A Comparative Study of Different Machine Learning Models for Landslide Susceptibility Prediction: A Case Study of Kullu to Rohtang Pass Transport Corridor, India

Nirbhav[1,*], Anand Malik[2], Maheshwar[3], Mukesh Prasad[4], Atul Saini[5], Nguyen Thanh Long[6]

[1] Department of Geography, Delhi School of Economics, University of Delhi, Delhi, India

[2] Swami Shraddhanand College, University of Delhi, Delhi, India

[3] TGT Computer Science, Directorate of Education, Delhi Government, Delhi, India

[4] School of Computer Science, Australian Artificial Intelligence Insitute, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia

[5] Delhi School of Climate Change & Sustainability, Institution of Eminence, University of Delhi, Delhi, India

[6] Vietnam Institute of Geosciences and Mineral Resources, Economic Geology and Geomatics Department, Hanoi, Vietnam

*Corresponding author

E-mail addresses: captainnirbhav@gmail.com (Nirbhav)

Abstract

Landslide susceptibility prediction can be considered a crucial step in landslide risk assessment. This prediction helps in planning the land use properly. The primary aim of the study is to investigate different machine learning methods and develop anatomy to train and validate the landslide susceptibility prediction models with the help of various statistical techniques. The Kullu-Rohtang pass transport corridor has been selected as the study area. Initially, a landslide inventory was prepared using different sources and nine landslide triggering features were used for further study. All landslide locations in the study area were arbitrarily divided into a ratio of 67:33 to train and test various landslide susceptibility prediction models. The best-triggering features were chosen with the help of the information gain ratio (IGR) defining the predictive capability of different triggering features. Afterwards, five landslide susceptibility prediction models were constructed using a decision tree, k-nearest neighbour (KNN), Gaussian naïve Bayes, support vector machine (SVM) and multilayer perceptron (MLP). The comparison and validation study of different resulting models was done by applying the receiver operating characteristic (ROC) curve, the kappa index and other statistical methods. Results show that the different models have the outstanding predictive capability with the decision tree model (100 %), the Gaussian naïve Bayes model (100 %), the SVM model (100%), and the MLP model (100 %) and the KNN model (99.9 %). The result indicates statistical differences among various models. The validation results demonstrate the perfect agreement between the expected and predicted landslides along the transport corridor.

Keywords: Landslide Susceptibility Prediction, Machine Learning, Statistical Measures, Landslide Inventory, Transport Corridor

1. Introduction

Landslides are among the most severe natural hazards that possess a menace to human lives and the economy of the prone regions. The Kullu-Rohtang Pass transport corridor is highly exposed to landslides and the vulnerability has been further increased due to road construction activities and deforestation. The increasing demand for land and the rapid urbanization has further escalated the threats of landslide occurrence (Saha et al. 2005; Achour et al. 2017). The key to lessen the landslides is to predict the frequency and the locations accurately. Identifying the site and type of landslide is generally conducted through landslide susceptibility analysis considering different meteorological and geo-environmental factors. There are several different procedures for predicting landslide susceptibility. In this era of computers, machine learning techniques have emerged as a boon for solving various scientific as well as engineering problems. These machine learning techniques are data-driven and are followed to predict landslide susceptibility accurately and successfully (Brenning 2005; Marrapu and Jakka 2013). These predictions help to take adequate actions to mitigate the landslide hazards and also provide a base for the administration to take appropriate steps for the development of the area (Lian et al. 2013; Pourghasemi et al. 2013; Bai et al. 2014; Trigila et al. 2015; Rokach 2016; Wang et al. 2016). Therefore, landslide susceptibility prediction is crucial for better land use planning and raising the economy of the region.

