Shahrood University of Technology

Journal of Mining and Environment (JME)

Journal homepage: www.jme.shahroodut.ac.ir

Iranian Society of Mining Engineering (IRSME)

Investigating the Applicability of Stacked Generalization Technique for the Prediction of Hard Rock Pillar Stability Status

Festus Kunkyin-Saadaari*, Jude Baah Offei, Sadique Ibn Sadique, Victor Kwaku Agadzie, and Ishmael Abeiku Forson

Faculty of Mining & Minerals Technology, Mining Eng, University of Mines and Technology, Tarkwa, Ghana

Article Info

Received 26 November 2024

Received in Revised form 28 January 2024

Accepted 21 February 2024

Published online 21 February 2024

DOI: 10.22044/jme.2025.15335.2940

Keywords

Matthews correlation coefficient

Hard rock pillar

Stacked generalisation

Underground mining

Extra trees

Abstract

The underground mining operations at the Obuasi Gold Mine rely heavily on the stability of hard rock pillars for safety and productivity. The traditional empirical and numerical methods for predicting pillar stability have limitations, prompting the exploration of advanced machine learning techniques. Hence, this work investigates the applicability of stacked generalization techniques for predicting the stability status of hard rock pillars in underground mines. Four stacked models were developed, using Gradient Boosting Decision Trees (GBDTs), Random Forest (RF), Extra Trees (ET), and Light Gradient Boosting Machines (LightGBMs), with each model taking turns as the meta-learner, while the remaining three models acted as the base learners in each case. The models were trained and tested on a dataset of 201 pillar cases from the AngloGold Ashanti Obuasi Mine in Ghana. Model performance was evaluated using classification metrics, including accuracy, precision, recall, F1-score and Matthews Correlation Coefficient (MCC). The RF-stacked model demonstrated the best overall performance, achieving an accuracy of 93.44%, precision of 94.27%, recall of 93.44%, F1-score of 93.59%, and MCC of 88.90%. Feature importance analysis revealed pillar depth and pillar stress as the most influential factors affecting pillar stability prediction. The results indicate that stacked generalization techniques, particularly the RF-stacked model, offer promising capabilities for predicting hard rock pillar stability in underground mining operations.

1. Introduction

Today, underground mining techniques, like self-supported, supported, and caving methods, are widely employed as mineral deposits are found at greater depths. However, these methods can lead to ground subsidence, and significant damage to surface structures. Therefore, choosing the right underground mining method to minimise subsidence is essential. Among self-supported techniques, partial extraction methods play a key role in controlling subsidence [1]. Room and pillar mining, a partial extraction method, is effective for deposits that are shallow to deep and flat, containing durable ores like coal, metals, and building stones [2]. In this approach, some ore is left as pillars to support the back. The stability of these pillars is crucial for their design, as instability can lead to serious incidents in underground mines such as pillar failures, collapses, injuries, fatalities, equipment damage, and lost work hours [3].

In the past few years, hard rock pillar stability has attracted much attention in geotechnical engineering and mining. As mineral resources get harder to reach, pillar failure can have severe economic and environmental effects. The problem is not just in predicting the stability of these geological structures but also in finding ways to increase prediction accuracy. Proper management of

* Corresponding author: fsaadaari@outlook.com (F. Kunkyin-Saadaari)

pillar stability helps keep the surrounding rock solid, preventing dangerous collapses that could endanger miners and equipment. Hard rock pillars can be considered as in situ rock between two or more underground openings, which may be of any shape and size, and splayed at the top and bottom to increase the support area on the roof and floor [4]. Martin and Maybee [4] reported that if a single pillar accidentally fails, the load carried by that pillar is distributed to the adjacent pillars, causing them to be overloaded. They further suggested that this continuous overloading process could give rise to an unstable dissemination of the load to the pillars, causing areas of the mine to collapse. It can now be said that most underground mining methods use pillars to extract ore without harming employees, thereby deeming it necessary to predict the stability of hard rock pillars. Hard rock pillar stability can be predicted using different methods, which are grouped into empirical, numerical, and machine-learning techniques. Different empirical techniques have been developed over time to predict the stability of pillars made of hard rock. Several researchers have discussed such empirical methods. Some of the reviewed articles are the linear shape effect formula discussed by Bieniawski *et al*.[5] and York [6], the size effect formula discussed by Hustrulid [7], the Hoek–Brown formula by Hoek *et al.* [8], and the power shape effect formula discussed by Salamon *et al*. [9], Hedley and Grant [10] and Bieniawski [11]. Pillar stability can be assessed by calculating the factor of safety (FoS), which is defined as the ratio of the average strength to the average stress experienced by the pillar [12]. Theoretically, if the FoS exceeds one, the pillar is regarded as stable. Research has indicated that pillars exhibiting a safety factor exceeding 1 may still experience failure due to irregular shapes, unpredictable material characteristics, and variations in mining practices [13]. Empirical approaches rely on the interpretation of data gathered from current or finished projects found in existing databases. As a result, generalising the findings beyond the characteristics of the initial site proves to be difficult [12].

Although empirical formulas are used to estimate pillar stress, they consider fewer factors, and have been validated at only a limited number of engineering sites. As a result, they may not be applicable in environments that differ from the original conditions [14].

When numerical methods are applied to assess pillar stability, they facilitate consideration of complex boundary conditions and material attributes. These techniques offer an in-depth analysis of deformation, stress distribution, and likely failure mechanisms that occur within the rocks and pillars. Numerical methods include the Finite Element Method (FEM) [15], Discrete Element Method (DEM) [15], Boundary Element Method (BEM) [16], and Finite Difference Method (FDM).

Numerical modelling techniques are more advantageous than the empirical methods in complex stress conditions. Some researchers have used numerical simulation methods to examine pillar stress, stability, and other properties [17-19]. However, despite being cost-effective and relatively simple to use. These methods involve several assumptions such as simplified boundary conditions and material properties, which can lead to idealised results that may not accurately reflect reality. Furthermore, the results and their accuracy can vary depending on the structural discretisation methods used, which means the predictions may not be reliable. Hence, applying these specially developed models to different situations can be challenging [14]. Also the anisotropic characteristics of rock masses, along with their complex non-linear behavior, pose significant challenges in thoroughly analysing the model inputs, and constitutive equations in numerical simulation approaches, thereby restricting the effectiveness of the results generated through this method [20]. Recently, machine learning techniques have shown substantial capability in predicting the stability of pillars with higher precision than that of traditional methods. This improvement is owing to an increase in data accessibility involving pillars [20]. Machine learning comprises modifications of systems involved in tasks associated with Artificial Intelligence (Al) including recognition, diagnosis, planning, robot control, and prediction, which may involve improvements to existing systems or the creation of entirely new ones [21]. The stability of crown pillars in large excavations was predicted by Tawadrous *et al.* [22] using Artificial Neural Networks (ANN). Their findings showed that the cloud model discussed by Ding *et al.* [23], presents a workable and trustworthy method for a thorough assessment of pillar stability. While Artificial Intelligence (AI) has been utilised in this domain, alternative

techniques remain underexplored. One such approach is Stacked Generalization, introduced by Wolpert [24]. Stacked generalization is a meta-learning technique that integrates multiple base-level models with a higher-level model to enhance predictive accuracy. This method has been successfully applied across various machine learning tasks, including classification, regression [25], and unsupervised learning [26]. The fundamental principle of Stacked generalization lies in its ability to analyse and mitigate the biases of base models concerning a given dataset [24]. By leveraging the complementary strengths of multiple well-performing models, this approach improves overall predictive performance. In classification and regression tasks, Stacked generalization effectively combines the predictive capabilities of different models, yielding superior results compared to any single model within the ensemble. This study employs four advanced stacked models using Gradient Boosted Decision Trees (GBDT), Random Forest (RF), Extra Trees (ET), and LightGBM as meta-learners. Each of these models takes turns serving as the meta-model, while the remaining models function as base learners. These models were chosen for their complementary strengths: GBDT's iterative error minimisation, RF's robustness against overfitting, ET's variance reduction, and LightGBM's scalability for large datasets. By employing this ensemble learning approach, the study aims to enhance predictive accuracy and reliability, ultimately contributing to improved decision-making in underground mining stability assessments. By leveraging these models in a stacked generalization framework, this study aims to optimise predictive performance by combining their individual strengths and mitigating potential biases inherent in single-model approaches.

