Landslide Susceptibility Prediction based on Decision Tree and Feature Selection Methods

Abstract

Landslide hazards give rise to considerable demolition and losses to lives in hilly areas. To reduce the destruction in these endangered regions, prediction of landslides incidents with good accuracy remains a key challenge. Over the years, Machine learning models have been used to increase the accuracy and precision of landslide predictions. These machine learning models are sensitive to the data on which they are applied. Feature selection is a crucial task in applying machine learning as meticulously selected features can significantly improve the performance of the machine learning model. These selected features decrease the learning time of the model and increase comprehensibility. In this paper, we have considered three feature selection methods namely chi-squared, extra tree classifier and heat map.The paper substantiates that feature selection can significantly increase the performance of the model. The study was carried out on the landslide data of Kullu to Rohtang Pass transport corridor in Himachal Pradesh, India. The classification score and receiver operating characteristics (ROC) curves were used to evaluate the model performance. Results exhibited that eliminating one or more features using different feature selection methods increased the comprehensibility of the model by reducing the dimensionality of dataset. The model achieved an accuracy of 90.74% and area under the ROC curve (AUROC) value of 0.979. Furthermore, it can be deduced that with reduced number of features model learns faster without affecting the actual result.

Keywords Feature selection methods, machine learning, landslide susceptibility prediction, receiver operating characteristics.

1. Introduction

Landslides are crucial natural hazards in hilly areas throughout the world (Pourghasemiet al., 2018). Even though landslides primarily happen restrictedly, substantial damage can happen to natural and human infrastructures at distinct level in mountainous regions (Holbling et al., 2012; Achu et al., 2022). Besides the tangible damage, landslides have wide ranging impact on the economy and human habitation (Hong et al., 2017). Various studies have been conducted and evaluated different landslides reduction strategies and landslide susceptibility mapping (Solway, 1999; Martire et al., 2012; Pradhan, 2013; Svalova, 2018; Pham et al., 2021). All studies carried out this using different knowledge based methods (Myronidis et al., 2016; FeizizadehandGhorbanzadeh, 2017) and machine learning methods (Sezer et al., 2011; Aghdam et al., 2016; PaurghasemiandKerle, 2016; Chen et al., 2017; Ghorbanzadeh et al., 2018b;

Achu et al., 2020). Over the few years, using machine learning for landslide susceptibility mapping and prediction has been increased rapidly.

Machine learning methods have been experienced in diverse research domains for example data mining and swarm intelligence (Lavrac, 1999; Maheshwar et al., 2015), pattern recognition (Narayanan, 2016), medical diagnosis (Goyal and Maheshwar, 2019; Maheshwar and Kumar, 2019) and artificial intelligence (Ghahramani, 2015) and have revealed prominent outcomes. All most all machine learning methods work into two phases: 1. Learning phase of the model and 2. Testing the model against test dataset. So, it is requisite to first understand the dataset and its different features causing the landslide. Feature selection methods can be used to analyse the correlation among the triggering features and occurrence of landslide events. It is an important step in machine learning which extremely influence the performance of a machine learning model (Premakanthan and Mikhael, 2001; Tirelli and Pessani, 2011). Features that are used to train the model have high effect on the accuracy of model. The unrelated features or partly relevant features can adversely results the performance. Feature selection is a technique where only those features are selected which put up highly to the prediction variable or output in which we are intended in. Having unrelated or not relevant features in data can decline the accuracy of model and make model learn on the basis of unimportant features. Feature selection comes with many advantages like minimizing over fitting by selecting on the important features, lessening the model complexity making it easily understandable, enhancing accuracy of model by making it to work on important features and decreasing the training time of model.

Features selection is broadly categorized into three categories: 1. Filter based methods 2. Wrapper based methods and 3. Embedded methods. These methods provide powerful techniques to select the major triggering features for landslide susceptibility prediction.

Filter based methods utilize a measure apart from error rate to deduce whether the feature is useful (Lee et al., 2011; Xue et al., 2012; Porkodi, 2014). Instead of regulating model, a subspace of the features is selected by ranking them through a convenient expressive measure. Wrapper based methods quantify models with a definite subset of features and assess the significance of each feature (Somol et al., 2005; Qiaoet al., 2006; Liang et al., 2015). Afterwards, they recapitulate and go for another distinct subset of features until the optimal features subset is attained. Embedded methods amalgamates the characteristics of both filter based and wrapper based methods (Ge et al., 2009; Windeatt et al., 2011; Chandershekar and Sahin, 2014; Guo et al., 2016; Lu, 2019). It is accomplished by algorithms having their own incorporated feature selection methods.Various most well-known examples of embedded feature selection methods include ridge and lasso regression which has their built in penalization functions for reducing overfitting.

The whole paper is organized as follows: After a brief introduction in section 1, section 2 describes the dataset and its general features. It sums up all the dataset resources and preparation of landslide inventory with major triggering features. Section 3 discusses the different methodologies used. Chi squared, extra tree classifier and heat map used as feature selection model and decision tree classifier model of machine learning for landslides prediction are detailed in this section. Section 4 examines the results and reveals the importance of eliminating features using different feature selection methods and its effect on the accuracy of the model. Different AUROCs exhibit that model achieve a good AUROC value of 0.979 in all three feature selection methods by reducing redundant features. Lastly, a brief discussion and conclusion is presented in section 5 followed by references.

2. Study Area

2.1. General features

The present study had been carried out along the transport corridor (NH-21) from Kullu-Rohtang Pass with a total length of 90 km. The study area has the latitude between 32° 0' 0'' N to 32° 20' 0'' N and longitude between 77° 5'0''E to 77° 15'0''E (Fig. 1). One kilometre buffer is considered on the each sides of the transport corridor for landslide susceptibility prediction.The chosen area lies in Himalaya ranges having a high elevation ranging between 1,279 m and 3,979 m from Mean Sea Level (MSL). The average rainfall is around 1,363 mm in the study region. Maximum number of landslides are primarily observed from July to September having a high rainfall during these months. The temperature fluctuates between 25^{o} Celsius and 4^{o} Celsius.

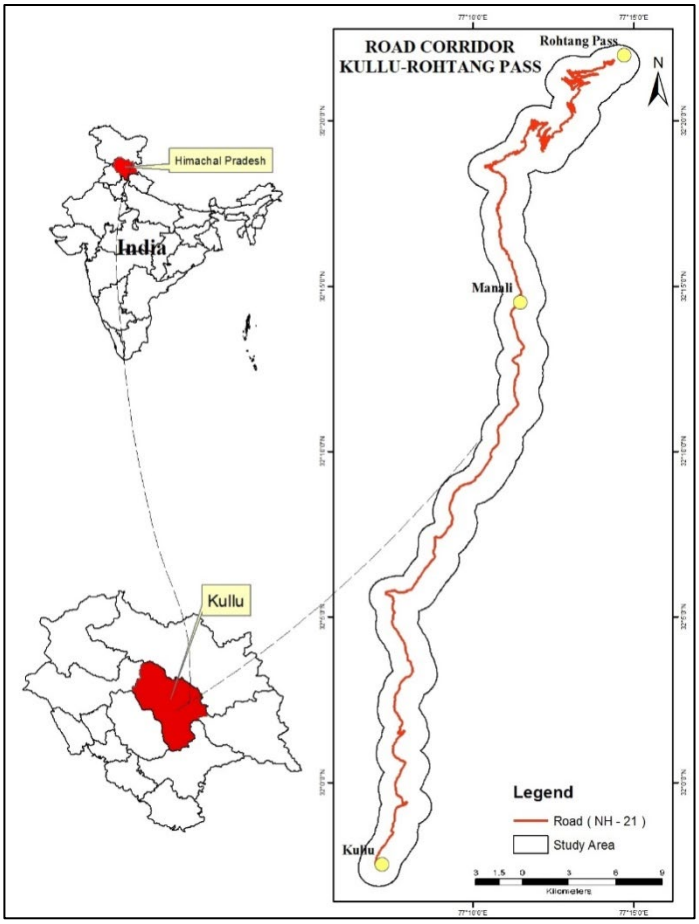

Fig.1. Study area.