The landslide susceptibility prediction has been done using different approaches including heuristic, statistical and different machine learning methods (Lee 2005; Bui et al. 2011; Pradhan 2013; Felicísimo et al. 2013; Costanzo et al. 2014; Achour et al. 2018). Various statistical approaches applied for landslide susceptibility prediction are comprised of multivariate regression, information value techniques, linear regression, and discriminant analysis (Lee 2005; Guzzetti et al. 2006; Budimir et al. 2015; Chen et al. 2017b; Arabameri et al. 2019). The evolution of machine learning has resulted in different models with upgraded outcomes. These models include decision tree (Nefeslioglu et al. 2010; Pradhan 2013; Maheshwar and Kumar 2019), support vector machine (SVM) (Brenning 2005; Yao et al. 2008; Yilmaz 2010; Yeon et al. 2010; Pradhan 2013; Pourghasemi et al. 2013; Hong et al. 2015; Chen et al. 2017a; Kalantar et al. 2018; Pourghasemi and Rahmati 2018), naïve Bayes (Pham et al. 2016, 2017; Goyal and Maheshwar 2019), multilayer perceptron models and entropy-based models (Zare et al. 2013; Pham et al. 2017; Esteves et al. 2019). These machine learning models have been used in various dimensions, for example, mapping flood-prone regions, gully erosion prediction and ecological modelling. These machine learning models have been used by many researchers to find the best model for landslide susceptibility prediction (Carrara and Pike 2008). Some bio-inspired algorithms like particle swarm optimization (PSO) and genetic algorithm have also been used (Maheshwar et al. 2015; Jahed Armaghani et al. 2017; Chen et al. 2017b; Mohamad et al. 2018).

The primary objective of this study is to learn about different landslide triggering features to predict the landslide susceptibility in a region with different machine learning models. The transport corridor faces huge losses due to landslides occurring during the monsoon season leading to the requirement of appropriate landslide susceptibility prediction to take prevention measures. It has been observed that many times the nearby regions along the transport corridor face a complete cut-off because of the massive landslide occurrences. In this study, five machine learning models such as the decision tree, the k-nearest neighbour (KNN), the naïve Bayes, the support vector machine (SVM) and the multilayer perceptron (MLP) have been used and a comparative study of their results has been done to find out the best suitable model for landslide susceptibility prediction in the study area.

The entire research work is compiled as follows: After a brief introduction in section 1, section 2 discusses the inventory preparation and the dataset with triggering features. It summarizes all the resources of the dataset and landslide inventory with significant triggering features. Section 3 enlightens the various methodologies along with statistical measures used in the study. Different machine learning models and accuracy assessment measures have been discussed in this section. Section 4 explores the results by evaluating each machine learning model on training and validating the dataset separately using different quality parameters. Different tests conducted to analyze the statistical differences among the various machine learning models are discussed in this section. The area under the receiver operating characteristic (AUROC) curve for different models is examined to evaluate their performances. Thereafter, a brief discussion and conclusion are given in section 5 followed by references.

2. Study Area and database

2.1 Study Area

The study area selected for carrying out research is along the Kullu-Rohtang Pass transport corridor with a length of 90 km. The latitudinal extent of the study area is from 32° 0’ 0’’ N to 32° 20’ 0’’ N and the longitudinal is from 77° 5’0’’E to 77° 15’0’’E (Fig. 1). To perform the landslide susceptibility prediction, a buffer of one kilometre on both sides of the corridor is considered. The study area lies in the state of Himachal Pradesh of India in the Himalayas ranges with elevations ranging between 1,279 m and 3,979 m from mean sea level (MSL). The average rainfall observed in the study area is about 1,363 mm. Most of the landslides primarily occurred from July to September because of heavy rainfall during these months. The temperature range lies from 4ºC to 25ºC.

The study area is comprised of various types of soil such as red loamy soil, brown hill soil and mountain meadow soil. The river Beas flows along the transport corridor. Different engineering activities like urbanization, road construction etc. along with deforestation have scaled up the landslide frequencies. The landslide incidences have led to escapable effects on transportation because many times the transportation facilities completely disconnect the region and impact the region's economy adversely.

2.2 Database

To carry out the study, ASTER DEM with 30m spatial resolution was used to perform topographic analysis. The survey of India (SoI) topographic sheet no. 52 H/3 and 52 H/4 on the scale of 1:50, 000 were used to digitize the benchmarks. ASTER DEM, USGS was used to analyse the topographic and geographic features such as elevation, slope, distance to drainage and distance to the road. Google Earth and Landsat 8 OLI, USGS was used to prepare land use and land cover (LULC) data. Geological quadrangle maps and GSI were used to prepare Geological and geomorphological data while ground water prospects maps by NRSA were used for preparing lineament density data. Landslide occurrence locations, their types, frequency and occurrence year were cumulated from BRO, Manali and PWD, and Kullu (Table 1).