2. Materials and Methods

2.1. Overview of research location

The Obuasi Gold Mine of AngloGold Ashanti is an underground mine engaged in gold mining and recovery activities in the Ashanti Region of the Republic of Ghana. The mining process began in 1897, and reached its end in the last quarter of 2014. Although production was stopped, some sections of the mine remained operational under restricted conditions, which included the development of an underground decline [27]. It is approximately 60 km south of Kumasi [27] and 200 km northwest of Accra. The primary means of accessing the mine are through shafts and a single access decline consisting of inter-level development ranging from 15 to 30 m. This decline, situated at the southern extremity of the mine, is designed to extend to a depth of approximately 1500 m [27]. Figure 1 displays the mine's location on a Ghana map.

2.2. Dataset description

This study utilised secondary data, comprising 201 datasets sourced from AngloGold Ashanti Ghana Limited, located at the Obuasi. The data consisted of the Pillar depth, width (W),height (H), width height ratio (W/H), Uniaxial Compressive Strength (UCS), Rock Mass Rating (RMR), pillar stress, and Pillar status. The Pillar depth, H, and W were measured in metres, W/H, and RMR was dimensionless, UCS and Pillar stress were in MPa, and the status can be classified as failed, stable, or unstable. Also since the study involves classification with a discrete output variable, a correlation analysis was performed only between input parameters to understand their relationships and potential multicollinearity. Figure 2 presents the correlation matrix of the input parameters, while Figure 3 shows the distribution of pillar stability statuses in the dataset. Each of the 201 data points was classified into one of three categories, failed, stable, or unstable, which were collectively termed the status of the pillars. For the 201 data points, 70% (140) were used to train the stacked models and 30% (61) were used to test the model. Tables 1 and 2 provide an overview of the collected dataset from AngloGold Ashanti. Table 1 presents a sample of the collected database, while Table 2 summarises the statistical description of the dataset, offering key insights into its distribution and characteristics. The pillar stress values used in this study were obtained from AngloGold Ashanti Obuasi Mine's geotechnical database, where stress is determined through a combination of tributary area theory and numerical modelling verification. At the mine, vertical pillar stress (σp) is calculated using $\sigma p = \gamma h(W + B)(L + B)/WL$, where γ is the unit weight of overburden rock (0.027 MN/m^3), h is depth below surface, W is pillar width, L is pillar length, and B is mining room width. These theoretical calculations are verified using FLAC3D numerical modelling that incorporates the actual mining geometry, rock mass properties, and sequential excavation effects.

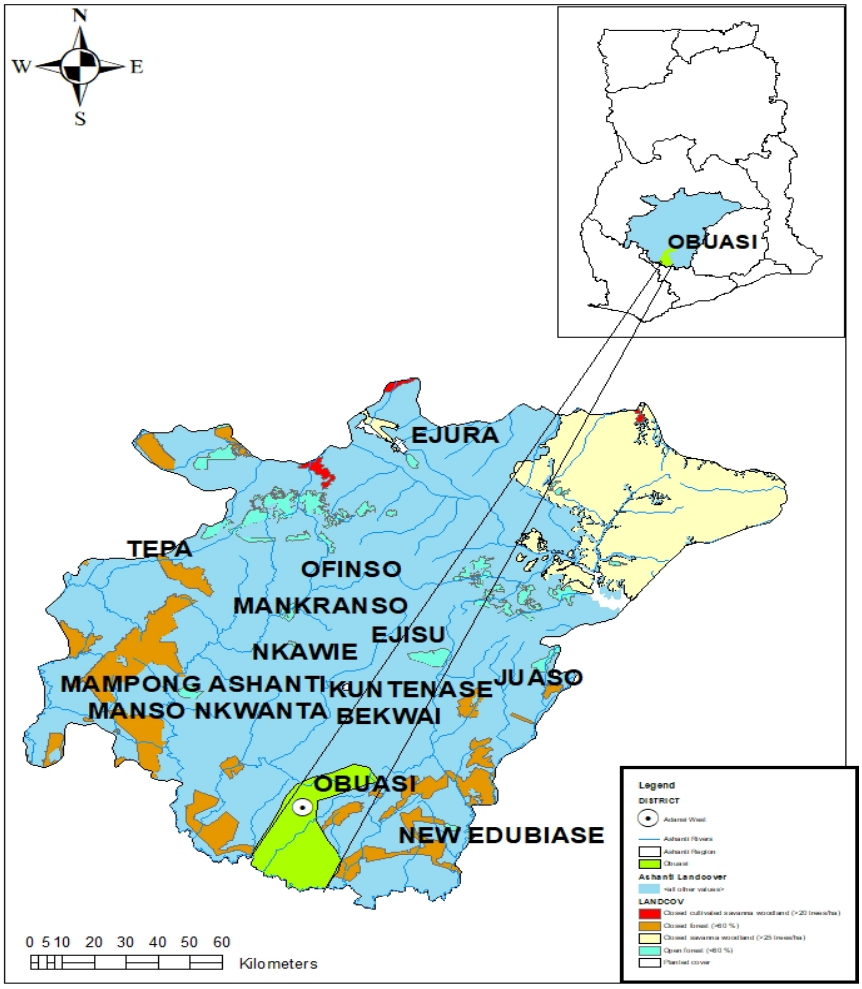

Figure 1. Location of the AngloGold Ashanti Obuasi Mine.

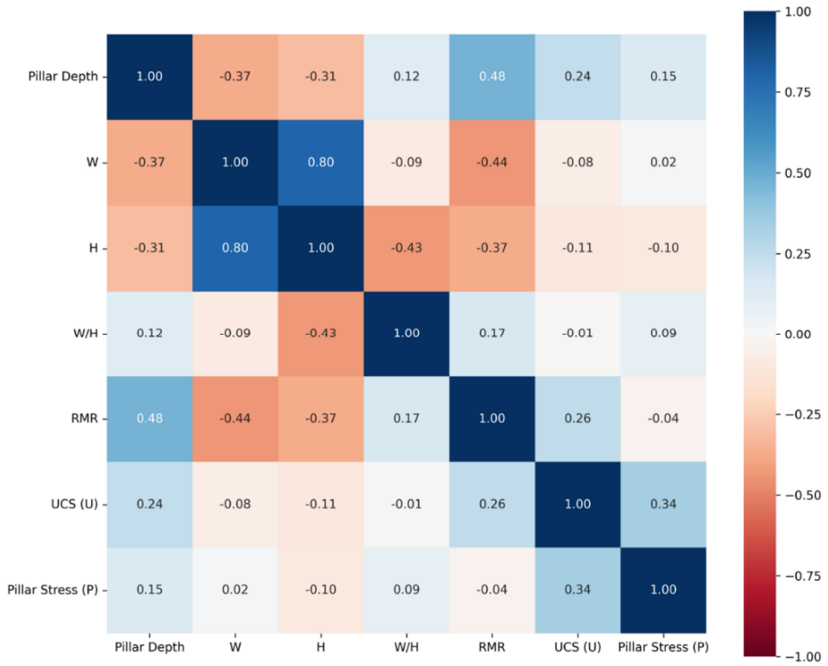

Figure 2. Statistical summary of the dataset.

