A comparative study of supervised machine learning approaches for slope failure production

Ashanira Mat Deris[1], *Badariah* Solemon[2*], and *Rohayu* Che Omar[2]

[1]Fakulti Teknologi Kejuruteraaan Kelautan dan Informatik, Universiti Malaysia Terengganu 21030 Kuala Nerus, Terengganu
[2]Institute of Energy Infrastructure, Universiti Tenaga Nasional, Jln IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia

Abstract. Over the years, machine learning, which is a well-known method in artificial intelligent (AI) field has become a new trend and extensively applied in various applications to solve a realworld problem. This includes slope failure prediction. Slope failure is among the major geo-hazard phenomenon which gives the significant impact to the environment or human beings. The estimation of slope failure in slope stability analysis is a complex geotechnical engineering problem that involves many factors such as geology, topography, atmosphere, and land occupancy. Generally, slope failure can be estimated based on traditional methods such as limit equilibrium method (LEM) or finite equilibrium method (FEM). However, beside the methods are quite tedious and time consuming, LEM and FEM have their own limitations and do not guarantee the effectiveness when dealing against problem with various geometry or assumptions. Hence, the introduction of machine learning approaches provides the alternative tools for the prediction of slope failure. Current study applies two mostly used supervised machine learning approaches, support vector machine (SVM) and decision tree (DT) to predict the slope failure based on classification problem using historical cases. 148 of slope cases with six input parameters namely "unit weight, cohesion, internal friction angle, slope angle, slope height and pore pressure ratio and factor of safety (FOS) as an output parameter", was collected from multinational dataset that has been extracted from the literature. For development of the prediction model, the slope data was divided into 80% training data and 20% testing data. The prediction result from testing data was validated based on statistical analysis. The result shows that SVM model has outperformed DT model by giving the prediction accuracy of 97%. ith the advent of technology and the introduction of computational intelligent methods, the prediction of slope failure using the machine learning (ML) approach is rapidly growing for the past few decades. This study employs an "artificial neural network" (ANN) to predict the slope failures based on historical circular slope cases. Using the feed-forward back-propagation algorithm with a multilayer perceptron network, ANN is a powerful ML method capable of predicting the complex model of slope cases. However, the prediction result of ANN can be improved by integrating the statistical analysis method, namely grey relational analysis (GRA), to the ANN model. GRA is capable of identifying the influencing factors of the input data based on the correlation level of the reference sequence and comparability sequence of the dataset. This statistical machine learning model can analyze the slope data and eliminate the unnecessary data samples to improve the prediction performance. Grey relational analysis-artificial neural network (GRANN) prediction model was developed based on six slope factors: unit weight, friction angle, cohesion, pore pressure ratio, slope height, and slope angle, with the factor of safety (FOS) as the output factor. The prediction results were analyzed based on accuracy percentage and receiver operating characteristic (ROC) values. It shows that the GRANN model has outperformed the ANN model by giving 99% accuracy and 0.999 ROC value, compared with 91% and 0.929.

1 Introduction

Landslide is among the most destructive popular geological hazards worldwide. It poses a great danger to the environment as well as safety of human life and damaging the property and resources [1]. According to Korup and Stolle [2], landslide is generally defines as forms and processes that generated from the downward and outward movement of hillslide-forming components, such as dirt, rock or debris, caused by the gravity force, and usually which the existence of water. Landslides not only occurred in steep terrain, but also may triggers in gently sloping to almost flat terrain. The major causes of landslides are due to the slope failure. In order to mitigate the impacts and consequences of landslides, researchers are progressively investigating multivariate data analysis approaches in the fields of machine learning (ML) or data mining to estimate possible occurrences of slope failure from the historical distribution patterns. Traditionally, geotechnical engineers employ various approaches for analyzing the stability of slope, including limit equilibrium method (LEM), finite element method (FEM), upper bound limit analysis, maximum probability, genetic programming, etc [3,4]. However due to the some limitation of the traditional methods, the ML method was introduced to

* Corresponding author: badariah@uniten.edu.my

predict the slope failure in slope stability analysis. One of the limitations of the traditional methods is, it is difficult to mathematically describe the relationship between the significant factors of slope due to the complex mechanism of slope failure [5].