Landslide inventory encompasses important data for predicting landslide occurrence (Pandey et al. 2022). The landslide inventory has been constructed using Geographic Information System (GIS) software (ArcGIS 10.2.2) with the help of satellite imageries and GPS waypoints. An overarching field inspection of 54 landslides eventuated along the transport corridor has been done using the Global Position System (GPS) and Google earth images. The spatial-temporal map has been prepared with the help of landslide data cumulated from BRO, Manali and PWD, Kullu. The major types of landslides that occurred along the transport corridor are rock falls and debris slides (Fig. 2). The landslide types are as per the records of BRO, Manali and PWD, Kullu. The largest and the smallest landslides mapped along the transport corridor are 4000 m^3 and 120 m^3 respectively.

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

2.3 Methodology

2.3.1. Training and testing dataset preparation

Landslide prediction can be considered as a binary classification problem where each landslide index is classified into landslide class and non-landslide class. Landslide locations are represented with a value of "1" and the non-landslide locations with a value of "0". Based on original triggering features, the dataset is pre-processed first. All the features are normalized using min-max normalization with Analytical Hierarchy Process (AHP). For landslide prediction, the landslide locations datasets are grouped into training and validating datasets. The training dataset is used for training the different machine learning models while the validating dataset is used for testing the models' performance and accuracy. All the landslide locations are divided into these two subsets randomly with a ratio of 67:33. The 36 landslide locations are used to train the models while 18 landslide locations are used to test the models' performance. The same amount of non-landslide locations are randomly sampled from the study area and added to the training and test datasets. In the next step, the nine landslide triggering features are analysed to train and test datasets.

2.3.2. Landslide triggering features analysis

Multicollinearity refers to the situation when there is a dependency among the triggering features in the dataset owing to a high correlation that results in errors during analysis (Bui et al. 2011). To analyse multicollinearity, different methods have been suggested such as the conditional index, the variance decomposition proportions, Pearson's correlation coefficients and variance inflation factors (VIF) and tolerances. The VIF and tolerances are the most widely used method to detect the multicollinearity among the triggering features while Pearson's correlation coefficients are also commonly used in different fields (O'brien 2007).

To spot the multicollinearity, the standard errors' variation of the triggering features are measured using VIF and tolerances. The higher the value of VIF the higher the multicollinearity. A VIF of a value of 10 or above and tolerance with a value less than 0.1 indicate the multicollinearity problem present in the dataset. Pearson's correlation coefficient method is used to assess the correlation coefficient of two landslide triggering features. Pearson's correlation coefficient can be stated as the covariance of two triggering features divided by the product of their standard deviations. A value of Pearson's correlation coefficient higher than 0.7 shows high collinearity among the features.

All landslide triggering features present in the initial dataset may not contribute equally to the prediction of a landslide. Some features may act as noise and reduce the landslide predictive ability of the machine learning model. Consequently, the predictive ability of landslide triggering features should be analysed and the features with low predictiveness should be eliminated from the dataset. This will help in attaining more accuracy of the machine learning model.

Different methods have been used to quantify the predictive ability of features, for example, Information Gain, Relief, Fuzzy-Rough Set and Information Gain Ratio (IGR). Information gain ratio is a very popular and commonly used method for selecting triggering features (Quinlan 1996; Tien Bui et al. 2016). Attributes having a higher value of the IGR have high predictive ability.

2.3.3. Machine learning models for predictions

To predict the landslides in the selected region, five machine learning models, Decision Tree (Bhardwaj and Pal 2012; Bertsimas and Dunn 2017; Dinov 2018), K-Nearest Neighbour (KNN) (Bröcker and Smith 2007; Pedregosa et al. 2011), Gaussian Naïve Bayes (Tien Bui et al. 2012), Support Vector Machine (SVM) (Tien Bui et al. 2012) and Multilayer Perceptron (MLP) (Kavzoglu and Mather 2003) are used. The research has been carried out using Python language with IDLE Python 3.8 version and ArcGIS 10.

