Rock Mechanics Bulletin 4 (2025) 100146

Full Length Article

Exploring machine learning techniques for open stope stability prediction: A comparative study and feature importance analysis

Alicja Szmigiel [a,*], Derek B. Apel [a,**], Yashar Pourrahimian [a], Hassan Dehghanpour [a], Yuanyuan Pu [b]

[a] University of Alberta, School of Mining and Petroleum Engineering, Edmonton, Alberta, T6G 2R3, Canada
[b] Chongqing University, Chongqing, 400044, China

ARTICLE INFO

Keywords:
Machine learning
Stope stability
Feature importance
Artificial neural network

ABSTRACT

The stability of underground excavations is essential for ensuring the safety of mining operations. Classical stability assessment methods, established in empirical formulas and rock mass classification systems, have long been employed for evaluating stope stability in underground mining. Stability graphs, a popular empirical approach, utilize factors like rock stress, joint orientation, and surface orientation to calculate stability numbers critical for stope design. However, modern advancements in machine learning present new opportunities for enhancing predictive capabilities and understanding complex relationships influencing stope stability. Building upon research demonstrating the feasibility of using machine learning for stability prediction, our study investigates and compares several machine learning algorithms. By analyzing a dataset comprising stope dimensions and geomechanical properties, we explore the potential of machine learning models such as Random Forest, Support Vector Machine, AdaBoost, XGBoost, LightGBM, and Artificial Neural Network in predicting stope stability. Evaluation metrics including accuracy, precision, recall, and F1 score are employed to assess model performance, with the Artificial Neural Network emerging as the most effective. Furthermore, SHapley Additive exPlanations (SHAP) analysis enhances interpretability by explaining the contribution of individual features to model predictions.

1. Introduction

Understanding the stability of underground excavations is important for ensuring the safety, durability, and longevity of mining operations. Over the years, classical stability assessment methods have played a crucial role in evaluating the stability of stopes in underground mining. These methods, rooted in well-established principles and empirical formulas derived from extensive field observations and laboratory tests, have provided valuable insights into rock mass behavior. Among the various empirical approaches utilized in assessing stope stability, stability graphs, as proposed by Mathews et al. (1980), have emerged as particularly popular. These graphs are built upon rock mass classification systems such as the *Q* value system by Barton et al. (1974) and the rock mass rating (RMR) introduced by Bieniawski (1973). By integrating factors like the rock stress factor (*A*), joint orientation adjustment factor (*B*), and surface orientation factor (*C*), stability graphs facilitate the calculation of the stability number N, a crucial parameter for designing stope dimensions and support.

However, in addition to these classical methods, modern advancements in machine learning have opened up new solutions for stope stability assessment. Research conducted by Adoko et al. (2022) demonstrates the feasibility of utilizing feed forward neural network classifiers to predict stope stability, achieving an average accuracy of 91 %. This study, based on a database comprising 225 cases from three mines in Ghana, West Africa, highlights the potential of machine learning in enhancing predictive capabilities and understanding complex relationships among various parameters influencing stope stability.

In our study, we analyze this dataset, by exploring and comparing several machine learning algorithms to further improve predictive accuracy. By incorporating parameters such shape dimensions and rock mechanical properties, we aim to gain a comprehensive understanding of the factors influencing stope stability. The comparative analysis includes machine learning models like Random Forest, Support Vector Machine (SVM), AdaBoost, XGBoost, LightGBM, and Artificial Neural Network

https://doi.org/10.1016/j.rockmb.2024.100146
Received 24 May 2024; Received in revised form 27 June 2024; Accepted 18 July 2024
Available online 31 August 2024

(ANN). Through rigorous evaluation and examination of metrics such as accuracy, precision, recall, and F1 score, we identify the strengths and weaknesses of each model in predicting stope stability. Furthermore, we employ SHapley Additive exPlanations (SHAP) analysis to gain interpretability into the predictions of each model. By analyzing the contribution of individual features to the model's predictions, SHAP analysis enhances transparency and reliability, facilitating improvements in stope design strategies and engineering efforts aimed at ensuring mining safety and efficiency.

2. Background of stope stability assessment

In underground mining operations, excavation stability is influenced by many geological and mining factors, each contributing to the overall integrity and safety. Open stope mining, while enhancing productivity and reducing exposure to unsafe conditions, carries the risk of significant overbreak, especially due to the characteristics of hangingwalls. Overbreak refers to the displacement of unstable rock beyond the intended stope design, often caused by sloughing (Capes, 2009). Such incidents lead to substantial operational costs, production disruptions, and safety hazards.

Several key factors influence overall stability in underground mining operations. Stress relaxation significantly impacts excavation by affecting rock mass behavior (Diederichs and Kaiser, 1999), while mining techniques, extraction rates, the mechanical and geological properties of surrounding rock, and stope dimensions also play crucial roles. Another incidents in mining operations involve faults, which are fractures or discontinuities in the Earth's crust where movement has occurred along the fracture plane. These movements can be horizontal, vertical, or diagonal, leading to the displacement of rock layers on either side of the fault. Faults are categorized based on their movement direction, the fault plane angle, and the type of stress causing the movement (Zhou et al., 2022). They are particularly common in metalliferous mines and are recognized as a cause of instability in mining excavations. In mining operations, managing the voids left after mining activities, known as post-mining space liquidation, is another critical issue. In open stope mining operations, this process often involves backfilling voids with various materials, commonly using a cemented rock fill or paste backfill (Lingga and Apel, 2018; Skrzypkowski, 2021a, 2021b). Additionally, the impact of adjacent stopes must be carefully assessed to control risks related to stability and stress changes from nearby mining activities. Managing mining edges, the boundaries of active excavation areas, is also crucial to prevent overbreak and maintain particular stope dimensions. Techniques such as controlled blasting and installing support systems are important in minimizing overbreak and ensuring safety.

Understanding the factors that influence stability is crucial for ensuring the safety, durability, and longevity of underground excavations. Classical stability assessment methods have long been employed to evaluate the stability of stopes in mining operations. These methods are based on well-established principles and empirical formulas derived from extensive field observations and laboratory tests.

Stability graphs, initially proposed by Mathews et al. (1980) represent one of the most popular empirical approaches utilized in assessing the stability of stopes in underground mining excavations. The method was developed based on popular and widely used rock mass classification systems such as *Q* value system presented by Barton et al. (1974) and rock mass rating (RMR) proposed by Bieniawski (1973). Matthew stability graph method revolves around determining crucial factors that influence the stability of the rock mass by utilizing specifically developed graphs that relate various characteristics and properties of the rock. These graphs facilitate factors such as the rock stress factor (*A*), joint orientation adjustment factor (*B*), and surface orientation factor (*C*). These factors are then combined to calculate stability number *N*, which is critical parameter developed for designing stope dimensions and support in underground mining openings. It serves as a quantitative indicator of the physical conditions and stability of the stopes and it is calculated as follows:

$$N = Q' \cdot A \cdot B \cdot C \tag{1}$$

The crucial factor for the successful assessment of opening stability in underground mining is the shape factor, which relates the dimensions of the opening and is commonly referred to as the hydraulic radius (*HR*). The hydraulic radius is a critical geometric parameter that characterizes the shape of the opening and is essential in determining its stability.

The term hydraulic radius is commonly understood as the ratio of the area of exposure of the hanging wall to its perimeter. In the context of inclined stopes, where the stope is not in a vertical position, the most critical aspect for calculating the *HR* is the exposure of the hanging wall (Fig. 1). The calculation of *HR* takes into account the spans of the stopes along the dip (*h*) and along the strike (*w*) (Tishkov, 2018), as shown of Fig. 1.

The final graph developed by Matthew for stope assessment plots the stability number N against the hydraulic radius *HR*. Each case in the dataset has been categorized as either stable, unstable, or caved based on previous evaluations of real cases. As a result, the graph is divided into three distinct zones, representing the stability assessments for the stopes.

This graph serves as a valuable tool for classifying new stopes into one of these three stability categories. When a new stope is to be assessed, its corresponding values of *N* and *HR* can be plotted on the graph. By examining the location of these values on the graph, it can easily be determined whether the new stope falls into the stable, unstable, or caved zone. This classification aids in understanding the potential stability conditions of the new stope and guides the selection of appropriate support measures to ensure its safety and structural integrity (Suorineni et al., 2001).

Visual separation of zones in stability assessment graphs can pose challenges related to subjectivity and reproducibility. Human judgment may introduce variations in categorizing stopes into stable, unstable, or caved zones, leading to discrepancies in the analysis. Additionally, the inherent risk associated with potential errors in categorization raises concerns about the reliability of the assessments. Empirical design methods have the potential to continuously improve and evolve over time as more data becomes available and engineers gain increased experience with the method. The nature of empirical design allows for flexibility and adaptability, enabling engineers to refine and update the approach based on new observations and findings.

Since 1980, significant research in mining engineering have contributed to enhancing the reliability and effectiveness of stability graphs for assessing underground mining excavations. These developments have been crucial in refining the stability graph method and making it a more robust and trusted tool in mining practice. In 1988 Potvin's approach introduced a refinement to the stability graph, departing from the traditional three-zone classification system proposed by Mathews et al. Instead, Potvin proposed the stability graph with two main zones, stable and caved. However, he also incorporated an

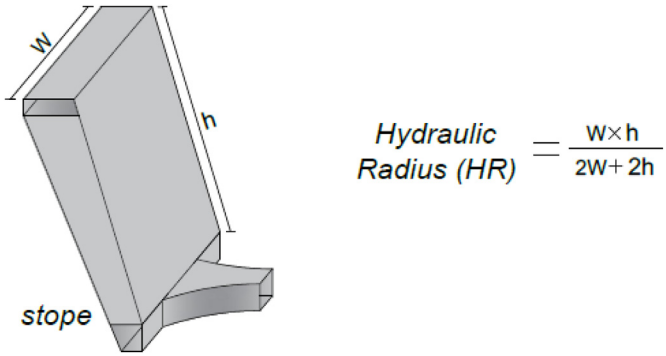

Fig. 1. Calculation of Hydraulic radius scheme.

additional transition zone, representing a critical boundary between the stable and caved regions (1988). Potvin acknowledged that the stability graph can be affected by human bias and unknown inherent errors when visually defining its zones and suggested utilizing statistical tools for zone definition instead.