ML approaches such as artificial neural network (ANN), support vector machine (SVM) and decision tree (DT), has been successfully employed in this field to simulate the geotechnical problems [6]. This is due to the ability of these methods to establish the nonlinear equations between input and output set of data [7]. ANN is a sophisticated ML approach that mimicking the brain neurons to generate a solution for a problem. It has shows capability in prediction of complex model and has been widely used by previous researchers to analyze the slope stability [8, 9, 10]. However, ANN has been reported to have certain limitations such as low convergence speed and less generalization performance Samui and Kothari [7].

SVM is an effective ML approach based on the principal of "structural risk minimization" (SRM) to create the decision planes to define decision boundaries. SVM can guarantee greater accuracy for a prediction process in many practical applications in various field compared to other ML approaches [6, 11, 12, 13]. SVM has been successfully applied by previous researchers in various fields including financial [14], health [15]2019), agriculture [16], manufacturing [17] etc. In geo-engineering, SVM also is widely applied for slope stability analysis.

Samui [6] developed prediction models for slope stability using SVM and ANN. The factor of safety (FOS) value was predicted using regression problem while slope stability status was predicted using classification problem. The result shows that for both FOS and stability status, SVM gives better prediction compared with ANN by giving the accuracy of 85.71%. Qi and Tang [18] studied the performance of six different ML approaches to predict the stability of slope. The ML approaches used to develop prediction models were SVM, random forest (RF), gradient boosting machine (GBM), decision tree, logistic regression, and multilayer perceptron ANN (MPNN). The result found that SVM was efficient in terms of the accuracy and the true negative rate. With the advent of the new technology, SVM also evolved significantly to improve the performance of the prediction result. The prediction result is proven to be improved with the introduction of new variant of SVM. Kumar et al [19] applied three different SVM variants to predict the landslide prediction of Mandakini River Basin in India. The models were developed using basic SVM, proximal SVM (PSVM) and L2-SVM-modified finite Newton (L2-SVM-MFN). 2009 cases of landslide were divided into 50% training dataset and 50% testing dataset. The result showed that L2-SVM-MFN has outperformed PSVM and SVM with prediction value of 0.829, compared with 0.807 and 0.79, respectively. Lin et al [20] has developed prediction model of slope stability using ML approaches including SVM, random forest (RF), gravitational search algorithm (GSA) and NB. Six slope factors with 107 of domestic and worldwide slope cases were analyzed and measured to develop the prediction model. The result shows that GSA and RF have outperformed SVM and BN by giving the accuracy of 88.89%.

DT is a well-known ML approach and an efficient tool for prediction and classification. For the past decade, DT has been applied in many applications including geology [21, 18, 22] manufacturing [23], medical [24], agriculture [etc]. Bui et al [25] investigate the landslide susceptibility assessment in province of Huo Binh, Vietnam based on SVM, DT and Naive Bayes (NB). Ten factors were selected to computes the indexes of landslide susceptibility. The factors were soil type, slope angle, relief amplitude, rainfall slope aspect, lithology, distance to rivers, distance to roads, distance to faults, and land use. 118 landslides cases were divided into 70% training and 30% testing for all the models. The result shows that SVM prediction model has outperformed DT and NB. Park et al [26] analyzed 548 landslides in Gangneung-si, Korea to predict the landslides susceptibility using three different approaches of decision tree namely Chi-square automatic interaction detection (CHAID), exhaustive CHAID and Quick, Unbiased, and Efficient Statistical Tree (QUEST). 20 landslide factors were considered as input parameters with area under the curve (AUC) were used as output factor to develop the prediction models. The result shows that CHAID model has outperformed exhaustive CHAID and QUEST by giving the prediction result of 87.1% compared with 86.9% and 82.8% respectively. From the review of the previous literature, it shows that SVM and DT are able to give good prediction result for slope stability analysis including landslide prediction. In the current work, the SVM and DT are applied to predict the slope failure based on classification problem.