2.3.4. Accuracy analysis and comparison

2.3.4.1. Area under receiver operating characteristics (AUROC) curve and Kappa index

The receiver operating characteristics (ROC) curve is used to assess the performance of machine learning models. ROC curve represents the graph between Sensitivity and Specificity with different thresholds. To perform the comparative study, the AUROC value can be used to compare the overall performance of machine learning models. AUROC represents the probability with which a model correctly classifies a randomly chosen landslide location as a landslide. An AUROC with a value of 1 indicates that the model has perfectly classified all landslide and non-landslide locations.

Kappa index can be used to measure the reliability of a machine learning model (Cohen 1960; Saito et al. 2009). Kappa index indicates the machine learning model's ability to classify the landslide locations and is evaluated as the proportion of observed agreement beyond that expected by chance (Guzzetti et al. 2006). The kappa index value shows an agreement with 0.8-1.0 meaning the perfect agreement, 0.6-0.8 meaning substantial, 0.4-0.6 meaning moderate, 0.2-0.4 means fair, 0.2-0.0 means slight and ≤ 0.0 means the poor agreement is there.

2.3.4.2. Accuracy measure and different quality parameters

To access the performance of the machine learning model, five statistical measures have been used namely *Positive predicted value, Negative predicted value, Sensitivity, Specificity and Accuracy*. The Positive predicted value shows the probability of the locations correctly classified as landslide locations and the Negative predicted value represents the probability of the locations correctly classified as non-landslide locations. Sensitivity represents the proportion of landslide locations that are correctly classified as landslide locations. Specificity represents the proportion of non-landslide locations that are correctly classified as non-landslide locations. Accuracy measures the proportion of landslide and non-landslide locations that the machine learning model has correctly classified.

$$Positive\ predicted\ value\ (\%) = \frac{TP}{TP+FP} \quad (1)$$

$$Negative\ predicted\ value\ (\%) = \frac{TN}{TN+FN} \quad (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

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.

2.3.4.3. Statistical evaluation measures

Different parametric and non-parametric statistical methods can be implemented to check if a machine learning model can be counted better than any other model or not. The Friedman test is a non-parametric method that can be used without having prior knowledge of the data and is used in this study. Friedman test can perform multiple comparisons at once to detect the significant differences among the models (Friedman 1937). The Friedman test is based on the null hypothesis that states that there is no significant difference in the performances of machine learning models. The p-value returned by the Friedman test represents the probability of null hypothesis rejection.

The Wilcoxon signed-rank test is also used to check the statistical significance of the difference between the two models. This model is also based on the null hypothesis stating that there is no statistical difference between the two models (Beasley and Zumbo 2003). A p-value less than 0.05 represent a significant difference between the models and greater than 0.05 means there is no significant difference between the models.

3. Results and discussion

3.1. Landslide triggering features analysis

In the presented research, multicollinearity among the landslide triggering features is determined by using VIF values along with tolerance (Table 3). Table 4 shows the Pearson's correlation coefficient of different features. It can be inferred from the result that the highest VIF value is 1.592 and the lowest tolerance is 0.628. For multicollinearity among the nine triggering features, the VIF value must be greater than 10 and tolerance must be less than 0.1 Geology & geomorphology and lineament density (0.69) showed the highest Pearson's correlation value. Nevertheless, this value is less than 0.7 which represents high collinearity (Table 4).

Following this, the step is to choose the best triggering features. In this research, the IGR value and its standard deviation for nine triggering features are calculated as shown in table 5. It could be noticed that elevation (0.603995) has the highest IGR value. It is succeeded by distance to drainage (0.135158), lineament density (0.094374), distance to road (0.073846), LULC (0.046241), slope (0.031531), geology & geomorphology (0.024267), aspect (0.001268) and relative relief (0.001179). It can be observed that the IGR value of aspect and relative relief is too low to contribute to the prediction of landslide events. Therefore, these two features can be treated as noise and are removed for further analysis.

3.2. Machine learning models results and analysis

After selecting the best-triggering features, five machine learning models Decision Tree, KNN, Gaussian Naïve bays, SVM and MLP are trained on the training dataset (Fig. 3). Tables 6 and 7 summarize the results. It can be inferred that the performance of the Decision Tree is highest in respect of classification accuracy at 100 %. This is succeeded by the MLP model (95.59 %), the SVM model (94.12 %), the KNN model (91.18 %) and the Gaussian Naïve Bays model (86.76 %).