Nickson (1992) was a pioneer in attempting to establish the boundary positions through statistical methods. He employed discriminant analysis on the multivariate stability database, a three-dimensional dataset and used Mahalanobis' distance to divide the data into distinct groups. He achieved this by deriving a linear separation between stable and caved unsupported scenarios, utilizing a logarithmic transformation. Notably, his analysis excluded unstable cases, and he did not determine separation lines for zones involving unstable or caving situations. Nickson also compared his statistically determined boundary, which separates stable and caved conditions, with Potvin's proposed transition zone. Based on his findings, Nickson recommended that Potvin's transition zone should be employed for designing unsupported stope surfaces. Hadjigeorgiou and colleagues (2025) gathered additional stability data and conducted a repeat analysis using discriminant methods, yielding comparable outcomes. Suorineni et al. (2001) introduced the Bayesian likelihood method as a powerful tool for statistically interpreting the stability graph. They employed an extended database based on the Potvin calibrated stability graph factors to illustrate the method's advantages. The Bayesian likelihood discrimination proved to be an optimal approach for statistically interpreting the stability graph due to its capability to reveal substantial overlap among the defined stability graph zones (stable, unstable, and caving). It also allowed for error rate estimation in the stability graph, delineated general transition boundaries between stable, unstable and caved stopes, estimated inherent predictive errors in stability graphs, evaluated the risk associated with using the stability graph for predictions, and introduced a multiple design curves stability graph based on the probability. The Bayesian likelihood discrimination's ability has been harnessed to provide deeper significance to the class boundaries in the stability graph and individual stope walls plotted within each class.

Numerical modeling stands as another widely adopted approach that has proven effective in addressing stope stability concerns. Henning and Mitri (2007), for instance, crafted a series of three-dimensional numerical models to explore the impacts of field stress, mining depth, stope configuration, and orientation on stope wall overbreak. Similarly, Purwanto et al. (2013) harnessed numerical modeling to establish the correlation between stope design and the stability of hanging wall. Hu and Cao (2009), by employing visual numerical simulation software, simulated and computed stress distribution and displacement variations within stopes during mining operations. They conducted an analysis of the stability of stope roofs and adjacent rock, as well as the alterations in sound emission associated with the mining process.

While classical stability assessment methods have been effective in many cases, the emergence of machine learning techniques has opened up new possibilities for enhancing their accuracy and efficiency. Machine learning algorithms can analyze large volumes of mining data, including shape of the opening, the properties of surrounding rock mass, underground conditions and historical stability records, to identify patterns and correlations that may not be easily discernible through traditional methods. Literature presents many successful applications of Machine Learning models in mining engineering and geotechnical challenges (Song et al., 2024; Liu et al., 2024). By training machine learning models on a dataset of known stability outcomes, engineers can develop predictive models that can assess the stability of new stopes.

Several studies have explored the integration of machine learning models for the prediction of stope stability. Erdogan Erten et al. (2021) introduced a hybrid artificial neural network (ANN) approach optimized through grid search. This method was compared with conventional techniques including Naive Bayes (NB), Decision Tree (DT), k-Nearest Neighbors (kNN), Support Vector Machine (SVM), as well as the traditional stability graph method. The findings of this study revealed that the performance of the stability graph method falls short of the capabilities exhibited by machine learning algorithms. Notably, the ANN model with hyper-parameters tuned using the grid search technique showcased superior performance in terms of accuracy, precision, recall, f-measure, and g-mean compared to other machine learning algorithms. Saadaari et al. (2020) investigated the viability of employing Ensemble Learning methods to categorize and predict the stability condition of stope surfaces. They introduced and evaluated four techniques—Random Forest (RF), Gradient Boosting (GB), Bootstrap Aggregating Classifier (BAC), and Adaptive Boosting (AB)—using widely accepted and effective assessment metrics. Upon analyzing the performance outcomes, it was evident that among the four machine learning models, Gradient Boosting (GB) and Bootstrap Aggregating Classifier (BAC) demonstrated the highest efficacy in accurately classifying and predicting the stability state of stopes, encompassing categories of caved, stable, or unstable. In a comparative investigation conducted by Qi et al. (2018a), five distinct artificial intelligence strategies based on machine learning and metaheuristic algorithms were explored for their potential in predicting the stability of open stope hangingwalls (HW). The assessed algorithms encompassed logistic regression (LR), multilayer perceptron neural networks (MLPNN), decision tree (DT), gradient boosting machine (GBM), and support vector machine (SVM). The optimization of hyperparameters was facilitated using the Firefly algorithm (FA), which yielded successful results for this purpose. Across the testing phase, the most favorable performance was exhibited by the optimized GBM model, closely followed by the SVM model and the optimized LR model. The study highlighted the remarkable predictive capabilities of these three machine learning models in forecasting HW stability. Several researchers have directed their efforts toward the refinement and customization of specific machine learning models for assessing stope stability. Qi et al. (2018b), in their study, concentrated on optimizing the Random Forest model for enhanced efficiency, while Santos and associates (2020) shifted their attention toward the utilization of Artificial Neural Networks.

It is significant to continue research in the area of stope stability in underground mining. The development of mining engineering is constantly moving towards increased exploitation and most favorable optimization, this is causing the industry to decide on larger sized opening thus reducing the quantity if the stopes. This approach, even though mostly efficient for rapid extraction and profit growth, could have a negative impact on the stability of the opening. When dimensions are exceeded to their maximum, the stability of the stopes is reduced causing them to shift toward unstable condition or even catastrophic failure.

At the stage of mine planning and stopes design, it is crucial to consider and evaluate all the parameters that have direct or non-direct influence on the stability condition. The integration of machine learning with classical stability assessment methods offers several advantages in that matter. It allows for more comprehensive and data-driven evaluations, considering a wider range of variables and their interactions, that might have crucial impact on the stability of each open stope. Machine learning models can also help identify complex relationships between shape factor, rock properties, and stability outcomes that may not be apparent using traditional approaches. Hence the necessity to investigate the stope stability data further, in order to determine which parameter has the most significant impact on the stability of a stope and how to use that knowledge to increase safety in mining environment.

3. Database analysis and pre-processing

Analyzing a database and proper preprocessing is a crucial step in the machine learning process and have significant impact on the success of any machine learning model. It is important to assess the quality and characteristics of the database before employing any predictive model. Real-world data often contains missing values, outliers, duplicates, and inaccuracies that have negative impact on the model performance. Analyzing the database helps identify and address these issues, ensuring that the data used especially for training is reliable and accurate. Proper

assessment provides insights into characteristics of the database and its features, it allows us to determine the type of each attribute, whether they are numerical or categorical, and understand the distribution of values. This understanding is crucial for choosing preprocessing techniques followed by implementation of appropriate machine learning algorithms. Different machine learning models create different assumptions about the database, analysis allows us to ensure that the chosen algorithm aligns with the characteristics and complexity of the data. Presented by Crone et al. (2006) investigation of the impact of preprocessing, have shown a strong evidence that preprocessing techniques have critical impact on the predictive performance of the models. The results presented encourage to implement analysis of the database and preprocessing in order to produce valid and accurate outcomes of the classification algorithms.

The database investigated, comes from research conducted by Adoko et al. (2022), tiled "A Feasibility Study on The Implementation of Neural Network Classifiers for Open Stope Design". The study utilized a feed forward neural network classifier to predict the stability of open stopes in underground mining obtaining an average accuracy of 91 %. The general misclassification of the model (less than 10 %) showed that FFNN outperformed the classic stability graph methods, which yielded the misclassification of almost 40 %.

Database consist of 225 cases and was collected from three different mines in Ghana, West Africa, in the period of over three months. It includes information such as height, span and length of each stope, as well as geomechanical properties: Q' value, rock stress factor (A), joint orientation adjustment factor (B) and surface orientation factor (C). The features that were passed to the algorithm were modified stability number N' and hydraulic radius HR determined from the database. Fig. 2 illustrates the distribution of cases. A plot of stability number N versus hydraulic radius was created, where different colors were employed to signify whether a specific case was classified as stable, unstable, or caved. It provided us with a visual representation of how the data is spread across different classes. Analysis of the plot revealed that cases with grater hydraulic radius tend to exhibit a higher tendency for instability and caving. As for the stability number, its influence on stability is somewhat less pronounced. However, instances with exceptionally low stability number values were predominantly categorized as caved. This observation also provided us with valuable insight into the relationship between weak rock conditions and low stability number values. In such scenarios, decreasing the dimensions of the opening can have a notably favorable effect on stability.

Adoko et al. (2022) investigated the Feed Forward Neural Network algorithm, where the input parameters were stability number N and shape factor HR. While the outcomes of this investigation yielded satisfactory accuracy in predicting stope stability, there remains untapped potential to delve deeper into the dataset. This further exploration holds the promise of enhancing the predictive capabilities of the model. Additionally, it offers an opportunity to conduct a comprehensive examination of all the constituent features that contribute to the computation of the shape factor HR and the stability number N. This extended analysis can provide a more in-depth understanding of the complex relationship among these features and their collective influence on stope stability. Therefore, in our study we investigated and compared several machine learning algorithms where the input data were all the parameters that combine into stability number N and HR. These parameters were: shape parameters—height of the stope (H), Span (S), Length (L) and Rock mechanical properties—Q' value, rock stress factor (A), joint orientation adjustment factor (B) and surface orientation factor (C). All the cases had an stability assessment determined: stable, unstable, caved. The example of 10 cases with all the parameters from the database is shown in the Table 1.

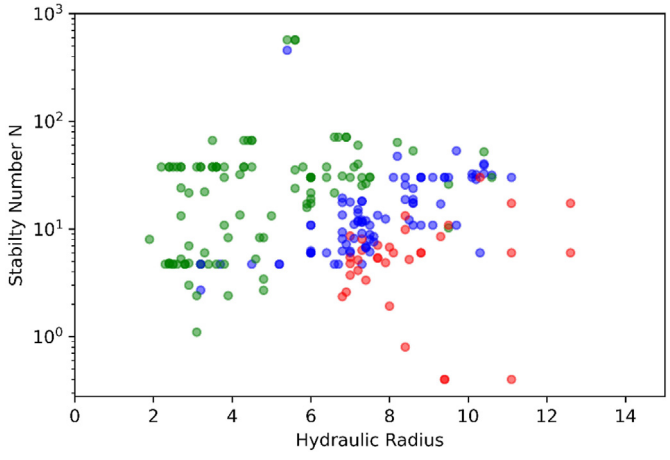

Fig. 2. Distribution of cases in database. Green—stable cases, blue—unstable and red—caved.

3.1. Database analysis

Before proceeding with data pre-processing the investigation of data investigation is a reasonable to step to ensure data quality and to help in making informed decisions about preprocessing technique and set the foundation for building effective and accurate machine learning models.

To identify the distribution of the data points the histograms of each feature were plotted (Fig. 3). Histograms provide a visual representation of the distribution of data across different values. This helps in understanding how the data is spread and whether it follows a normal distribution, skewed distribution, or has other patterns.