Different varieties of soil are present in the study area for example mountain meadow soil, red loamy soil and brown hill soil. The study region lies along the banks of the river Beas. The deforestation, engineering activities like road construction and urbanization are key activities escalating the landslide frequencies (Saha et al., 2005). The landslide incidence along this corridor has unavoidable impacts on the transportation andat times there is a complete disconnection of the transportation facilities influencing the economy of the region.

2.2. Dataset

2.2.1. Data preparation

In this study, the topographic analysis has been done using ASTER DEM with 30m spatial resolution. Benchmarks have been digitized using the survey of India (SoI) topographic sheet no. 52 H/3 and 52 H/4 on the scale 1:50, 000. Geographic and topographic features such as slope, elevation, distance to road and distance to drainage are analysed using ASTER DEM, USGS. Land use and land cover (LULC) data is prepared using Google Earth and Landsat 8 OLI, USGS.Geological quadrangle maps, GSI is used to prepare Geological and geomorphological data and ground

water prospects maps by NRSA for preparing lineament density data. Landslide occurrence locations, their types, frequency and occurrence year have been cumulated from BRO, Manali and PWD, Kullu (Table 1).

Table 1

Dataset types and their sources

Data Type	Data Base	Resolution and Scale	Data Derivative
Topographic map	Survey of India (SoI)	RF 1: 50, 000	Boundary of the study area, transport route
Satellite data	Google Earth, Landsat 8	30 meter	Land use and land cover
ASTER DEM	USGS	30 meter	Slope, distance to Road, distance to Drainage, Elevation
Landslide data	BRO. Manali PWD, Kullu and NDMA govt. reports		Landslide locations, frequency, types of landslide, year of occurrence, road damage and cost
Ancillary data	Geological Quadrangle Map, GSI	1:250,000	Geology and Geomorphology
	GSI and Ground Water Prospects Map by NRSA	1:250,000 and 1:50,000	Lineament Density

USGS (United States Geological Survey)
BRO (Border Road Organization)
PWD (Public Work Department)
GSI (Geological Survey of India)
NDMA (National Disastrous Management Authority)
NRSA (National Remote Sensing Agency

2.2.2. Landslide inventory and triggering features

Landslide inventory is comprised of crucial and essential data for landslide prediction. In this study, the landslide inventory is prepared using Geographic Information System (GIS) with the help of satellite imageries and GPS way points (Fig. 2). A comprehensive field investigation of 54 landslidelocations along the transport corridor was carried out with the help of Global Position System (GPS) and Google earth images in 2018. Spatial temporal map are constructed using 18 years (2000 to 2018) landslide data collected from BRO, Manali and PWD, Kullu. All the landslide locations are divided into training and testing datasets randomly with a ratio of 70:30. The 38 landslide locations are used to train the models while 16 landslide locations are used to test the models' performance.

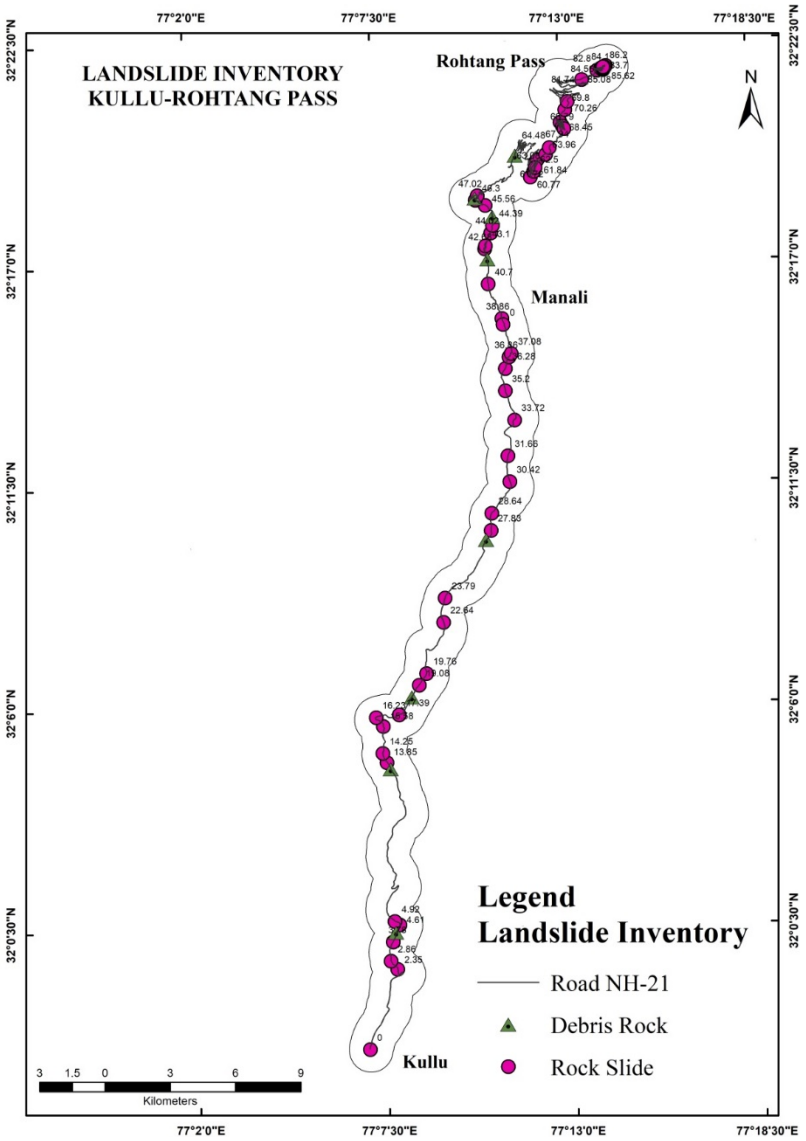

Fig.2. Landslide inventory map of Kullu-Rohtang pass transport corridor.

The types of landslide i.e. rock slide and debris slide are according to the records of BRO, Manali and PWD, Kullu. The largest and the smallest landslides mapped along the transport corridor were 4000 m^3 and 120 m^3 respectively.

For predicting landslide, the relation between geo-environmental factors and historical landslide events is carried out. A set of seven triggering features (slope, elevation, land use and land cover, geology & geomorphology, lineament density, distance to road and distance to drainage) have been considered for landslide prediction analysis. Continuous factors (slope, elevation, lineament density and so on) had been discretized using their normalized values which are calculated by using Analytical Hierarchical Process (AHP) model (Saaty, 1990a). Landslide triggering features are categorized into different classes. The detailed description of all triggering features is presented in Table 2.