The Decision Tree has a 100 % positive predicted value indicating that the model has correctly predicted the landslide in 100 % of cases. The positive predicted value for the SVM model 94.29 % is followed by the MLP model (91.89 %), the KNN model (91.18 %) and the Gaussian Naïve Bayes model (87.88 %). The negative predicted value is highest for the Decision Tree model (100 %) and the MLP model (100 %) shows that the probability of correctly classifying the non-landslide locations is 100 %. The SVM model has a negative predicted value (96.96 %) followed by the KNN model (91.18 %) and the Gaussian Naïve Bayes model (85.71 %).

The sensitivity is highest for the Decision Tree model (100 %) and the MLP model indicating that 100 % of the landslide locations are correctly classified under the landslide class. It is followed by the SVM model (94.18 %), the KNN model (91.18 %) and the Gaussian Naïve Bayes model (88.24 %). The Decision Tree model has the highest specificity (100 %) showing that 100 % of the non-landslide locations are correctly classified under the non-landslide class. This is accompanied by the SVM model (93.94 %), the KNN model (91.18 %) and the MLP model (91.18 %) and the Gaussian Naïve Bayes model (88.24 %).

The Decision Tree model has an outstanding AUROC value (1.000) indicating that the model has excellent prediction capability. The AUROC values for the KNN model (0.966), the Gaussian Naïve Bayes model (0.933), the SVM model (0.977) and the MLP model (0.894) show a very good predictive capability. The Kappa index values (Table 7) for the Decision Tree model (1.00), the KNN model (0.824), the SVM model (0.912) and the MLP model (0.912) show a perfect agreement between actual and predicted landslides. The Kappa index for the Gaussian Naïve Bayes model (0.735) shows a substantial agreement between the actual and predicted landslides.

3.3. Model validation and comparison

The performance and prediction probability of landslide models is evaluated using the test dataset (Fig. 4), AUROC value, Kappa index value and various statistical measures. The outcomes have been summarized in Tables 8 and 9. It can be visualised that outstanding performance with AUROC=1.000 except for the KNN model (AUROC =0.999) has been achieved by all models (Fig. 5). The Kappa index value for all models is 1.0 except for the KNN model (0.95) shows the perfect agreement between the expected and predicted landslides (Table 9). All models have outstanding predictive capability and correctly classify the landslides data into landslide class and non-landslide class (Table 8).

Sensitivity and specificity are also 100 % for the Decision Tree model, the Gaussian Naïve Bayes model, the SVM model and the MLP model. The KNN model has 100 % sensitivity and 95 % specificity. This indicates that all models have good capability to predict landslide locations as landslide class and all non-landslide locations as a non-landslide class.

Friedman's test at a 5 % significant level is used to evaluate whether landslide models are significantly different from each other or not. The p-value achieved is 0.00721 which is less than 0.05 and signifies that the null hypothesis can be rejected. The null hypothesis states that no significant difference between the two models at the significant level of α=0.05. The Friedman test merely signifies the performance distinction of all landslide susceptibility models; it is unable to compare any two models.

To perform this, Wilcoxon signed-rank test is used to evaluate pairwise statistical differences between landslide models with a significant level α at 0.05. The results are summarized in Table 10. It can be seen that the Decision Tree model has a statistical difference from the Gaussian Naïve Bayes model and the MLP model. The

MLP model is also statistically different from the KNN model, the Gaussian Naïve Bayes model and the SVM model.