2 Methodology

2.1 Case study

To develop prediction models, historical cases of slope stability collected from a multinational dataset that has been extracted from the literature [27, 28, 29, 30]. The dataset consists of 148 slope cases with six input parameters namely "unit weight, internal friction angle, cohesion, slope angle, slope height and pore pressure ratio", and "factor of safety (FOS)" as output parameter. Two distinctive classes of FOS which are 1 and 0 refer to the "stable" and "unstable" slope, respectively. Table 1 shows the basic statistical of the slope cases.

Table 1. Basic statistical of the slope cases.

Slope parameters	Statistic		
	Minimum	Maximum	Standard Deviation
Unit weight (γ)	13.97	31.3	4.0193
Cohesion, (c)	4.95	300	47.1177
Internal friction angle (φ)	0	45	10.9554
Slope angle (β)	16	59	10.1382
Slope height (H)	3.6	511	138.2752

Pore water pressure ratio (ru)	0	45	3.6844

Two types of slope condition were identified from 148 slope cases: stable and unstable. Figure 1 shows the pie chart of the slope cases.

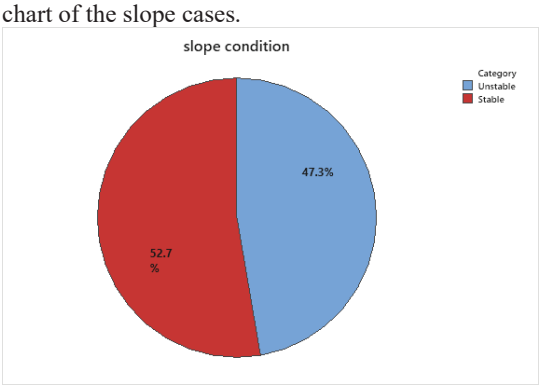

Fig.1. Pie chart of the slope cases.

From the figure, it can be seen that 52.7% of the slope are stable (78 cases) and 47.3% (70%) are unstable. Both types of slope condition are relatively balanced. To develop prediction models using ML approaches, the whole datasets need to be divided into two new subsets which are training and testing data. This is to ensure the generalization capability of the datasets. Basically, the training data is used to train the model and tuning the hyper-parameters while testing data is used for the prediction purpose to test its generalization capability. This study divides the datasets into 80:20 where 80% of the data is used for training and 20% is used for testing.

2.2 Support vector machine

SVM was founded by Vapnik [31], implementing the SRM principal is one of the most widely used ML approaches. The principal of SRM is an improvement of the Empirical Risk Minimisation (ERM) which is employed in neural network. The difference between SRM and ERM is where SRM minimise an upper bound of the expected risk while ERM minimise the error of the training data. This difference leads to the greater ability for SVM to generalize the goal of statistical learning to solve the problem. Generally SVM model consists of five major steps, which are illustrated in Figure 2.

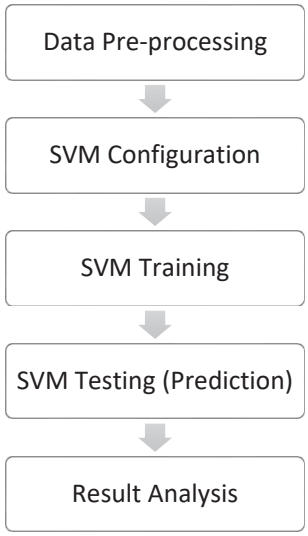

Fig. 2. Steps in SVM.

SVM problem can be solved based on classification or regression analysis [6]. For the classification problem, the objective for the algorithm is to obtain an optimal hyper plane with the maximum margin distance of two different classes in N-dimensional space that classifies the data points. Optimizing the gap from the margins offers some assurance so that potential data points can be identified with high levels of trust. Figure 3 shows the optimal hyper plane for two classifiers.

