Journal of Scientific & Industrial Research
Vol. 80, January 2021, pp. 66-74

Machine Learning Techniques to Predict Slope Failures in Uttarkashi, Uttarakhand (India)

Poonam Kainthura[1,2]* and Neelam Sharma[1]

[1]Banasthali University, Jaipur 304 022, India

[2]University of Petroleum and Energy Studies, Dehradun 248 007, India

Received 07 August 2020; revised 15 October 2020; accepted 09 December 2020

Uttarkashi region is highly prone to landslides because of its geological structure. The exact occurrence of these landslides events is difficult to predict due to its complex mechanism and dependence on the number of triggering factors. Moreover, the behavior and prediction of unstable slopes are of high importance failing of which otherwise can have a devastating impact. This research work aims at studying and modeling slopes with the help of supervised machine learning models: Support vector machine, Backpropagation, Random Forest, and Bayesian Network models. To train and test these models a total of 629 instances are taken. Moreover, the independence of individual features is studied with the help of multicollinearity analysis. The capability of considered methods was evaluated using various performance evaluation metrics. Evaluation and comparison of the results show that the performance of all classifiers is satisfactory for slope failure analysis (AUC=0.595–0.915). Based on the results Random Forest proved to be most efficient to predict slope failures (Accuracy=88%, AUC=0.915). These outcomes can be beneficial for government agencies in early-stage risk mitigation.

Keywords: Landslides, Multicollinearity analysis, Patterns, Risk mitigation, Triggering factors

Introduction

Landslips are frequent natural events that occur in hilly regions all over the world. They can be classified in various forms and influences more than 15 of the total area of India. These phenomena are more active in the young mountain belts of India such as the hilly regions of Uttarakhand which is frequently affected by the disastrous situation of landslides, particularly in the rainy season. There can be various sources that trigger landslides such as anthropogenic activities in the region, dormant landslides, deforestation, and movement of habitation. Anthropogenic activities belong to man-made actions causing damage to the eco-system, increasing pollution, constructing unscientific roads in the hills, *etc*. Moreover, drilling activities in the mountains without understanding the consequences also lead to the slope instability which can ultimately result in the slope failure. Slope failure results in damages to settlements and habitats. Based on their formation, slopes can be natural or artificial. Therefore, understanding the dynamics of the stability of slopes is a crucial task to manage landslide events.[1] In general, rainfall is considered as the most influencing factor of landslides. Alternatively, stable slopes can also turn to unstable slopes because of underlying factors such as earthquakes, toe cutting. In practice, it is a difficult task to identify and predict unstable slope areas where landslides never occurred because of the huge resources required, accessibility issues, and complex topography. Slopes have a significant impact on vegetation, road transmission, agriculture, habitats, and settlements. Therefore, it is paramount to evaluate areas that are susceptible to the landslides and to predict it. In this scenario, it is important to identify the triggering factors that influence landslides. Although, it is not simple to evaluate the magnitude of multiple factors that influence slope properties.[2] Weathering of rocks is also considered an important factor that triggers landslide conditions. The weathering state of rocks results in a reduction in shear strength of slope.[3] Vegetation also plays a major role in slope failures due to mechanical factors such as plant roots system and overloading of trees.[4] The exploration and prediction of slope failure conditions are very important and challenging due to complex slope structures and its mechanism.[5] Landslides can be grouped into various ways based on failure mechanisms and influencing factors. In recent times, numerous studies have been presented on slope failure analysis and by using different sets of influencing

Author for Correspondence
E-mail: poonamkainthura@gmail.com

factors. However accurate predictions of slope failures are challenging as key causative factors are hidden. In spite, of these difficulties, many researchers conducted slope stability analysis in different countries by applying different methods. The methods include field investigations and analysis, qualitative and quantitative analysis.[6,7] The study and investigations cleared the concept of instable slopes without giving any robust solution. The current trend is focused on machine learning techniques to analyze geological and geomorphological conditions of hilly regions.[8,9] Machine learning methods displayed acceptable results for slope failure analysis and predictions. Lin *et al.*[10] described various supervised machine learning techniques to analyze and predict slope stability integrating six causative factors. Researchers applied machine learning methods to predict slope failures based on the factor of safety considering six causative factors. Various machine learning models were applied for the construction of prediction models that later evaluated and compared for slope failure predictions.[11] Multivariate analysis was conducted by Bui *et al.*[12] to analyze slope stability applying the concept of a neural network. Qi & Tang applied machine learning algorithms based on metaheuristic concepts on multiple datasets to analyze the stability of slopes.[13] Hybrid model using artificial neural network based on metaheuristic technique was proposed by Hoang and Pham to assess the stability of the slope.[14] Geotechnical and geometrical conditions were taken as input parameters for neural network scheme and estimated factor of safety for classifications. Integrated models were developed by Qian *et al.*[15] using artificial neural networks and limit analysis to analyze and predict soil slope stability with a high accuracy rate. Further, the accuracy of the above machine learning models can be improved using ensemble techniques.[16] These slope failures lead to landslides that damage lives and habitats. The identification of these landslides can be done using machine learning techniques.[17]