From the histograms we can notice that shape parameters are mostly concentrated around similar values. For the heigh it is especially prominent that most of the cases are clustered around 30 m, there are only a few instances where the height deviates significantly from this value. Similar tendency can be noticed in rock mechanics parameters, especially for Q' values, factors A and C. Addressing this issue might involve considering strategies such as standardization and/or rescaling the data to provide the model with a more varied and informative set of input features. Plotting histograms assisted us in understanding the distribution of these parameters, which is crucial for making informed decisions during the preprocessing stage and optimizing the model's performance.

For further investigation the boxplots, also known as box-and-whisker plots were created (Fig. 4). A boxplot is a graphical representation that displays the distribution and spread of a dataset. It provides a visual summary of the central tendency, spread, and potential outliers in the data. Boxplots offer a concise and informative determination of key statistical measures, making them a valuable tool in the exploratory data analysis phase before applying machine learning algorithms.

Box plots describe a sample by utilizing the 25th, 50th, and 75th percentiles—referred to as the lower quartile ($Q1$), median (m or $Q2$), and upper quartile ($Q3$)—along with the interquartile range ($IQR = Q3-Q1$), encompassing the central 50 % of the data. Quartiles demonstrate resilience to outliers and keep information about both the center and spread. As a result, they are favored over the mean and standard deviation for database with asymmetry or irregular shapes and for samples containing extreme outliers (Krzywinski and Altman, 2014). The median is defined by a line separating the box, marking the midpoint of the dataset. This line indicates that 50 % of the data is surpassing the median. The top of the box plot represents the upper quartile ($Q3$), which means that 25 % of the data exceeds this value, while the lower quartile ($Q1$) is represented at the bottom of the box, where 25 % of the data is less than this value. The top "whisker" illustrates values higher than the median, and outliers are represented by dots above the top "whisker". A similar interpretation applies to the bottom "whisker" and outliers. Box plots can also show the skewness in the dataset, with the position of the median on the box indicating how much data falls above or below it.

Creating individual box plots for each stability assessment

Table 1
Adoko et al. (2022) Database (first 10 examples).

Case number	Height (*H*)	Span (*S*)	Length (*L*)	*Q*'	*A*	*B*	*C*	Stability Asses.
1	30	11.9	30	4.7	1	1	8	STABLE
2	30	30	43	1.5	1	0.5	8	CAVED
3	30	11.9	20	4.7	1	1	1	UNSTABLE
4	30	20	30	1.5	1	0.5	8	UNSTABLE
5	30	20	30	4.7	1	0.8	8	STABLE
6	30	22.3	43	4.7	1	1	1	UNSTABLE
7	30	22.3	30	4.7	1	1	8	STABLE
8	30	30	43	4.7	1	0.8	8	UNSTABLE
9	30	10	20	4.7	1	1	4.7	STABLE
10	30	10	30	4.7	1	1	8	STABLE

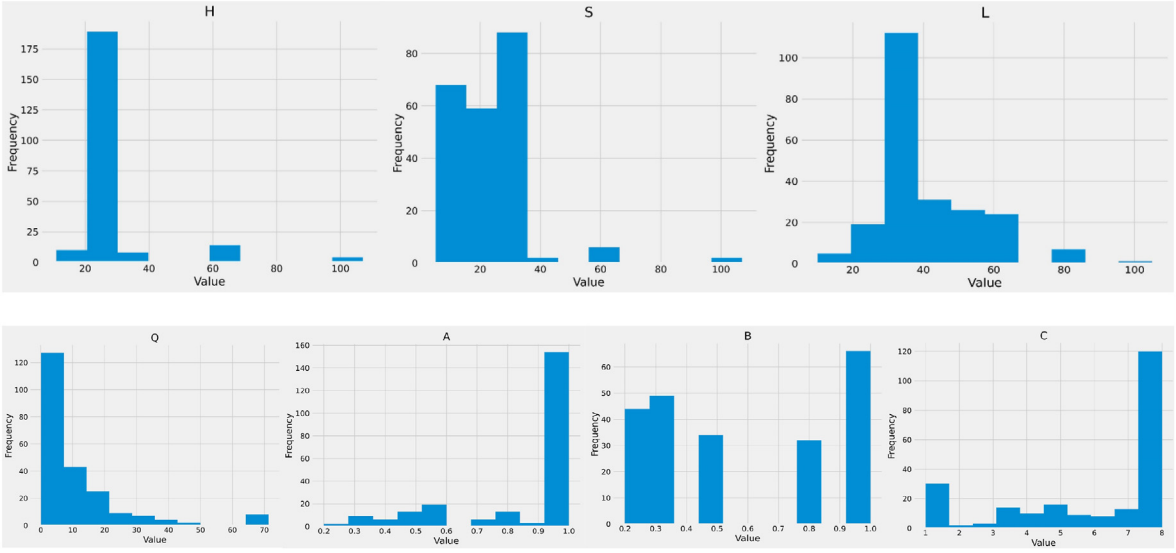

Fig. 3. Histograms of the parameters.

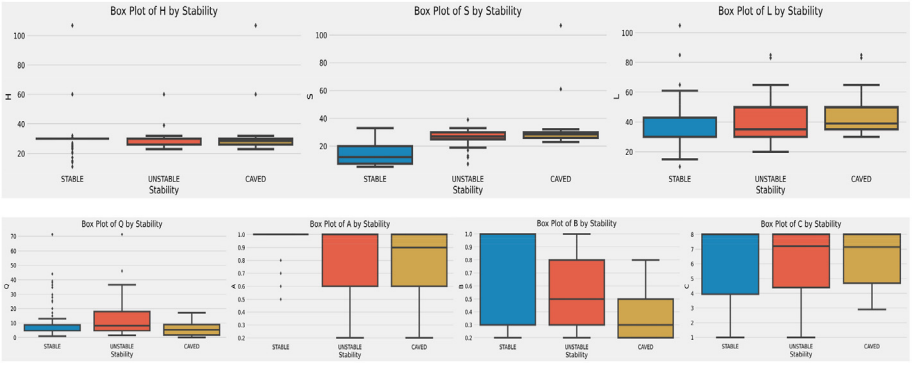

Fig. 4. Box—plots of the parameters.

category—stable, unstable, and caved—offers an extensive visual representation of how the features are distributed within specific classes. This approach enables a detailed analysis of the tendencies and variations in each feature relative to the different stability conditions. Boxplots with extended lengths indicate a greater dispersion of data. Upon closer examination, it becomes evident that factors *A*, *B*, and *C* exhibit a more scattered distribution compared to the shape values and factor *Q*'. However, an exception is observed in the case of factor *A* for stable cases, where a substantial number of instances cluster around similar values, contributing to a distinct pattern within this stability category. This observation shows that, for these specific factors, the data tends to vary widely, especially in unstable and caved scenarios, with an exemption for stable cases of factor *A*, where a more concentrated distribution is apparent. A limited number of outliers are evident, particularly related to shape parameters, within all stability classes. These outliers show instances that noticeably deviate from the range of the central parts of the data. In our dataset, an intentional choice was made to keep these outliers, as they capture natural variations within the dataset. This decision acknowledges that these outliers contribute valuable information and reflect diversity in the data, rather than being treated as anomalies or errors.

The final step in analyzing our database involved the generation of a

correlation heatmap (Fig. 5), a graphical representation that shows the interconnections between all the features. The heatmap provides a comprehensive overview of how each feature relates to every other feature in the dataset. The color in the heatmap signifies the strength and direction of the correlation: red shades denote stronger positive correlations, while darker blue shades indicate weaker or negative correlations.

A higher correlation between features implies a statistical relationship where changes in one feature are associated with changes in another. In some cases, this can lead to redundant information, where one feature might provide similar insights as another. In machine learning, dealing with highly correlated features can be crucial because it may introduce bias into predictive models (Srivastava, 2023). However, in our specific dataset, we observed that the features do not exhibit a strong correlation. This lack of interconnection is a positive sign for the performance of machine learning models. When features are not highly correlated, it suggests that each feature contributes unique information and insights to the model. In such cases, the model is less likely to be influenced by unnecessary or overlapping information, which can lead to more accurate and unbiased predictions.

The presented analysis of the data allowed us to fully understand its structure, complexity, and potential. The fact that the database has three distinct classes: stable, unstable, and caved, implies that we had to specifically look for machine learning models that have ability to manage multiclass classification problems. Another factor to consider is the size of the database, which consists of 225 cases. Small databases often contain a limited range of examples, which may not fully represent the diversity of the real-world scenarios the model is meant to handle. This limited size was one of the reasons we decided not to remove the outliers in our data, as they can provide valuable information and show diversity. The lack of correlation between the features signifies that the model might be less prone to overfitting in the training data and could generalize better to new, unseen data. In order to effectively handle the complexities and challenges posed by our dataset, we made a decision to employ and compare several Machine Learning models. The objective was to assess the performance of each model and determine which one could achieve the highest accuracy and most correct predictions, while being suitable solution for our specific problem. The Machine Learning models investigated are: Random Forest, Support Vector Machine (SVM), AdaBoost, XGBoost, LightGBM, and Artificial Neural Network (ANN).

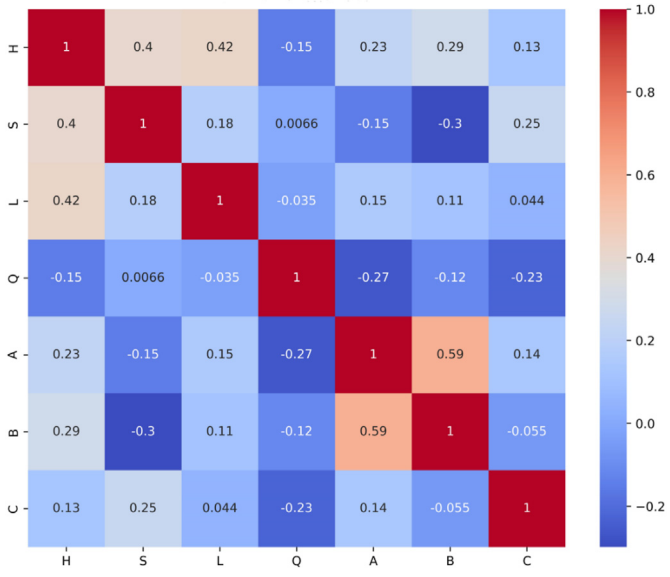

Fig. 5. Correlation heatmap of all parameters.