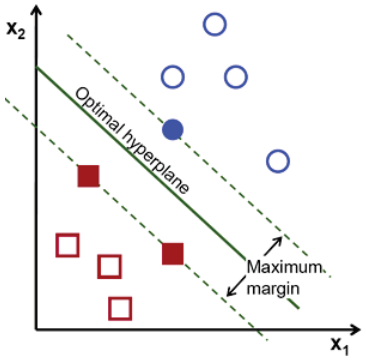

Fig. 3. Optimal hyperplane.

From Figure 3, it shows that the data point is linearly divided by the hyperplane with a feature class is defined by y= {1, -1}. SVM attempts to search for the largest margin between the two classes. Considers the training data as follows:

$$y_i\,(x_i w + b) - 1 \geq 0 \quad \forall_i, \tag{1}$$

then the data points for the above equality hold lie on the hyperplanes $x_i w + b = 1$and $x_i w + b = -1$. The margin can be expressed as follows:

$$margin = \frac{2}{||w||} \tag{2}$$

The maximum decision boundary margin can be calculated as in Eq (3).

$$max_{w,b} = \frac{2}{||w||} \tag{3}$$

Where x_i is the input parameters, y_i is the output parameter, *w* and *b* are the model parameters. The kernel

function is applied in SVM to measure the correlation between two inputs when the classification problem is non-linearly separable. It can describe the inner product of pair wise samples to reduce the dimensions of the feature vector.

To develop SVM model, the dataset need to be divided into training and testing dataset. Among 148 cases, 80% or 118 slope case of the data was selected as training and 20% or 30 slope cases was selected as testing dataset. The selection of training and testing datasets also considers the 5-fold cross validation. The K-fold Cross Validation (KCV) is a widely used, efficient, simple and reliable re-sampling method [32]. KCV divides the dataset into k individual substances so that all cases are taken from each subset for training and testing. SVM performance generally depends on the choice of the kernel function that satisfies the Mercer's theorem. Kernel function enables the operation to be carried out in the input space instead of the potentially high dimensional future space. Therefore, the inner product does not need to be evaluated in the future space. There are four kernel functions that are normally used by researchers namely, radial basis function (RBF) also known as Gaussian kernel, linear kernel, sigmoid kernel and polynomial kernel [1].

i. RBF kernel
RBF is a universal kernel function and the most commonly used kernel function in SVM [1]. RBF kernel function is applied in the nonlinear mapping of SVM. RBF kernel function is given as follow:

$$K(x,y) = \exp(-||x-y||2/(\ 2) \quad (4)$$

ii. Linear kernel
Linear kernel is generally described as:

$$K(x,y) = x\ y \quad (5)$$

iii. Polynomial kernel
The polynomial function is directional where the output is dependent on the direction of the two vectors in low dimensional space. It is because of the kernel dot product. Polynomial kernel function is given by:

$$K(x,y) = (x\ y + 1)p \quad (6)$$

iv. Sigmoid kernel
Sigmoid kernel function is also defined as Multi Layer Perception Kernel or Hyperbolic Tangent Kernel. Sigmoid kernel function is given as in Eq. (7).

$$K(x,y) = \tanh(kx\ y- \ ) \quad (7)$$

2.3 Decision Tree (DT)

DT is a supervised ML approach that adapts the tree-like graph or model for a decision making. DT consists of two types of tree, classification and regression trees. The basic idea of DT is to develop a model based on learning process of several decision rules, which is called decision tree, from the overall data. The total population of the data is divided into two or more homogeneous sets depending on the most important input variables divider. DT process could save the overall processing time as there is no need for variable transformation due to the fact that the tree structure will remain the same either with or without the transformation [33]. Using DT, a complex modeling relationship between the variables can be easily interpreted by the decision makers.