This research work is focused on the scattering of unstable slopes and the identification of the triggering conditions of slope failures in the Uttarkashi district of Uttarakhand, India. Few studies were conducted in the past for the Himalayan region of India to understand various scenarios of slope failures and related landslides.[18] There are many causative factors behind these failures and factors are correlated to each other. Heavy precipitation, the impact of the 1991 Uttarkashi earthquake, flash floods, and extreme rainfall is a most influencing factor of slope failures and occurrences of landslides.[19–22] Accurate analysis and prediction of slope failures are very crucial for Geotechnical engineers and risk mitigation team.

Machine Learning techniques have proven efficient in several domains. It can play a key role in the efficient analysis and prediction of disasters. For the area under study, Machine Learning techniques can be beneficial to recognize hidden patterns of slope failure and identification of critical parameters. It explores the dataset using different concepts that can be useful to predict possible future slope failures. Machine Learning methods can be grouped into supervised, unsupervised, and semi-supervised learning. Classification and regression are supervised learning models while clustering is an unsupervised model, semi-supervised is the mixture of both supervised and unsupervised learning. The major part of this research is to predict slope failure conditions by exploring and analyzing the hidden patterns generated by input factors. Therefore, classification and regression techniques can be useful to construct models that are capable to predict outcomes based on historical data. However, the prediction of slope failures is a challenging task due to complex slope structures and the dynamics of the parameters involved.[23] It is a challenging task to identify the precise input related to important geotechnical factors.

Study Area

The study region Uttarkashi is shown in (Fig. 1) and is situated in the hilly northern portion of the

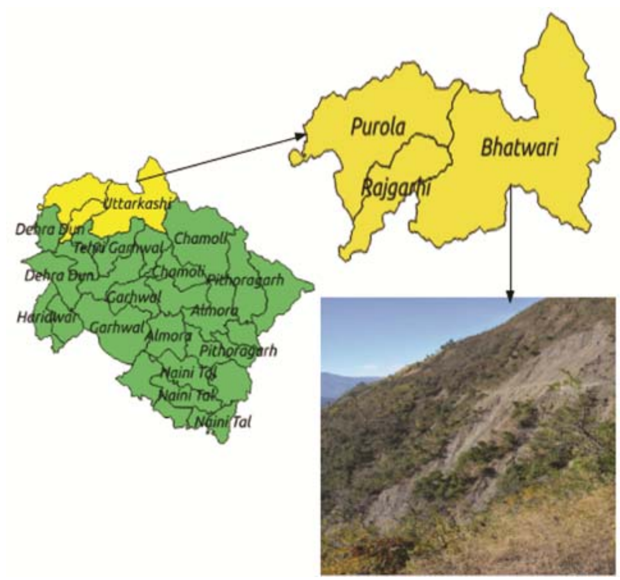

Fig. 1 — Study Area (Uttarkashi)

Uttarakhand state in India which lies within longitudes 78° 26'E and latitude 30° 44'N. It is a border district bordering with four other districts: Chamoli and Rudraprayag, Tehri and Dehradun in east, south and west directions, and Himachal Pradesh in the northwest. Two National highways NH-134 and NH-108, pass through the region. Geologically the region contains Quaternary deposits alongside the valleys. The components of the study area are mainly quartzite, limestone, chlorite schist, and Meta basics with quartzite. The formation of prominent high hill range on both sides of the river occurred due to the presence of these rocks which are well exposed in the road part and along the tributaries of the Bhagirathi River.[24] The slopes in the region are very steep (45°) and moderate (30°–35°) in most of the areas. The elevation range of the terrain is 700 m to 6319 m respectively. Geo-morphologically, the terrains are grouped into glacial, glacial-fluvial, fluvial, and denudation landforms. The vegetation varies from sparse to thick and erosion rate and sediment load is of higher magnitude.[25] The climatic condition of the study region is moderate and rainfall starts from June to September. In the region, more than 90% of landslides are shallow translational failures in nature and are still active.

