Predictive Modeling of Slope Stability Using Hybrid Machine Learning Approaches: PCA- GMM based WOA-XGBoost

Muhammad Kamran (✉ m.kamran@utas.edu.au)
University of Tasmania

Research Article

Keywords: Slopes stability; whale optimization algorithm, XGBoost, Predictive modelling, gaussian mixture model

Posted Date: December 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3726217/v1

License: ☺️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Predictive Modeling of Slope Stability Using Hybrid Machine Learning Approaches: 
PCA-GMM based WOA-XGBoost

Muhammad Kamran1*

*School of Engineering, University of Tasmania, Hobart, TAS 7001, Australia, Email: m.kamran@utas.edu.au

Abstract:
Slope instability is a widespread environmental concern that affects all geographical regions. It is crucial to consider slope instability in mining, civil, tunnelling, and geotechnical engineering projects. A catastrophic event involving slope instability can lead to severe economic losses, casualties, and property damage. This study introduces the predictive modelling of slope stability using hybrid machine learning approaches. Firstly, a total of 404 events of slope stability were gathered using easily accessible input attributes. Secondly, principal component analysis (PCA) has been applied to reduce the dimension of the slope stability dataset. Thirdly, gaussian mixture model (GMM) was employed to categorize the PCA obtained dataset into distinct clusters. Fourthly, the parameters of extreme gradient boosting (XGBoost) were optimized using whale optimization algorithm (WOA). Lastly, the performance of the suggested model was assessed using several performance indices. The results indicate that the proposed hybrid PCA-GMM based WOA-XGBoost model exhibits the highest level of reliability. The proposed unsupervised-supervised machine learning method using a metaheuristic algorithm technique can considerably reduce slope-related disastrous incidents while rendering mining operations safer and more sustainable.

Keywords: Slopes stability; whale optimization algorithm, XGBoost, Predictive modelling, gaussian mixture model

1. Introduction:
Slope instability, in combination with earthquakes and volcanoes, ranks among the three primary environmental concerns that have an impact on all geographic regions. When it comes to in mining, civil, tunnelling, and geotechnical engineering projects, slope instability becomes an essential consideration. Devastating economic losses, fatalities, and property damage may result from a catastrophic event involving slope instability. A catastrophic event occurred in the Hong'ao landfill located in China. The catastrophe resulted in the demolition of 33 structures, the demise of 77 individuals, and direct financial losses amounting to 880 million RMB [1].
Likewise, a landslide occurred in the Koshe dump in Yah, the nation's capital of Ethiopia, causing the disappearance of more than eighty individuals [2]. The incident resulted in the death of eleven individuals. A total of twenty-two troops lost their lives in a landslide that occurred in the Quang Tri province of Vietnam, as a result of heavy rainfall [3]. An essential component of disaster mitigation is the capacity to predict the potential of slope instability, which can lead to severe and disastrous outcomes.

Prediction systems for slope stability fall into four primary categories. Technological instruments take precedence in instrumental monitoring. There are currently a variety of techniques in use for monitoring potential hazard indicators of slope instability by tracking slope deformation on-site. Zhang et al. [4] used dispersed fibre optic strain monitors to measure the shear displacement and discovered two possible circular surfaces for sliding to illustrate their argument. A number of researchers utilised acoustic emissions (AE) technologies to observe the signals produced by soil and rock fractures across the slope [5-7]. The stability of slopes was evaluated over time using criteria derived from AE through an examination of the correlation between AE characteristics and deformation. In addition to remote sensing [8], terrestrial laser scanning [9], synthetic aperture radar [10], and time domain reflectometry [11], additional techniques were employed for slope stability monitoring. These technologies offer a high level of prediction accuracy due to the direct availability of data on slope stability beforehand. However, they are costly and necessitate a complicated process of installation.