Table 2

Landslide triggering features

S. No	Landslide triggering factors	Class
1	Slope (degree)	(1) very gentle (0-15°); (2) gentle (15° -30°); (3) moderate (30° -45°);(4) steep (45° -60°); (5) very steep (>60°)
2	Elevation (m)	(1) <1000; (2) 1000-2000; (3) 2000-3000; (4) >3000
3	Land use and land cover	(1) dense forest; (2) agriculture; (3) sparse forest; (4) settlement;(5)barren land; (6) snow cover
4	Geology & geomorphology	(1) highly dissected hill and valley; (2) snow cover; (3) schist and quartzite; (4) granitic gneiss and granitoid; (5)glacio-fluvial deposits and quaternary alluvium; (6) quartzite schist; (7) carbonaceous slate and limestone; (8)Biotite schist and kynite gneiss
5	Lineament density (km/km^2)	(1) low; (2) medium; (3) high
6	Distance to road (m)	(1) <200; (2) 200-400 (3) >400
7	Distance to drainage (m)	(1) <200; (2) 100-200; (3) >200

3. Methodology

In our work, chi square, extra tree classifier and heat map were used as feature selection methods. Each of these methods, uses a different approach of selecting the relevant features e.g. chi square is a filter method of feature selection and extra tree classifier is an ensemble method for selecting a certain number of features. All these feature selection methods are discussed along with their mathematical formulations.

3.1. Chi Squared method

Chi-squared is broadly used feature selection method in machine learning for assessing the morality of an attribute. Chi squared has been used for feature selection in different domains (Pal et al., 2015; Bahassine et al., 2016; Sun et al.,2017; Jie et al., 2019). Chi-squared is primarily applied on categorical features of data. Chid-square between each feature and the target is calculated and the required numbers of features having best Chi-squared scores are considered. It computes the extent of independence of between categorical features.To calculate the Chi-squared score, let n_1 is the number of times feature x and class c come together, n_2 is the number of times feature x comes without class c, n_3 is the number of times class c comes without feature x, n_4 be the number of times neither x nor c occurs together. Let the size of the training set is N. Then the Chi-squared score is given by (Rajab, 2017).

$$score(x,c)=\frac{N\times(n_1n_4-n_2n_3)}{(n_1+n_2)\times(n_2+n_3)\times(n_3+n_4)\times(n_4+n_1)} \quad (1)$$

Features with low score are generally considered as irrelevant and not considered during the training phase of the machine leaning model.

3.2. Extra tree classifier method

Extra tree classifier (also known as Extremely Randomized Tree Classifier) is another largely used method for feature selection. It is an ensemble learning approach which builds up the results of many dissimilar decision trees to show its final result (Geurtset al.,2006; Pinto et al.,2015; Zafariet al.,2019). The feature importance class of extra tree classifier can be used to compute the feature importance of each feature. Feature importance assign score to all features present in the data. The higher the score the admissible is the feature concerning to the output variable.

All decision trees in the extra trees forest is built from the native training data. Afterwards, at each test node, every decision tree is supplied with arbitrary set of k features from the feature-set. Each decision tree must pick the finest feature to split the data grounded on certainbasis (typically Gini index (Chandra and Varghese, 2009; Jin, 2009). To compute the Gini index, let X is a feature having n distinct values, ($x_1, x_2, ..., x_n$), present in the data*N*. Let split on X divides the data N into N_1 and N_2. The Gini index for X is computed using equation (4).

$$Gini_X(N)=\frac{|N_1|}{|N|}Gini(N_1)+\frac{|N_2|}{|N|}Gini(N_2) \quad (2)$$

The $Gini(N)$ is computed using

$$Gini(N)=1-\sum_{i=1}^{m} p_i^2 \quad (3)$$

Where p_i is the probability that a row in N lies to class C_i and is calculated using $\frac{|C_{i,N}|}{|N|}$.

The Gini importance (feature importance) is then computed using

$$\Delta Gini(X)=Gini(N)-Gini_X(N) \quad (4)$$

To carry out feature selection, all features are sorted in descending order with respect to the Gini importance (feature importance) of each feature and the top features can be picked up.

3.3. Heat map

Heat Map is very prominent and widely used method for feature selection (Lin et al.,2013; Mengmeng et al., 2019). Heat Map uses the correlation that expresses how the features in a dataset linked to each other or target variable. This correlation might be positive (increment in the value of feature increments in the target variable's value) or negative (increment in the value of feature decrements in the target variable's value).

Using Heat Map it becomes easier to recognize the features that are most correlated to the target variable. The value of the correlation coefficient lies between [-1, 1].

— Value nearer to 0 means the correlation is weak.

— Value nearer to -1 means strong negative correlation

— Value nearer to 1 means strong positive correlation

All the features that have strong positive correlation to the target feature are selected to training the machine learning model.

3.4. Decision tree classifier

Classification (Friedl and Brodley, 1997; Kamber et al.,1997; Aggarwal,2004;Bhardwaj and Pal, 2011; Bertsimas and Dunn, 2017) comes under supervised learning techniques (Caruana and NiculescuMizil, 2006; Garciaet al.,2013) of machine learning. The primary objective of decision tree learning is to construct predictive model that very precisely and accurately predict the class of a given testing sample (Bradley et al.,1998; Carvalho and Freitas, 2004). All internal nodes of the decision tree specify a test on the dataset feature while the leaf represents the class. The branch represents result of the test directed on the datasetfeature at each internal level.

At each level of the decision tree, the triggering feature is decided by the splitting principle. The splitting principle uses the attribute selection method to find the best triggering feature to be used as split point. A triggering feature with maximum information gain ratio is used as the split-point.

So, if a sample, X, is given for which the landslide type class is unknown, the triggering features of the sample are tested for the decision tree. A path from the root of the decision tree to the leaf is detected. The leaf node notifies the class of landslide type to which the given sample belongs.

4. Experimental Results and Analysis

The step by step procedure followed to compute the accuracy of machine learning model using different feature selection methods is shown in Fig. 3. All the experiments have been conducted on system with 4 GB RAM, 500 GB hard disk and INTEL core i3 processor. The model is developed using Python language on IDLE 3.8 32-bit version.

4.1. Multi collinearity problem analysis

Multi collinearity (Mansfield and Helms, 1982; Allen, 1997; Alin, 2010) is a situation where one independent feature is greatly correlated to one or more independent features. That is to say, one independent feature can easily be predicted by other independent feature with considerable degree of validity. There are various strategies to inspect multi collinearity in data. We have used variance inflation factor (VIF) and tolerance for detecting multi collinearity problem.