Material and Methods

Slope failure analysis and prediction in the present study has been carried out in five steps as shown in Fig. 2: (1) data collection and preprocessing, (2) using attribute selection method for the selection of factors that influence slope, (3) preparation of training and testing datasets, (4) construction of slope failure prediction models, and (5) evaluation and comparison of slope failure models. The steps are discussed as follows:

Data Collection and Pre-Processing

Slope inventory contains necessary and crucial data for the prediction of slope failures. The data for this research work is collected from historical landslide reports supported by the Geological Survey of India (GSI). The collected data included extra attributes and missing values that required proper pre-processing for efficient analysis. The process is completely carried out by using open-source python libraries. Firstly, data were encoded into binary values using encoding function fit transform which is available in the python library. Secondly, data cleaning was done to fill missing data using the K-nearest neighbour method. Next, relevant attributes were selected using a dimensionality reduction technique: information gains attribute evaluation. This method can calculate the gain value of each parameter in contrast to predictor class and assigns a rank to each parameter. The highest rank parameter gives more information and considered more relevant for constructing classifier models for predictions. Finally, 629 instances were selected for the analysis and prediction purpose as input for selected models.

A sample of these instances is shown in Table 1. For model building purposes, the dataset was split into two subsets: 1) training samples and 2) test samples in a ratio of 70:30 percentage.

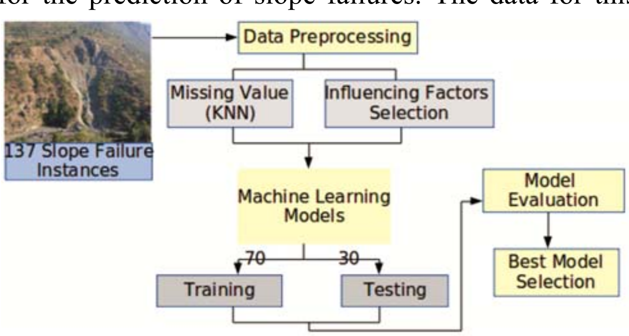

Fig. 2 — Workflow of the Proposed Methodology

Influencing Factors

Many geological and geo-morphological factors and seismic activities are responsible for slope failures. This section describes the failure process of stable slopes by analyzing the factors that trigger the failures. This mechanism starts when the deposit accumulates at the top of the rock and due to heavy rainfall fracture starts forming on the accumulated deposit at the top and through which water can penetrate. If there is rainfall for a long time than the surface water will continuously flow through these channels. This process results in erosion and ultimately losing the strength of soil which will increase the width and depth of the fractures. Gradually this process will increase the fracture line

Table 1 — Sample Slope failure data

Parameter	Values			
Susceptibility	Stable	Metastable	Unstable	Stable
Upslope (degree)	20	30	70	10
Downslope (degree)	60	70	60	40
Weathering	High	Low	Moderate	Low
Erosion	Yes	No	Yes	No
Rainfall	Yes	No	No	No
Road Influence	Yes	Yes	Yes	No
Slope Failure/ Decision	Yes	No	Yes	No

resulting in slides in some cases. Due to other internal blockages the pore water pressure increases, impacting the stress of the accumulated deposit. This put more weight on the slope resulting in poor strength of the shear. The accumulated deposit keeps on deforming which reduces the mass of soil and ultimately joints results in slides. For the study area the following major triggering conditions are discussed below:

Slope Angle (Upslope/Down Slope)

It is the most influential factor in inducing in-stable slope and assessment of instable slopes is the basic step to evaluate the risk of landslides. Shear strength and stress are two terms always associated with slopes that are directly affected by internal and external activities on slopes. In the study region slope angle varies from 1° to 90° and in some areas negative slopes are also marked. The slopes ranging from moderate to steep (25° to 40°) directly causes stress in the slope and are more susceptible to landslides.

Erosion

Toe erosion due to toe cutting and road constructions along with rainfall on massive and fractured slopes can be the cause of slope instability. Also, anthropogenic activities play a major role in triggering erosion by turning stable slopes to unstable slopes.

Rainfall

The identification of rainfall induced landslides zones is a complicated task as the increase in hydro meteorological instances due to climatic changes. Rainfall is the most critical parameter that triggers landslides in Uttarkashi region and it has a clear effect over the stable slopes. The average rainfall it receives is 1693 mm.The monsoon period in Uttarkashi starts from June to September and 75% of it is received in these months. Rainfall and surface runoff can change surface morphology.[26] Gullies are formed due to unstoppable surface runoff if rainfall happens for a long time. When these gullies are eroded continuously, it deforms the structure of the slope which may arise the failure condition of the slope resulting in slope failure helpful in identifying a correlation between crucial variables that can be useful in decision making.

Road Cut Influence

Hills in the northern region of India is constituted of unstable, fractured, and massive slopes. Road construction and other anthropogenic activities along with rainfall on such hills play a major role in the failure of these slopes. The major anthropogenic activity in slopes isthe construction of roads and buildings without considering the slope cutting manner.

