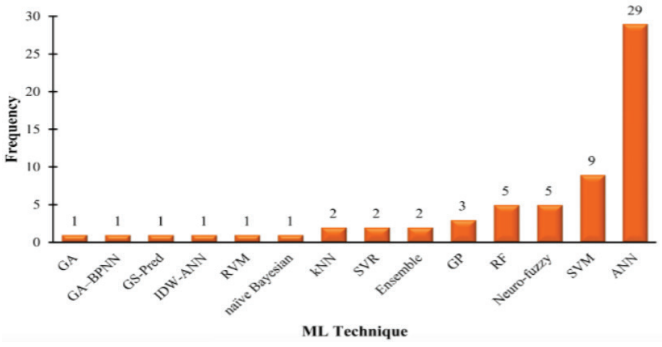

Figure 1: Distribution of machine learning techniques for mineral resource estimation. Artificial neural network (ANN) is the most applied technique.

Table 1 details the type of data used for the reviewed ML techniques. Most (over 80%) were based on field data from exploration drilling (drill cores) or trenching. The significant usage of field data demonstrates the industrial applicability of the proposed ML techniques, suggesting that these techniques can be incorporated into resource estimation tools in the industry. Historical, laboratory, and simulated data use accounts for less than 20% of the implementation. This study is based on the geochemical compositions of stream sediments. Interpreting the geochemistry of stream sediment samples requires a different approach than rock samples. Not only do the stream sediment samples contain secondary geochemical halos subject to the effects of multiple factors, but it is also recommended that the upstream catchments that supplied the samples be identified and that the values of the samples be attributed to these catchments for spatial interpretation.

Table 1: Summary of implementation of machine learning techniques. Most (over 80%) were based on field data from exploration drilling (drill cores) or trenching.

Implementation	Count
Historical and field data	1
Laboratory-scale data	1
Experimental and laboratory-scale data	2
Historical data	3
Simulated and field data	3
Field data	42

2. Conventional Technology:

We have referred to conventional resource estimation techniques as geometric and geostatistical methods. These estimation techniques are widely applied in mineral deposit evaluation. They have been used in the mining industry for a long time and form the basis of most mineral deposits being exploited today. Geometric techniques include classical polygon methods, square blocks, rectangular uniform blocks, triangular blocks, and polygonal blocks. These methods, particularly the polygonal approach, have been applied to estimate ore for a wide range of mineral deposits[12] and are most useful at the early exploration stage when collected data is relatively small.[13]

This technique's basis is that grades are allocated a definite area of influence generated by constructing the vertical bisectors between adjacent samples or intersections to obtain a regional estimate.[14] It is conceptually simple and fast, can classify irregular data, and is adaptable to narrow orebodies, but it may not be able to model thick and non-surface orebodies.[14]

Geostatistics is a subfield of statistics concerned with predicting random variables associated with spatial and spatiotemporal datasets.[15] It has a set of statistical methods widely applied in the mining industry, especially for predicting ore grades in mining operations. Other geoscience fields that apply geostatistics are petroleum, hydrology, and agriculture. Among the various geostatistical techniques, inverse distance weighting (IDW) and kriging are the most common methods used in mineral resource estimation.

3. ML (Machine Learning) Technology:

ML techniques cover geoscience applications, from identifying geochemical anomalies to evaluating mineral resources. ML techniques are not new in the mining industry; however, their application has surged during the last decade due to promising results in solving complex problems in the industry. ML is the study and application of computer algorithms to create intelligent systems that improve automatically through the experience without being explicitly programmed.[16] It is classified as a subfield of artificial intelligence (AI), which is

the science and engineering of making intelligent machines. ML applies computer algorithms to analyze and learn from data to determine or predict outcomes in various fields based on the structure of available data being analyzed.

ML models are categorized as supervised learning, unsupervised learning, and reinforcement learning.[17] Each of these classes is further categorized based on the nature of the problem being solved, using various corresponding learning algorithms or applications to handle such problems. In practice, many ML algorithms are applied in engineering; however, this study only considers algorithms commonly used for mineral resource estimation problems.

Figure 2 shows a generalized ML implementation process. The first step in the ML model development cycle (problem definition) deals with understanding the problem, characterizing it, and eliciting knowledge to acquire the relevant data. The second step (data collection) involves collecting all relevant data based on the features defined in the first step, followed by data preparation, pre-processing, and transformation.

Feature selection deals with automatic or manual selection parameters or variables that contribute most to the prediction variable or output of interest. Next, the data is divided into training, validation, and testing sets based on the data partition. Typically, 60~70% of the data is used for training, 20% for validation, and 10~20% for testing. An ML model is trained, validated, and tested using the partitioned datasets (train model). Model evaluation involves re-validating the model performance using new sample data. The model parameters (e.g., number of training steps, learning rate, initialization values) can be modified until a satisfactory performance is achieved, and then the model can be deployed for prediction.