3.2. Database pre-processing

To prepare our data to be passed to machine learning models, we have undergone a series of processing steps. This data preparation process involves stages aimed at optimizing the dataset's format and structure to ensure compatibility with machine learning algorithms. The goal is to enhance the models' effectiveness by providing them with a well-organized and informative input, enabling them to learn and generalize patterns effectively during the training phase. Pre-processing involves employing various techniques and is widely recognized as a crucial task, establishing a significant portion, potentially up to 80 %, of the overall model development effort (Fan et al., 2021). In practical terms, it is almost always beneficial to apply pre-processing techniques to input data before presenting it to a model (Nawi et al., 2013).

The initial pre-processing technique applied to our data involved feature scaling, also known as standardization. This was specifically necessary because the input variables in our dataset exhibit varying ranges, leading to each feature having a different scale. Such differences across the dataset can pose challenges in developing an accurate model. This need for standardization becomes evident after examining histograms and boxplots, showcasing the diverse value ranges of different features. For instance, factors A and B share a range from 0 to 1, while factor C spans from 2 to 8, the Q' value extends to 70, and shape parameters reach even up to 100. It's essential to note that scale relates to the variation in the value ranges, not the distribution shape. Standardization is essential for ensuring uniformity in scale across features, crucial in the proper development of the model.

For our database we decided to employ StandardScaler tool as a standardization method, which follows standard normal distribution. StandardScaler operates by expecting the data to be normally distributed within each feature and subsequently scales these values to center the distribution around 0, with a standard deviation of 1 (Raju et al., 2020). The process involves determining the mean and standard deviation for each variable. Subsequently, the scaled feature is calculated by adjusting each value based on these statistics. This standardized approach ensures that the features have consistent scales, facilitating a more uniform and effective analysis of the data during the modeling process. The StandardScaler is considered the best option as it effectively handles data with varying scales and ensures that the machine learning algorithms perform optimally, preventing issues such as biased weighting of features (Han et al., 2012).

The Machine Learning models under consideration required numerical values for all input variables. However, the feature labels in our database were categorical, classified as stable, unstable, and caved. To ensure optimal Machine Learning performance and obtain reliable outputs, a crucial step involved converting these categorical labels into numerical values. This transformation was executed as follows: stable was assigned the numerical value 1, unstable was assigned 2, and caved was assigned 0. This categorization enables the models to effectively process and interpret the labels during the learning process, ensuring compatibility with the numerical expectations of the machine learning algorithms.

Another important step in the stage of data preparation is splitting it into training, testing and validation sets. The primary reason for splitting the database is for evaluation of the performance of the model (Hastie et al., 2017). The model adjusts its internal parameters during the training process to minimize the difference between its predicted outputs and the actual values in the training set (Birba, 2020). The validation set is used to fine-tune the model during the training phase and make decisions about its architecture and hyperparameters. The validation set provides an unbiased evaluation which helps to detect and reduce overfitting (Xu and Goodacre, 2018). Finally, the testing set serves as an independent subset of the dataset to assess how well the model generalizes to new, unseen data. The testing set is used to calculate various performance metrics which provide informative measures of the model's effectiveness in making predictions (Jain et al., 2022).

For our specific problem, we decided to randomly split the dataset into those mentioned before three subsets, where 80 % of the samples was allocated as training, 10 % as validation and 10 % as testing set. We applied the same splitting for all the evaluated machine learning models to maintain the consistency and ensure fairness in the comparative analysis of the models.

In summary, investigating the data is essential to ensure data quality and make decisions and setting the foundation for building effective and accurate machine learning models. By plotting histograms (Fig. 3), we identified the distribution of data points across various features, revealing that shape parameters are mostly concentrated around similar values, particularly height, which clusters around 30 m. Rock mechanics parameters, such as Q' values and factors A and C, also show similar tendencies. This observation suggests the need for strategies like standardization to provide the model with a more varied and informative set of input features. Further analysis using boxplots (Fig. 4) offered insights into the central tendency, spread, and potential outliers within the dataset, highlighting that factors A, B, and C exhibit greater dispersion compared to shape values and Q'. A correlation heatmap (Fig. 5) showed that features in our dataset do not exhibit strong correlations, indicating that each feature contributes unique information to the model. This detailed analysis underscores the importance of employing and comparing multiple machine learning models to handle our multiclass classification problem effectively and to determine which model performs best given the dataset's structure and complexity. Fig. 6 presents all the pre-processing steps for machine learning workflow.

4. Comparative study overview

In the scope of our study, our objective was to explore and compare a selection of the most widely adopted machine learning models that have demonstrated effectiveness in addressing problems similar to ours. We specifically looked for models with established success in handling multiclass classification tasks, aligning with the nature of our problem where instances needed to be categorized into one of three classes (stable, unstable, caved). An important consideration was the suitability of the chosen models for problems without significant correlation between the features. We wanted to ensure that they could effectively capture diverse relationships. The decision-making process also involved taking into account the number of instances in our dataset, which consists of 225 cases. Additionally, given the supervised nature of our problem, where each case is labeled with a known class, we focused on machine learning models designed for supervised learning.

Conducting a comparative study and examining different machine learning models offers several advantages and insights that contribute to the effectiveness of the overall analysis. Different models have unique strengths and weaknesses. A comparative study allows for an evaluation of their performances, helping to identify which models are more suitable and effective for the given problem. Some problems may be inherently complex, and certain models may handle it better than others. Comparative studies assist in determining which models are better equipped to deal with such relationships and patterns within the data. Different models may also emphasize different features in making predictions. Comparing different ML algorithms can shed light on which features are considered most important across various models, providing valuable insights into the factors influencing the predictions. Relying on a single model may create risks, especially if that model is not well-suited to the specific characteristics of the data. A comparative study addresses this risk by exploring a range of models and offering a more comprehensive understanding of their performance.

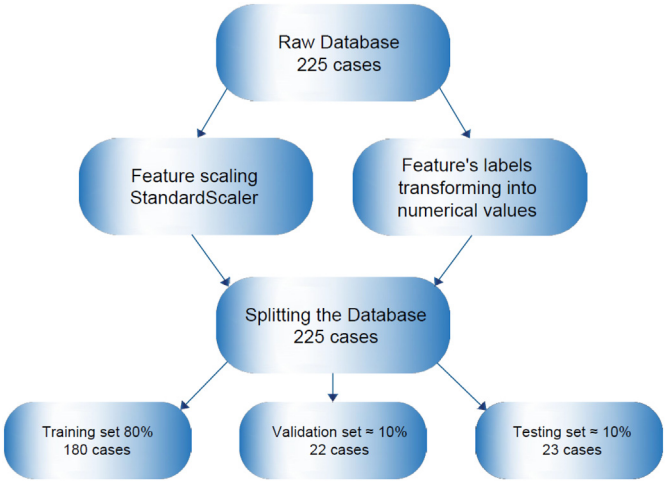

Fig. 6. Pre-processing workflow for the database preparation.

4.1. Investigated machine learning models

To conduct this comparative study the following machine learning models were investigated: Random Forest, Support Vector Machine (SVM), AdaBoost, XGBoost, LightGBM and Artificial Neural Network (ANN). The inclusion of these diverse models ensures a thorough examination, offering insights into their respective strengths and weaknesses in the context of our study.

4.1.1. Random forest

Random Forest is a powerful and versatile machine learning model known for its satisfactory performance across a wide range of tasks. It belongs to the ensemble learning family, where multiple decision trees are combined to form a more flexible and accurate model. The strength of Random Forest lies in its ability to reduce overfitting, handle complex relationships in data, and provide insightful feature importance rankings. Breiman's (2001) innovative work on Random Forests laid the foundation for this machine learning technique, and subsequent research and practical applications have endorsed its significance in the machine learning field.

By combining the predictions from number of individual trees, each trained on a random part of the data, the model achieves a higher degree of accuracy and generalizes well. Random Forest has demonstrated excellence in both classification and regression problems, making it a popular choice in many research fields, including mining engineering. Several researchers have successfully applied RF in their study focused around area of stoping mining methods and stability predictions. Qi et al. (2018a) conducted a study to predict the stability of hanging wall with random forest, Szmigiel and Apel (2022) investigated the feasibility of employing random forest to asses the stability of open stopes, and Jorquera et al. (2023) applied random forest to predict the dilution in sublevel stoping mining method. Random forest have been also successfully applied in related mining engineering disciplines, such as predicting mining induced stresses in underground openings (Vinay et al., 2023), rock pillar stability (Zhou et al., 2015) or ground settlement predictions (Zhou et al., 2017).

These successful applications have built our confidence in including random forest model in our study, to observe its performance on our specific data compared with other models, as well as determine the effects of particular features on its predictions.

4.1.2. Support vector machine

The Support Vector Machine (SVM) is a powerful machine learning model known for its efficacy in classification tasks, particularly excelling in scenarios with high-dimensional data. Developed by Cortes and Vapnik, SVM constructs an optimal hyperplane that efficiently separates different classes in the feature space, maximizing the margin between them (Cortes and Vapnik, 1995).

SVM's has an ability to handle complex decision boundaries, making it well-suited for non-linear classification tasks. It is possible due to the

use of kernel functions that transform input data into higher-dimensional spaces. This is especially advantageous attribute, that should effectively handle our database which exhibits a low correlation index between the features. Importantly, SVM proved to be effective in multi-class classification problems by employing strategies like one-vs-one or one-vs-all, demonstrating adaptability and satisfying performance across diverse applications (Evgeniou and Pontil, 2001). Some applications of SVM model in the area of mining engineering focus on open stopes, Jorquera et al. (2023). study about predicting dilution in sublevel stoping evaluates SVM along with other models, similarly few comparative open stope stability studies conducted (Erdogan Erten et al., 2021; Qi et al., 2018b). However most of the research with SVM models focuses around different mining engineering challenges, such as predicting the stability of hard coal pillars (Li et al., 2023), slope stability analysis (Samui, 2008), modeling displacement time series of geomaterials (Feng et al., 2004), or predictions of mining subsidence (Li et al., 2011). All these mentioned advantages of SVM model, as well as not that many applications in the area of stope stability, convinced us in evaluating this particular machine learning technique along others.

4.1.3. Adaptive Boosting—AdaBoost

Adaptive Boosting, is an ensemble learning algorithm introduced by Freund and Schapire (1996). This effective method operates by combining the predictions of weak learners, typically decision trees, to form a powerful and accurate classifier. AdaBoost assigns higher weights to misclassified examples, forcing following weak learners to focus on these challenging cases. The final prediction is then determined through a weighted majority vote of the individual weak learners. AdaBoost's adaptability and capability to handle complex relationships in data, and its emphasis on learning from misclassifications and continuously elevating its prediction, make it particularly suitable for multiclass classification problems and uncorrelated features (Wang, 2012). Numerous application of AdaBoost algorithm can be found in the area of rockbursts predictions in mining engineering (Ahmad et al., 2022; Wang et al., 2023; Li et al., 2022), but not many applications of this particular model can be found in the area of stope stability assessment, except Saadaari et al. (2020) study. Given that this particular model isn't popular in mining stoping method applications, although it has demonstrated efficiency in addressing similar challenges in engineering, we made a decision to incorporate and assess its performance in our study.