The second strategy is utilised within the domain of theoretical analysis. This statement is presented from the standpoint of a mechanical component. Analytical and theoretical approaches such as the limit analysis method, strength reduction method (SRM), and limit equilibrium method (LEM) have been employed to investigate slope stability, among others [12-14]. The stability of a slope is determined by calculating the factor of safety (FOS), which is calculated using the ratio of the propelling force to the resisting force. When FOS is greater than 1, slope stability is attained; when FOS is less than 1, instability is observed [15]. Faramarzi et al. [16] employed LEM to ascertain the FOS and evaluate the stability of the granite slope in the Chamshir dam hole. To ascertain the FOS of the slope model that had been previously constructed, Liu [17] employed the SRM. To assess the dependability of circular-type failure on homogeneous gradients, Mbarka et al. [18] combined the Monte Carlo method, LEM, and SRM. Despite the simplicity of the methods, slopes characterised by complex conditions are inapplicable to the simplified formulas and assumptions of the theoretical and analytical approaches.
The method of numerical simulation is ranked third. Due to the rapid development of numerical simulation tools, methods such as the discrete element method [21], numerical manifold approach [22], boundary element method [20], and finite element method (FEM) [19] have become extensively used in slope stability research. The combined finite-discrete element method was employed by Sun et al. [23] to simulate the progressive failure of slopes composed of joined rocks. Ma et al. [24] addressed the issue of slope stability when saturated and unsaturated seepage was present through the implementation of rapid Lagrangian analysis of continua under complex stress conditions. We investigated the kinetic elements of slope instability by employing a particle flow code [25]. Haghnejad et al. [26] investigated the effect of blast-induced vibration on slope stability through the application of dynamic pressure in three-dimensional unique element codes. Song et al. [27] employed the enhanced smoothed-particle hydrodynamics method to ascertain the slope safety factor. In their investigation, Zhang et al. (28) utilised an appropriate failure process evaluation to assess the stability of the steep rock slope and determine the mode of collapse during excavation operations. Additionally, numerical simulations have been employed by some researchers in conjunction with mathematical methodologies to assess the stability of the slope. Dyson and Tolooiyan [29] computed slope stability probability and FOS utilising random FEM and Monte Carlo simulation. Although numerical simulation methods are user-friendly, constitutive models and mechanical factors have a significant impact on the accuracy [30].

The machine learning (ML) algorithm constitutes the fourth strategy. In an effort to improve slope stability prediction models, scientists have endeavoured to integrate ML algorithms as the number of slope cases has grown. Forecasts for both the stability status and FOS are accessible. An artificial neural network was utilised by [31] to estimate the FOS and slope stability (SS) from 46 slope instances obtained by Sah et al. [32]. FOS was computed employing genetic programming and support vector machines, respectively [33-34]. Amirkiyaei and Ghasemi [35] constructed two tree-based models to assess circular failure gradients utilising data from 87 cases. Zhou et al. [36] predicted the SS from 221 slope examples using the gradient-boosting machine. The predictive performance of the optimum-path forest in combination with k-nearest neighbor provides satisfactory outcomes to predict slope stability [37]. Extreme learning machines including several data driven frameworks were utilised in slope stability prediction study [38]. Mahmoodzadeh et al. [39] employed a diverse range of techniques in order to identify FOS, such as deep neural networks, support vector regression, k-nearest neighbours (k-NN), and long-short-term memory. In terms of slope
stability prediction, each of the ML methods mentioned above performed exceptionally well. However, to enhance its credibility, an extensive variety of slope stability scenarios are required.

To the best of the author's knowledge, the application of data dimension reduction algorithm in conjunction with both unsupervised and supervised machine learning algorithms, optimized through metaheuristic algorithms, has not been explored for predicting coal pillar stability. Hence, the objective of this research is to assess the effectiveness of the PCA-GMM based WOA-XGBoost algorithm in slope stability prediction. An extensive dataset was compiled, consisting of 404 historical slope instances that exhibited various levels of collapse risk. This dataset considered both the geological conditions and the process of slope deterioration. The results suggest that the hybrid PCA-GMM with WOA-XGBoost model provides the most dependable predictions pertaining slope stability.

2. Data Acquisition

There is a high probability that the slope failure surfaces will be located close to the possible slide surface. Excessive shear stress on the possible slip surface, as a result of excavation at the slope's base or water seepage at its peak, leads to local slope instability. The physical-mechanical characteristics of the possible slide surface, the fundamental geometrical parameters, and the external triggering variables are the three primary determinants of slope stability, according to several engineering instances and theoretical investigations. Considering their independent behavior and the ease with which their values can be obtained, the six attributes were taken into consideration for this study as depicted in Table 1.

Table 1. Attributes considered in determining slope stability.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Symbol</th>
<th>Unit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit weight</td>
<td>$\gamma$</td>
<td>kN/m$^3$</td>
<td>Unit weight is the weight of soil or rock expressed as a number per unit volume.</td>
</tr>
<tr>
<td>Cohesion</td>
<td>$c$</td>
<td>kPa</td>
<td>The portion of shear strength that is not influenced by the</td>
</tr>
</tbody>
</table>
typical effective tension in soil or rock motion is determined by cohesion.

The internal friction angle is a parameter used to quantify the shear strength of soil or rock.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal friction angle</td>
<td>( \phi )</td>
<td>-</td>
</tr>
<tr>
<td>Slope angle</td>
<td>( \beta )</td>
<td>-</td>
</tr>
<tr>
<td>Slope height</td>
<td>( H )</td>
<td>m</td>
</tr>
<tr>
<td>Pore pressure ratio</td>
<td>( r_u )</td>
<td>-</td>
</tr>
</tbody>
</table>

The angle formed by the inclined slope plane and the slope base plane is referred to as the slope angle.