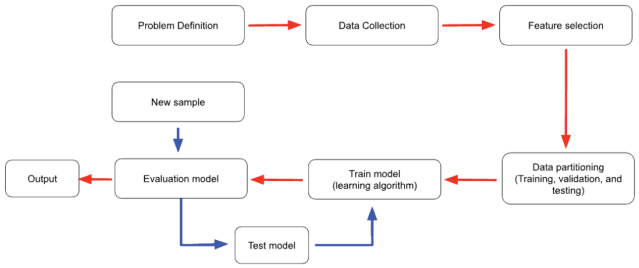

Figure 2: Machine learning implementation stages. It's worth noting that the data is divided into training(60~70%), validation(20%), and testing sets(10~20%) based on the data partition.

4. Application to Dataset:

The raw data of stream sediments based on the preliminary study were collected from the Korean geochemistry map project conducted by KIGAM (Korea Institute of Geology and Mineral) and other studies (Table 2). Geochemical data (29 elements) from sediment samples (N=23,563) collected throughout Korea (94,756 km^2) were analyzed (Figure 3).

Table 2: Chemical data of three stream sediments in South Korea. The sampling point of stream sediment is shown in latitude and longitude. The raw data of stream sediments based on the preliminary study were collected from the Korean geochemistry map project conducted by KIGAM and other studies.

Latitude	Longitude	SiO_2	Al_2O_3	Fe_2O_3	CaO	Na_2O	K_2O	MgO	P_2O_5	MnO	TiO_2
128.17	35.81	57.76	20.46	4.68	2.31	1.06	2.60	1.37	0.07	0.10	0.60
128.17	35.79	56.37	21.37	5.64	2.84	1.57	2.65	1.55	0.15	0.08	0.77
128.16	25.80	59.84	18.82	4.71	1.77	1.43	2.80	1.15	0.13	0.06	0.65
Latitude	Longitude	Ba	Cu	Li	Ni	Pb	Sr	V	Zr	Co	Cr
128.17	35.81	1491.00	12.76	20.31	17.76	30.05	265.50	50.96	89.26		
128.17	35.79	1771.00	19.47	21.14	20.76	32.61	277.30	71.65	55.66	11.50	95.70
128.16	25.80	1725.00	13.79	26.64	18.30	30.09	207.60	49.75	56.25		
Latitude	Longitude	Rb	Zn	Ce	Cs	Sc	Eu	Yb	Th	Hf	
128.17	35.81										
128.17	35.79	89.60	95.60	145.00	3.93	8.33	1.35	1.82	14.40	7.78	
128.16	25.80										

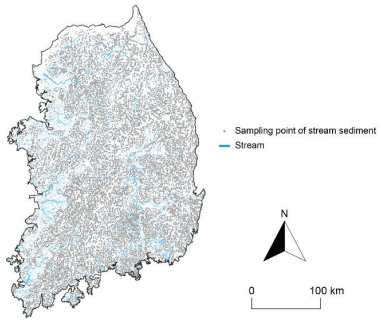

Figure 3: Sampling point of stream sediment in South Korea. Sediment samples (N=23,563) were collected throughout Korea (94,756 km^2).

Ten of the sediment's major elements - SiO_2, Al_2O_3, Fe_2O_3, CaO, Na_2O, K_2O, MgO, P_2O_5, MnO, TiO_2 - were identified by XRF, and 19 trace elements - Ba, Cu. Li, Ni, Pb, Sr, V, Zr, Co, Cr, Rb, Zn, Ce, Cs, Sc, Eu, Yb, Th, Hf were analyzed by ICP-AES, and the results of the missing element data for each sample are shown in Table 3.

Table 3: Descriptive statistical analysis of raw data in sediment. The missing element data for each sample is approximately 35.84% of the total.

Category	Information
Data Structure	23,563 rows, 35 columns
Key Information	Location information (UTM-X, UTM-Y, latitude, longitude), chemical composition of minerals (e.g., SiO_2, Al_2O_3, etc.)
Number of Rows with Missing Values	8,445 rows (approximately 35.84% of the total)
Total Number of Missing Values	92,548 (approximately 11.22% of total data)

The process of extracting 1) Exploratory Data Analysis, 2) data preprocessing, and 3) feature selection for mineral exploration using sediment geochemical data was performed with Python (version 3.12) for Windows.

The dataset includes location information 'UTMX' 'UTM-Y,' and information on the mineral resources at those points. The mineral properties comprise 29 attributes, including Ni, Co, Li, SiO_2, Al_2O_3, Fe_2O_3, etc.

To use the geochemical data for machine learning, we determined element thresholds(98%) through PCA of 29 elements and categorized the data into two groups based on this thresh-

DOI: 10.36838/v6i12.3

old. Data with values above the threshold were categorized as anomalies, and values below the threshold were categorized as normal. We also categorized them into 17 major rock types and added geochemical data, locations, and rock types to the training data.