4.1.4. Extreme Gradient Boosting—XGBoost

Extreme Gradient Boosting, is another powerful and efficient machine learning algorithm introduced by Chen and Guestrin (2016). In comparison to AdaBoost, XGBoost stands out for its scalability, fast processing, and ability to handle extensive datasets. While both are an ensemble learning methods that combine weak learners to create a successful model, XGBoost employs a more advanced optimization approach, incorporating regularization terms and parallel processing to improve performance. XGBoost is well-suited for multiclass classification, it implements a one-vs-all strategy, creating individual classifiers for each class and then combining their outputs. The algorithm's ability to capture complex relationships between the features and effectively manage them, makes it an excellent candidate to include in our study. Some mining engineering applications of XGBoost can be found in literature, including subsidence predictions (Gu et al., 2022), rock fragmentation and ground vibration predictions in blasting operations (Chandrahas et al., 2022) or hard rock pillar and underground entry-type excavations stability predictions (Liang et al., 2020; Zhou et al., 2023).

4.1.5. LightGBM

LightGBM, model developed by Microsoft, is a gradient boosting framework designed for distributed and efficient training on large datasets. Compared to other gradient boosting algorithms like AdaBoost and XGBoost, LightGBM employs a histogram-based approach, accumulating data points to accelerate the training process. Its leaf-wise growth strategy focuses on minimizing loss, resulting in a more accurate and adaptive model. LightGBM stands out for its ability to handle categorical features efficiently and it is suitable for multiclass classification problems, employing, similarly to other models, a one-vs-all strategy (Ke et al., 2017). The area of mining engineering research doesn't show many LightGBM model applications, except study on predicting hard rock pillar stability (Liang et al., 2020) and mineral grade estimation (Kaplan et al., 2021).

Given its ability to handle complex relationships in data, LightGBM is a compelling choice for our study, where predicting the stability of open stopes involves dealing with diverse and complex patterns within the dataset.

4.1.6. Artificial neural network

The last model that we have decided to employ in our study is a popular widely used artificial neural network. This effective machine learning model is inspired by the structure and functioning of the human brain and it's built of interconnected nodes organized into layers. Unlike ensemble methods such as presented before Radom Forest, AdaBoost, XGBoost, and LightGBM, ANN operates on a fundamentally distinct architecture, leveraging connected neurons to capture complex, non-linear patterns within the input data. While boosting models excel in combining weak learners to form a strong classifier, ANN stands out for its capacity to automatically learn hierarchical representations of features. Unlike decision tree-based models, ANN does not rely on predetermined splits but dynamically adjust weights during training, making them well-suited for capturing challenging relationships. Including ANN in our comparative study offers a comprehensive exploration of diverse modeling approaches, ensuring a thorough understanding of how different models handle the challenge of predicting the stability of open stopes. There is numerous research focusing on ANN applications in mining engineering, with stability of stopes included. These include presented by Wang et al. (2002) application of ANN model in designing underground excavation spans or a feasibility study conducted by Adoko et al. (2022) on implementing ANN classifiers for open stope design.

5. Evaluation and results

5.1. Training and evaluation

The first model subjected to training and testing on our dataset was the Random Forest (RF). RF offers an adaptable framework by allowing the adjustment of hyperparameters to optimize model performance and address overfitting concerns. Two crucial hyperparameters in RF are the number of estimators (decision trees), and the maximum depth of each individual tree. Calculating accuracy for validation and training set allowed us to notice and address any overfitting problems displayed by the RF model by adjusting those parameters. When the accuracy for training set is much higher compared to validation, the model is learning too well from the training set and struggles to properly generalize for new unseen data. When the number of decision trees and the depth of each tree were set to higher values, the accuracy for training set was reaching close to 100 %, while the validation accuracy was dropping significantly. Thus, the optimal number of decision trees was set to 100, with the depth of each tree equal to 5. This hyperparameter setting achieved the accuracy of training and validation set equal to 88 % and 90 % respectively.

Second model that was trained was Support Vector Machine. The SVM model has two parameters, C and Gamma that serve as adjustable hyperparameters. The C parameter, often referred to as the regularization parameter, impacts the trade-off between achieving a smoother decision boundary and correctly classifying training points. A smaller C value results in a softer margin, allowing for more points to be classified correctly but potentially leading to overfitting, while a larger C value enforces a stricter margin, prioritizing a simpler separation between classes but potentially sacrificing the correct classification of some training points. The gamma parameter is associated with the radial basis

function (RBF) kernel, a common choice in SVMs models for handling non-linear relationships in data. A smaller gamma value results in a wider Gaussian kernel, leading to smoother decision boundaries and potentially underfitting the training data. In contrast, a larger gamma value narrows the Gaussian kernel, allowing the model to capture more complex patterns in the data but possibly leading to overfitting (Liu et al., 2006). Grid search approach was performed to find the optimal combination of C and gamma values for a SVM model. Grid search is an optimization technique that systematically searches through a predefined set of hyperparameter values for a model. It involves evaluating the effectiveness for each combination of C and gamma values using a cross-validation, in order to find the ones that result in the best performance without overfitting.

The same grid search method was uniformly employed across all ensemble learning algorithms—AdaBoost, XGBoost, and LightGBM—during the training phase to tune their hyperparameters. Specifically, the parameters under investigation were the number of estimators (representing weak learners or trees) and the learning rate, which controls the contribution of each weak learner to the final combined model. Cross-validation was performed to find the most efficient values of those parameters, ensuring the most accurate results while avoiding overfitting.

In the final step of the modeling process, an Artificial Neural Network (ANN) was built and tailored to our dataset. The exploration of various ANN model structures was performed, with an evaluation on the validation set. The finalized architecture consists of four layers: an input layer, two hidden layers, and an output layer. The input layer's configuration aligns directly with the number of features in the dataset. Subsequently, the first and second hidden layers consist of 32 and 16 neurons, respectively and the output layer's structure is determined by the number of classes within the dataset. Activation functions played a crucial role in shaping the model, with the first three layers employing the swish activation function, known for its non-linearity and smoothness. The output layer, responsible for producing the final predictions, employed the softmax function to ensure appropriate class probabilities. The optimization of weights during each epoch and the minimization of the algorithm's loss were facilitated by the Adam optimizer. Categorical cross-entropy was adopted as the loss function, aligning with the multiclass nature of the problem, contributing to the model's ability to determine and classify instances accurately.

5.2. Results

All the investigated Machine Learning models were assessed with the same evaluation metrics, to demonstrate a clear contrast between their results and conduct a proper comparative analysis. The most popular and basic metric is the accuracy score, which can be calculated for both validation and testing sets (Fig. 7). Upon the examination of the model performance, it becomes clear that the Artificial Neural Network (ANN) stands out as the leading performer, achieving a great accuracy score of 91 % on the testing set. Following closely is the Random Forest model, demonstrating an accuracy score of 86 %. Notably, both the Support Vector Machine and LightGBM classifiers showed identical accuracy scores for both the validation and testing sets. However, it is crucial to highlight that, at this point of evaluation, the Adaptive Boosting classifier appeared as the least effective, displaying the lowest accuracy among all the models assessed.

In the next step of our analysis the averaged classification report metrics such as precision, recall and F-1 score were determined and compared (Fig. 8). This provides aggregated performance evaluation across all classes, presents summary of how well the model performs on average. These metrics provide a single numerical value that represents the overall performance, simplifying the interpretation of results. Precision measures the accuracy of positive predictions. It calculates the ratio of true positive predictions to the total predicted positives, high precision values indicate fewer false positives predictions. Recall measures calculate the ratio of true positive predictions to the total actual positives, higher recall values indicate fewer false negatives predictions. The F1 score is a harmonic mean of precision and recall that combines them into a single value, higher F1 score is generally desirable because it indicates a better trade-off between these two metrics. These metrics collectively provide a comprehensive understanding of a model's strengths and weaknesses, aiding in the selection of the best-performing model for a given task (Powers, 2008; Sokolova and Lapalme, 2009). Similarly as with accuracy, the ANN model achieved the highest score for all three metrics, while RF's results declined for recall and F-1 score. Among all the models, AdaBoost is again showing the lowest results, therefore it exhibits poor predicting capabilities for stope stability data. The significantly better performance of ANN can be caused by its ability to capture complex, non-linear relationships and handle multiclass classification tasks effectively. In contrast, AdaBoost struggled because it is less suited for such complexity, is sensitive to noisy data, and lacks the flexibility in feature representation and parameter tuning that ANN offers.

In the context of stope stability predictions, precision appears as a crucial metric, especially in the case of stable class, as it directly addresses the accuracy of positive predictions, holding particular significance when the cost associated with false positives is high. These false positive predictions could potentially lead to significant safety concerns. When an unstable or caved condition is incorrectly classified as stable, it creates a major risk to the safety and integrity of the mining environment. Precision in this context becomes the crucial metric because it emphasizes the correctness of the model in identifying actual conditions. The precision scores across all classes are shown on Fig. 9. The recall score on the other hand (Fig. 10), has a significant importance in the case of caved predictions. When the recall score for caved class declines, it means that there are instances that are incorrectly classified as stable or unstable. This misprediction are also dangerous as they are causing some potentially dangerous conditions to be presented as either unstable or stable.

When analyzing the precision and recall graphs, a crucial realization

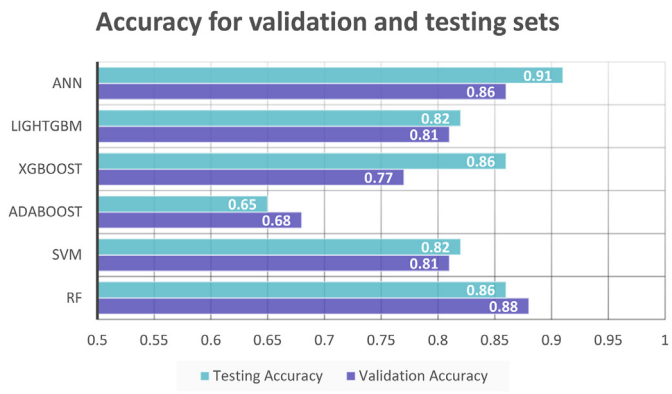

Fig. 7. Accuracy score for all the Machine Learning models.