The vertical distance between the base of the slope and the highest point of the slope defines slope height.

When the pressure in the pores is divided by the pressure in the overburden, the resulting ratio is called the pore pressure ratio.

The 404 slopes with failure risk from different nations were compiled into a database for this investigation [32,36,57,63-72]. Slope stability can be classified as either stable (207 instances) or failed (197 instances). Table 2 displays the findings of the statistical assessment conducted.
on the slope stability dataset. Figure 1 displays the distribution of slope stability over the entire dataset.

### Table 2. Statistical analysis of slope stability data

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Minimum</th>
<th>Median</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$ (kN/m$^3$)</td>
<td>10.06</td>
<td>21.38</td>
<td>31.3</td>
<td>21.69</td>
<td>3.84</td>
</tr>
<tr>
<td>$c$ (kPa)</td>
<td>0</td>
<td>29.7</td>
<td>300</td>
<td>39.38</td>
<td>40.54</td>
</tr>
<tr>
<td>$\phi$ (°)</td>
<td>0</td>
<td>28.27</td>
<td>57.36</td>
<td>27.74</td>
<td>9.63</td>
</tr>
<tr>
<td>$\beta$ (°)</td>
<td>4.24</td>
<td>34.03</td>
<td>59.35</td>
<td>34.19</td>
<td>10.86</td>
</tr>
<tr>
<td>$H$ (m)</td>
<td>3.45</td>
<td>51</td>
<td>565</td>
<td>84.26</td>
<td>94.97</td>
</tr>
<tr>
<td>$r_u$</td>
<td>0</td>
<td>0.2</td>
<td>0.75</td>
<td>0.18</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Fig 1. Distribution of slope stability over the entire dataset.

A violin plot is a robust data visualisation technique that integrates the information from a box plot and a kernel density plot. It efficiently demonstrates the pattern of distribution and general information of a dataset. The breadth of the violin at a specific point reflects the concentration of data at that value. Violin plots are useful for comparing different levels in a dataset, as they offer a full depiction of both the average and the variability of the data. This visual depiction is especially valuable in the process of examining data, enabling researchers to promptly
comprehend the distributional attributes of their data. As a result, they are extensively utilised in exploratory data analysis and statistical reporting. Figure 2 displays the violin plots of six attributes. It can be seen from Figure 2 that $\gamma$, $\phi$, and $\beta$ were distributed in a balanced manner, with medians located roughly in the centre of the box plots. However, there are a few outliers in the case of $c$, $H$, and $r_u$.

Fig 2. Violin plot of the input attributes

A heatmap is a graphical representation of data in a matrix style, with values represented by a colour spectrum. By means of this graphical method, patterns, trends, and variations within a dataset are notably revealed. By associating the colour intensity of the data, it represents with the magnitude of the cell in the matrix, it becomes possible to rapidly distinguish between high and low values. Heatmaps are frequently employed in the field of data analysis to offer valuable insights into the preponderance of values, correlation structures, and relationships between variables. Figure 9 illustrates the heatmap representing the Pearson correlation coefficient ($R$) among each attribute. As shown in Figure 3, the correlation between all variables was below 0.5, with the maximum correlation measuring 0.41. This finding suggests that the degree of correlation between the attributes was inadequate. As a result, each attribute was reasonably autonomous and significant in predicting slope stability.
Figure 4 shows the slope stability dataset's correlation pair plots, which help to understand the entire dataset's distribution. The diagonal line displayed the distribution plots of the six qualities, whereas the non-diagonal line showed the correlation scatter plots between indicators. The disparities in the attribute distributions between the two slope conditions were insignificant, and no significant relationship could be found between the attributes.

Thus, it proved challenging to categorise the slope stability status using a single attribute; for improved accuracy, it would be beneficial to combine the effect of all attributes.
Fig 4. Scatter plots illustrating the correlation between six attributes.