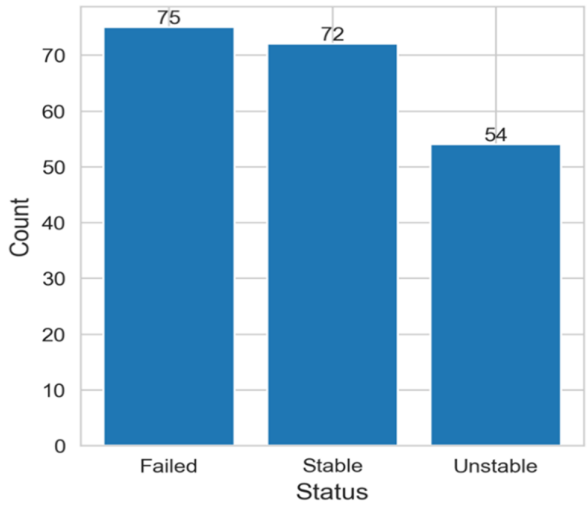

Figure 3. Bar chart showing the distribution of the pillar status.

Table 1. Sample of the collected database.

Pillar depth	W	H	W/H	RMR	UCS (U)	Pillar stress (P)	Status
621	4	3	1.3	88	239	82	Failed
163	5	3	2.0	88	93	39	Unstable
283	23	32	0.7	73	99	30	Unstable
448	10	5	1.8	88	309	45	Stable
859	5	5	1.0	88	229	36	Stable
602	3	1	2.4	68	171	98	Failed
188	13	22	0.6	53	93	47	Failed
354	7	4	1.7	68	93	34	Stable
362	9	3	2.7	88	169	55	Stable
457	4	4	1.0	88	93	58	Failed

Table 2. Statistical description of the dataset from AngloGold Ashanti.

Parameter	Unit	Statistic				
		Min	Max	Mean	Std dev	Mode
Pillar depth	m	91	922	383.82	209.80	279
Width (W)	m	1.9	45	10.8	8.03	6.1
Height (H)	m	1.47	112.9	12.03	14.94	3
RMR	-	50	98	79.69	12.25	90
UCS	MPa	70	316	164.72	64.015	94
Width(H)/Height(H)	-	0.31	4.51	1.25	0.66	1
Pillar Stress (P)	MPa	25	127.6	57.81	24.02	93.5

2.3. Data preparation

The dataset consists of 7 columns and 201 rows. A thorough investigation was performed to ensure the presence of entries in all the rows and columns. To avoid errors, commas were removed from all the values within the dataset. To avoid interference, missing data points were eliminated before the models were developed.

2.4. Methods Used

2.4.1. Gradient-boosting decision trees

Ke *et al.* [28] indicated that Gradient Boosting Decision Trees (GBDT) enhance a model's predictive accuracy by repeatedly applying learning techniques to reduce the loss function, which quantifies the discrepancy between the predicted and actual target values. Each iteration of the decision tree involves modifying the coefficients, biases, or weights associated with the input variables that are used to estimate the target value. The ultimate prediction was derived from the aggregate of the outputs from all decision trees. Stefanos *et al.* [29] also mentioned that, unlike decision trees, every regression tree features a continuous score at each leaf node, and for a specific dataset, the decision rules within the trees are employed to categorise it into leaves. Subsequently, the final prediction is calculated by summing the scores found in the relevant leaves. They also stated that, to acquire the set

of functions utilised in the model, they reduced the following objective:

$$L(\phi)=\sum_{i} l(\hat{y}_i,y_i)+\sum_{k}\Omega(f_k) \quad (1)$$

Li *et al.* [27] also mentioned how to form an approximate loss function by minimising the following objective function at the t-th iteration:

$$\tilde{L}^{(t)}=\sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}f_t^2(x_i)\right]+\Omega(f_t) \quad (2)$$

This process continues until either a pre-determined number of iterations is reached, or the convergence criteria are satisfied [30]. However, Zhang *et al.* [31] mentioned a limitation of current GBDT implementations, as the output of each decision tree consists of a single variable. This is because each leaf in the decision tree produced only one variable.

2.4.2. Extra trees

A subset of candidate features selected at random is used in the extra trees technique. Rather than determining which thresholds work best, it generates thresholds at random for every potential characteristic and uses the best of these random thresholds as the splitting criterion. This generally leads to a reduction in the model variance, although it may slightly elevate the bias [32]. This increases the randomness of the Extra Trees (ET) by eliminating the preference associated with choosing the most optimal feature. As a result, it reduces the possibility of the dataset being overfitted, but it might have a little bit more bias [33].

2.4.3. Random forest

Tree-based models are the basis of the Random Forest (RF) method. Until a pre-determined stopping requirement is met, a tree-based model systematically splits a given dataset into two subsets according to a particular criterion. They are referred to as leaf nodes or leaves at the end of the decision trees [34]. Random forest is mostly used in regression or classification tasks. Hastie *et al.* [35] explained that making a prediction at a new point x for the regression, the formula is as follows:

$$f_{rf}^{B}(x)=\frac{1}{B}\sum_{b=1}^{B}T_b(x) \quad (3)$$

and for classification purposes, the *b-th random-forest-tree* class prediction is $\hat{C}_b(x)$ next.

$$\hat{C}_{rf}^{B}(x)=\text{majority vote}\left\{\hat{C}_b(x)\right\}_1^B \quad (4)$$

2.4.4. Light gradient boosting machines

Microsoft's machine learning ensemble algorithm LightGBM provides a solid implementation of the gradient boosting method. This framework makes use of tree-based learning techniques and is built using two cutting-edge strategies for distributed efficiency: Exclusive Feature Bundling (EFB) and Gradient-based One-Side Sampling (GOSS) [28]. According to Ke *et al*. [28], GOSS analyses only the remaining data instances to obtain the information obtained, excluding a substantial fraction of those with minor gradients. GOSS can accurately estimate the information gained on a smaller dataset because data records with larger gradients are crucial to the computation of information gain. To reduce the total number of features, mutually exclusive features were consolidated, using the EFB. Because the LightGBM operates at high speed, it is prefixed with 'light'.

Tingting *et al*. [36] stated that if we aim to build a LightGBM model comprising T trees, the additive training procedure for a dataset containing n examples can be outlined as follows:

$$\hat{y}_i^{(t)}=\sum_{k=1}^{t}f_k(x_i)=\hat{y}_i^{(t-1)}+f_t(x_i) \quad (5)$$

where f_t represents the function obtained from the t-th decision tree, and $\hat{y}_i^{(t)}$ is the estimated value for the i-th example at the t-th iteration. He also mentioned that by reducing the following goal, the fs of every iteration can be discovered:

$$L^{(t)}=\sum_{i}^{n} l(y_i,\hat{y}_i^{(t)})+\sum_{t=1}^{T}\Omega(ft) \quad (6)$$

2.4.5. Stacked generalisation technique

Stacked generalisation, or stacking, is a significant ensemble learning method that constructs a new model by combining

predictions from multiple base models, thereby enhancing overall performance [26]. According to Dietterich [37], ensemble methods are learning algorithms that generate a group of classifiers and classify new data instances by amalgamating their predictions through a weighted voting process. Ensemble techniques are broadly classified into simple and advanced methods, with simple techniques like bagging and boosting forming the foundation of ensemble machine learning. Bagging, as described by Breiman [25], reduces variance by training several models on various subsets of the data and then combining their predictions by averaging. According to Freund and Schapire [38], boosting involves training models sequentially, where each model corrects the errors of its predecessor, thereby reducing bias. These simple ensemble methods improve model robustness and performance by leveraging multiple models. Advantages of stacked generalisation include improved accuracy, reduced risk of overfitting and underfitting, the ability to utilise diverse models, and adaptability to specific problems [39].