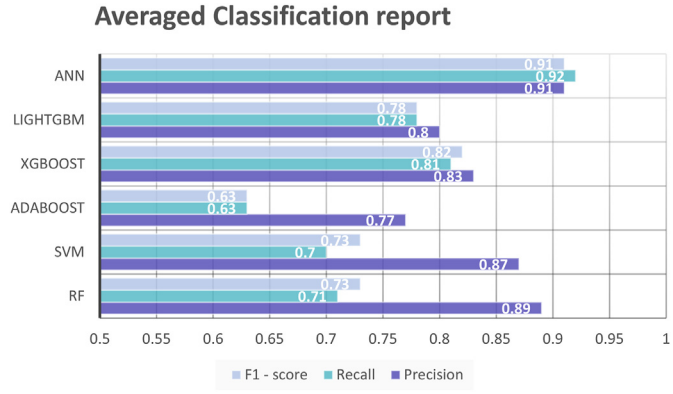

Fig. 8. Classification report for all the Machine Learning models.

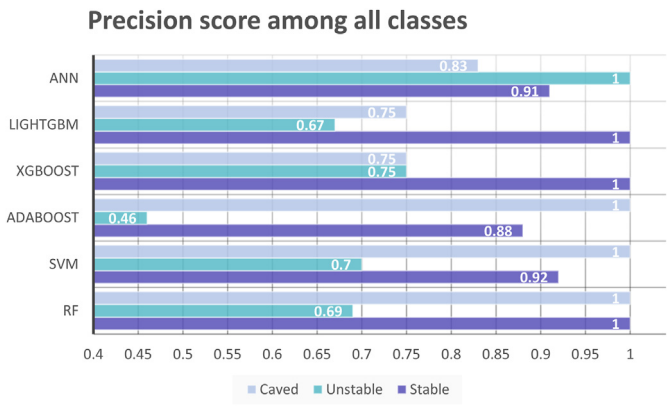

Fig. 9. Precision scores for all Machine Learning models.

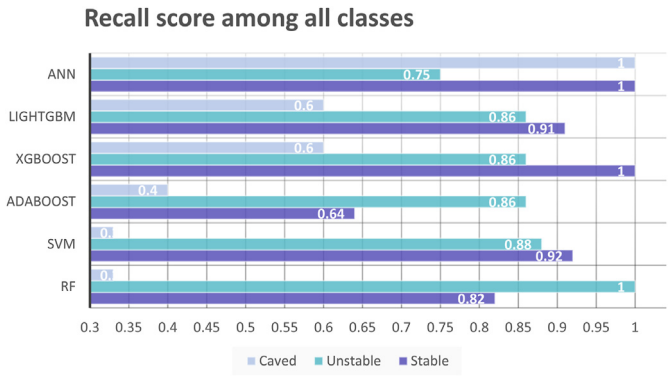

Fig. 10. Recall scores for all Machine Learning models.

was evident, emphasizing the limitations of the Random Forest model despite its apparent efficiency based on overall accuracy. While accuracy is a comprehensive metric, a more careful examination exposed a critical concern related to the recall score for the caved class. The recall value of 0.33, indicated a notable deficiency in the RF model's ability to accurately identify instances of caved conditions. This score showed that a substantial portion of the actual caved cases were being misclassified as either stable or unstable. Such misclassifications create a significant risk to mine safety, as they imply that potentially hazardous conditions associated with caved areas are not being properly identified.

The Artificial Neural Network once again demonstrated its exceptional capability in producing the best classification results across all classes, standing out in comparison to all other models evaluated in this study. The ANN achieved a perfect recall score of 1 for both stable and unstable cases, indicating that none of these cases were incorrectly assigned to a different class. However, a challenge appears in the classification of unstable cases, which represent conditions that are between stable and caved, posing potential risks in a mining environment. The precision score of 0.82 for caved class and 0.91 for stable class, means that some of those unstable cases were assigned to one of those two classes. However, the lower precision for the caved class indicates that a larger proportion of these unstable cases was classified as caved. This aligns with a safer approach, as misclassifying unstable conditions as caved is a more reasonable and careful choice in terms of mine safety.

Artificial Neural Networks demonstrated superior performance across all evaluation metrics in our study due to their inherent advantages in handling complex and nonlinear relationships within the data. ANNs are capable of capturing intricate patterns and dependencies through multiple layers of neurons and non-linear activation functions, which enables them to learn and generalize from the data more effectively than simpler models. Their ability to adjust weights through backpropagation allows them to optimize the model for accuracy, precision and recall scores.

5.3. *Feature importance analysis with SHAP*

In the context of our study on open stope stability, conducting feature importance analysis is important for a comprehensive understanding of the factors influencing stability outcomes. Feature importance analysis allows us to determine the relative contribution of each parameter to the predictive performance of our machine learning models. By identifying the most influential features, we gain insights into the key determinants of stope stability, enabling engineers to prioritize and focus their attention on critical factors during the design and planning stages. This knowledge not only enhances the interpretability of the model but also assists in making informed decision-making in real-world mining scenarios. Additionally, feature importance analysis helps in the optimization of data collection efforts, guiding engineers to gather more detailed information on crucial parameters. Ultimately, this process regulates the machine learning model with domain-specific knowledge, ensuring that the predictions are not only accurate but also reflect the nuances of open stope stability and safety assessments.

To perform this feature analysis a SHapley Additive exPlanations (SHAP) analysis was performed for all machine learning models investigated in this study. SHAP is a framework for explaining the output of machine learning models outcomes. It assigns a value (Shapley value) to each feature, indicating its contribution to the model's prediction (Lundberg and Lee, 2017). In the context of our study on open stope stability assesment, performing SHAP analysis is crucial for several reasons. It provides interpretability to complex machine learning models, helping us understand how each input feature influences the stability predictions. This interpretability is essential for gaining insights into the driving factors behind stability conditions. The information from SHAP analysis is invaluable for improving stope design strategies and focusing engineering efforts on the aspects that have the greatest influence on stability. Moreover, SHAP analysis aids in the validation and verification of the model's predictions. By understanding the rationale behind each prediction, engineers can assess the model's reliability and identify potential areas of improvement in both the model and the underlying data. SHAP analysis enhances the transparency, interpretability, and reliability of our machine learning model, making it an essential step in ensuring the practical applicability of the model's predictions in real-world mining scenarios.

SHAP value summary plots for RF, SVM, AdaBoost, XGBoost, LightGBM and ANN are shown on Figs. 11–16. The chart's horizontal axis represents SHAP values, while the vertical axis displays all the features in our data, with the most influential one being on top. Each point on the chart corresponds to a SHAP value for a prediction and a feature. Red indicates a higher feature value, while blue indicates a lower one. By observing the distribution of the red and blue dots, we can gain a general understanding of the impact of feature directionality. The positive SHAP values (horizontal axis) are indicating a more positive impact on the

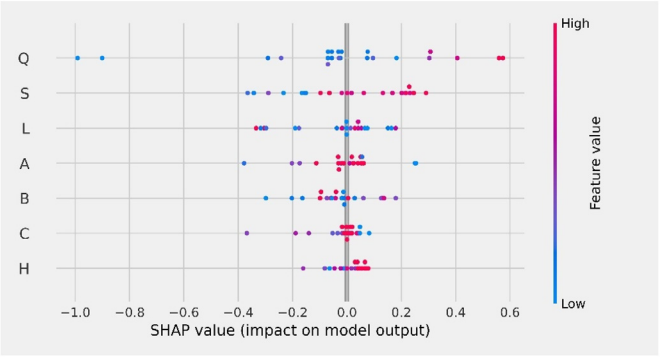

Fig. 11. Random Forest—SHAP value summary plot.

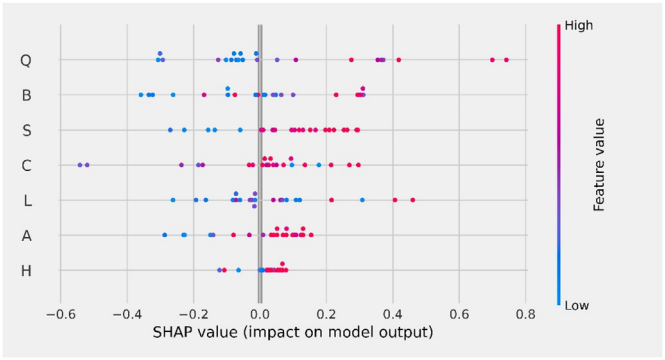

Fig. 12. SVM—SHAP value summary plot.

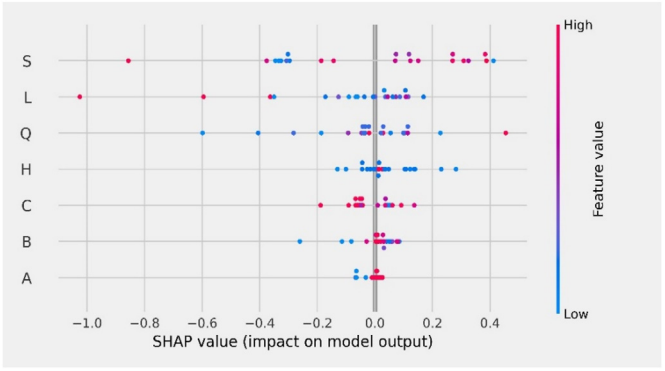

Fig. 13. AdaBoost—SHAP value summary plot.

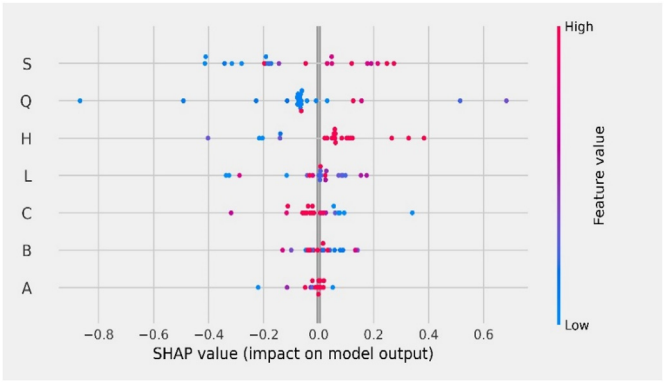

Fig. 14. XGBoost—SHAP value summary plot.