Weathering

It is the process where rocks disintegrate, losing the surface of rocks and which are transported away by rainfall and erosion. This process results in the reduction of shear strength inside slopes. Weathering of rocks can be natural or artificial and it is a crucial factor that can turn stable slopes to unstable ones. Due to the geological conditions of Uttarkashi rocks in the slopes are highly weathered and disintegrated. The strength of the rocks is affected by the degree of weathering such as low moderate and high.

Susceptibility

It is a state where a surface can be influenced by any internal or external factor resulting in slopes failure that can lead to landslides. In this study region, susceptibility is divided into three classes: stable, metastable, and unstable. There are lesser cases of failures when susceptibility is stable. The number of failures increases when susceptibility is meta-stable and unstable along with fractured and massive rock. There exist other external factors that are responsible to convert stable slopes to unstable slopes.

Multicollinearity Analysis

In multicollinearity analysis two or more predictor variables can be correlated with others increasing the standard error of the coefficient i.e. applying multicollinearity process to some significant attributes turns statistically insignificant. The results of this matrix can help identify the correlation between crucial variables that can be useful in decision making. This co-relation matrix represents which variables affect the target variable the most to predict future instances. It is also helpful in attribute selections. These results can be plotted in colors to visualize results that range in between -1 to +1, where -1 represents minimum correlation and +1 represents the maximum correlation between two variables. The result of multicollinearity analysis (Fig. 3) indicates when slope is unstable then rainfall becomes triggering factor for failure of slopes. Along with that when roads are constructed in this stage the probability of slope failures increases.

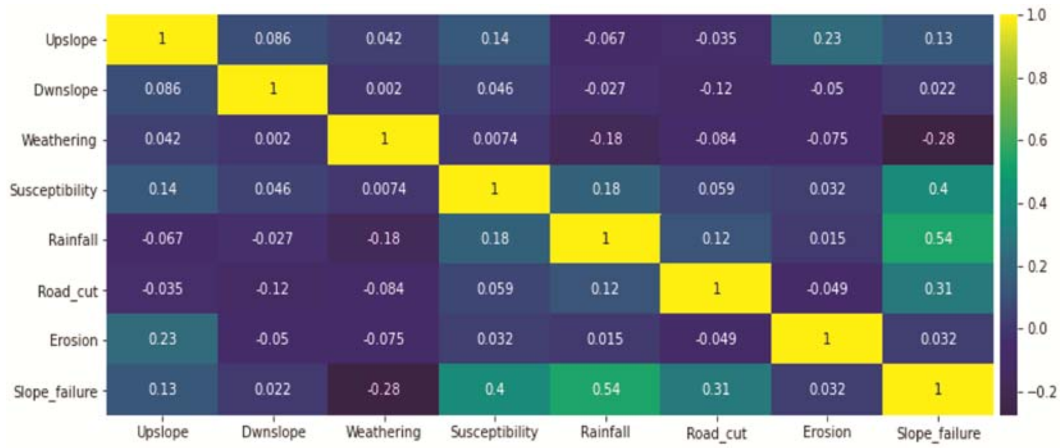

Fig. 3 — Multicollinearity Analysis of Causative Factors (0: Worst Collinearity, 1: Best Collinearity)

Classification Methods

This research approached slope failure analysis like a machine learning classification problem using supervised learning. The slope failure attributes were added with labels yes or no values which need to be predicted by the model. The development phase includes sci-kit-learn: a Python-based module to implement models using machine learning. A total of 629 slopes failure conditions in the study area is taken into consideration in the analysis. The data has been provided by the Geological Survey of India which was carried out by field investigations. The other factors in combination with unstable slopes show that these slopes have a high probability to fail due to any external factor and which may lead to the landslides in the area. The dataset was divided into training and testing dataset. The training dataset is applied to train classifiers for prediction while testing dataset is used to test the strength of the model. In this study, a random split method was used in a proportion of 70:30 as a training set from the actual data-set and remaining proportion for testing purpose. The following supervised Machine learning models: Bayesian Network (BN), Support Vector Machine (SVM), Random Forest (RF), Backpropagation (ANN) are trained using a training set. The trained classifiers are then used to predict slope failures based on the specific patterns and conditions on a given test data. In the following subsections, these models are discussed.