5. Geological Review of Formation of Ni, Co, and Li:

As shown in the geological map of Korea (Figure 4), the geological structure of South Korea originated from the Precambrian tectonic plate and is composed of the following complex structures. Its complex structure includes continental interiors, active continental margins, and fault zones. These geological features have been instrumental in the formation of mineral resources such as Ni, Co, and Li.[18]

Ni and Co mineral resources in South Korea are often associated with ultrabasic and bare rocks. These rocks are closely related to mantle-derived molten rock upwelling from continental rift zones, which enrich Ni and Co. Korea's active continental margins, and rift zones provide favorable conditions for forming these mineral resources.[19] Li mineral resources are mainly found in pegmatite veins associated with granitic bodies. Pegmatites are associated with granitic bodies formed by Precambrian tectonic activity and contain high concentrations of Li. This is likely due to the enrichment of Li in the last minerals, which crystallize as the granite cools.[20]

Mineral resources such as Ni, Co, and Li found in Korea are directly related to the complexity of the Precambrian geotectonic domains. The stability of the continental interior, along with the dynamic geological activities of active continental margins and rift zones, enables the formation of these valuable mineral resources. Understanding these geological characteristics provides crucial information for establishing exploration and development strategies for mineral resources in Korea.

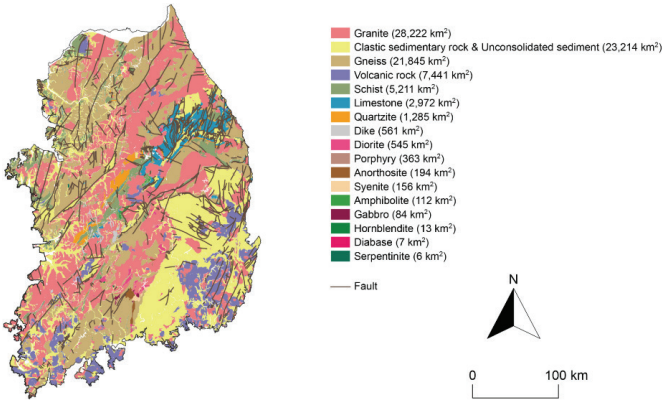

Figure 4: Geology map of South Korea classified with rock type. Lines explain the fault. Ni and Co mineral resources are often associated with ultrabasic and bare rocks. In addition, Li mineral resources are mainly found in pegmatite veins associated with granitic bodies.

■ Result and Discussion

1. Data Analysis and Preprocessing:

1.1. Exploratory Data Analysis (EDA):

EDA is crucial for understanding and analyzing data, uncovering hidden patterns, and examining basic features, distributional shapes, correlations, outliers, and visualizations.[21] EDA plays an important role in the initial data analysis, helping extract meaningful information from complex datasets.[22] In this study, EDA was employed to investigate the essential characteristics of the data and visualize the distribution of minerals by region. This process lays the foundation for deeper analysis based on the observed patterns in the data.

1.2. Descriptive Statistics:

Descriptive statistics aim to describe the characteristics of a data set, focusing on identifying the overall characteristics, variability, and outliers by summarizing the data distribution. This step is essential in the early stages of data analysis, providing insights into understanding data features and extracting critical information for solving problems. The primary focus of descriptive statistics includes identifying and summarizing central tendencies, distribution shapes, variance, mean, median, and mode of a dataset. Table 4 presents the statistical analysis results for the major minerals. The data shows no missing values for Ni and Li, while Co had 8,409 (23563-15154=8409) missing values. For Ni, the total data count is 23,563, with an average of 49 ppm and a standard deviation of 15.96. In addition, most Ni values are distributed between 0 and 50, with a high frequency around 12. For Co, there is a total of 15,154 data with an average of 14.14 ppm and a standard deviation of 8.48. With 429 outliers, most values are distributed between 0 to 40, with a high frequency of around 13.8. The total data count for Li is 23,563, with an average value of 71 ppm and a standard deviation of 25.46. With 894 outliers identified, most values are between 0 and 100, with a high frequency of around 36.

Table 4: Results of data preprocessing for statistical analysis of Ni, Co, and Li. The data shows no missing values for Ni and Li, while Co had 8,409 missing values. For Ni, with an average of 49 ppm(with 791 outliers). For Co, there is an average of 14.14 ppm (with 429 outliers); for Li, there is an average value of 71 ppm (with 894 outliers).

Mineral	Total number of data	Eigenvalue	The most common value (count)	Mean	Standard Deviation	Number of outliers	Range of distribution
Ni	23,563	6,891	12 (259 times)	49	15.96	791	0-50
Co	15,154	1,374	8 (116 times)	14.14	8.48	429	0-40
Li	23,563	7,975	36 (166 times)	71	25.46	894	0-100

1.3. Data Distribution Analysis:

Data distribution analysis is essential for identifying and understanding the data characteristics. Knowing the data distribution allows for selecting appropriate subsequent analysis methods and statistical techniques by considering the characteristics of the data. Additionally, data distribution can detect outliers and incorporate them in data preprocessing or modeling. Standard tools for representing data distribution visually include histograms, density plots, and box plots. The histogram displays the number of data for each section, the density plot shows the data density as a curve, and the box plot indicates the minimum, maximum, and median values. Statistical indicators

such as normal distribution, skewness, and kurtosis evaluate the data form. Here, the normal distribution refers to a symmetrical distribution where data is concentrated around the center. Skewness measures the degree of asymmetry in data (whether skewed to the right or left). At the same time, kurtosis indicates the length of the distribution's tails, with higher values representing longer-tailed distributions.