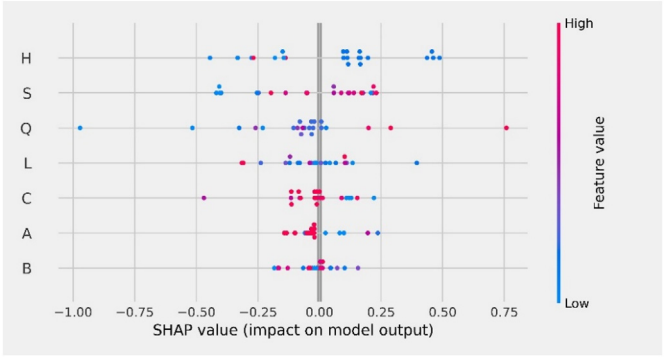

Fig. 15. LightGBM—SHAP value summary plot.

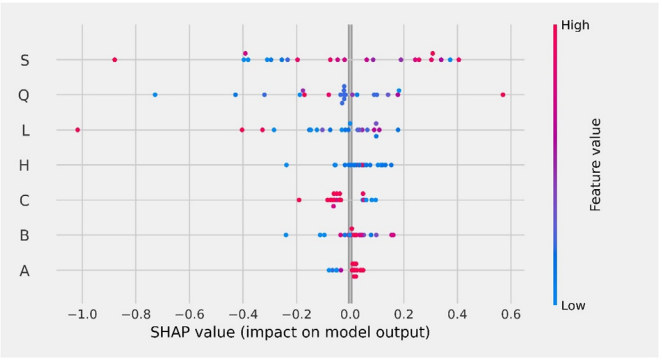

Fig. 16. ANN model—SHAP value summary plot.

model, in our case those positive values mean that there is higher probability of predicting that case into stable or unstable class. However, the more negative the SHAP value, the higher the likelihood that the model will assign the case to the unstable class.

In the case of a RF model, the *Q* value has the most influence on the final prediction, followed by a shape value *S* which is the span of the opening and then *L* (length). The higher value of the *Q* factor, the higher probability that the stope will have a stable condition, similar trend can be noticed with the *S* factor. In case of the SVM model, it can be noticed that *Q* value is also the most influential, with the higher values having more positive impact on the stability of the stope. Surprisingly for the SVM model, the *B* factor has a second highest impact, where in other models its influence is less significant. For other models, the shape parameters, especially span, have the most significant impact. These results are good for a mine designing stage, as it is usually more feasible to adjust the size of the opening (such as a stope length or span) than to change the overall surrounding rock conditions. Engineers have greater control over the dimensions of the opening and can adapt them to suit specific requirements, such as equipment accommodation and stability considerations. Therefore, mine design prioritizes optimizing the size and shape of openings while working within the limitations of the surrounding rock conditions. However models such as AdaBoost, XGBoost and LightGBM, that have prioritized the shape of the opening in their predictions, are also models that exhibited less accurate overall performances. These models also have a lack of consistency when it comes to the value of the feature impacting the predictions, with some having higher values influencing the model positively and others doing the opposite.

The Artificial Neural Network (ANN) model exhibited the best performance in terms of predictive capabilities, consistently delivering high classification scores across all stability classes. Upon closer examination of the feature importance analysis, it became apparent that the span factor (*S*) exhibited a high degree of influence on the model's predictions. Despite being the most influential factor, there was a notable lack of consistency in the impact of higher span values on stability predictions. While higher span values generally appeared to have a positive influence on stability predictions, there were instances where this trend was not maintained. In fact, there was one case where a high span value resulted in a significantly negative impact on stability. This inconsistency in the influence of span highlights the complexity of the relationship between this factor and stope stability.

Since the variability observed in the influence of span on stability predictions is not consistent, it is advisable to prioritize the second most influential factor, which is the rock quality factor (*Q*). The analysis revealed a clearer trend with *Q*, where lower values consistently exhibited either highly negative or neutral influences on stability predictions. This reliability in the relationship between *Q* and stability suggests that it may be a more dependable factor to consider when making predictions. Additionally, the length factor (*L*) was observed to have a more consistently negative impact on stability predictions for larger values. Although not as influential as factors *S* or *Q*, the consistent

trend with length reinforces its importance in stope stability assessment.

6. Conclusions

In this study, we explored the application of various machine learning models to predict the stability of open stopes in underground mining operations. The investigation aimed to compare the performance of different models and provide insights into their effectiveness in addressing the complex problem of stope stability assessment.

Upon evaluating the performance of the investigated machine learning models, several key findings emerged. The Artificial Neural Network (ANN) demonstrated exceptional predictive capabilities, outperforming all other models with an accuracy score of 91 % on the testing set. This result underscores the effectiveness of ANN in accurately categorizing instances into stable, unstable, or caved classes. Following closely behind, the Random Forest model exhibited promising performance, achieving an accuracy score of 86 %. Despite its lower accuracy compared to ANN, Random Forest demonstrated competitive results, highlighting its potential as a reliable predictive tool for stope stability assessment. In contrast, the Support Vector Machine (SVM), LightGBM, and AdaBoost classifiers displayed comparable accuracy scores, although lower than ANN and Random Forest. Notably, the Adaptive Boosting classifier exhibited the least effective performance, demonstrating the lowest accuracy among all models assessed.

Precision and recall metrics provided further insights into the performance of the models in the context of stope stability predictions. Precision, which measures the accuracy of positive predictions, emerged as a crucial metric, especially for the stable class. The ANN model demonstrated high precision scores across all classes, indicating its capability to accurately identify stable conditions with minimal false positives. The recall score for the caved class emerged as a significant concern, particularly for the Random Forest model. Despite its high overall accuracy, Random Forest exhibited a notable deficiency in accurately identifying instances of caved conditions. This limitation poses significant safety risks in mining environments.

The SHapley Additive exPlanations (SHAP) analysis provided valuable insights into the factors influencing the stability predictions of each machine learning model. For instance, the Q value emerged as the most influential factor in the predictions of the Random Forest and SVM models, emphasizing the importance of rock mass quality in determining stope stability. Interestingly, models such as AdaBoost, XGBoost, and LightGBM prioritized the shape parameters, particularly span, in their predictions. While this prioritization aligns with mine design principles, these models exhibited less accurate overall performances. The ANN model exhibited factor S as the most influential on the model predictions. However, in that case, there is also lack of consistency for the higher values, which seem to have rather positive influence on the stability except few cases. In the case of ANN it is better to look at the second most influential factor which is Q, where it is more clear that lower values have either highly negative or neutral influence on the stability predictions. It is also worth noticing that the factor L, has clearly more negative impact for the larger values. Due to the inconsistency of factor S, the analysis suggests relying on the second most influential factor Q, for more accurate stability predictions.

Despite the promising results obtained in this study, several paths for future research and improvement exist. Further exploration of the dataset and refinement of the feature selection process could enhance the predictive capabilities of the models. Additionally, investigating ensemble methods and hybrid approaches combining machine learning with classical stability assessment methods may yield even more accurate predictive models.

In conclusion, this study explores the potential of machine learning techniques in addressing the complex problem of stope stability assessment. By leveraging advanced algorithms and extensive data analysis, mining engineers can make informed decisions to enhance safety and optimize the design of underground excavations.

CRediT authorship contribution statement

Alicja Szmigiel: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation. **Derek B. Apel:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization. **Yashar Pourrahimian:** Writing – review & editing, Supervision. **Hassan Dehghanpour:** Writing – review & editing. **Yuanyuan Pu:** Writing – review & editing, Data curation.

Ethical statement

The authors state that the research was conducted according to ethical standards.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements and Funding body

The Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant is financially supported by this project: NSERC RGPIN-2019-04572 Apel. The authors are grateful for their support.

References

Adoko, A.C., Saadaari, F., Mireku-Gyimah, D., Imashev, A., 2022. A feasibility study on the implementation of neural network classifiers for open stope design. Geotech. Geol. Eng. 40 (2), 677–696. https://doi.org/10.1007/s10706-021-01915-8.

Ahmad, M., Katman, H.Y., Al-Mansob, R.A., Ahmad, F., Safdar, M., Alguno, A.C., 2022. Prediction of rockburst intensity grade in deep underground excavation using adaptive boosting classifier. Complexity 1–10. https://doi.org/10.1155/2022/6156210.

Barton, N., Lien, R., Lunde, J., 1974. Engineering classification of rock masses for the design of tunnel support. Rock Mech. 6 (4), 189–236. https://doi.org/10.1007/BF01239496.

Bieniawski, Z.T., 1973. Engineering Classification of Jointed Rock Masses, vol. 15. Transaction of the South African Institution of Civil Engineers, pp. 335–344.

Birba, D.E., 2020. A Comparative study of data splitting algorithms for machine learning model selection, Computer and Information Sciences. KTH, School of Elect. Eng. Comp. Sci. (EECS).

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Capes, G.W., 2009. Open Stope Hangingwall Design Based on General and Detailed Data Collection in Unfavourable Hangingwall Conditions. The University of Saskatchewan, Canada. NR62618 Ph.D.

Chandrahas, N.S., Choudhary, B.S., Teja, M.V., Venkataramayya, M.S., Prasad, N.S.R.K., 2022. XG boost algorithm to simultaneous prediction of rock fragmentation and induced ground vibration using unique blast data. Appl. Sci. 12 (10), 5269. https://doi.org/10.3390/app12105269.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. https://doi.org/10.48550/ARXIV.1603.02754.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. https://doi.org/10.1007/BF00994018.

Crone, S.F., Lessmann, S., Stahlbock, R., 2006. The impact of preprocessing on data mining: an evaluation of classifier sensitivity in direct marketing. Eur. J. Oper. Res. 173 (3), 781–800. https://doi.org/10.1016/j.ejor.2005.07.023.

Diederichs, M.S., Kaiser, P.K., 1999. Tensile strength and abutment relaxation as failure control mechanisms in underground excavations. Int. J. Rock Mech. Min. Sci. 36 (1), 69–96. https://doi.org/10.1016/S0148-9062(98)00179-X.

Erdogan Erten, G., Bozkurt Keser, S., Yavuz, M., 2021. Grid search optimised artificial neural network for open stope stability prediction. Int. J. Min. Reclamat. Environ. 35 (8), 600–617. https://doi.org/10.1080/17480930.2021.1899404.

Evgeniou, T., Pontil, M., 2001. Support vector machines: theory and applications. In: Paliouras, G., Karkaletsis, V., Spyropoulos, C.D. (Eds.), Machine Learning and its Applications. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp. 249–257.

Fan, C., Chen, M., Wang, X., Wang, J., Huang, B., 2021. A review on data preprocessing techniques toward efficient and reliable knowledge Discovery from building operational data. Front. Energy Res. 9, 652801. https://doi.org/10.3389/fenrg.2021.652801.

Feng, X.-T., Zhao, H., Li, S., 2004. Modeling non-linear displacement time series of geomaterials using evolutionary support vector machines. Int. J. Rock Mech. Min. Sci. 41 (7), 1087–1107. https://doi.org/10.1016/j.ijrmms.2004.04.003.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. In: International Conference on Machine Learning.