Support Vector Machine

Support Vector Machine model focused on finding the optimal decision boundaries that divide classes. The advantage of using this model is that it is highly capable to handle non-separable and high dimensional data. The prediction of slope failures due to multiple external factors cannot be identified theoretically. SVM attempts to find the line or margin that best divides the two slope failure classes efficiently. In SVM, multiple lines separate the classes, but the largest margin line is selected as a final boundary. The various points located on this line are known as support vectors.

$$D = \{(U_i, V_i)\}_{i=1}^{N} \quad \ldots (1)$$

where U_i represents input variables of slope failure analysis, $V_i \ \epsilon \ \{+1,-1\}$ represents the number of outcomes (in this work two classes are considered). N represents the total number of instances. The optimal decision boundary that separates the data when classes are linearly separable can be expressed as:

$$f(U) = w.U + b = 0 \quad \ldots (2)$$

where w denotes weight and b states bias value. To handle non-linear classification SVM uses its kernel function which can be expressed as:

$$f(U) = sign\left(\sum_{j=1}^{N} \alpha_i V_i K(U_i, U_j)\right) + b \quad \ldots (3)$$

where, $K(U_i, U_j)$ is a kernel function.

Random Forest

Random Forest is a popular ensemble method that can be used to increase the overall accuracy of the model. The individual decision trees are constructed depending on the values of an independent random vector. The classification models predict the value returned by individual trees on the principle of voting. The decision tree identifies significant rules and patterns from a given input dataset automatically in the form of a tree.[27] Using these rules slope failure probabilities can be easily identified for any new set of data. The most informative factor of slope failures can be ranked using the gain ratio method. In the later stage, statistical measures like pessimistic pruning are applied to overcome the problem of over-fitting. The parameters with high impact for the study area were identified as slope angle, rainfall, erosion, susceptibility, and road cut influence. In the following steps the process of constructing a tree network is defined:

Step 1: Given a dataset D, the average information can be calculated as:

$$info(D) = -\sum_{i=1}^{m} p_i \, log_2(p_i) \quad \ldots (4)$$

where info (D) or entropy is the average information to identify target class, p_i denotes non-zero probability, m denotes the number of classes.

Step 2: The expected amount of information for individual attributes $info_A(D)$ is measured by:

$$info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} info(D_j) \quad \ldots (5)$$

where $|D_j|/|D|$ denotes the weight of j_{th} partition, v denotes the total number of partitions.

Step 3: Information gain $Gain(A)$ is computed by the following expression:

$$Gain(A) = info(D) - info_A(D) \quad \ldots (6)$$

Step 4: The gain value shows biasing with many testing outcomes. To overcome this problem gain ratio method is used which used split information $Splitinfo_A(D)$ of an attribute to normalize information gain as shown in the following expression:

$$Splitinfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} log_2(\frac{|D_j|}{|D|}) \quad \ldots (7)$$

Step 5: The *Gain Ratio* is expressed as:

$$GainRatio(A) = \frac{Gain(D)}{Splitinfo_A(D)} \quad \ldots (8)$$

Random forest are more robust to outliers and errors.

Backpropagation

Artificial neural networks (ANN) mimic the behavior of a biological neural network that can be used for predictions once trained sufficiently accurately on input data. A neural network is constructed in a set of input layers, hidden layers, and output layers connected by some suitable weights. A Backpropagation is a feed-forward neural network that contains at least one hidden layer. Backpropagation is used to train a model for classification using a sigmoid function. The classification of slope failures can be described in three stages. Firstly, triggering factors are identified that converts stable slopes to an unstable state. These are provided as input values to the input layer of the network. Secondly, a network is trained until the total error meets the precision requirement of the model. In the final stage, the slope failure state is classified and which is considered as a result of the output layer of the network. Considering the influencing slope failure parameters, the following is the architecture adopted for the neural network: input layer (n) = 6, hidden layer = 2n+1 (Kolmogorov theorem), and one neuron in the output layer.[28] ANN can be implemented in the following steps:

Step 1: Initialize weights *(w)* and bias *(b).*

Step 2: Propagate the input forward

$$H_i(in) = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_ix_i + b_i \quad \ldots (9)$$

where $H_i(in)$ is input to hidden layers, x is input from input layers, w is adjusted weight assigned to each edge and b is bias value. Similarly, $H(out)$ is the output of the hidden layers.

$$H_i(out) = \frac{1}{(1+exp^{-H_i(in)})} \quad \ldots (10)$$

Similarly, input $O_i(in)$ for an output layer is computed and output $O_i(out)$ for an output layer is computed using the sigmoid function. In the above equation, $H_i(out)$ is output from hidden layers which is computed using a sigmoid function.

Step 3: Compute errors and repeat the process until the target value is achieved:

$$E_{total} = \sum \frac{1}{2}(Target(out) - O_i(out))^2 \quad \ldots (11)$$

Step 4: Back propagate the errors by readjusting the weights

$$\frac{\partial E_{total}}{\partial w_i} \quad \ldots (12)$$

where ∂E_{total} is the error to w_i.