2.5. Model development process

Building a strong predictive model involves completing a number of basic steps in the model development process. The loading of the dataset became the starting point, and preparing the data to guarantee its purity and preparation for analysis was successful. After that, the data was used to train several base models, each of which found a different pattern. After being trained to effectively combine these outputs, a meta-model is trained using the generated predictions as inputs. Lastly, the performance of the stacked model was examined to guarantee correctness and dependability before it was used to create predictions for fresh data. The processes in the development of the stacked model are shown in Figure 4.

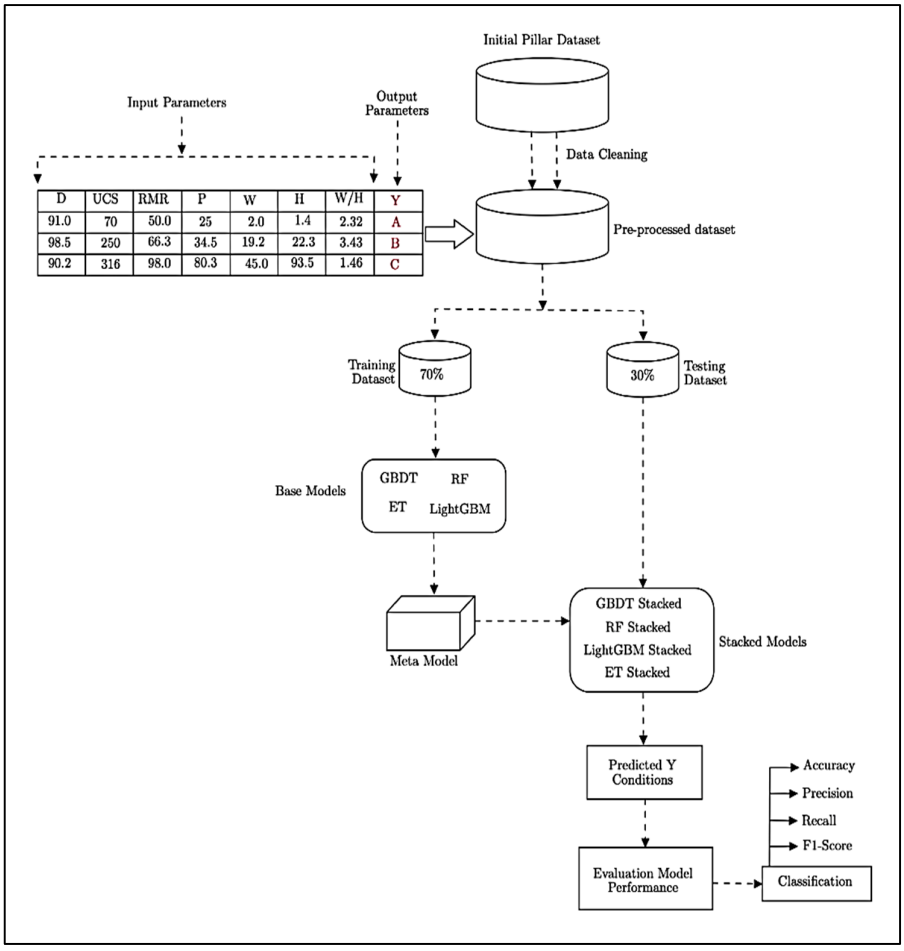

Figure 4. Model development process.

2.5.1. Data cleaning

The process of data cleaning is essential for discovering and correcting inaccuracies, inconsistencies, and mistakes in a dataset. This step is important in machine learning since the quality of the data impacts both the performance and trustworthiness of the machine learning models. In this study, the data was meticulously cleaned. This process entails rectifying missing values, discarding duplicates, and fixing errors like typographical mistakes and formatting issues. Additionally, the thorough handling of outliers ensures a strong dataset, which improves the reliability and accuracy of the ensuing machine learning analyses.

2.5.2. Data splitting

The dataset, consisting of 201 pillar cases, was divided into training and testing sets. From these cases, 140 samples (70%) were used as the learning dataset, where the models learn the relationships between the input parameters (pillar depth, width, height, UCS, RMR, W/H, and pillar stress) and the output (pillar status - failed, unstable, or stable). The remaining 61 samples (30%) were reserved as independent testing data. This split ensures sufficient data for model training while enabling evaluation of the models' ability to predict pillar stability status for unseen cases, simulating real-world applications.

2.5.3. Hyperparameter tuning

The optimal hyperparameter values were determined through grid search optimization with cross-validation on the training dataset. For each model, a range of potential values was tested for each hyperparameter, and the combination that yielded the highest classification accuracy was selected. The learning rate of 0.1 was found to provide the best balance between model convergence and training time across all models (Table 3).

Table 3. Critical hyperparameters and their optimal values.

Model	Hyperparameters	Optimal values
GBDT-Stacked	Learning_rate	0.1
	n_estimaters	100
	max_depth	3
RF-Stacked	Learning_rate	0.1
	Max_depth	None, 10, 20
	n_estimaters	50, 100, 20…0
ET-Stacked	Learning_rate	0.1
	n_estimaters	100
	Max_depth	-1
LightGBM-Stacked	Learning_rate	0.1
	Max_depth	-1
	n_estimators	100
	Min_samples_split	20

2.5.4. Evaluation

A confusion matrix was used to assess the stacked classification models' performance, as seen in Figure 5. Precision, accuracy, FI-score, Mathews Correlation Coefficient (MCC), and confusion matrix were used to calculate these metrics. After that, ranking was applied to identify the top-performing model.

Confusion matrix

A confusion matrix is a size $n \times n$ square matrix that shows how well a categorisation model works. The confusion matrix for this classification task was a 3x3 matrix, where each row represented the actual class, and each column indicated the projected class. The results were classified as stable, failed, or unstable. An illustration of a confusion matrix is presented in Figure 5.

The confusion matrix is made up of diagonal predictions C_{ii}, representing the number of instances correctly predicted for each status, that is, C_{AA} for status A, C_{BB} for status B, C_{CC} for status C and for the predictions that are not diagonal C_{ij}, which indicate misclassifications where the actual Status "i" is predicted as Status "j". With example being C_{AB} for instances, where status A is predicted as status B, C_{AC} for status A predicted as Status C, C_{BA} for Status B predicted as Status A, C_{BC} for Status B predicted as Status C, C_{CA} for Status C predicted as status A and C_{CB} for status C predicted as status B. The confusion matrix was used to determine performance parameters like

accuracy, precision, F1-score, recall and Mathews Correlation Coefficient (MCC). These can be expressed numerically using Equations (7) through (15). The components for each class are shown in a confusion matrix in Table 4.

		Predicted Values (j)		
		Stable (A)	Failed (B)	Unstable (C)
Actual Values (i)	Stable (A)	C_{AA}	C_{AB}	C_{AC}
	Failed (B)	C_{BA}	C_{BB}	C_{BC}
	Unstable (C)	C_{CA}	C_{CB}	C_{CC}

Figure 5. Confusion matrix.

Accuracy:

$$\text{Accuracy} = \frac{C_{AA}+C_{BB}+C_{CC}}{C_{AA}+C_{AB}+C_{AC}+C_{BA}+C_{BB}+C_{BC}+C_{CA}+C_{CB}+C_{CC}} \quad (7)$$

Precision:

Precision for status "A" is calculated as:

$$\text{Precision}_A = \frac{C_{AA}}{C_{AA}+C_{BA}+C_{CA}} \quad (8)$$

The precision of statuses B and C follow the same format.