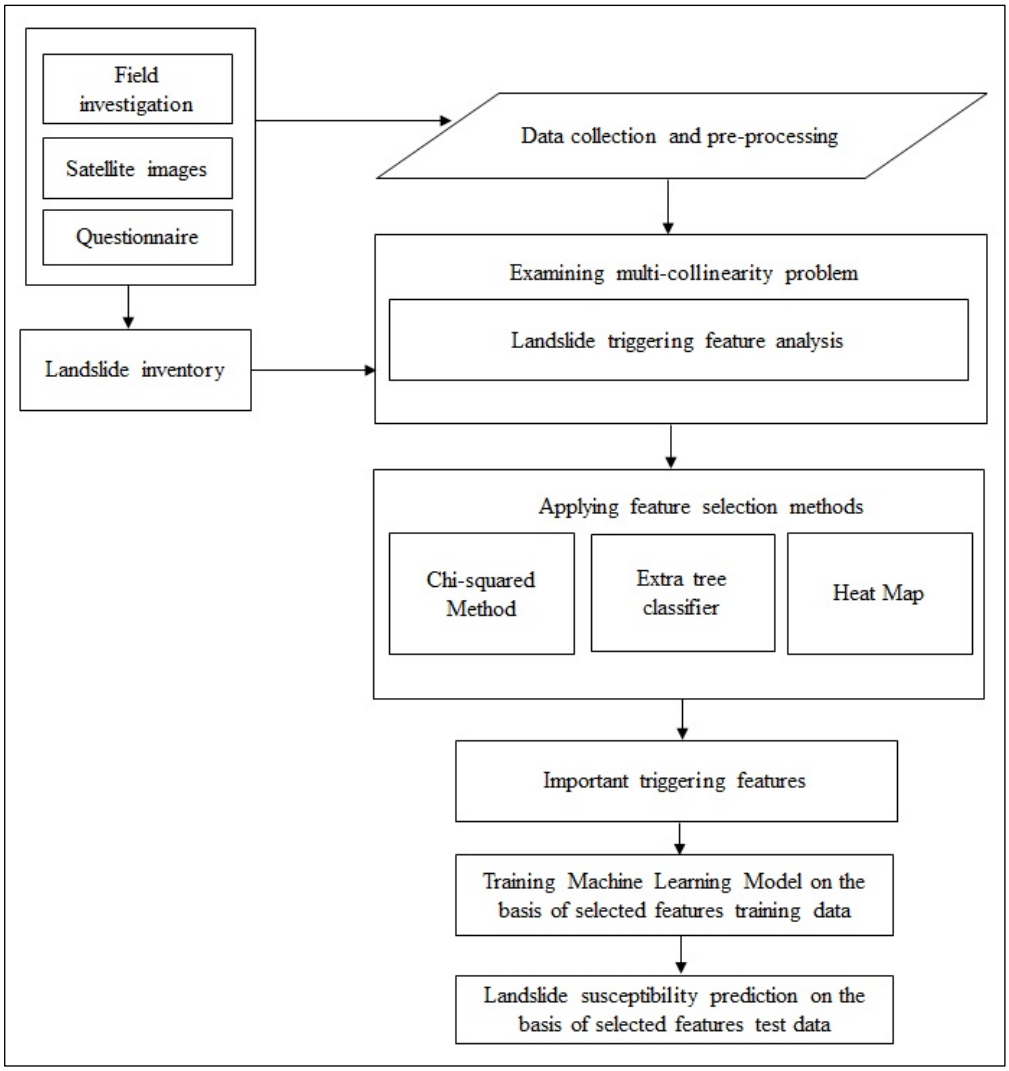

Fig.3. Flowchart of landslide prediction model.

A VIF of 10 or above specifies cause to be bothered about multi collinearity problem (Wang, 1996; Lin, 2008; Abdalla and Almghari, 2011). Tolerance is sharply linked to VIF and is inverse of it. The value of tolerance for different features should not to be less than 0.1 (Miles, 2005). In our study, multi collinearity is analysed among the features and the result is summarized in Table 3.

Table 3
Multi collinearity among features

Features	VIF	Tolerance
Slope	1.252	0.799
Elevation	2.961	0.338
LULC	1.266	0.790
Distance to road	2.171	0.461
Distance to drainage	1.513	0.661
Geology & geomorphology	3.005	0.333
Lineament density	3.025	0.331

It is comprehensible that the features are having VIF value much smaller than 10 and the tolerance value is also above 0.1. Lineament density is having maximum VIF value 3.025 and least tolerance value 0.331.

4.2. Triggering features selection and elimination

Features selection plays a critical role in predictive modelling. Table 4 and Table 5summarize the feature score using chi-square and extra tree classifier for all features present in the dataset. From Table 4, it is apparent that Geology & geomorphology and slope are two features having very less values of chi-square score with 0.024267 and 0.031531 respectively.

Table 4
Chi-square score of all features

Features	Score
Elevation	0.603995
Distance to drainage	0.135158
Lineament density	0.094374
Distance to road	0.073846
LULC	0.046241
Slope	0.031531
Geology & geomorphology	0.024267

Eliminating these two features can significantly decrease the complexity of the model and also help in reducing the learning time during training phase. FromTable 5 it is evident that lineament density and geology & geomorphology features are having insufficient values to be selected as important features. The feature importance score for lineament density and geology & geomorphology are 0.035189 and 0.069843 respectively.

Table 5
Score of all features using Extra Tree Classifier

Features	Score
Slope	0.272454
LULC	0.213393
Distance to drainage	0.201469
Elevation	0.118316
Distance to road	0.089336
Geology & geomorphology	0.069843
Lineament density	0.035189

If these features are not considered while training the model then the performance of the model can increase substantially. Fig. 4 and Fig. 5 summarize the feature selection result in more depictive manner and can be easily visualized to check the percentage of each feature that it contributes to its selection as an important feature. In Fig. 4, slope and geology & geomorphology have 2.4% and 3.1% of feature importance respectively.

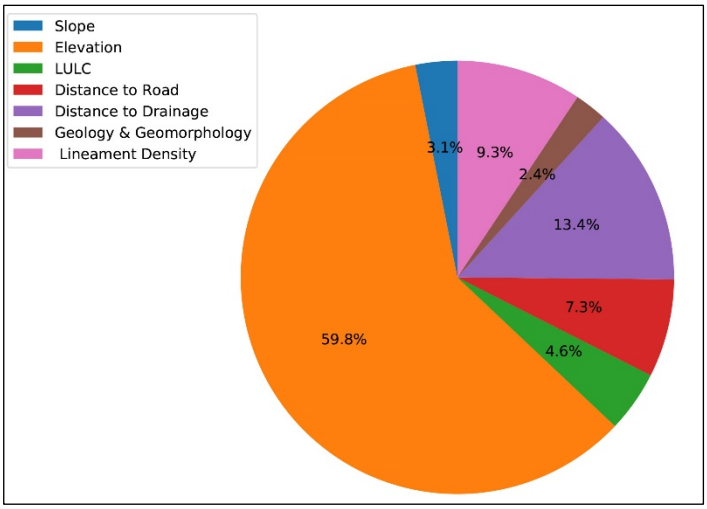

Fig.4. Feature importance using chi-square score.

Therefore, when chi-squared method is used as feature selection, these features contribute negligible to model performance and can be discarded. In Fig. 5, lineament density and geology & geomorphology have feature importance 3.9% and 6.7% respectively.

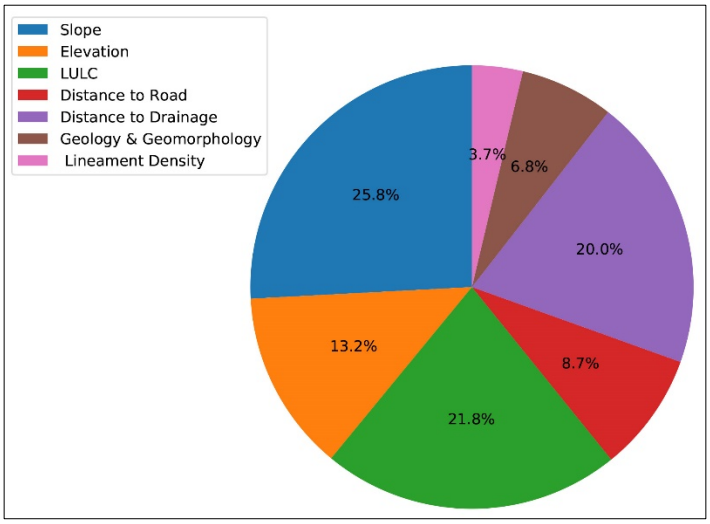

Fig.5. Feature importance using extra tree classifier.