3. Methodology

3.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely recognised technique that generates a series of optimal linear approximations for a given set of high-dimensional data [51]. It has garnered significant attention in numerous academic works. PCA is widely utilised throughout various fields of investigation, ranging from neuroscience to computer graphics, due to its simplicity and nonparametric nature in extracting pertinent information from complex data sets [52-53]. PCA offers a clear strategy for reducing the dimensionality of a complex data collection with no extra effort.
Figure 5 (a) illustrates a set of values containing two parameters that have been evaluated within the X-Y coordinate scheme. The U axis represents the primary trend of data variation, while the V axis, which is orthogonal to it, represents the second significant orientation. By transforming each (X, Y) position into its equivalent (U, V) value, the information becomes asymmetrical, suggesting that there is no covariance between the U and V variables. PCA establishes the coordinate system that is defined by the primary patterns of variability for a specific dataset. The U-V axis system depicted in Figure 5 (b) represents this particular axis arrangement. The vectors U and V are commonly known as the principal components. In this new context, it is crucial to note that the level of variation is greater along the U axis than the V axis. PCA generates novel variables by combining the existing variables in a linear manner.

### 3.2 Gaussian mixture model (GMM)

A mixed distribution of several Gaussian distributions, the Gaussian mixture model (GMM) is a popular tool for characterising the mixed density distribution. When machine learning is applied to classification problems, the distribution characteristics of intra-class data in the feature space can be effectively characterised [54]. In recent years, the GMM has found widespread application in rock engineering, geotechnical engineering, and structural design [55-56]. Figure 6 represents the flowchart of GMM.
The GMM comprises $k$ distributions with Gaussian characteristics. These Gaussian distributions are used to create distinct classifiers. The probability density function of a GMM is formed by linearly combining these Gaussian distributions:

$$ p(X | \beta) = \sum_{i=1}^{K} a_i f(x_i | \mu_i, \Sigma_i) $$

(1)

In Equation (5), the parameter $\beta = (a_1, \ldots, a_k, u_1, \ldots, u_k)$ represents a Gaussian mixture distribution. The $a_i$ variable is the weighting coefficient, which indicates the likelihood of selecting this component. Whereas $f(x_i | \mu_i, \Sigma_i)$ is the Gaussian probability density distribution of the component defined as follows:

$$ f(x_i | \mu_i, \Sigma_i) = \frac{1}{(2\pi)^d |\Sigma|^{1/2}} e^{-\frac{1}{2}(x_i - u_i)^T \Sigma_i^{-1}(x_i - u_i)} $$

(2)

In Equation (6), $\mu_i$ represents the mean, while $\Sigma_i$ is the covariance matrix.
The Gaussian distributions of the GMM are shaped like ellipsoids, with their centres located at the mean vector \( \mu_i \). The covariance matrix determines further geometric properties of these distributions, such as their volume, shape, and orientation. The covariance matrix are able to be configured by the eigen-decomposition of the form \( \Sigma_i = \lambda_i D_i A_i D_i^T \). The \( \lambda_i \) and diagonal matrix \( A_i \) are utilised to ascertain the configuration of the volume and density contours, respectively. In addition, the orthogonal matrix \( D_i \) is employed to manipulate the orientation of the ellipsoid. The three parameters can be assigned as either constant or variable for distinct Gaussian distributions. Fiueiredo and Jain (2002) introduced an unsupervised learning technique known as the F–J algorithm, which has gained significant popularity for obtaining the GMM [57]. Additional information can be discovered in the publications of [57-58].

3.3 Whale Optimisation Algorithm (WOA)

The Whale Optimisation Algorithm was initially proposed by Mirjalili and Lewis [59]. This method is a type of metaheuristics that is based on swarms. It mimics the exploration and exploitation phases of humpback whales' behaviours when they appear for and feed on prey. Humpback whales employ a strategic technique in nature where they first locate their food, typically a group of small fish or plankton. They then dive to a depth of approximately 12 metres and initiate an upward spiral trajectory towards the surface. During this ascent, they release bubbles to create a virtual nett, effectively herding and corralling the prey. The technique used for this purpose is known as bubble-net feeding [60-61]. The mechanism of WOA is shown in Figure 7.

![Fig 7. Humpback whales exhibit bubble-net feeding behaviour.](image_url)
To be more precise, the algorithm takes into account a group of \( n \) whales that are actively searching for food. The positions' coordinates, \( \vec{X}_i \) serve as possible choices for the parameters that require optimisation. During the exploitation phase, the most optimal position, \( \vec{X}^* \) is regarded as the target prey. In order to imitate the behaviour of prey that is shrinking and being encircled, the position of the \( i \)th whale in the \((t + 1)\)th iteration is updated using the following formula:

\[
\vec{X}_i(t + 1) = \vec{X}^*(t) - AD^-
\]

(3)

\[
\vec{D} = |C\vec{X}^*(t) - \vec{X}_i(t) |
\]

(4)

Whereas \( A \) and \( C \) being coefficients that can be selected as follows:

\[
A = 2a . r \text{ and } C = 2r
\]

(5)

where \( r \) is an arbitrary number between 0 and 1, and \( a \) decreases linearly from 2 to 0.