Bayesian Network

Bayesian network is a machine learning technique that combines graph theory (directed acyclic graph, nodes, and edges) and statistical methods (uncertainties, probabilities) and provides a probabilistic approach of reasoning under uncertainty. The Bayesian network is a powerful decision supporting system for many real-life problems. In a Bayesian network, nodes represent a set of variables (events) in a directed acyclic manner connected by arcs. These nodes represent *[cause (parent) – effect (child)]* relationship possessing conditional dependencies among them which is maintained by conditional probability tables (CPT) given by Equation 13. The joint probability function is used to calculate the probability of the outcome node as shown in Equation 14. This model uses slope parameters as random variables and compute uncertainty under reasoning as a probability distribution that results in relative likelihoods of slope failures.[29] The formulation of the Bayesian network is as follows:

Step 1: Construct directed acyclic graph (DAG)

Step2: Compute CPT for each variable:

$$P(X_i|Y) = P(Y|X_i) \times \frac{P(X_i)}{P(X)} \quad \ldots (13)$$

Step 3: Compute Joint probability function to generate predict the outcome:

$$\mathbf{P(X_i) = \prod_{i=1}^{n} P(X_i|predecessor(Y_i))} \quad \ldots (14)$$

where Xi is the number of instances, Y_i is the number of attributes, $P(X_i|\mathrm{predecessor}(Y_i))$ denotes X_i are the parents ofY_i, and $P(X_i)$ is the probability of combined values of X.

Step 4: Model Training.

Results

Model evaluation and its performance measures are the important criteria to adopt any model. The various evaluation metrics are discussed below:

Evaluation metrics

Confusion matrix: the performance of models is evaluated using a confusion matrix. In this, the actual slope failures instance correctly predicted are denoted by true positive (TP), the actual non-slope failures instance correctly predicted as true negative (TN), actual non-slope failure instances predicted as false positive (FP) and actual slope failures instances correctly predicted as false negative (FN). A good model should always have a high percentage of true positive rate (TPR) which is also known as Sensitivity. It is termed as Specificity When compared to False Positive Rate (FPR). Finally, the confusion matrix shows true positive rates and false-positive rates. The formulation of the above-discussed metrics is as follows:

$$Specificity = \left(\frac{TN}{N}\right),\ Sensitivity = \left(\frac{TP}{P}\right) \quad \ldots (15)$$

$$Accuracy = \left(\frac{TP+TN}{P+N}\right) \quad \ldots (16)$$

Precision is defined as a measure of exactness i.e. proportion of positive slope failure instances that are correct. Precision will have a value of 1.0 if the model results in no false positive case. A recall is defined as a measure of exactness, i.e., the proportion of actual slope failures predicted correctly. A recall will have a value of 1.0 if a model results in no false negative case. It can be expressed as:

$$Precision = \left(\frac{TP}{TP+FP}\right),\ Recall = \left(\frac{TP}{TP+FN}\right) \quad \ldots (17)$$

Area under curve (AUC) values are produced from sensitivity (TPR) and specificity (FPR) values using a receiver operating characteristic (ROC) curve. A good performing model will cover the maximum area under curve such that the value is always close or equal to 1. The best graph between TPR versus FPR scores will be close to the upper left corner resulting in 100 percent TPR and 0 percent FPR value. A value of AUC equal to 1 represents that all slope failures instances are correctly classified. In general, the performance of any model is categorized based on the AUC values generated as poor (< 0.60), moderate (0.60–0.70), good (0.70–0.80), very good (0.80–0.90), and excellent (>0.90).

Landslide Evaluation

In this section, results are presented from analysis for all considered classifiers. The results from evaluation metrics such as Sensitivity, Specificity, and Area under Curve (AUC) are compared in Table 2. The ROC curves plot Sensitivity against Specificity for the different classifiers are compared in (Fig. 4).

Table 2 — Performance of Bayesian Network (BN), Support Vector Machine (SVM), Backpropagation (BP), Random Forest (RF) Classifier

Performance Measure	Results			
	BN	SVM	BP	RF
Sensitivity	0.590	0.909	0.867	0.963
Specificity	0.558	0.767	0.720	0.720
Precision	0.620	0.850	0.800	0.890
Recall	0.580	0.840	0.790	0.840
ROC	0.595	0.890	0.874	0.915
Accuracy (%)	58%	86%	82%	88%

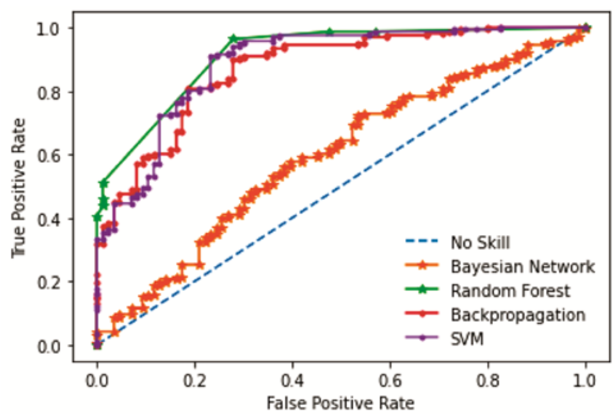

Fig. 4 — Performance comparison of Bayesian Network (BN), Support Vector Machine (SVM), Backpropagation (ANN), Random Forest (RF) classifier Using ROC curve