These two features impart insignificant to performance of model and are discarded when extra tree classifier is used as feature selection method. Fig. 6 reveals that elevation and lineament density are least correlated to the target feature class with mere -0.21 and 0.095 correlation values.

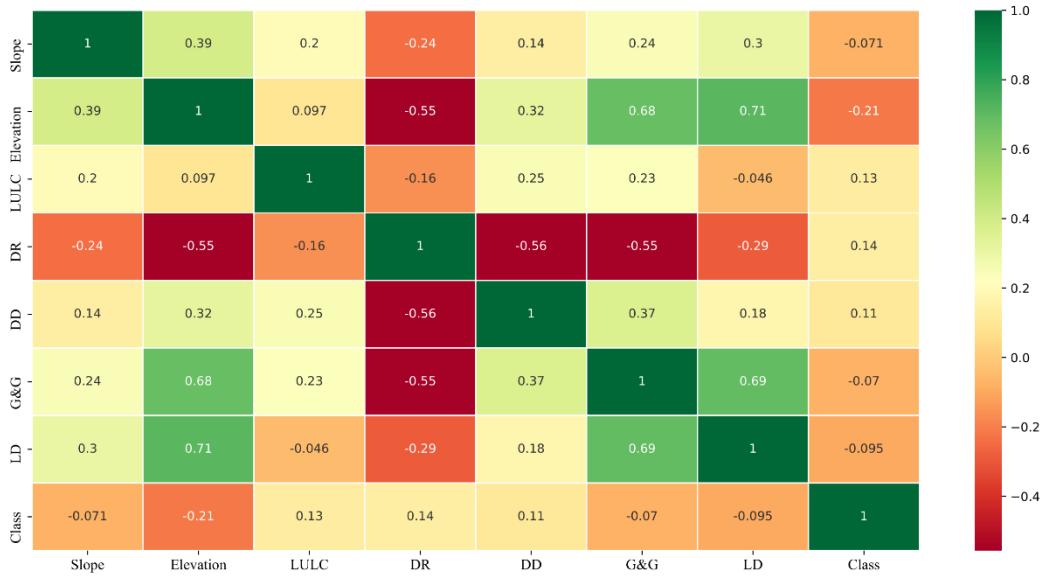

Fig.6. Heat map of different features.

*LULC (Land Use and Land Cover), DR (Distance to road), DD (Distance to drainage), G&G (Geology and geomorphology), LD (Lineament density)

So, these two feature are irrelevant when Heat Map is used as feature selection method. The performance of machine learning model is evaluated on the rest of the features.

From the above discussion, it can be inferred that selection and elimination of features is distinctly dependent on the feature selection method used. For example, slope may be an irrelevant feature for chi-squared method but it is a decisive feature for extra tree classifier. In the same way, elevation may be insignificant for Heat Map but it is an influential feature for chi squared method.

4.3. Validation and comparison

The landslide susceptibility prediction results of individual models have been validated using distinct test dataset. This dataset slice was not a part of training phase of the model. The whole dataset was split into two categories in a ratio of 70-30. The training dataset comprised of 70% and test data of 30% of the original dataset.Thereafter, the landslide susceptibility prediction model is trained used training dataset and its results validation is carried out on the test dataset. The accuracy of the landslide susceptibility prediction model is evaluated against different feature selection methods using the formula given below

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (5)$$

Where, *TP* (*True positive*) represents the number of landslide locations that are correctly classified, *TN* (*True negative*) represents the number of non-landslide locations that are correctly classified, *FN* (*False negative*) represents the

number of landslide locations that are classified as non-landslide locations and *FP* (*False positive*) represents the number of non-landslide locations that are classified as landslide locations.

4.3.1. Using accuracy measure

Table 6 shows the accuracy of the model when chi-squared method is used for feature selection to train the model. To begin with, the accuracy of the model is evaluated without eliminating any feature from the dataset and the model achieves a good accuracy of 90.74%.

Table 6

AUROC and Accuracy in case of Chi Square

Model	Eliminating triggering features	AUROC	Accuracy (%)
Model 1	Without eliminating any feature	0.979	90.74
Model 2	Eliminating geology & geomorphology	0.979	90.74
Model 3	Eliminating geology & geomorphology, slope	0.942	85.19
Model 4	Eliminating geology & geomorphology, slope, LULC	0.864	77.78

After this, features are eliminated one at a time depending on their chi-squared score and performance of the model is assessed. It is apparent that eliminating geology & geomorphology feature does not alter the accuracy of model. Therefore, geology & geomorphology can be discarded reducing the training time and complexity of the model. However, when slope is also eliminated the accuracy of the model is reduced significantly to 85.19%. So, it can be inferred that geology & geomorphology is the only unimportant feature when chi-squared method is used for feature selection.

Table 7reveals the effect of eliminating the unimportant features during the training phase of the model on the model's accuracy. The features are eliminated using feature importance score calculated using extra tree classifier method of feature selection. Lineament density being having the least scoreis eliminated first and the accuracy of the model is examined. Eliminating lineament density has no effect on the accuracy.

Table 7

AUROC and Accuracy in case of Extra Tree Classifier

Model	Eliminating triggering features	AUROC	Accuracy (%)
Model 1	Without eliminating any feature	0.979	90.74
Model 2	Eliminating lineament density	0.979	90.74
Model 3	Eliminating lineament density, geology & geomorphology	0.979	90.74

Model 4	Eliminating lineament density, geology & geomorphology, distance to road	0.968	88.89
Model 5	Eliminating lineament density, geology & geomorphology, distance to road, elevation	0.938	85.19

Therefore, it can be eliminated safely. Geology & geomorphology is eliminated next and no change in the accuracy is observed. Distance to road feature is removed further and the accuracy of the model is declined to 88.89%. So, it can be concluded that lineament density and geology & geomorphology are two irrelevant features when extra tree classifier is used as feature selection method. This will help in reducing the dimensionality of the model.

Table 8 shows the accuracy of machine learning model using Heat Map as a feature selection method. To evaluate the model performance, the features are eliminated based on their correlation value in the Heat Map. The maximum performance was accomplished by the model when two least significant features were eliminated. Thus, elevation and lineament density features were removed without affecting the accuracy of the model.

Table 8

AUROC and Accuracy in case of Heat Map

Model	Eliminating triggering features	AUROC	Accuracy (%)
Model 1	Without eliminating any feature	0.979	90.74
Model 2	Eliminating elevation	0.979	90.74
Model 3	Eliminating elevation, lineament density	0.979	90.74
Model 4	Eliminating elevation, lineament density, slope	0.936	85.19
Model 5	Eliminating elevation, lineament density, slope, geology & geomorphology	0.906	83.33

Although, slope and geology & geomorphology features also had small correlation values but eliminating these features decrease the accuracy remarkably. Therefore, when Heat Map is used as feature selection method, the elevation and lineament density were two irrelevant features and were removed.