Furthermore, the position spiral update technique is mimicked using the following equation:

\[
\vec{X}_i(t + 1) = \vec{D}'E^{bl} \cos2\pi l + \vec{X}^*(t)
\]

(6)

Where \( \vec{D}' = |\vec{X}^*(t) - \vec{X}_i(t)| \), \( b \) and \( l \) are both predefined parameters and a random number between -1 and 1.

A random number \( p \in [0, 1] \) can be utilised to ascertain the specific behaviour exhibited by a whale in a given iteration, given that whales simultaneously swim in decreasing circles and a spiral trajectory.

The algorithm has made the assumption that the prey is situated at the present optimal position during the exploitation phase. This is not always the case, as whales frequently conduct arbitrary searches based on the location of other whales in their cohort.

The exploration is represented by the subsequent equations:

\[
\vec{X}_i(t + 1) = \vec{X}_{rand}(t) - A\vec{D}
\]

(7)

\[
\vec{D} = |\vec{X}_{rand}(t) - \vec{X}_i(t) |
\]

(8)

where \( \vec{X}_{rand} \) represents the position of a whale at random within the population, excluding the \( i \)th whale.
If $A \leq 1$, then the determination of whether to conduct a local search (exploitation) or a global search (exploration) is reached. Further details regarding WOA, including its latest developments, applications, and variants, can be found in [62-64] and the cited sources.

3.4 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an algorithm for ensemble learning in machine learning techniques [65-66]. The incorporation of statistical boosting techniques allows it to incorporate basic classification and regression trees (CART's) [67-68]. By using boosting to produce a "consensus prediction model", the model's estimation precision is improved by creating many trees instead of just one. XGBoost builds the tree by adding each previous tree's residuals to the current tree in an orderly manner. Therefore, by displaying the mistakes made by previous trees, the result tree builds the overall prediction. When the loss function is small, this interaction between successive model structures can be described as a gradient descent that improves prediction at each stage via emergence of a new tree, ultimately reducing the fall. When the training error reaches a certain threshold or when the maximum number of trees has been reached, the new tree will not be expanded any further. Adding random sampling to gradient boosting significantly improves its estimation precision and execution speed; this all-encompassing method is called "stochastic slope boosting". Figure 8 depicts the organisational structure of a Tree model in the XGBoost method.

$$\bar{u}_i = u_i^0 + \eta \sum_{k=1}^{E} f_k(V_i)$$  \hspace{1cm} (9)
Whereas \( \bar{u}_i \) shows the predicted result for \( i \)th data, \( u_i^0 \) represents the starting point, which is the average of the actual attributes in the training dataset, and \( V_i \) is the attribute vector. The estimator’s number is \( E \), and for every estimator \( f_k \), which ranges from 1 to \( E \), there is an independent tree structure. In order to prevent overfitting and ensure that the model runs smoothly when adding a new tree, the parameter \( \eta \) is used as the learning rate. An important point to keep in mind while discussing machine learning algorithms is the possibility of overfitting.

\[
\begin{align*}
 u_i^{-k} & = u_i^{-(k-1)} + \eta f_k
\end{align*}
\]

(10)

In relation to Eq. 4, during the \( k \)th phase, an estimator \( k \)th is coupled to the framework. The \( k \)th anticipated outcome, \( u_i^{-k} \) is determined from the predicted result at the previous stage, \( u_i^{-(k-1)} \).

The developed \( f_k \) of the supplement \( k \)th estimator is provided in Eq 6.

\[
\text{obj} = \gamma N + \sum_{a=1}^{N} [G_a \omega_a + \frac{1}{2} (H_a + \lambda) \omega_a^2]
\]

(11)

Whereas \( N \) represents the \( k \)th tree's leaves and \( \omega_a \) represents the weight of each leaf from 1 to \( N \), \( \gamma \) and \( \lambda \) are regularisation parameters that are employed to avoid overfitting by ensuring structural coherence. The characteristics \( G_a \) and \( H_a \) represent the total data associated with a leaf of the first loss function gradient and the previous one, respectively.

The process of creating the \( k \)th tree involves dividing a single leaf into several segments. Utilising the GAIN characteristics offered by Equation 7 facilitates comprehension of this process.

\[
\text{Gain} = \frac{1}{2} \left[ \frac{C_L^2}{D_L + \lambda} + \frac{C_R^2}{D_R + \lambda} + \frac{(C_L + C_R)^2}{D_L + D_R + \lambda} \right]
\]

(12)

Whereas \( C_R \) and \( D_R \) represent the aligned right leaf, and \( C_L \) and \( D_L \) represent the aligned left leaf that follows the split. Typically, the splitting criterion is applied when the gain attribute is close to zero. The gain parameter indirectly affects the regularisation parameters \( \gamma \) and \( \lambda \); a greater regularisation parameter will cause a drop in the gain parameter, which will prevent the convolution of the excruciating leaf. However, this will reduce the framework's adaptability to the training dataset.