The process of DT consists of two steps which are making the trees and pruning the trees [33]. To avoid the unnecessary nodes in the tree, pruning process may required in most cases. The generation of DT consists of a "root node", "intermediate nodes" and "leaf nodes" with the branches that connect all the nodes. The samples from the root nodes must comply according to the growing rules where all samples in the node are divided into the same group of subsets. This is to ensure that the samples features in the same subset are as homogeneous as possible and two distinct subsets samples are as heterogeneous as possible. The intermediate nodes also need to abide with this rule. Each of the leaf nodes will presents the samples class. The iteration of the nodes will be completed when the maximum depth is reached with the characteristic of all the nodes are in similar group [34]. Figure 4 shows the example of schematic view of a portion of decision tree.

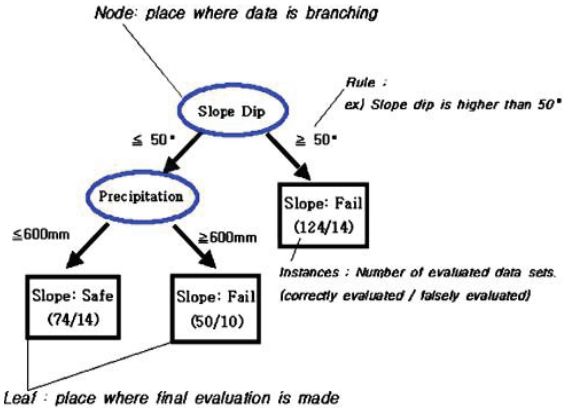

Fig. 4. Schematic view of decision tree [35].

3 Result and discussion

3.1 Confusion matrices

Confusion is a summary that associated with a classifier to explain the actual and predicted results. Confusion matrix is a basic tool to evaluate the confidence of classification result. The number of correct and incorrect prediction result are summarized for each class. Considers there are two classes for a classification problem, the confusion matrix as shown in Table 2.

Table 2. Confusion matrix.

		Predicted	
		Class 1	Class 2
Actual	Class 1	TP	FN
	Class 2	FP	TN

where
"TP = True Positive: The number of correct positive prediction

TN = True Negative: The number of correct negative prediction
FP = False Positive: The number of incorrect positive prediction
FN = False Negative: The number of incorrect negative prediction"

3.2 Prediction result

Table 3 shows the accuracy comparison of the prediction result for different kernel functions in SVM; Linear, RBF, Sigmoid and Polynomial. The accuracy for classification problems is calculated as in Eq. (8):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FB} \quad \text{x } 100\% \tag{8}$$

Table 3. Prediction results.

Kernel function	Accuracy
Linear	90.00
RBF	96.67
Sigmoid	63.34
Polynomial	93.34

From Table 3, it can be seen that RBF kernel function has outperformed Linear, Polynomial and Sigmoid function with the accuracy of 96.77%, compared with 90%, 93.33% and 63.33%, respectively. Table 4 shows the comparison prediction result of SVM-RBF and DT in terms of confusion matrices and Figure 5 shows the schematic view for the DT prediction.

Table 4. Confusion matrices of the prediction results.

		Actual		Accuracy
SVM-RBF		Positive	Negative	
Predictive	Positive	10	1	96.67
	Negative	0	19	
DT				
Predictive	Positive	15	0	83.34
	Negative	5	10	

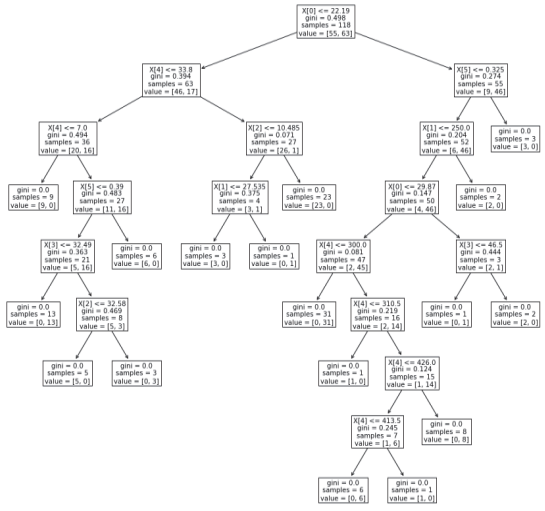

Fig. 5. Decision tree of slope prediction.