Discussion

In comparison to traditional methods such as domain expert opinion through field investigations and analysis, machine learning techniques show higher efficiency to analyze and predict the slope failures.[30] In this work, following machine learning methods namely Random Forest, Support Vector Machine, Backpropagation, and Bayesian Networks were applied to assess and compare to predict slope failures. Selected machine learning methods are widely used in slope stability analysis. The results show that the performance of the three classifiers Random forest, Back propagation and SVM was good considering selected causative factors. Analysis of comparison results shows that the Random Forest model outperformed in comparison to other classifiers in terms of accuracy. The accuracy rate of the Random Forest model is 88%. The area under curve (AUC) value of the Random Forest model is 0.915 which is higher than other models. These results can be classified into several rules. The first rule explains that the probability of slope failure when existing condition represents the presence of unstable slopes in combination with slope angle 55° and rainfall amount is high. Cutting roads on steep slopes cause the failure of slopes. Heavy rainfall on any slope angle and low weathering turn to erosion resulting in slope failure. These are commonly accepted patterns observed in the slope failure analysis and which are common in several classifiers. Moreover, the predicted class slope failure is highly interactive and dependent on the various triggering factors of slope failures. The scores generated by the Random Forest are significantly better in comparison to other models.

Based on the performance comparison on the evaluation metrics: ROC curve, accuracy, precision, and recall, the performance of the Random Forest model can be considered as the best to predict slope failures for the selected study area. The research also shows the importance of machine learning classifiers by extracting hidden patterns of failures in a dynamic dataset.

Conclusions

This research adopted machine learning techniques as an effective solution for the prediction of unstable slopes and hidden interactive patterns related to slope failures in the Uttarkashi region. Overall, all four classification models show good prediction capability for slope failures in the Uttarkashi district. Based on the evaluation metrics results and comparison, it can be concluded that the Random Forest is highly capable to predict slope failures. Thus, the model can be applied for the assessment and development of slope failure prediction models. The receiver operating characteristics curve helps in comparing and selecting a highly capable model. The factors considered for the analysis are sensitive to the stability of slopes. In practice, Random forest can be used as an intelligent model to predict slope failure patterns on the availability of real time data. This model can also be useful for slope failure analysis and prediction in other regions of the state that have similar slope descriptions and which are susceptible to slope failures and landslides. As a scope of improvement, the predictive capability of considered classifiers can be enhanced using optimization schemes. Also, not all classification models are suitable for every slope conditions. Therefore, the selection of a classification model should be done with a complete understanding of triggering factors. The construction of roads is the main reason for non-stop slope failures along national highways in Uttarkashi district. Finally, the machine learning models can be successfully applied for slope stability analysis and predictions that may be used as decision making tool by government agencies for planning developmental activities such as road constructions on moderate and steep slopes. The model will also help the disaster management team in risk planning and mitigation on time.

Acknowledgments

The authors are grateful to the Geological Survey of India, Dehradun, India for providing the landslide dataset.

References

1 Zhang L L, Zhang J, Zhang L M & Tang W H, Stability analysis of rainfall induced slope failure: A review, *Proc I Civil Eng-Geotech*, 2011 299–316.

2 Pilgrim A T, Puvaneswaran P & Conacher A J, Factors affecting natural rates of slope development, *CATENA*, **13(1–2)** (1986) 169–180.

3 Miščević P & Vlastelica G, Impact of weathering on slope stability in soft rock mass, *J Rock Mech Geotech*, **6** (2014) 240–250.

4 Kokutse N K, Temgoua A G T & Kavazović Z, Slope stability and vegetation: Conceptual and numerical investigation of mechanical effects, *Ecol Eng*, **86** (2016) 146–153.

5 Boutrup E & Lovell C W, Searching techniques in slope stability analysis, *Eng Geol*, **16(1–2)** (1980) 51–61.

6 Wang Z, Xia Z, Zhang H L, Gao & Cheng J, Field Geological Investigations and Stability Analysis of Duanjiagou Landslide, *ISPRS Int J Geo-Info*, **9(1)** (2020).

7 Han Z, Su B, Li Y, Yangbans M, Weidong W & Guangqi C, Comprehensive analysis of landslide stability and related countermeasures: a case study of the Lanmuxi landslide in China, *Sci Rep*, **9** (2019).