When it comes to efficiently and effectively executing XGBoost—a non-proprietary technique—is the best option. Ensembles are constructed using decision tree frameworks. The trees are sequentially appended to the ensemble and adjusted to enhance the prediction.
accuracy of the previously generated framework. Ensemble machine learning techniques of this nature are commonly referred to as boosting. The frameworks are constructed using a variety of stochastic gradient descent optimisation algorithms and differentiable loss functions. This technique is called "gradient boosting" because the gradient loss decreases as the framework is fitted.

4. Results and Discussion

4.1. Establishing a comprehensive predictive model

This study introduces PCA-GMM Based WOA-XGBoost model to predict slope stability. Firstly, a total of 404 events of slope stability were gathered using easily accessible input parameters. Secondly, PCA has been applied to reduce the dimension of the slope stability dataset. Thirdly, gaussian mixture models (GMM) was employed to categorize the PCA dataset into distinct clusters. Additionally, a training set comprising around 75% of the GMM dataset is selected, while a testing set comprising around 25 % of the GMM data is allocated. Fourthly, the parameters of extreme gradient boosting (XGBoost) were optimized using whale optimization algorithm (WOA). Lastly, the performance of the suggested model was assessed using several performance indices. Figure 9 represent the flow chart of the study.

![Flow chart of the study.](image)

4.2 Performance assessment metrics
A confusion matrix is an essential tool for assessing the performance of a classification-based machine learning model. It offers a comprehensive analysis of the model's predictions by comparing them to the actual classification labels of a dataset. The matrix is structured into four fundamental elements: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). TP depicts cases accurately classified as positive, TN denotes instances accurately classified as negative, FP indicates instances inaccurately classified as positive, and FN means instances inaccurately classified as negative. The confusion matrix is an integral tool for computing crucial performance metrics, including accuracy, precision, recall, and F1-score [69-72]. These measurements provide a detailed comprehension of the model's strengths and weaknesses, especially in situations when there are disparities in class distribution.

Accuracy serves as an indicator of the model's overall correctness. It represents the proportion of instances that are accurately predicted in relation to the total number of instances.

\[
\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (13)
\]

Accuracy levels span from 0 to 1, with 1 denoting perfect exactness and 0 denoting an inaccurate model.

Precision denotes the degree of reliability with which the model generates positive predictions. It signifies the proportion of accurately anticipated positive instances in relation to the overall number of predicted positives.

\[
\text{Precision} = \frac{\text{True Positives}}{\text{True Positive} + \text{False Positives}} \quad (14)
\]

Recall (Sensitivity or True Positive Rate) quantifies the model's capability to identify and retain every positive instance. It signifies the proportion of accurately predicted positive observations in relation to the total number of true positives.

\[
\text{Recall} = \frac{\text{True Positives}}{\text{True Positive} + \text{False Positives}} \quad (15)
\]

The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall, especially when there is an uneven class distribution.

\[
\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{True Positive} + \text{False Positives}} \quad (16)
\]

4.3 Establishment of the proposed predictive model
Python is widely recognised for its adaptability, intelligibility, simplicity, and substantial community backing. It is a high-level, multipurpose programming language. Python has achieved significant popularity and is now a prominent programming language utilised in a multitude of fields, such as automation, data science, web development, and artificial intelligence. Multiple programming paradigms, such as object-oriented, procedural, and functional programming, are supported by the language, allowing developers to select the approach that best suits their requirements. Python also supports a robust ecosystem comprising third-party frameworks, libraries, and tools, including TensorFlow for machine learning, Django for web development, and NumPy for scientific computation. Python has experienced substantial growth in prominence within the domains of civil and structural engineering, specifically in relation to data-driven problems involving classification, regression, and clustering.