From the Table 4, the confusion matrix shows that for RBF-SVM, there are 11 cases of stable slope with 10 has correct prediction and 1 case is wrongly predicted while 19 cases of unstable slope with 19 cases are correctly predicted and zero case of inaccurately predicted. For DT, there are 15 cases of stable slope with all are correctly predicted while 15 cases of unstable slope with 10 cases are correctly predicted and 5 cases are wrongly predicted. As can be seen, the SVM model has higher accuracy compared to DT model which is 96.67%.

4 Conclusion

This study has implemented a novel approach for slope failure prediction using SVM and DT classification problem. Six slope parameters are identified ("Unit weight, cohesion internal friction angle, slope angle, slope height, pore pressure ratio, slope condition") from the 148 slope cases extracted from the historical worldwide slope data. The conclusions are as follows:

i. Among four kernel functions that are tested in SVM, RBF give better accuracy with 96.67%.
ii. The comparison between SVM-RBF and DT was found that SVM-RBF performed better than DT for slope failure prediction.

It should be noted that this research extends the body of knowledge in two aspects. First, the capability of SVM and DT as classifier to predict the slope stability and second, to verify that the performance of SVM model highly depends to the selection of kernel function.

This project was funded by UNITEN under Fundamental Research Grant Scheme FRGS/1/2018/STG08/UNITEN/02/1.

References

1. B. Feizizadeh, M.S. Roodposthi, T. Blaschke, J. Aryal, Arabian J. of Geosci.**10**, 122 (2017)
2. O. Korup, A. Stolle, Geology Today **30**, 1:26–33 (2014)
3. A. Mouyeaux, C. Carvajal, P. Bressolette, L. Peyras, P. Breul, C. Bacconnet, Compt. And Geotechnics **101**, 34−47 (2018)
4. L. He, Q. Tian, Z. Zhao, Q. Zhang, J. Zhao, Intr.J. Geomevhanics **186**, 04018044 (2018)
5. J. Zhou, E. Li, S.Yang, M. Wang, X. Shi, S. Yao, H. S. Mitri, Safety Sci. **118**, 505−518 (2019)
6. P. Samui, Env. Geology **56**, 2:55 (2008)
7. P. Samui, D. P. Kothari, Scientia Iranica **18**, 1:53–58 (2011)
8. M. Fakir, M. Ferentinou, *A Holistic Open Pit Mine Slope Stability Index Using Artificial Neural Networks*, in International Society for Rock Mechanics and Rock Engineering, January 2017 (2017)
9. M. J. Sulewska, Comp. Assisted Methods in Eng. and Sci. **18**, 4:231–241.
10. Z. G. Qian, A. J. Li, W. C. Chen, A. V. Lyamin, J. C. Jiang, Soils and Foundations **59**, 2:556–569 (2019)
11. A. M. Deris, A. M. Zain, R. Sallehuddin, Procedia Eng. **24**, 308–312 (2011)