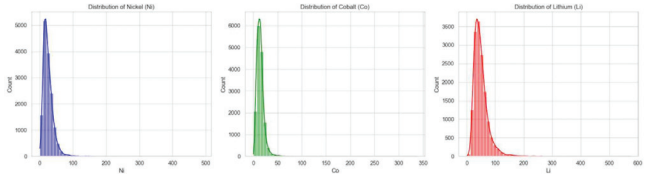

Figure 5: Histogram of Ni, Co, and Li. The Ni, Co, and Li distribution shows data concentrated at lower values. All three minerals exhibit high skewness and kurtosis values, indicating asymmetric and unbalanced distributions.

Figure 5 represents the distribution histogram of major minerals. The Ni distribution shows data concentrated at lower values, with a decreasing frequency of data as the value increases, indicating an asymmetric distribution. Also, the graph's skewness of 5.37 and the kurtosis of 91.29 suggest that the distribution has greater asymmetry and concentration at specific values compared to a normal distribution. Similarly, the data distribution of Co is also concentrated on lower values, showing a decreasing frequency of data as value increases. This indicates that the distribution also has the characteristics of an asymmetric distribution in which the skewness is 8.07 and kurtosis is 219.94. Again, the distribution of Co is a distribution with greater asymmetry and concentration of specific values compared to a normal distribution. The Li distribution follows the same trend, with a skewness of 2.83 and a kurtosis of 22.19.

In summary, the main characteristics observed in the data distribution analysis of Ni, Co, and Li are the presence of outliers and their influence. All three minerals exhibit high skewness and kurtosis values, indicating asymmetric and unbalanced distributions. The outliers present can distort the analysis results, making it necessary to identify and appropriately handle them for accurate data interpretation. Furthermore, the data distribution of each element significantly differs from a normal distribution, especially for the distribution of Ni and Co, which showed extreme values. This suggests the potential for these three elements to exhibit very high concentrations in specific regions or conditions.

1.4. Outlier Detection:

Outlier detection identifies extreme values or unusual observations in a dataset, revealing potential issues that may affect the analysis through values that deviate from the typical data patterns. Outliers can arise from various causes, such as data collection errors, measurement equipment faults, or natural phenomena. Detecting outliers is crucial for enhancing data accuracy and preventing distortion in analysis results.

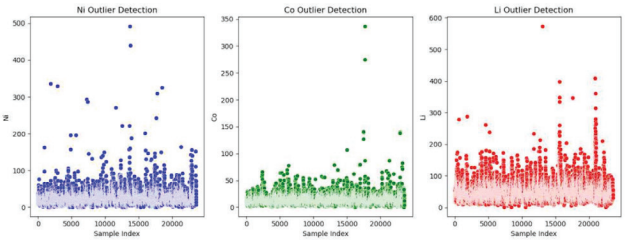

Figure 6: Visualization of density distributions by Ni, Co, and Li index for outlier detection. Values exceeding 50 for Ni, 40 for Co, and 100 for Li are considered outliers.

Figure 6 shows a density distribution chart by data index for outlier detection. Primary data containing outliers are visualized and expressed as a scatterplot. According to analysis, values exceeding 50 for Ni, 40 for Co, and 100 for Li are considered outliers. In machine learning model training, outliers can lead to instability, making it crucial to either remove or appropriately handle them during preprocessing. However, outliers might indicate significant geological events or mineral deposits in mineral exploration rather than statistical errors. Therefore, analyzing them for mineral exploration and predictive modeling is important instead of simply removing the outliers.

1.5. Map Visualization:

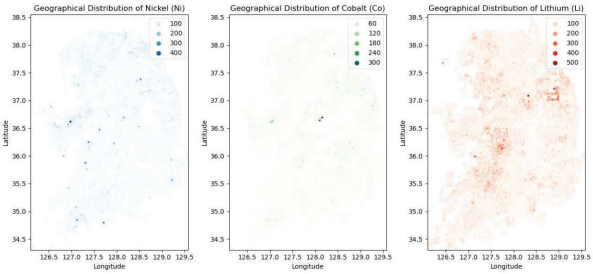

Figure 7: Spatial distribution characteristics of Ni, Co, and Li. Ni appears relatively low in most areas, but certain regions show high concentrations (up to 400 ppm). Co is less widespread than Ni, concentrated in specific areas with high concentrations (300 ppm). On the other hand, Li is distributed over the broadest area among the three minerals.

Figure 7 displays a map visualization of mineral concentrations in South Korea. The darker the color, the higher the mineral concentration, indicating that the mineral is more concentrated in that area and, therefore, more likely to be a valuable resource. Ni appears relatively low in most areas, but certain regions show high concentrations, with values reaching up to 400 ppm. Co, found in specific areas with high concentrations, is less widespread than Ni, showing a maximum value of 300 ppm. On the other hand, Li is distributed over the broadest area among the three minerals, with many areas showing high concentrations, and its concentrations reach values up to 500 ppm.