A set of observations of potentially correlated variables is transformed into a set of values of linearly uncorrelated variables via PCA, which employs an orthogonal transformation. These are referred to as "principal components". The optimal matrix transformation is specified with the maximum variance feasible in the first principal component. It indicates that the model attempts to incorporate the maximum amount of variability present in the data. After the initial component, each succeeding component strives to achieve the maximum variance feasible while remaining orthogonal to the antecedent components. Based on the three principal components, the six actual attributes of 404 patterns of slope stability data utilised in this study were plotted. Furthermore, three principal components established space can identify an adequate visualisation while preserving the original data characteristics after data reduction. Figure 10 displays the three-dimensional structure of PCA. The principal components of the entire slope stability are presented in Table 3 after incorporating PCA. A conversion is performed from the original database (404 × 6 matrix structure) to a (404 × 3) matrix structure.
Table 3. Overview of principal components after incorporating data dimensionality reduction

<table>
<thead>
<tr>
<th>Serial Number</th>
<th>Principal component 1</th>
<th>Principal component 2</th>
<th>Principal component 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.80954</td>
<td>-0.20187</td>
<td>0.270577</td>
</tr>
<tr>
<td>2</td>
<td>-0.8729</td>
<td>-0.57937</td>
<td>1.276173</td>
</tr>
<tr>
<td>3</td>
<td>-0.58628</td>
<td>-0.54707</td>
<td>0.980239</td>
</tr>
<tr>
<td>4</td>
<td>-0.55693</td>
<td>0.519179</td>
<td>1.427929</td>
</tr>
<tr>
<td>5</td>
<td>0.160969</td>
<td>-0.2662</td>
<td>-0.44713</td>
</tr>
<tr>
<td>…..</td>
<td>…..</td>
<td>…..</td>
<td>…..</td>
</tr>
<tr>
<td>400</td>
<td>-0.12934</td>
<td>-0.96184</td>
<td>-0.11344</td>
</tr>
<tr>
<td>401</td>
<td>0.726892</td>
<td>-1.07633</td>
<td>-1.72919</td>
</tr>
<tr>
<td>402</td>
<td>-0.84932</td>
<td>-0.83652</td>
<td>-0.71781</td>
</tr>
<tr>
<td>403</td>
<td>-0.32026</td>
<td>-0.96985</td>
<td>2.409304</td>
</tr>
<tr>
<td>404</td>
<td>-0.0696</td>
<td>-0.9818</td>
<td>0.572807</td>
</tr>
</tbody>
</table>

GMM are probabilistic models that distribute the data elements into distinct clusters via the soft clustering technique. The data is assumed to be generated from a combination of various cluster-specific Gaussian distributions. With GMM, data points can be assigned probabilities
that let them belong to numerous clusters at once. Every cluster is represented by one of a fixed number of Gaussian distributions, according to GMM. Therefore, it is common practise for a GMM to cluster data points from the same distribution. Two distinct categories have been established for slope stability based on the current literature: stable (depicted by 0) and failed (depicted by 2). The Python Scikit-learn library was utilised in the construction of this GMM model [73-74].

To assess the clustering impact of GMM, the "Silhouette coefficient" is utilised. This value is determined by combining the two elements of cohesion and separation degree and observing the variation in the silhouette coefficient. The silhouette coefficient is bounded within the range of -1 to 1. A greater silhouette coefficient signifies an improved clustering effect; for samples joined in closer proximity, the silhouette coefficient is smaller; for samples separated further apart, the silhouette coefficient increases in magnitude.

In order to prepare the dataset for GMM, we utilised a preprocessing approach that involved the use of standardisation technique on the PCA obtained dataset. A series of iterations have been executed throughout the GMM analysis, as depicted in Figure 10. The reliability and consistency of the clusters is evidenced by the silhouette score of 0.53, which was acquired during the eighty-ninth iteration of the slope stability dataset derived from PCA. hence, upon conducting a statistical evaluation of the performance, it is evident that the GMM exhibits the capacity to precisely categorise slope stability failures and stable conditions.
It is crucial to identify the optimal hyperparameters for XGBoost and other machine learning models [75-76]. This procedure is carried out so that the models can be as generalizable as possible to the greatest extent possible. In most cases, suboptimal performance results as a consequence of the adoption regarding default hyper-parameters, reliance on experience-based model selection, or trial-and-error methodologies for hyper-parameter selection. When formulating machine learning-based slope stability prediction models, it is imperative to implement systematic and automated methodology for model selection.

As a result, there has been significant progress in the application of metaheuristic techniques, such as heuristic optimisation algorithms (e.g., differential flower pollination, history-based...
adaptive differential evolution, and others), to determine the optimal parameter values for machine learning models [77-80]. The utilisation of the WOA to optimise the XGBoost parameter configuration search is a component of the slope stability database. The foraging and prey-seeking behaviours of humpback whales serve as inspiration for WOA. The justification for selecting the WOA is that it is a contemporary and state-of-the-art metaheuristic that has been effectively implemented in engineering applications. The whale population (pop size) and maximum number of iterations (epoch) are both specified as 50 and 150 for WOA. The XGBoost parameters were applied with the established configuration. When PCA-GMM are combined with the GWO- XGBoost machine learning approach, the prediction error for slope stability events in the testing datasets is limited to a single instance as displayed in Figure 12. The statistical outcomes of the proposed mechanism are presented in Table 4. The combined accuracy of the proposed mechanism is 99 percent. Hence, the outcomes suggest that the proposed hybrid PCA-GMM-based WOA-XGBoost model proposed in this study demonstrates the utmost reliability. By combining a metaheuristic algorithm technique with an unsupervised-supervised machine learning approach, slope-related catastrophes may be substantially reduced, thereby improving the sustainability and safety of mining operations.