12. V. H. Quej, J. Almorox, J. A. Arnaldo, L. Saito, J. of Atmospheric and Solar-Terrestial Physics **155**, 62–70 (2017)
13. S. Hosseini, B. M. H. Zade, Comp. Network, 107168 (2020)
14. L. Chao, J. Zhipeng, Z. Yuanjie, Expert Systems with Applications **123**, 283–298 (2019)
15. Y. Uchida, T. Funayama, Y. Kogure, *SVM classification of data obtained from a health conditon monitoring system using flexible force sensing resistors*. In Proceedings of the 5th Int. Conf. on Sensors Engineering and Electronics Instrumentation Advances (SEIA'1019), September 2019, Tenerife (Canary Islands), Spain, (2019)
16. J. Ruan, H. Jiang, X. Li, Y. Shi, F.T. Chan, W. Rao, IEEE Transactions on Industrial Informatics, **15**, 12:6510–6521 (2019)
17. J. Lu, X. Liao, S. Li, H. Ouyang, K. Chen, B. Huang, Complexity (2019).
18. C. Qi, X. Tang, Comp. & Indust. Eng. **118**, 112–122 (2018)
19. D. Kumar, M. Thakur, C. S. Dubey, D. P. Shukla, Geomorpholgy **295**, 115–125 (2017)
20. Y. Lin, K. Zhou, J. Li, IEE Access, 2169–3536 (2018)
21. X. N. Bui, H. Nguyen, Y. Choi, T. Nguyen-Thoi, J. Zhou, J. Dou, Scientific reports **10**, 1:1–17 (2020)
22. Y. Wu, Y. Ke, Z. Chen, S. Liang, H. Zhao, H. Hong, Catena **187**, 104396 (2020)
23. A. Kim, K. Oh, J. Y. Jung, B. Kim, Inter. J. of Computer Integrated Manufacturing **31**, 8:701–717 (2018)
24. S. J. Narayanan, R, Soundrapandiyan, B. Perumal, C. J. Baby, *Emphysema medical image classification using fuzzy decision tree with fuzzy particle swarm optimization clustering*, in Smart Intelligent Computing and Applications (pp. 305–313), 2019, Springer, Singapore (2019)
25. D. Tien Bui, B. Pradhan, O. Lofman, I. Revhaug, Mathematical problems in Engineering (2012)
26. S. J. Park, C. W. Lee, S. Lee, M. J. Lee, Remote Sensing **10**, 10:1545 (2018)
27. N. K. Sah, P. R. Sheorey, L.N. Upadhyaya, Intr. J. of Rock Mehanics and Mining Sci. & Geomechanics Abstracts **31**, 47–53 (1994)
28. K. P. Zhou, Z. Q. Chen, *Stability prediction of tailing dam slope based on neural network pattern recognition*, in 2009 Second International Conference on Environmental and Computer Science (pp. 380-383). December 2009, IEEE (2009)
29. J. Li, F. Wang, *Study on the forecasting models of slope stability under data mining*, in Earth and Space 2010: Engineering, Science, Construction, and Operations in Challenging Environments (pp. 765-776) (2010)
30. Y. Xiaoming, L. Xibing, *Bayes discriminant analysis method for predicting the stability of open pit slope*, in 2011 International Conference on Electric Technology and Civil Engineering (ICETCE) (pp. 147-150), April 2011, IEEE (2011)
31. V. Vapnik, The nature of statistical learning theory (Springer, New York, 1995)
32. D. Anguita, A. Ghio, S. Ridella, D. Sterpi, *K-Fold cross validation for error rate estimate in support vector machines*. In DMIN, July 2009, (pp. 291-297) (2009)
33. J. Dou, A. P. Yuus, D. T. Bui, A. Merghadi, M. Sahana, Z. Zhu, C. W. Chen, K. Khosravi, Y. Yang, B. T. Pham, Sci. of The Total Envi. **662**, 332–346 (2019)
34. C. Qi, X. Tang, Comp & Indust. Eng. **118**, 112–122 (2018)
35. S. Hwang, I. F. Guevarra, B. Yu, Eng. Geology, **104**, 1–2:126–134 (2009)
36. S. Rajeswari, K. Suthendran, Comp. and Electronics in Agriculture **156**, 530–539 (2019)
37. B. Choubin, E. Moradi, M. Golshan, J. Adamowski, F. Sajedi-Hosseini, A. Mosavi,Sci of The Total Env.**651**, 2087-2096 (2019)
38. Y. Radhika, M. Shashi, Inter. J. of Computing Theory Eng **1**, 1793–8201 (2009)