1.6. Data Preprocessing:

The dataset contains a significant amount of missing values. As for Co, there are 8,409 (35.69%) missing values in the Co

DOI: 10.36838/v6i12.3

alone (Table 4), with 23 in latitude and longitude, and a total of 8,422 rows in the total dataset have at least one missing value (Table3). Two methods can handle the missing values in the dataset: either removing or replacing missing values. Removing rows with many missing values can help the completeness of the data, but this may lead to a loss of too much information when the missing values constitute a large portion of the dataset. On the other hand, replacing missing values by utilizing statistical methods such as mean, median, and mode can preserve the size of the dataset but may increase its complexity. This study aims to improve the data quality by removing missing values. Considering the nature of missing values and how they do not occur randomly but due to specific patterns or circumstances, simply replacing them could ignore the underlying causes of the problem. Removing missing values accurately prevents data distortions and maintains the interpretability of models.

2. Feature Engineering:

2.1. Definition:

Feature engineering involves extracting domain knowledge from data to create features for operating machine learning algorithms. This process entails creating or choosing data features for machine learning models. Feature engineering significantly influences model performance and is a crucial step in learning model evaluation. It requires considerable expertise, time, and cost. In this study, we have focused on effective feature engineering, using location and chemical data of stream sediment.

Machine learning is a function of the input data and can be linear or non-linear. While extensive data can produce accurate outcomes, it might also result in errors. Therefore, having independent variables in a linear function does not necessarily increase the accuracy of the dependent variable in statistical analysis. Ideal input data consists of accurate information that is neither too insufficient nor too abundant. The appropriate approach involves collecting sufficient data and then identifying which features are useful. This process, known as feature selection or extraction, determines whether a feature is useful. Feature engineering can be broadly categorized into statistical methods and machine learning methods. In this study, we employed statistical (Correlation) and machine learning-based methods (Decision Tree and Random Forest) for feature engineering.

2.2. Types of Feature Engineering:

2.2.A. Correlation Analysis:

When choosing the main characteristics of Ni, Co, and Li, correlation analysis is crucial in understanding the strength and direction of correlations and linear connections between variables.[23]

2.2.B. Decision Tree:

Decision Tree is a standard predictive modeling method used in data analysis and machine learning (ML).[24] The technique involves analyzing the data characteristics to form a decision path and classifying the data by evaluating the importance of each variable. A decision path is first created to build a decision tree by choosing the variables that provide the most important information at each step. A decision tree algorithm based on information theory chooses key characteristics for target variables such as Ni, Co, and Li. The primary objective of this process is to identify the input variables that have the greatest influence on predicting the target variables.

2.2.C. Random Forest:

Random Forest belongs to an ensemble technique that integrates multiple decision trees to improve the prediction accuracy of a target variable.[25] This method is used in mineral exploration to predict the content of specific minerals such as Ni, Co, and Li. It facilitates efficient learning of complex structures while mitigating the risk of model overfitting. In mineral exploration, employing Random Forest can synthesize the predictions from numerous decision trees generated from bootstrap samples to derive a final prediction of mineral content.

2.3. Evaluating and Validating Performance:

A total of Four indicators Were Used for Evaluation.

2.3.A. Mean Squared Error(MSE):

MSE compares the average of the squared differences between the predicted and actual values of the dependent variable. As the amount of training data increases, the discrepancy in the calculation process's loss rate diminishes, making it easier to estimate outliers by continually squaring the differences. A smaller value indicates better performance.

2.3.B. Root Mean Squared Error (RMSE)

Compared to MSE, RMSE is calculated robustly against outliers in the data, so it is less affected by outliers. The smaller the value, the better the performance.

2.3.C. Mean Absolute Error (MAE):

Mean Absolute Error (MAE) is a metric to measure the average absolute difference between the estimated values and the actual values.

2.3.D. R^2 Score:

R^2 serves as an indicator of how well the independent variables explain the dependent variables. A coefficient close to 1 indicates a better correlation between the independent and dependent variables.

This was performed with Python (version 3.12) for Windows. The model was trained, as shown in Figure 8, and inferred, as shown in Figure 9.

Figure 8: The screen that trains the Model.

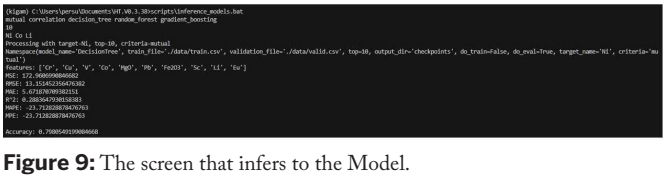

Figure 9: The screen that infers to the Model.

2.4. Results and Discussion:

In this study, we use a regression tree-based model, which is known to perform effectively in mineral prediction. The Base-

line model predicts mineral performance using all available features. We then compare the baseline results with those obtained using only the top 10 features selected through feature engineering.[26]

Table 5: Feature engineering performance of Ni, Co, and Li. With an MSE and an R^2 score, the performance improvement was achieved by statistical-based correlation for Ni. For Co, feature engineering using a Decision Tree proves to be effective in enhancing performance. In the case of Li, when applying feature engineering using Random Forest, these values represent improvements of 28.807 and 0.052 compared to the baseline.