Gu, Z., Cao, M., Wang, C., Yu, N., Qing, H., 2022. Research on mining maximum subsidence prediction based on genetic algorithm combined with XGBoost model. Sustainability 14 (16), 10421. https://doi.org/10.3390/su141610421.
Hadjigeorgiou, J., Leclair, J., Potvin, Y., 1995. An update of the stability graph method for open stope design. CIM Rock Mechanics and Strata Control Session. Nova Scotia, Halifax, pp. 14–18.
Han, J., Kamber, M., Pei, J., 2012. In: Data mining: concepts and techniques, 3rd ed. Elsevier/Morgan Kaufmann, Amsterdam.
Hastie, T., Tibshirani, R., Friedman, J.H., 2017. In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed. Springer, New York, NY. corrected at 12th printing 2017.
Henning, J.G., Mitri, H.S., 2007. Numerical modelling of ore dilution in blasthole stoping. Int. J. Rock Mech. Min. Sci. 44 (5), 692–703. https://doi.org/10.1016/j.ijrmms.2006.11.002.
Hu, H., Cao, Y., 2009. Numerical simulation modeling and calculation analysis on stope roof stability under the complex geological conditions in deep mining. In: 2009 International Conference on Engineering Computation. IEEE, Hong Kong, China, pp. 175–177.
Jain, E., Neeraja, J., Banerjee, B., Ghosh, P., 2022. A Diagnostic Approach to Assess the Quality of Data Splitting in Machine Learning. https://doi.org/10.48550/ARXIV.2206.11721.
Jorquera, M., Korzeniowski, W., Skrzypkowski, K., 2023. Prediction of dilution in sublevel stoping through machine learning algorithms. In: IOP Conf Ser: Earth Environ Sci, vol. 1189, 012008. https://doi.org/10.1088/1755-1315/1189/1/012008, 1.
Kaplan, U.E., Dagasan, Y., Topal, E., 2021. Mineral grade estimation using gradient boosting regression trees. Int. J. Min. Reclamat. Environ. 35 (10), 728–742. https://doi.org/10.1080/17480930.2021.1949863.
Ke, G., Meng, Q., Finely, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30 (NIP 2017).
Krzywinski, M., Altman, N., 2014. Visualizing samples with box plots. Nat. Methods 11 (2), 119–120. https://doi.org/10.1038/nmeth.2813.
Li, C., Zhou, J., Du, K., Dias, D., 2023. Stability prediction of hard rock pillar using support vector machine optimized by three metaheuristic algorithms. Int. J. Min. Sci. Technol. 33 (8), 1019–1036. https://doi.org/10.1016/j.ijmst.2023.06.001.
Li, D., Liu, Z., Armaghani, D.J., Xiao, P., Zhou, J., 2022. Novel ensemble intelligence methodologies for rockburst assessment in complex and variable environments. Sci. Rep. 12 (1), 1844. https://doi.org/10.1038/s41598-022-05594-0.
Li, P., Tan, Z., Yan, L., Deng, K., 2011. Time series prediction of mining subsidence based on a SVM. Mining Science and Technology (China) 21 (4), 557–562. https://doi.org/10.1016/j.mstc.2011.02.025.
Liang, W., Luo, S., Zhao, G., Wu, H., 2020. Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. Mathematics 8 (5), 765. https://doi.org/10.3390/math8050765.
Lingga, B.A., Apel, D.B., 2018. Shear properties of cemented rockfills. J. Rock Mech. Geotech. Eng. 10 (4), 635–644. https://doi.org/10.1016/j.jrmge.2018.03.005.
Liu, L., Song, Z., Li, X., 2024. Artificial intelligence in tunnel construction: a comprehensive review of hotspots and frontier topics. Geohazard Mech. 2 (1), 1–12. https://doi.org/10.1016/j.ghm.2023.11.004.
Liu, R., Liu, E., Yang, J., Li, M., Wang, F., 2006. Optimizing the hyper-parameters for SVM by combining evolution strategies with a grid search. In: Huang, D.-S., Li, K., Irwin, G.W. (Eds.), Intelligent Control and Automation. Springer Berlin Heidelberg, pp. 712–721.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777.
Mathews, K.E., Hoek, E., Wyllie, D.C., Stewart, S., 1980. Prediction of Stable Excavation Spans for Mining at Depths below 1000 Metres in Hard Rock. Ottawa, ON.
Nawi, N.M., Atomi, W.H., Rehman, M.Z., 2013. The effect of data pre-processing on optimized training of artificial neural networks. Procedia Technol. 11, 32–39. https://doi.org/10.1016/j.protcy.2013.12.159.
Nickson, S.D., 1992. Cable Support Guidelines for Underground Hard Rock Mine Operations. https://doi.org/10.14288/1.0081080.
Potvin, Y., 1988. Empirical Open Stope Design in Canada. https://doi.org/10.14288/1.0081130.
Powers, D., 2008. Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation. Mach. Learn. Technol. 2.
Purwanto, Shimada H., Sasaoka, T., Wattimena, R.K., Matsui, K., 2013. Influence of stope design on stability of hanging wall decline in cibaliung underground gold mine. IJG 04 (10), 1–8. https://doi.org/10.4236/ijg.2013.410A001.
Qi, C., Fourie, A., Du, X., Tang, X., 2018a. Prediction of open stope hangingwall stability using random forests. Nat. Hazards 92 (2), 1179–1197. https://doi.org/10.1007/s11069-018-3246-7.
Qi, C., Fourie, A., Ma, G., Tang, X., Du, X., 2018b. Comparative study of hybrid artificial intelligence approaches for predicting hangingwall stability. J. Comput. Civ. Eng. 32 (2), 04017086. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000737.
Raju, V.N.G., Lakshmi, K.P., Jain, V.M., Kalidindi, A., Padma, V., 2020. Study the influence of normalization/transformation process on the accuracy of supervised classification. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, Tirunelveli, India, pp. 729–735.
Saadaari, F.S., Mireku-Gyimah, D., Olaleye, B.M., 2020. Development of a stope stability prediction model using ensemble learning techniques - a case study. Ghana Min. J. 20 (2), 18–26. https://doi.org/10.4314/gm.v20i2.3.
Samui, P., 2008. Slope stability analysis: a support vector machine approach. Environ. Geol. 56 (2), 255–267. https://doi.org/10.1007/s00254-007-1161-4.
Santos, A.E.M., Amaral, T.K.M., Mendonça, G.A., Silva, D.D.F.S.D., 2020. Open stope stability assessment through artificial intelligence. REM, Int Eng J 73 (3), 395–401. https://doi.org/10.1590/0370-44672020730012.
Skrzypkowski, K., 2021a. Determination of the backfilling time for the zinc and lead ore deposits with application of the BackfillCAD model. Energies 14 (11), 3186. https://doi.org/10.3390/en14113186.
Skrzypkowski, K., 2021b. 3D numerical modelling of the application of cemented paste backfill on displacements around strip excavations. Energies 14 (22), 7750. https://doi.org/10.3390/en14227750.
Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45 (4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.
Song, Z., Li, X., Huo, R., Liu, L., 2024. Intelligent early-warning platform for open-pit mining: current status and prospects. Rock Mech. Bull. 3 (1), 100098. https://doi.org/10.1016/j.rockmb.2023.100098.
Srivastava, M., 2023. Addressing spurious correlations in machine learning models: a comprehensive review. Open Sci. Framework.
Suorineni, F.T., Kaiser, P.K., Tannant, D.D., 2001. Likelihood statistic for interpretation of the stability graph for open stope design. Int. J. Rock Mech. Min. Sci. 38 (5), 735–744. https://doi.org/10.1016/S1365-1609(01)00033-8.
Szmigiel, A., Apel, D.B., 2022. Predicting the stability of open stopes using Machine Learning. J. Sustain. Min. 21 (3), 241–248. https://doi.org/10.46873/2300-3960.1369.
Tishkov, M., 2018. Evaluation of caving as a mining method for the Udachnaya underground diamond mine project. In: Proceedings of the Fourth International Symposium on Block and Sublevel Caving. Australian Centre for Geomechanics, Perth, pp. 835–846.
Vinay, L.S., Bhattacharjee, R.M., Ghosh, N., Kumar, S., 2023. Machine learning approach for the prediction of mining-induced stress in underground mines to mitigate ground control disasters and accidents. Geomech. Geophys. Geoenerg. Georesour. 9 (1), 159. https://doi.org/10.1007/s40948-023-00701-5.
Wang, J., Milne, D., Pakalnis, R., 2002. Application of a neural network in the empirical design of underground excavation spans. Min. Technol. 111 (1), 73–81. https://doi.org/10.1179/mnt.2002.111.1.73.
Wang, R., 2012. AdaBoost for feature selection, classification and its relation with SVM, A review. Phys. Procedia 25, 800–807. https://doi.org/10.1016/j.phpro.2012.03.160.
Wang, R., Chen, S., Li, X., Tian, G., Zhao, T., 2023. AdaBoost-driven multi-parameter real-time warning of rock burst risk in coal mines. Eng. Appl. Artif. Intell. 125, 106591. https://doi.org/10.1016/j.engappai.2023.106591.
Xu, Y., Goodacre, R., 2018. On splitting training and validation set: a comparative study of cross-validation, Bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J. Anal. Test 2 (3), 249–262. https://doi.org/10.1007/s41664-018-0068-2.
Zhou, J., Huang, S., Tao, M., Khandelwal, M., Dai, Y., Zhao, M., 2023. Stability prediction of underground entry-type excavations based on particle swarm optimization and gradient boosting decision tree. Undergr. Space 9, 234–249. https://doi.org/10.1016/j.undsp.2022.08.002.
Zhou, J., Li, X., Mitri, H.S., 2015. Comparative performance of six supervised learning methods for the development of models of hard rock pillar stability prediction. Nat. Hazards 79 (1), 291–316. https://doi.org/10.1007/s11069-015-1842-3.
Zhou, J., Shi, X., Du, K., Qiu, X., Li, X., Mitri, H.S., 2017. Feasibility of random-forest approach for prediction of ground settlements induced by the construction of a shield-driven tunnel. Int. J. GeoMech. 17 (6), 04016129. https://doi.org/10.1061/(ASCE)GM.1943-5622.0000817.
Zhou, Q., Liu, D., Lin, X., 2022. Pre-evaluation of fault stability for underground mining based on geomechanical fault-slip analysis. Geomat. Nat. Hazards Risk 13 (1), 400–413. https://doi.org/10.1080/19475705.2022.2032401.