The equation for Weighted Average Precision (WAP):

$$\text{WAP} = \frac{\text{Precsion}_A \times (tp_A + fn_A) + \text{Precsion}_B \times (tp_B + fn_B) + \text{Precsion}_C \times (tp_C + fn_C)}{\text{Total number oi nstances}} \quad (9)$$

Recall:

Recall for status "A" is calculated as:

$$\text{Recall}_A = \frac{C_{AA}}{C_{AA}+C_{AB}+C_{AC}} \quad (10)$$

The precision of statuses B and C follow the same format:

The equation for Weighted Average Recall (WAR):

$$\text{WAR} = \frac{\text{Recall}_A \times (tp_A + fn_A) + \text{Recall}_B \times (tp_B + fn_B) + \text{Recall}_C \times (tp_C + fn_C)}{\text{Total number of instances}} \quad (11)$$

F1-Score:

The F1-Score for Status "A" is calculated as:

$$F1_A = 2 \times \frac{\text{Precision}_A \times \text{Recall}_A}{\text{Precision}_A + \text{Recall}_A} \quad (12)$$

The F1-Score of Statuses B and C follows the same format.

The Equation for Weighted Average F1-Score (WAF):

$$F1_A = 2 \times \frac{\text{Precision}_A \times \text{Recall}_A}{\text{Precision}_A + \text{Recall}_A} \quad (13)$$

Matthews Correlation Coefficient (MCC):

The MCC for status "A" is calculated as:

$$MCC_A = \frac{tp_A \times tn_A - fp_A \times fn_A}{\sqrt{(tp_A + fp_A)(tp_A + fn_A)(tn_A + fp_A)(tn_A + fn_A)}} \quad (14)$$

The equation for average MCC:

$$\text{Average MCC} = \frac{MCC_A + MCC_B + MCC_C}{3} \quad (15)$$

Table 4. Confusion matrix components for each class.

	Class A	Class B	Class C
True Positive (tp)	C_{AA}	C_{BB}	C_{CC}
False Positive (fp)	$C_{BA} + C_{CA}$	$C_{AB} + C_{CB}$	$C_{AC} + C_{BC}$
False Negative (fn)	$C_{AB} + C_{AC}$	$C_{BA} + C_{BC}$	$C_{CA} + C_{CB}$
True Negative (tn)	$C_{BB} + C_{BC} + C_{CB} + C_{CC}$	$C_{AA} + C_{AC} + C_{CA} + C_{CC}$	$C_{AA} + C_{AB} + C_{BA} + C_{BB}$

Ranking

The ranking method of evaluation is a performance appraisal technique, in which the performances of stacked models are compared against each other and then ranked in order of performance. In this method, each model is placed in a hierarchy based on its performance in classification metrics. This approach helps in identifying the top-performing model as well as those who may need improvement by creating a clear distinction between different levels of performance.

2.6. Selection of features and dependent variable

The input parameters were selected at this point. Since these variables are thought to have the biggest effects on pillar stability, the study's parameters were Depth, Width (W), Height (H), Rock Mass Rating (RMR), Uniaxial Compressive Strength (UCS), and Pillar Stress (P). Additionally, the pillars' condition was chosen as the target variable and divided into three groups: failed, unstable, or Stable. The independence of input parameters was considered during feature selection. Although pillar depth and pillar stress are theoretically correlated, both parameters were retained in the model development because local variations in geology, mining sequence, and pillar geometry can cause significant deviations in their relationship. The inclusion of both parameters improved the model's predictive capability by capturing these site-specific variations.

3. Results and Discussion

3.1. Model performance analysis

The performance evaluation of the four stacked generalisation models revealed distinct capabilities in predicting pillar stability. The GBDT-stacked model demonstrated strong predictive performance, particularly in identifying unstable cases, as shown in its confusion matrix in Figure 6. It achieved an average accuracy of 91.80% and an MCC of 85.51%, ranking second overall with a total rank of 13, as detailed in Table 11. In contrast, the RF-stacked model emerged as the top performer, surpassing other models with a precision of 94.27%, recall of 93.44%, F1-score of 93.59%, and MCC of 88.90%. Its robustness, highlighted in Figure 7 and Table 6, secured it the highest total rank of 19, reflecting its superior generalization across metrics.

The ET-Stacked model exhibited moderate performance, with an average accuracy of 86.89% and precision of 88.49%. Despite its confusion matrix in Figure 8 showing reasonable correct predictions, it ranked lowest overall with a total rank of 5, attributed to its lower recall of 86.89% and MCC of 78.61%, as summarized in Table 7. Conversely, the LightGBM-Stacked model delivered competitive results, achieving the highest testing accuracy of 93.77% alongside balanced recall of 90.16%, precision of 91.36%, and F1-score of 90.48%. Its performance, visualized in Figure 9 and Table 8, earned it a joint second-place ranking with a total rank of 13.

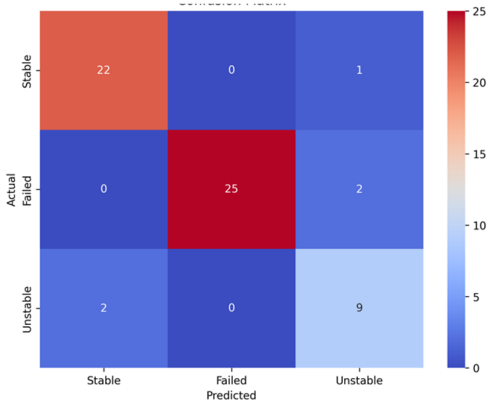

Figure 6. Confusion matrix of GBDT-Stacked model

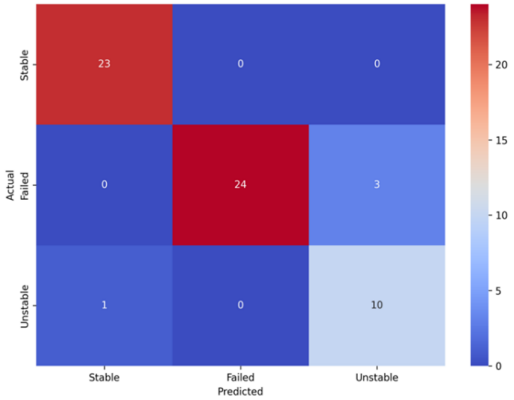

Figure 7. Confusion matrix of RF-Stacked model

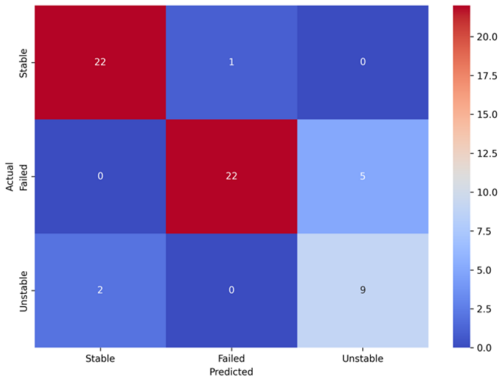

Figure 8. Confusion matrix of ET-Stacked model

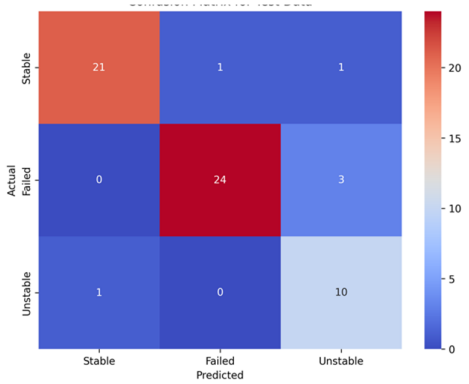

Figure 9. Confusion matrix of LightGBM-Stacked model

Table 5. Summary of GBDT-Stacked model.