Models	MSE	RSME	MAE	R^2 Score
Decision Tree Baseline (Ni)	181.220	13.462	5.966	0.254
w/Correlation	172.602	13.138	5.648	0.290
w/Decision tree	177.209	13.312	5.853	0.270
w/Randon forest	177.181	13.311	5.844	0.271
Decision Tree Baseline (Co)	24.076	4.907	2.485	0.400
w/Correlation	24.458	4.946	2.499	0.390
w/Decision tree	23.746	4.873	2.461	0.408
w/Randon forest	23.887	4.887	2.438	0.405
Decision Tree Baseline (Li)	512.380	22.636	16.206	0.075
w/Correlation	484.701	22.016	15.456	0.125
w/Decision tree	529.047	23.001	15.428	0.045
w/Randon forest	483.573	21.990	15.396	0.127

Table 5 shows the prediction performance for Ni, Co, and Li. For Ni, the MSE is 181.220, and the R^2 score is 0.254. Performance improvement has been achieved due to the application of feature engineering. Specifically, the feature engineering using the correlation method yielded the greatest performance improvement, with an MSE of 172.602 and an R^2 score of 0.290. Hence, the performance improvement achieved by statistical-based correlation surpasses machine learning-based feature engineering, indicating the need for diverse methodological approaches to feature analysis to depend on the target mineral type and characteristics. For Co in the Decision Tree Baseline, MSE is 24.076, and the R^2 score is 0.400. Applying feature engineering to correlation methodologies results in a decrease in performance. However, feature engineering using a Decision Tree proves to be effective in enhancing performance. Applying feature engineering through Decision Tree improves performance with an MSE of 23.746 and an R^2 score of 0.408 compared to the baseline. In the case of Li, the value of MSE is 512.380, and the R^2 score is 0.075. Feature selection through Decision Tree has a negative impact on performance.

On the other hand, feature selection through Correlation and Random Forest has a positive effect. When applying feature engineering using Random Forest, the MSE and R^2 scores are 483.573 and 0.127, respectively. These values represent improvements of 28.807 and 0.052 compared to the baseline, indicating the most significant improvement among all target minerals. This underscores the positive impact of applying feature selection on performance, even for minerals with high prediction difficulty. The reason for the decreased performance of Ni and Co after applying feature selection is predicted to be the broader distribution of the raw data, which makes model prediction difficult.

Overall, performance differences depend on the target mineral and its model. This proves that feature engineering is an important step in mineral exploration. Considering the escalating challenge of mineral resource discovery, adopting a data-driven AI approach incorporating feature engineering results derived from comprehensive feature exploration is advantageous.

Based on the spatial information (latitude and longitude) and geochemical data, the optimal feature engineering method was selected to extract the feature importance required for model training. The results are shown in Table 6.

Table 6: Feature performance summary. From the correlation analysis, the correlation coefficient between Ni and Cr is 0.60, Co and Sc is 0.69, and Li and Cs is 0.32, indicating that Cr, Sc, and Cs are the most important variables in predicting the presence of Ni, Co, and Li. Based on the decision tree, the presence of Ni minerals is predicted as Cr (0.41), Co minerals Sc (0.42), and Li minerals Cs (0.16), respectively. Based on random forest, the presence of Ni is Cr (0.43), Co minerals Sc (0.43), and Li minerals Cs (0.16).

Method	Mineral	Feature
Correlation	Ni	Cr, Co, MgO, V, Fe_2O_3, Sc, Cu, TiO_2, Eu, P_2O_5
	Co	Sc, Fe_2O_3, Cr, Ni, Eu, V, MgO, TiO_2, Cs, MnO
	Li	Cs, Rb, Al_2O_3, K_2O, Cu, Zr, Ba, Ni, P_2O_5, Co
Decision Tree	Ni	Cr, Cu, MgO, Cs, V, Pb, Co, SiO_2, Ce, Zr
	Co	Sc, Ni, MnO, Fe_2O_3, Eu, Cr, V, Hf, Na_2O, Rb
	Li	Cs, Ba, Zr, Cu, Sr, K_2O, Pb, MgO, Al_2O_3, Yb
Random Forest	Ni	Cr, Cu, V, MgO, Co, Cs, Pb, Zr, Ba, CaO
	Co	Sc, Ni, MnO, Eu, Cr, Fe_2O_3, Rb, MgO, Na_2O, Hf
	Li	Cs, Ba, Zr, Sr, Cu, K_2O, SiO_2, MgO, Pb, Ni

First, we used the correlation analysis method to select important features based on the correlation between the elements Ni, Co, and Li and the properties of other elements or compounds. Then, we evaluated the correlation coefficients of these features. The elemental symbols (importance) of the top 10 extracted minerals are shown below. The top elements and compounds highly correlated with Ni were Cr (0.60), Co (0.48), MgO (0.40), Fe2O3 (0.40), and V (0.40). In particular, the correlation coefficient between Ni and Cr is 0.60, indicating that Cr plays an important role in predicting the presence of Ni. The top elements and compounds that correlate well with Co include Sc (0.69), Fe_2O_3 (0.56), Cr (0.51), Ni (0.48), and V (0.42). The correlation between Co and Sc is 0.69, indicating that Sc is the most important variable in predicting the presence of Co. The top elements and compounds that are strongly correlated with Li include Cs (0.32), Rb (0.20), Al_2O_3 (0.18), K_2O (0.15), and Cu (0.09). In particular, the correlation coefficient between Li and Cs is 0.32, indicating the important role of Cs in predicting the presence of Li.