4.3.2.Using ROC curve

The receiver operating characteristics (ROC) is a curve to appraise the performance of a machine learning model (Hanley and McNeil,1983). A ROC curve is a plot between true positive rate (TPR) and false positive rate (FPR). The true positive rate (TPR) and false positive rate (FPR) are calculate using the formulas given below:

$$True\ Positive\ Rate = \frac{TP}{TP+FN} \quad (6)$$

$$False\ Positive\ Rate = \frac{FP}{FP+TN} \quad (7)$$

The area under the ROC curve (AUROC) can be used as a measure to evaluate the performance of the model. The higher the area the better the model is. AUROC has been used in various domains for prediction (Kannanand Vasanthi, 2019), selecting features (Hiroshi, 2006), evaluating machine learning models (Andrew, 1997; Quentin, 1997), and in decision making.

The ROC curves for different feature selection methods have been shown in Fig. 7. From ROC curve of chi-squared method (Fig. 7a), it is clear that AUROC is 0.979 without eliminating any feature. The AUROC values remain unchanged when geology & geomorphology feature is eliminated to access model performance. The AUROC value is 0.942 when slope is also eliminated indicating that the model performance is reduced. The AUROC value is further declined to 0.864 when LULC is eliminated. Therefore, from AUROC also, it can be judged that in case of chi-squared method for feature selection, geology & geomorphology is only unimportant feature and can be discarded for training and assessing the performance of model.

In case of Extra Tree Classifier, the AUROC value without eliminating any feature is 0.979 (Fig. 7b). Based on the features' score, lineament density and geology & geomorphology are considered as unimportant features and eliminated before training the model over the dataset. AUROC value the model still remained to 0.979 indicating that eliminating these features does affect the accuracy of the model. Instead, the model comprehensibility has increased and learning time increased due to reduction in dimensionality of the dataset.

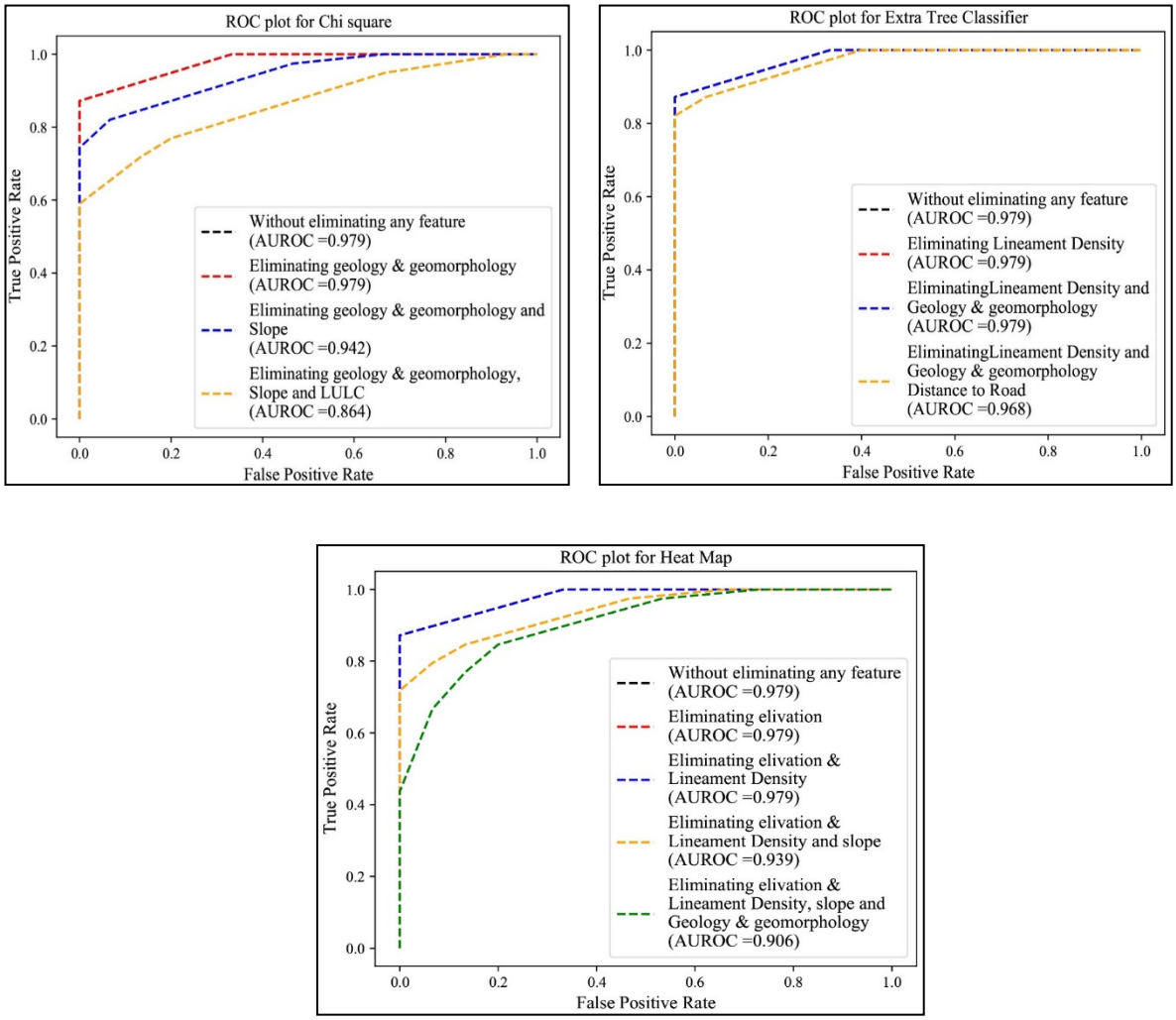

Fig.7. (Clockwise direction from top left) a. ROC plot for chi-square b. ROC plot for extra tree classifier c. ROC plot for heat map.

From Fig. 7c, it can be revealed that the Heat Map method of feature selection treats elevation and lineament density as the irrelevant features and model achieved and AUROC value of 0.979 after eliminating these unimportant features. Further elimination of features decrease the AUROC which in turn decrease the model performance e.g. eliminating slope slides down the value of AUROC to 0.939 which further reduced to 0.906 when geology & geomorphology is eliminated. Therefore, it can be signified that elevation and lineament density are irrelevant features and eliminating these before training the model would result a decrease in training time of model and eventually increase the model comprehensibility.

5. Discussions and conclusions

The paper discusses the use of three feature selection methods for landslide susceptibility prediction from a set of topographical and geological features. The landslide inventory of Kullu to Rohtang Pass transport corridor, Himachal

Pradesh, India was prepared to train and test the machine learning model. The features selection was carried out using chi squared, extra tree classifier and heat map methods to reveal the features which contribute most in landslide susceptibility prediction. Each feature selection method provides different set of features to train the model. Chi squared method of feature selection treats geology & geomorphology as an irrelevant feature and eliminates it before training the model on it and yields an AUROC with value 0.979. Extra tree classifier revealed lineament density and geology & geomorphology as unimportant features and the model exhibits AUROC value of 0.979 after eliminating these irrelevant features. On the other hand, heat map produces the lineament density and elevation as unimportant features and AUROC value of 0.979 is achieved by the model after eliminating these features. The relationship between landslide and different conditioning factors can differ from area to area, but there can be similarities in some areas. For instance, the elevation and slope is found more relevant with landslides which are also found crucial in other studies also (Chen et al., 2020; Akgun et al., 2012). Moreover, feature selection methods except heat map shows slope and elevation as important features which are also authenticated by previous studies (Wang et al., 2017; Hong et al., 2018c). All the feature selection methods reduce the dimensionality of the dataset and consequently decreases the training time of the model. This also in turn decreases the model complexity and makes it more comprehensible. It must be noted that the model training time is decreased in each of the feature selection method, comparing the training time of model using different feature selection methods could be an interesting direction for future research. Using swarm intelligence methods of feature selection and training model is another direction for future research.