Status/Metric	Accuracy	Precision	Recall	F1-score	MCC
Stable	0.95082	0.91667	0.95652	0.93617	0.89674
Failed	0.96721	1.00000	0.92593	0.96154	0.93514
Unstable	0.91803	0.75000	0.81818	0.78261	0.73327
Average	0.91803	0.92350	0.91803	0.91971	0.85505

Table 6. Summary of RF-Stacked model.

Status/Metric	Accuracy	Precision	Recall	F1-score	MCC
Stable	0.98361	0.95833	1.00000	0.97872	0.96598
Failed	0.95082	1.00000	0.88889	0.94118	0.90378
Unstable	0.93443	0.76923	0.90909	0.83333	0.79716
Average	0.93443	0.94268	0.93443	0.93589	0.88897

Table 7. Summary of ET-Stacked model.

Status/Metric	Accuracy	Precision	Recall	F1-score	MCC
Stable	0.95082	0.91667	0.95652	0.93617	0.89674
Failed	0.90164	0.95652	0.81481	0.88000	0.80493
Unstable	0.88525	0.64286	0.81818	0.72000	0.65660
Average	0.86885	0.88493	0.86885	0.87233	0.78609

Table 8. Summary of LightGBM-Stacked model.

Status/Metric	Accuracy	Precision	Recall	F1-score	MCC
Stable	0.95082	0.95455	0.91304	0.93333	0.89496
Failed	0.93443	0.96000	0.88889	0.92308	0.86803
Unstable	0.91803	0.71429	0.90909	0.80000	0.75800
Average	0.93765	0.91363	0.90164	0.90475	0.85834

Analysis of training and testing metrics in Tables 9 and 10 underscores the RF-Stacked model's consistency, with minimal overfitting and superior generalization, maintaining an MCC of 88.90% in both training and testing phases. While the LightGBM-Stacked model achieved the highest testing accuracy, its slightly lower MCC of 85.83% compared to the RF-Stacked model indicates marginally weaker class-balanced performance. The GBDT-Stacked model showed stable results across datasets, whereas the ET-Stacked model lagged in critical metrics, including an F1-score of 87.23% and MCC of 78.61%, as shown in Tables 7 and 9.

Table 9. Models Performance Results for the Testing Dataset.

Stacked generalization Model	Classification metrics				
	Accuracy	Precision	Recall	F1-score	MCC
GBDT	0.91803	0.92350	0.91803	0.91971	0.85505
Random Forest	0.93443	0.94268	0.93443	0.93589	0.88897
ET	0.86885	0.88493	0.86885	0.87233	0.78609
LightGBM	0.93765	0.91363	0.90164	0.90475	0.85834

Table 10. Models performance Results for the Training Dataset.

Stacked generalization Model	Classification Metrics				
	Accuracy	Precision	Recall	F1-score	MCC
GBDT	0.97857	0.97981	0.97857	0.97850	0.96817
Random Forest	0.93443	0.94268	0.93443	0.93589	0.88897
ET	0.97143	0.97142	0.97143	0.97135	0.95656
LightGBM	0.93765	0.91363	0.90164	0.90475	0.85834

Table 11. Ranking of models.

Model	Rank					
	Accuracy	Precision	Recall	F1-score	MCC	Total rank
GBDT	2	3	3	3	2	13
Random Forest	3	4	4	4	4	19
ET	1	1	1	1	1	5
LightGBM	4	2	2	2	3	13

3.2 Practical implications and limitations

The comparative performance analysis reveals several important implications for practical applications in underground mining operations. The RF-Stacked model's superior performance, with an MCC of 88.90%, precision of 94.27%, and F1-score of 93.59%, makes it particularly suitable for critical stability assessments where high prediction accuracy is essential. The GBDT-stacked model, achieving a recall of 91.80% for failed pillars, suggests its potential use in early warning systems, though its unstable-class F1-score of 78.26% limits standalone utility. The LightGBM-Stacked model, with the highest testing accuracy of 93.77% and balanced precision of 91.36%, is well-suited for general stability monitoring. However, the ET-Stacked model, with the lowest total rank of 5 and an MCC of 78.61%, is not recommended for deployment without refinement.

A persistent limitation across all models is their reduced efficacy in classifying "Unstable" pillars. For example, the RF-stacked model achieves an unstable-class F1-score of 83.33%, while the LightGBM-stacked and GBDT-stacked models score 80.00%, and 78.26%, respectively. This highlights the need for improved feature engineering or expanded datasets capturing intermediate stability states, while the traditional Factor of Safety (FoS) approach remains a practical baseline, the RF-Stacked model's MCC of 88.90% demonstrates machine learning's ability to capture nuanced stability factors that FoS might overlook in

complex environments like deep mining. It is proposed that integrating FoS with advanced models, rather than relying solely on empirical thresholds, to enhance decision-making.

Again, some key limitations include sensitivity to data distribution, as evidenced by the GBDT-stacked model's MCC drop from 96.82% (training) to 85.51% (testing), and the lightGBM-stacked model's MCC of 85.83%, which reflects slightly weaker class-balanced performance. The ET-Stacked model's low recall of 86.89% further emphasises risks in deploying underperforming models without validation. These limitations underscore the importance of external dataset validation to ensure generalisability across geological conditions. Careful model selection, prioritising the RF-Stacked model for critical tasks and LightGBM-stacked for efficiency, remains vital for real-world mining applications.

This study goes on to demonstrate the potential of stacked generalisation techniques for pillar stability prediction, rather than providing a universal model; the methodology presented here serves as a framework that mining operations can adapt to develop their own site-specific models using their historical pillar performance data. This approach is particularly valuable since each mine has unique geological and operational conditions that influence pillar stability. Mining operations can implement this methodology to create customised prediction models that account for their specific rock mass characteristics, mining methods, and depth of operations.

3.3. Interpretations of sensitivity analysis

The sensitivity analysis, conducted through feature importance evaluation of the RF-stacked model, revealed significant insights into the factors influencing pillar stability prediction. As illustrated in Figure 10, pillar depth emerged as the most influential parameter, followed closely by pillar stress, while other parameters demonstrated relatively lower importance in the prediction process. This finding aligns with fundamental rock mechanics principles, where increasing depth typically correlates with higher in-situ stresses and more complex ground conditions. The dominant influence of pillar depth underscores the increasing challenges faced in deeper mining operations, particularly relevant to the AngloGold Ashanti Obuasi Mine context, where mining depths extend to approximately 1500 m, as depicted in Figure 1.

The substantial influence of pillar stress as the second most significant parameter emphasizes the critical role of stress distribution in pillar stability assessment. This finding is particularly pertinent given the geological characteristics of the Obuasi Mine, where the relationship between depth and pillar stress may be complicated by varying rock mass conditions, as evidenced by the range of Rock Mass Rating values shown in Figure 2. The identification of these key parameters through sensitivity analysis provides valuable guidance for prioritizing monitoring efforts and resource allocation in underground mining operations, particularly in deep environments where stress-related challenges are pronounced.

These insights from the sensitivity analysis have significant implications for practical applications and future research. The clear prioritization of depth and pillar stress as primary factors suggests these parameters should be central to monitoring systems and stability protocols. Furthermore, this understanding could guide the development of streamlined predictive models that focus on these critical inputs while maintaining high accuracy, as demonstrated by the RF-stacked model's performance metrics in Tables 6 and 9 including an MCC of 88.90% and F1-score of 93.59%. However, while parameters such as pillar width and rock mass quality showed lower importance in Figure 10, their collective contribution remains essential to the model's overall predictive capability, reinforcing the need for comprehensive data integration in stability assessments.