8 Yacine A & Pourghasemi H R, How do machine learning techniques help in increasing accuracy of landslide susceptibility maps?, *Geosci Front*, **11(3)** (2020) 871–883.

9 Goetz J N, Brenning A, Petschko H & Leopold P, Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling, *Comput Geosci*, **81** (2015) 1–11.

10 Lin Y, Zhou K & Li J, Prediction of Slope Stability Using Four Supervised Learning Methods, *IEEE Access*, **6** (2018) 31169–31179.

11 Bui D T, Moayedi H, Gör M, Jaafari A & Foong L K, Predicting Slope Stability Failure through Machine Learning paradigms, *ISPRS Int J Geo-Info*, **8** (2019).

12 Sakellariou M G, Michael & Ferentinou M D, A study of slope stability prediction using neural networks, *Geotech Geol Eng*, **23** (2005) 419–445.

13 Qi C & Tang X, Slope stability prediction using integrated metaheuristic and machine learning approaches: A comparative study, *Comput Ind Eng*, **118** (2018) 112–122.

14 Hoang N & Pham A, Hybrid artificial intelligence approach based on metaheuristic and machine learning for slope stability assessment: A multinational data analysis, *Expert Syst Appl*, **46** (2016) 60–68.

15 Qian Z G, Li A J, Chen W C, Lyamin A V & Jiang J C, An artificial neural network approach to inhomogeneous soil slope stability predictions based on limit analysis methods, *Soils Found*, **59(2)** (2019) 556–569.

16 Liu Z, Gilbert G, Cepeda J M, Lysdahl A O K, Piciullo L, Hefre H & Lacasse S, Modelling of shallow landslides with machine learning lgorithms, *Geosci Front*, **12(1)** (2021) 385–393.

17 Wang H, Zhang L, Yin K, Luo H & Li J, Landslide identification using machine learning, *Geosci Front*, **12(1)** (2020) 351–364.

18 Sah N, Kumar M, Upadhyay R & Dutt S, Hill slope instability of Nainital City, Kumaun Lesser Himalaya, Uttarakhand, India, *J Rock Mech Geotech*, **10** (2018) 280–289.

19 Jamir I, Gupta V, Thong G T & Kumar V, Litho-tectonic and precipitation implications on landslides, Yamuna valley, N W Himalaya, *Phys Geography*, **41(4)** (2019)235–388.

20 Cotton F, Campillo M, Deschamps A & Rastogi B K, Rupture history and seismotectonics of the 1991 Uttarkashi, Himalaya earthquake, *Tectonophysics*, **258(1–4)** (1996) 35–51.

21 Gupta V, Nautiyal H, Kumar V, Jamir I & Tondon R, Landslide hazards around Uttarkashi Township, Garhwal Himalaya, after the tragic flash flood in June 2013, *Nat Hazards*, **80** (2016) 1689–1707.

22 Martha T R, Roy P, Govindha Raj K B, Kumar K V, Diwakar P G & Dadhwal V K, Landslides triggered by the June 2013 extreme rainfall event in parts of Uttarakhand state, India, *Landslides*, **12** (2014) 135–146.

23 Valdiya K S, Himalayan transverse faults and folds and their parallelism with subsurface structures of North Indian plains, *Tectonophysics*, **32** (1976) 353–386.

24 Ramola R C, Yadav M & Gusain G S, Distribution of natural radionuclide along Main Thrust in Garhwal Himalaya, *J Radiat Res Appl Sc*, **7** (2014) 614–619.

25 Singh S, Malik Z A & Sharma C M, Tree species richness, diversity, and regeneration status in different oak (Quercus spp.) dominated forests of Garhwal Himalaya, India, *J Asia Pac Biodivers*, **9** (2016) 293–300.

26 Fustos I, Abarca-del-Río R, Mardones M, González L, & Araya L R, Rainfall-induced landslide identification using numerical modelling: A southern Chile case, *J S Am Earth Sci*, **101** (2020).

27 Hwang S, Guevarra I F & Yu B, Slope failure prediction using a decision tree: A case of engineered slopes in South Korea, *Eng Geol*, **104(1–2)** (2009) 126–134.

28 Wang H B, Xu W Y & Xu R C, Slope stability evaluation using Backpropagation Neural Networks, *Eng Geol*, **80(3–4)** (2005) 302–315.

29 Li X, Zhang L & Zhang S, Efficient Bayesian networks for slope safety evaluation with large quantity monitoring information, *Geosci Front*, **9(6)** (2018) 1679–1687.

30 He X, Xu H, Sabetamal H & Sheng D, Machine learning aided Stochastic reliability analysis of spatially variable slopes, *Comput Geotech*, **126** (2020).