![Fig 12. Confusion matrix of the proposed PCA-GMM in combination with WOA-XGBoost on slope stability database](image-url)
Table 4. Statistical finding of the proposed algorithm on slope stability database

<table>
<thead>
<tr>
<th></th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1-score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstable</td>
<td>99</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td>Stable</td>
<td>100</td>
<td>93</td>
<td>96</td>
</tr>
<tr>
<td>Final Accuracy</td>
<td></td>
<td></td>
<td>99</td>
</tr>
</tbody>
</table>

5. Contribution and limitations

The main advantage of this study is the proposal of integrated data-driven techniques for predicting slope stability. This paper provides the following advancements to slope analysis:
(a) By utilising confusion matrices, precision, recall, and F1-score, it is possible to comprehensively assess the stability and resilience of classification models. (b) The integration of artificial intelligence techniques using machine learning algorithms holds great potential for solving issues related to classification. (c) This study offers a reliable model recommendation for predicting slope stability, which can serve as a valuable reference for future researchers engaged in similar investigations. (d) The methodology presented in this paper exhibits considerable potential for wider application.

An evident flaw of the current analysis is the disregard for other elements that could impact slope stability, such as the amount of vegetation, rock structure, and engineering disruption. The very small size of the slope dataset poses an additional limitation. The effectiveness of AI approaches that have been suggested will enhance when additional data becomes available. Another limitation is that the assessment of slope stability is more precisely classified as a regression or multiclass classification issue, with the latter being the focus of current research.

There are numerous physical and geometric variables that must be precisely estimated in order to develop a method for predicting slope stability, which is a very difficult task. These prediction methods must be intuitive to use and deliver a substantial degree of precision. Additionally, it is crucial that the prediction be generated within a brief computational timeframe, given the engineering application's requirement for prompt estimations. The complexity associated with devising prediction approaches for slope stability has been heightened by these requirements. On the contrary, dependable, and precise slope stability prediction has the capability to pinpoint areas susceptible to collapse, ascertain the most suitable retaining structures, and devise effective excavation strategies.

6. Conclusion
Predicting slope stability is an essential task in mining, civil, tunnelling, and geotechnical engineering projects. This study examined the effectiveness of the PCA-GMM when coupled with the WOA-XGBoost algorithm for the prediction of slope stability. A comprehensive collection of 404 historical slope instances with varying degrees of collapse risk was obtained, taking into account both the slope deterioration process and geological circumstances. The performance of the suggested model was assessed using a number of performance assessment indicators. The results suggest that the proposed hybrid PCA-GMM driven WOA- XGBoost model demonstrates the highest level of reliability in the prediction of slope stability.

Slope instability is a comprehensive environmental issue that impacts every geographic area. Slope stability has to be carefully accounted for geotechnical, mining, civil, and tunnelling engineering projects. Severe economic losses, casualties, and property damage may result from a calamitous event involving slope instability. The stability of slopes affects the safety of numerous construction projects, including mountain roads, earth dams, and retaining walls. Consequently, slope failures are highly undesirable occurrences that give rise to a number of catastrophic events in many different countries. As the economy grows and the population expands, an increasing number of man-made structures are being built in response to slope failures. Consequently, the rapid estimation of slope stability is of the utmost importance prior to making any critical decisions concerning slope design and remedial support.

Ensuring the safety of mining operations and preventing catastrophic events require an early warning system for slope stability in the mining industry. Slope stability pertains to the capacity of an embankment or slope to endure the applied forces without experiencing substantial deformation or failure. Slopes utilised in mining operations may be susceptible to a number of hazards, including weathering, geological conditions, seismic activity, and human activities. The implementation of a predictive early warning system necessitates the fusion of diverse methodologies and technologies in order to monitor and evaluate slope stability conditions.

Mining operations necessitate the cooperation of data scientists, geotechnical engineers, and mining specialists to implement a comprehensive early warning system for slope stability. It is imperative to conduct routine training and exercises to ensure that personnel are adequately equipped to react efficiently in the case of a warning. By implementing the suggested strategy, the probability of catastrophic incidents will be reduced, and the sustainability and safety of mining operations will be improved.
7. Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

8. References