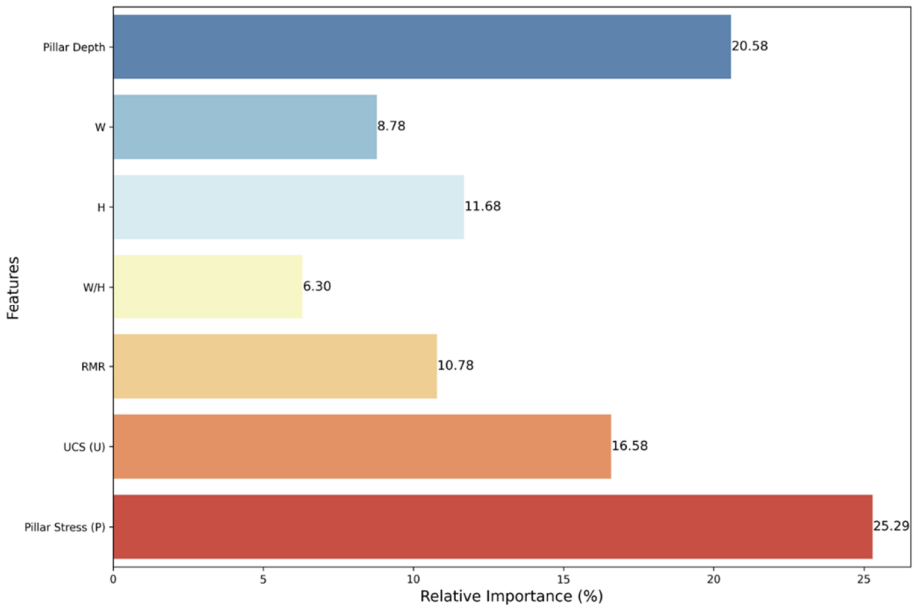

Figure 10. Feature importance of RF-stacked model.

4. Conclusions

The current study has successfully demonstrated the applicability and effectiveness of stacked generalisation techniques in predicting hard rock pillar stability in underground mining operations. Four advanced stacked models were developed using GBDT, RF, ET, and LightGBM as meta-learners, with each taking turns as the meta-model while the others served as base learners. Rigorous evaluation using classification metrics—including Matthews Correlation Coefficient (MCC), precision, accuracy, F1-score, and recall—provided robust validation of their predictive capabilities.

Among the tested models, the RF-Stacked Model demonstrated superior performance, achieving an accuracy of 93.44%, precision of 94.27%, recall of 93.44%, F1-score of 93.59%, and MCC of 88.90%. This outstanding performance, coupled with its highest overall ranking of 19 in the comparative analysis (Table 11), establishes the RF-Stacked Model as the most reliable choice for pillar stability prediction in critical applications. The LightGBM-Stacked Model achieved the highest testing accuracy of 93.77%, making it suitable for general monitoring, while the GBDT-Stacked Model showed strong recall of 91.80% for failed pillars, supporting its use in early warning systems. However, the ET-Stacked Model, with the lowest total rank of 5 and MCC of 78.61%, requires significant refinement for practical deployment.

Feature importance analysis (Figure 10) identified pillar depth and pillar stress as the most influential factors, aligning with rock mechanics principles and the geological context of the AngloGold Ashanti Obuasi Mine, where depths exceed 1500 m (Figure 1). These insights underscore the need to prioritize these parameters in monitoring and risk assessment strategies, particularly in deep mining environments with complex stress distributions, as reflected in the Rock Mass Rating variability.

The findings advance intelligent systems in geotechnical engineering by offering a machine learning-driven alternative to traditional empirical methods like the Factor of Safety (FoS). However, the models' reduced efficacy in classifying "unstable" pillars (e.g., RF-stacked F1-score of 83.33%) highlights opportunities for refinement through improved feature engineering or hybrid approaches integrating FoS with machine learning. Future research should focus on temporal data integration, validation across diverse geological conditions, and addressing class imbalance challenges. This study provides a foundation for enhancing safety and efficiency in underground mining, bridging empirical practices with next-generation predictive analytics.

Conflicts of interest

The authors declare no conflict of interest concerning this work.

Ethics statement

This study adhered to strict ethical protocols.

Funding body

The authors did not receive support from any organisation regarding the submitted work.

Acknowledgements

The authors wish to extend their sincere appreciation to the management of AngloGold Ashanti Obuasi Mine for their significant support in providing the data that facilitated this research.

Data access statement

Research data can be acquired by reaching the corresponding author with a valid request.

References

[1]. Yu, Y., Deng, K. Z., & Chen, S. E. (2018). Mine size effects on coal pillar stress and their application for partial extraction. *Sustainability*, *10*(3), 792.

[2]. Kadkhodaei, M. H., Ghasemi, E., Zhou, J., & Zahraei, M. (2024). Evaluation of underground hard rock mine pillar stability using gene expression programming and decision tree-support vector machine models. *Deep Underground Science and Engineering, 1,* 53-73.

[3]. Wessels, D. G., & Malan, D. F. (2023). A limit equilibrium model to simulate time-dependent pillar scaling in hard rock bord and pillar mines. *Rock Mechanics and Rock Engineering*, *56*(5), pp.3773-3786.

[4]. Martin, C. D., & Maybee, W. G. (2000). The strength of hard-rock pillars. *International Journal of Rock Mechanics and Mining Sciences*, *37*, 1239–1246.

[5]. Bieniawski, Z. T., & Van Heerden, W. L. (1975). The significance of in situ tests on large rock specimens. *International Journal of Rock Mechanics and Mining Sciences and Geomechanics Abstracts*, *12*(4), 101–113.

[6]. York, G. (1998). Numerical modelling of the yielding of a stabilising pillar/foundation system and a new design consideration for stabilising pillar foundations. *The Journal of The South African Institute of Mining and Metallurgy*, *98*(6), 281-298.

[7]. Hustrulid, W. (1976). A review of coal pillar strength formulas. *Rock Mechanics*, *8*, 115–145.

[8]. Hoek, E., & Brown, E. T. (1980). Underground Excavations in Rock. *Institution of Mining and Metallurgy*, 527.

[9]. Salamon, A., & Munro, M. (1967). A study of the strength of coal pillars. *Journal of The South African Institute of Mining and Metallurgy*, *68*, 55–67.

[10]. Hedley, D. G. F., & Grant, F. (1972). Stope-and-pillar design for the Elliot Lake Uranium Mines. *Canadian Institute of Mining, Metallurgy and Petroleum*. 74-78.

[11]. Bieniawski, Z. T. (1968). The effect of specimen size on compressive strength of coal. *International Journal of Rock Mechanics and Mining Sciences and Geomechanics Abstracts*, *5*(4), 325–335.

[12]. 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. *Natural Hazards*, *79*(1), 291–316.

[13]. Ahmad, M., Al-Shayea, N.A., Tang, X. W., Jamal, A., M. Al-Ahmadi, H., & Ahmad, F. (2020). Predicting the pillar stability of underground mines with random trees and C4.5 decision trees. *Applied Sciences*, *10*(18), 6486.

[14]. Zhou, J., Chen, Y., Chen, H., Khandelwal, M., Monjezi, M., & Peng, K. (2023). Hybridising five neural-metaheuristic paradigms to predict the pillar stress in bord and pillar method. *Frontiers in Public Health*, *11*, 1119580.