The results of the feature engineering based on the decision tree show that the most important attributes for predicting the presence of Ni minerals are Cr (0.41), Cu (0.11), MgO (0.08), V (0.05), and Cs (0.04). In particular, Cr is the most important variable for predicting the presence of Ni minerals, with

DOI: 10.36838/v6i12.3

an importance score of 0.41. These results suggest that Cr has a strong association with the presence of Ni minerals. For Co minerals, Sc (0.42), MnO (0.10), Al_2O_3 (0.09), Fe_2O_3 (0.08), and Eu (0.06) were found to be the most important attributes with an importance score of 0.42. The importance of Sc means that it plays a large role in predicting the presence of Co minerals. For Li minerals, Cs (0.16), Zr (0.08), Ba (0.07), Cu (0.05), and Sr (0.05) followed. Cs is the most important attribute with an importance score of 0.16, which can be interpreted to mean that Cs plays a crucial role in predicting the presence of Li.

The result of feature selection based on random forest, the attributes that are important in predicting the presence of Ni are Cr (0.43), Cu (0.11), MgO (0.06), V (0.06), and Co (0.04). In particular, Cr has an importance score of 0.43 with NI, suggesting that it is closely related to the presence of Ni. For Co minerals, the order of importance is Sc (0.43), Ni (0.99), MnO (0.06), Cr (0.05), and Eu (0.05). The most important variable for predicting the presence of Co is MnO, which has an importance score of 0.361, which means that MnO plays a significant role in predicting the presence of Co. For Li, Cs (0.16), Ba (0.06), Zr (0.06), Cu (0.05), and Sr (0.05) were selected. The most important variable for predicting the presence of Li minerals was selected as Cs, with an importance score of 0.16, which indicates that Cs plays a decisive role in predicting the presence of Li.

■ Conclusion and Further Study

Due to the increasing demand and insufficient supply of strategic minerals, along with rising uncertainty in international affairs and continued inefficiencies in resource exploration, new methods for resource development are required. Advanced research using machine learning (ML) methodology is being pursued to predict mineral resources, especially in regions with complex geological environments and intricate mineral distributions. Notably, this study presents the first application of feature engineering using a sediment geochemical dataset in South Korea. This study explored three feature engineering methodologies (correlation, decision tree, random forest) for three minerals (Ni, Co, Li) in South Korea. Preliminary results demonstrate that using features highly correlated with each target mineral enhances the accuracy of mineral exploration. A data-driven AI perspective that applies feature engineering results obtained through detailed feature exploration appears advantageous. In addition, data visualization and model analysis results indicate that additional data collection and a more precise understanding of geological characteristics are necessary to improve predictions. Future research will analyze the effectiveness of feature engineering in depth. On the other hand, we are preparing the following paper, "Al model application and Resource Prediction Mapping in Korea".

One major challenge encountered in this preliminary study is reflecting the complexity of the geological processes in forming geochemical patterns near the surface. Thus, further research is necessary to determine whether stream sediments have been contaminated by upstream facilities or construction activities (e.g., abandoned mines) and to assess the usability of the data. Before applying the Al model, feature engineering should be preceded by weighting, limiting, or excluding data by evaluating the geological implications. For example, specifying the 90 percentile as the geochemical threshold, and extracting feature importance for samples containing Li concentration above the 90 percentile would be expected to improve the accuracy of the prediction.

Another challenge identified is the application of deep learning algorithms under geological constraints. While deep learning algorithms are powerful tools for extracting geological features and integrating multi-source data to discover geochemical patterns or delineate geochemical anomalies that conventional methods cannot detect, some detected anomalies/patterns may lack geographical interpretation. This is because the results obtained by deep learning algorithms depend solely on the data without considering geological conditions. Mineralization is a multifactorial process involving numerous geological processes and controlling factors. Therefore, detected geochemical patterns/anomalies related to mineralization should be interpreted within a geological framework. Incorporating sample catchment basins (i.e., area of influence around stream sediment samples) in the deep learning of sediment geochemical data could provide a more robust analysis of geochemical anomalies than conducting deep learning and weighted drainage catchment basin analysis separately.

This approach should reflect an understanding of the geological constraints, such as bedrock-centered mineralization processes and geostructural settings (i.e., faults).

The third challenge lies within the deep learning algorithm itself. Research suggests that the analysis of geochemical compositional data using original SOM (self-organizing map) may not yield coherent results.[27] It is thus necessary to propose novel ideas, such as new deep learning architectures, dropout regularization, compositional problem-solving strategies, and new geochemical and mineral exploration sampling strategies.[4,10,28] Further studies should aim to obtain stream sediment analyses from actual Li-bearing areas and use them as answer sets to perform supervised learning, thereby probabilistically determining the presence of minerals in existing data.