References

Abdalla, M., Almghari, K.I., 2011. Remedy of multicollinearity using ridge regression. Journal of Al Azhar University Gaza (Natural Sciences) 13, 119-134.

Achu, A.L., Aju, C.D., Pham, Q.B., Reghunath, Rajesh., Anh, Duong Tran 2022. Landslide susceptibility modelling using hybrid bivariate statistical-based machine-learning method in a highland segment of Southern Western Ghats, India. *Environ Earth Sci* 81(13), 360 .

Achu, A.L., Aju, C.D., Reghunath, Rajesh., 2020. Spatial modeling of shallow landslide susceptibility: a study from the southern western ghats region of Kerala, India, Annals of GIS, 26(2), 113-131.

Aggarwal, C.C., 2004. On demand classification of data streams. Proceedings ACM SIGKDD International Conference Knowledge Discovery Data Mining, pp. 503-508.

Aghdam, I.N., Varzandeh, M.H.M., Pradhan, B., 2016. Landslide susceptibility mapping using an ensemble statistical index (wi) and adaptive neuro-fuzzy inference system (ANFIS) model at Alborz mountains (Iran). Environ. Earth Sci. 75, 553.

Akgun, A., Sezer, E. A., Nefeslioglu, H. A., Gokceoglu, C., and Pradhan, B., 2012. An easy-to-use MATLAB program (MamLand) for the assessment of landslide susceptibility using a Mamdani fuzzy algorithm. Land Degradation & Development 38(1), 23-34.

Alin, A., 2010. Multicollinearity. Wiley Interdisciplinary Reviews: Computational Statistics 2: 370-374.

Allen, M.P., 1997. The problem of multicollinearity.Understanding regression analysis.Springer, Boston, MA.

Andrew P.B., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7), 1145-1159.

Bahassine, S., Madani, A., Kissi, M., 2016.An improved Chi-sqaure feature selection for Arabic text classification using decision tree. 11th International Conference on Intelligent Systems: Theories and Applications (SITA), pp. 1-5.

Bertsimas, D., Dunn, J., 2017. Optimal classification trees.Machine Learning 106, 1039-1082.

Bharadwaj, B.K., Pal, S., 2011. Data Mining: A prediction for performance improvement using classification. International Journal of Computer Science and Information Security 9(4), 136-140.

Bradley, P.S., Fayyad, U.M., Reina, C., 1998.Scaling clustering algorithms to large databases.Knowledge Discovery and Data Mining, 9-15.

Caruana, R., Niculescu-Mizil, A., 2006.An empirical comparison of supervised learning algorithms. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, pp. 161-168.

Carvalho, D.R., Freitas, A.A., 2004. A hybrid decision tree/genetic algorithm method for data mining. Information Sciences 163(1-3), 13-35.

Chandra, B., Varghese, P.P., 2009. Fuzzifying Gini Index based decision trees.Expert Systems with Applications 36(4), 8549-8559.

Chandrashekar, G., Sahin, F., 2014.A survey on feature selection methods. Computers & Electrical Engineering 40(1), 16-28.

Chen, W., Li, Y., Xue, W., Shahabi, H., Li, S., Hong, H., Wang, X., Bian, H., Zhang, S., Pradhan, B., and Ahmad, B. B., 2020. Modeling flood susceptibility using data-driven approaches of naive Bayes tree, alternating decision tree, and random forest methods. Science of the Total Environment, 701, 134979.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D.T., Duan, Z. Ma, J. A., 2017. Comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena 151, 147-160.

Feizizadeh, B., Ghorbanzadeh, O., 2017. GIS-based interval pairwise comparison matrices as a novel approach for optimizing an analytical hierarchy process and multiple criteria weighting. GI_ Forum 1, 27-35.

Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sensing of Environment 61(3),399-409.

Ge, L., Li, G.Z., You, M.Y., 2009. Embedded feature selection for multi-label learning. Journal of Nanjing University (Natural Sciences) 45(5), 671-676.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3-42.

Ghahramani, Z., 2015. Probabilistic machine learning and artificial intelligence. Nature 521, 452-459.

Ghorbanzadeh, O., Blaschke, T., Aryal, J., Gholaminia, K., 2018b.A new GIS-based technique using an adaptive neuro-fuzzy inference system for land subsidence susceptibility mapping. J. Spat. Sci.1-17.

Goyal, S., Maheshwar., 2019. Naive bayes model based improved k-nearest neighbor classifier for breast cancer prediction. In: Luhach A., Jat D., Hawari K., Gao XZ., Lingras P. (eds.), Advanced Informatics for Computing Research, ICAICR, Communications in Computer and Information Science, Springer, Singapore. pp1075.

Guo, Y., Chung, F., Li, G., 2016. An ensemble embedded feature selection method for multi-label clinical text classification. IEEE International Conference on Bioinformatics and Biomedicine,pp. 823-826.

Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148 (3), 839-843.

Hiroshi Mamitsuka., 2006. Selecting features in microarray classification using ROC curves. Pattern Recognition 39(12), 2393-2404.

Holbling, D., Fureder, P., Antolini, F., Cigna, F., Casagli, N., Lang, S., 2012. A semi-automated object-based approach for landslide detection validated by persistent scatterer interferometry measures and landslide inventories. Remote Sens. 4, 1310-1336.

Hong, H., Tsangaratos, P., Ilia, I., Liu, J., Zhu, A. X., Chen, W., 2018c. Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China. Science of The Total Environment 625, 575-588.

Hong, H., Chen,W., Xu, C., Youssef, A.M., Pradhan, B., Tien Bui, D., 2017. Rainfall-induced landslide susceptibility assessment at the Chongren area (China) using frequency ratio, certainty factor, and index of entropy.Geocarto Int. 32, 139-154.

Jie Wang., Jing Xu., Chengan Zhao., Yan Peng., Hongpeng Wang., 2019. An ensemble feature selection method for high-dimensional data based on sort aggregation. Systems Science & Control Engineering 7(2), 32-39.

Jin, R., Breitbart, Y., Muoh, C., 2009. Data discretization unification. Knowl. Inf. Syst. 19(1), 1-29.

Kamber, M., Winstone, L., Wan G., Shan C., Jiawei H., 1997. Generalization and decision tree induction: efficient classification in data mining. Proceedings Seventh International Workshop on Research Issues in Data Engineering. High Performance Database Management for Large-Scale Applications, Birmingham, UK, pp. 111-120.

Kannan R., Vasanthi V., 2019. Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease. Soft Computing and Medical Bioinformatics.Springer Briefs in Applied Sciences and Technology.Springer, Singapore.

Lavrac, N., 1999. Machine learning for data mining in medicine.Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, pp. 47-62.

Lee, I.H., Lushington, G.H., Visvanathan, M., 2011.A filter-based feature selection approach for identifying potential biomarkers for lung cancer.Journal of Clinical Bioinformatics 1(1), 11.

Liang, D., Tsai, C.F., Wu, H.T., 2015. The effect of feature selection on financial distress prediction. Knowledge Based Systems 73, 289-297.