[15]. Li, X., Kim, E., & Walton, G. (2019). A study of rock pillar behaviours in laboratory and in-situ scales using combined finite-discrete element method models. *International Journal of Rock Mechanics and Mining Sciences*, *118*, 21–32.

[16]. Jaiswal, A., Sharma, S. K., & Shrivastva, B. K. (2004). Numerical modelling study of asymmetry in the induced stresses over coal mine pillars with the advancement of the goal line. *International Journal of Rock Mechanics and Mining Sciences*, *41*(5), 859–864.

[17]. Ahmed, S. S., Gunzburger, Y., Renaud, V., & AlHeib, M. (2017). The initialisation of highly heterogeneous virgin stress fields within the numerical modelling of large-scale mines. *International Journal of Rock Mechanics and Mining Sciences*, *99*, 50-62.

[18]. Li, X., Kim, E., & Walton, G. (2019). A study of rock pillar behaviours in laboratory and in-situ scales using combined finite-discrete element method models. *International Journal of Rock Mechanics and Mining Sciences*, *118*, 21-32.

[19]. Li, C., Zhou, J., Armaghani, D. J., & Li, X. (2021). Stability analysis of underground mine hard rock pillars via a combination of finite difference methods, neural networks, and Monte Carlo simulation techniques. *Underground Space*, *6*(4), 379-395.

[20]. Salih, A., & Abdul Hussein, H. (2022). Lost circulation prediction using decision tree, random forest, and extra trees algorithms for an Iraqi oil field. *Iraqi Geological Journal*, *55*(2E), 111–127.

[21]. Nilsson, J. N. (2005). Introduction to machine learning. *Department of Computer Science,* Stanford University Stanford, *CA 94305*, 209.

[22]. Tawadrous, A. S., & Katsabanis, P. D. (2007). Prediction of surface crown pillar stability using artificial neural networks. *International Journal for Numerical and Analytical Methods in Geomechanics*, *31*(7), 917–931.

[23]. Ding, H., Li, G., Dong, X., Lin, Y. (2018). Prediction of pillar stability for underground mines using the stochastic gradient boosting technique. *IEEE Access: Practical Innovations, Open Solutions*, *6*, 69253–69264.

[24]. Wolpert, D.H. (1992). Stacked Generalisation. *Neural Networks. 5*(2), 241-259.

[25]. Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

[26]. Smyth, P., & Wolpert, D. (1997). Stacked Density Estimation. *Neural Information Processing Systems*, *10*, 668–674.

[27]. Li, Q., Wu, Z., Wen, Z., & He, B. (2020). Privacy-preserving gradient boosting decision trees. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(1), 784–791.

[28]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.Y. (2017). LightGBM: A highly efficient Gradient Boosting Decision Tree. *Neural Information Processing Systems*, 3146–3154.

[29]. Fafalios, S., Charonyktakis, P., Tsamardinos, I. (2020). Gradient Boosting Trees. *Gnosis Data Analysis PC. 1,* 1–3.

[30]. Liang, W., Luo, S., Zhao, G., & Wu, H. (2020). Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. *Mathematics*, *8*(5), 765.

[31]. Zhang, Z., & Jung, C. (2019). GBDT-MO: Gradient boosted decision trees for multiple outputs. *Computer Vision and Pattern Recognition*, *32*, 3156-3167.

[32]. Eslami, E., Salman, A.K., Choi, Y., Sayeed, A., & Lops, Y. (2020). A data ensemble approach for real-time air quality forecasting using extremely randomised trees and deep neural networks. *Neural Computing and Applications*, *32*(11), 7563–7579.

[33]. Ghazwani, M., & Begum, M.Y. (2023). Computational intelligence modelling of hyoscine drug solubility and solvent density in supercritical processing: gradient boosting, extra trees, and random forest models. *Scientific Reports*, *13*(1), 10046.

[34]. Schonlau, M., & Zou, R.Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3–29.

[35]. Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83–85.

[36]. Chen, T., Xu, J., Ying, H., Chen, X., Feng, R., Fang, X., Gao, H., & Wu, J. (2019). Prediction of extubation failure for intensive care unit patients using light gradient boosting machine. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 150960–150968.

[37]. Dietterich, T.G. (2000). Ensemble Methods in Machine Learning. *International Workshop on Multiple Classifier Systems*, 1–15.

[38]. Freund, Y., & Schapire, R.E. (1997). A decision-theoretic generalisation of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

[39]. van der Laan, M., Polley, E., & Hubbard, A. (2008). Super Leaner. *UC Berkeley Division of Biostatistics Working Paper Series.*, *222*, 1-20

بررسی کاربرد تکنیک تعمیم پشته‌ای برای پیش‌بینی وضعیت پایداری ستون سنگ سخت

فستوس کونکین-ساداری*، جود باه اففی، صدیق بن صدیق، ویکتور کواکو آگادزی، و اسماعیل آبئیکو فورسون

دانشکده فناوری معدن و مواد معدنی، مهندس معدن، دانشگاه معادن و فناوری، تارکوا، غنا

ارسال ۲۰۲۴/۱۱/۲۶، پذیرش ۲۰۲۵/۰۲/۲۱

* نویسنده مسئول مکاتبات: fsaadaari@outlook.com

چکیده:

عملیات معدنکاری زیرزمینی در معدن طلای Obuasi به شدت به پایداری ستون‌های سنگ سخت برای ایمنی و بهره‌وری متکی است. روش‌های تجربی و عددی سنتی برای پیش‌بینی پایداری ستون دارای محدودیت‌هایی هستند که باعث اکتشاف تکنیک‌های پیشرفته یادگیری ماشین می‌شوند. از این رو، این کار به بررسی کاربرد تکنیک‌های تعمیم انباشته برای پیش‌بینی وضعیت پایداری ستون‌های سنگ سخت در معادن زیرزمینی می‌پردازد. چهار مدل پشته‌ای با استفاده از درخت‌های تصمیم‌گیری افزایش گرادیان (GBDTs)، جنگل تصادفی (RF)، درخت‌های اضافی (ET) و ماشین‌های تقویت گرادیان نور (LightGBMs) توسعه یافتند، که هر مدل به نوبت به عنوان فراآموز انتخاب می‌شد، در حالی که سه مدل باقی‌مانده به‌عنوان یادگیرنده پایه در هر مورد عمل می‌کردند. این مدل‌ها بر روی مجموعه داده‌ای از ۲۰۱ کیس ستونی از معدن AngloGold Ashanti Obuasi در غنا آموزش و آزمایش شدند. عملکرد مدل با استفاده از معیارهای طبقه بندی، از جمله دقت، دقت، یادآوری، امتیاز F1 و ضریب همبستگی متیوز (MCC) مورد ارزیابی قرار گرفت. مدل انباشته RF بهترین عملکرد کلی را نشان داد و دقت ۹۳.۴۴٪، دقت ۹۴.۲۷٪، یادآوری ۹۳.۴۴٪، امتیاز F1 ۹۳۵۹ ٪ و MCC ۸۸.۹۰ ٪ را به دست آورد. تجزیه و تحلیل اهمیت ویژگی عمق ستون و تنش ستون را به عنوان تأثیرگذارترین عوامل مؤثر بر پیش بینی پایداری ستون نشان داد. نتایج نشان می‌دهد که تکنیک‌های تعمیم انباشته، به ویژه مدل انباشته RF، قابلیت‌های امیدوارکننده‌ای را برای پیش‌بینی پایداری ستون سنگ سخت در عملیات معدنکاری زیرزمینی ارائه می‌دهند.

کلمات کلیدی: ضریب همبستگی متیوز، ستون هارد راک، تعمیم انباشته، معدن زیرزمینی، درخت اضافی.