■ References

1. Henckens, M.L.C.M.; Biermann, F.H.B.; Driessen, P.P.J. Mineral resources governance: A call for the establishment of an International Competence Center on Mineral Resources Management. Resour. Conserv. Recycl. 2019, 141, 255–263.
2. Crowson, P.C.F. Mineral reserves and future minerals availability. Miner. Econ. 2011, 24, 1–6.
3. Zerzour, O.; Gadri, L.; Hadji, R.; Mebrouk, F.; Hamed, Y. Geostatistics-Based Method for Irregular Mineral Resource Estimation.
4. 25. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine learning in agriculture: A review. Sensors 2018, 18, 2674.
5. Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V., 2018. Machine learning for the geosciences: challenges and opportunities. IEEE Trans. Knowl. Data Eng.
6. Zuo, R., Xiong, Y., 2018. Big data analytics of identifying geochemical anomalies supported by machine learning methods. Nat. Resour. Res. 27, 5–13.

7. Lin, D., Fu, K., Wang, Y., Xu, G., Sun, X., 2017. MARTA GANs: unsupervised representation learning for remote sensing image classification. IEEE Geosci. Remote Sens. Lett. 14, 2092–2096.
8. Zhu, L., Chen, Y., Ghamisi, P., Benediktsson, J.A., 2018. Generative adversarial networks for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 99, 1–18.
9. Mulder, V.L., De Bruin, S., Schaepman, M.E., Mayr, T.R., 2011. The use of remote sensing in soil and terrain mapping—a review. Geoderma 162, 1–19.
10. Cheng, Q., 2018. Mathematical geosciences: local singularity analysis of nonlinear earth processes and extreme geo-events. In: Handbook of Mathematical Geosciences. Springer, Cham, pp. 179–208.
11. Xiong, Y., Zuo, R., 2018. GIS-based rare events logistic regression for mineral prospectivity mapping. Comput. Geosci. 111, 18–25.
12. Haldar, S.K. Mineral resource and ore reserve estimation. In Mineral Exploration; Elsevier: Amsterdam, The Netherlands, 2018; pp.145–165.
13. Afeni, T.B.; Akeju, V.O.; Aladejare, A.E. A comparative study of geometric and geostatistical methods for qualitative reserve estimation of limestone deposit. Geosci. Front. 2020, 12, 243–253.
14. Glacken, I.M.; Snowden, D.V. Mineral resource estimation. In Mineral Resource and Ore Reserve Estimation—The AusIMM Guide to Good Practice; Edwards, A.C., Ed.; The Australasian Institute of Mining and Metallurgy: Melbourne, Australia, 2001; pp. 189–198.
15. Varouchakis, E.A. Geostatistics. In Spatiotemporal Analysis of Extreme Hydrological Events; Elsevier: Amsterdam, The Netherlands,2018; ISBN 9780128116890.
16. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine learning in agriculture: A review. Sensors 2018, 18, 2674.
17. Zuo, R. Machine Learning of mineralization-related geochemical anomalies: A Review of potential methods. Nat. Resour. Res.
18. O. J. Kim, "Mineral Resources of Korea1," in CircumPacific Energy and Mineral Resources. American Association of Petroleum Geologists, 01 1976.
19. H.-M. Cho, H. Kim, H.-T. Jou, J. K. Hong, and C.-E. Baag, "Transition from rifted continental to the oceanic crust at the southeastern Korean margin in the east sea (Japan Sea)," Geophysical Research Letters, vol. 31, 2004.
20. I.-H. Oh, S.-J. Yang, C.-H. Heo, J.-H. Lee, E.-J. Kim, and S.-J. Cho, "Study on the controlling factors of Li-bearing pegmatite intrusions for mineral exploration, Uljin, South Korea," Minerals, vol. 12, no. 5, 2022.
21. Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, USA-Turkey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, USA
22. J. T. Behrens, "Principles and procedures of exploratory data analysis." Psychological Methods, vol. 2, pp. 131–160, 1997.
23. R. Prematunga, "Correlational analysis." Australian Critical Care: Official Journal of the Confederation of Australian Critical Care Nurses, vol. 25 3, pp. 195–9, 2012.
24. F.-J. Yang, "An extended idea about decision trees," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 349–354, 2019.
25. Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32
26. H. Friedman, "Greedy function approximation: A gradient boosting machine." Annals of Statistics, vol. 29, pp. 1189–1232, 2001.
27. Cortés, J.A., Palma, J.L., 2008. Using Self-Organizing Map with geochemical compositional data. In: AGU Fall Meeting.
28. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–195

Authors

Sunghyun Yoon: Currently attending North London Collegiate School Jeju as a junior and interested in the field of environmental science.

Hyoju Jun: Currently attending Saint Paul Preparatory Seoul as a junior and interested in Engineering.

Minwoo Chung: Currently attending The Masters School as a sophomore and interested in the field of climate change and oceanography.

William Lee: Currently attending Mercersburg Academy as a sophomore and interested in the field of agronomy and soil sciences.

DOI: 10.36838/v6i12.3