Lin, F., 2008.Solving multicollinearity in the process of fitting regression model using the Nested estimate procedure. Qual. Quant. 42, 417-426.

Lin, W., Chu, H., Wu, J., Sheng, B., Chen, Z., 2013.A Heat-Map-Based algorithm for recognizing group activities in videos. IEEE Transactions on Circuits and Systems for Video Technology 23(11), pp. 1980-1992.

Lu, M., 2019. Embedded feature selection accounting for unknown data heterogeneity. Expert Systems with Applications 119, 350-361.

Maheshwar, Kaushik, K., Arora, V., 2015. A hybrid data clustering using firefly algorithm based improved genetic algorithm. Procedia Computer Science 58, 249-256.

Maheshwar, Kumar, G., 2019. Breast cancer detection using decision tree, naive bayes, KNN and SVM classifiers: A comparative study. International conference on smart systems and inventive technology (ICSSIT), Tirunelveli, India, pp. 683-686.

Mansfield, E.R., Helms, B.P., 1982. Detecting multicollinearity.The American Statistician 36(3), 158-160.

Martire, D., De Rosa, M., Pesce, V., Santangelo, M.A., Calcaterra, D., 2012.Landslide hazard and land management in high-density urban areas of Campania region, Italy. Nat. Hazards Earth Syst. Sci. 12, 905-926.

Mengmeng Li., Zhigang Shang., Zhongliang Yang., Yong Zhang., Hong Wan., 2019. Machine learning methods for MRI biomarkers analysis of pediatric posterior fossa tumors.Biocybernetics and Biomedical Engineering 39(3), 765-774.

Miles, J., 2005. Tolerance and variance inflation factor. In Encyclopedia of statistics in Behavioral Science; Everitt, B.S., Howell, D.C., Eds.; John Wiley and Sons: Hoboken, NJ, USA, pp. 2055–2056.

Myronidis, D., Papageorgiou, C., Theophanous, S., 2016. Landslide susceptibility mapping based on landslide history and analytic hierarchy process (AHP). Nat. Hazards 81, 245-263.

Narayanan, B.N., Djaneye B.O., Kebede, T.M., 2016. Performance analysis of machine learning and pattern recognition algorithms for Malware classification.IEEE National aerospace and electronics conference (NAECON)and Ohio innovation summit (OIS), Dayton, OH,pp. 338-342.

Pal, B., Sadia Zaman., Md. Abu Hasan., 2015. Chi-Square statistic and principal component analysis based compressed feature selection approach for Naive Bayesian Classifier. Journal of Artificial Intelligence Research & Advances 2(2), 16-23.

Pham, Quoc Bao., Achour, Yacine., Ali, Sk Ajim., Parvin, Farhana., Vojtek, Matej., Vojteková, Jana., Al-Ansari, Nadhir., Achu, A. L.., Costache, Romulus., Khedher, Khaled Mohamed., Anh, Duong Tran., 2021. A comparison among fuzzy multi-criteria decision making, bivariate, multivariate and machine learning models in landslide susceptibility mapping. Geomatics, Natural Hazards and Risk 12(1), 1741-177.

Pinto, A., Pereira, S., Correia, H., Oliveira, J., Rasteiro, D.M.L.D., Silva, C.A., 2015.Brain tumour segmentation based on extremely randomized forest with high-level features. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3037-3040.

Porkodi, R., 2014. Comparison of filter based feature selection algorithms: An overview. International journal of Innovative Research in Technology & Science 2(2), 108-113.

Pourghasemi, H., Gayen, A., Park, S., Lee, C.W., Lee, S., 2018. Assessment of landslide-prone areas and their zonation using logistic regression, logitboost, and naivebayes machine-learning algorithms.Sustainability 10, 3697.

Pourghasemi, H.R., Kerle, N., 2016. Random forests and evidential belief function-based landslide susceptibility assessment in western Mazandaran province, Iran. Environ. Earth Sci. 75, 185.

Pradhan, B.A., 2013. Comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput.Geosci. 51, 350-365.

Premakanthan, P., Mikhael, W.B., 2001. Speaker verification/recognition and the importance of selective feature extraction: review. Proceedings of the 44th IEEE 2001 Midwest Symposium on Circuits and Systems.MWSCAS 1, pp. 57-61.

Qiao, L.Y., Peng, X.Y., Peng, Y., 2006. BPSO-SVM wrapper for feature subset selection.DianziXuebao(Acta Electronica Sinica) 34(3), 496-498.

Quentin T.W., 1997. Targeting the poor using ROC curves. World Development 25(12), 2083-2092.

Rajab, K.D., 2017. New hybrid features selection method: a case study on websites phishing. Security and Communication Networks 2017, 1-10.

Garcia, S., Luengo, J., Saez, J.A., Lopez, V., Herrera, F., 2013.A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering25(4), pp. 734-750.

Saaty, T.L., 1990a. How to make a decision: the analytic hierarchy process. European Journal Operational Research 48(1), 926.

Saha, A.K., Gupta, R.P., Sarkar, I., Arora, M.K., Csaplovics, E., 2005. An approach for GIS-based statistical landslide susceptibility zonation-with a case study in the Himalayas. Landslides 2(1), 61-69.

Sezer, E.A., Pradhan, B., Gokceoglu, C., 2011. Manifestation of an adaptive neuro-fuzzy model on landslide susceptibility mapping: Klang valley, Malaysia. Expert Syst. 38, 8208-8219.

Solway, L., 1999. Socio-economic perspective of developing country megacities vulnerable to flood and landslide hazards. In Floods and Landslides: Integrated Risk Assessment. Springer, Berlin, Heidelberg. pp. 245-260.

Somol, P., Baesens, B., Pudil, P., Vanthienen, J., 2005. Filter-versus wrapper-based feature selection for credit scoring. International Journal of Intelligent Systems 20(10), 985-999.

Sun, J., Zhang, X., Liao, D., Chang, V., 2017.Efficient method for feature selection in text classification.International Conference on Engineering and Technology (ICET), pp. 1-6.

Svalova, V., 2018.Landslide risk management for urbanized territories.In Risk Management Treatise for Engineering Practitioners.IntechOpen.

Tirelli, T., Pessani, D., 2011.Importance of feature selection in decision-tree and artificial-neural-network ecological applications.Alburnusalburnusalborella: A practical example. Ecological Informatics 6(5), 309-315.

Wang, F., Xu, P., Wang, C., Wang, N., Jiang, N., 2017. Application of a GIS-Based Slope Unit Method for Landslide Susceptibility Mapping along the Longzi River, Southeastern Tibetan Plateau, China, ISPRS International Journal of Geo-Information, 6(6).

Wang, G.C.S., 1996. How to handle multicollinearity in regression modelling.The Journal of Business Forecasting Methods & Systems 15(1), 23-27.

Windeatt, T., Duangsoithong, R., Smith, R., 2011. Embedded feature ranking for ensemble MLP classifiers. IEEE transactions on neural networks 22(6), pp. 988-994.

Xue, B., Cervante, L., Shang, L., Browne, W.N., Zhang, M., 2012. A multi-objective particle swarm optimisation for filter-based feature selection in classification problems. Connection Science 24(2-3), 91-116.

Zafari, A., Zurita-Milla, R., Izquierdo-Verdiguier, E., 2019.Evaluating the performance of a Random Forest Kernel for land cover classification. Remote Sensing 11(5), 1-20.