4. Conclusions

In this paper, a comparative study of five machine learning models, decision tree, KNN, naïve Bayes, SVM and MLP for landslide susceptibility prediction is carried out. The Kullu-Rohtang Pass transport corridor in Himachal Pradesh, India has been selected as the study area. The landslide inventory has been prepared from different sources including landslide data collected from BRO, Manali and PWD, Kullu. A total of 54 landslide locations have been taken into consideration. These locations are described by nine triggering features namely; aspect, relative relief, slope, elevation, LULC, lineament density, geology and geomorphology, distance to road and distance to drainage. Information gain measure is used to identify the irrelevant features and thus eliminated them from the dataset. Therefore, aspect and relative relief do not take part in machine learning model training and validating purposes. Thereafter, the landslide samples are divided into training and test datasets in a ratio of 67:33. The same amounts of randomly selected non-landslide locations are added to the datasets. All five machine learning models are trained on the training data with the decision tree model (100 %), the MLP model (95.59 %), the SVM model (94.12 %), the KNN model (91.18 %) and the Gaussian Naïve Bays model (86.76 %) accuracy. The machine learning models' performance was then validated on the test dataset. All machine learning models have shown outstanding results with the decision tree (100%), the MLP model (100 %), the SVM model (100%), the KNN model (97.5 %) and the Gaussian Naïve Bays model (100%) accuracy. The AUROC curve values for the decision tree model (1.000), the KNN model (0.999), the naïve Bayes model (1.000), the SVM model (1.000) and the MLP model (1.000) indicate that all models have outstanding predictive capabilities. Different statistical measures are used to evaluate the significance of a machine learning model. The Friedman test and Wilcoxon signed-rank test analysis have shown that the machine learning models have significant differences. It must be noted that the performance of machine learning models is dependent on the data and the results may vary for different cases.

The false positive and false negative section of the confusion matrix shows the errors in landslide susceptibility prediction. The false positive results are sometimes acceptable though they may restrict the land use in the area by misclassifying the non-landslide locations as the locations of the landslides. The false negative section is the area to be studied carefully as if landslide locations are falsely classified as non-landslide locations then it may result in catastrophic outcomes. From table 8, it is clear that all models have both false positive and false negative values as 0. Thus, models are well trained to classify the landslide and non-landslide locations correctly except for the KNN model with a false positive value of to 1which is not a serious issue of concern and is affecting the economy faintly.

Statements & Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Authors Contributions

Data collection: [Nirbhav]; Writing-original draft preparation: [Nirbhav, Maheshwar]; Conceptualization: [Nirbhav, Maheshwar, Anand Malik, Atul Saini]; Methodology: [Nirbhav, Maheshwar]; Formal analysis and investigation: [Nirbhav, Maheshwar, Anand Malik, Atul Saini]; Writing-review & editing: [Nirbhav, Maheshwar, Anand Malik, Mukesh Prasad, Atul Saini, Nguyen Thanh Long]; Supervision: [Anand Malik, Mukesh Prasad]

Acknowledgements

We are thankful to anonymous reviewers for the insightful and constructive comments and suggestions, which helped to improve the overall quality of the manuscript. Nirbhav would like to thank the University Grants Commission (UGC), and the Government of India for providing a senior research fellowship to carry out this research. Also, the authors thank the National Disaster Management Authority (NDMA) Government of India, Border Road Organization (BRO), Manali and Public Work Department (PWD), and Kullu for providing various data sets used in this research.

References

Achour Y, Boumezbeur A, Hadji R, et al (2017) Landslide susceptibility mapping using analytic hierarchy process and information value methods along a highway road section in Constantine, Algeria. Arab J Geosci 10:194. https://doi.org/10.1007/s12517-017-2980-6

Achour Y, Garçia S, Cavaleiro V (2018) GIS-based spatial prediction of debris flows using logistic regression and frequency ratio models for Zêzere River basin and its surrounding area, Northwest Covilhã, Portugal. Arab J Geosci 11:550. https://doi.org/10.1007/s12517-018-3920-9

Arabameri A, Pradhan B, Rezaei K, et al (2019) GIS-based landslide susceptibility mapping using numerical risk factor bivariate model and its ensemble with linear multivariate regression and boosted regression tree algorithms. J Mt Sci 16:595–618. https://doi.org/10.1007/s11629-018-5168-y

Bai SB, Wang J, Thiebes B, et al (2014) Susceptibility assessments of the Wenchuan earthquake-triggered landslides in Longnan using logistic regression. Environ Earth Sci 71:731–743. https://doi.org/10.1007/s12665-013-2475-z

Beasley TM, Zumbo BD (2003) Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs. Comput Stat Data Anal 42:569–593. https://doi.org/10.1016/S0167-9473(02)00147-0

Bertsimas D, Dunn J (2017) Optimal classification trees. Mach Learn 106:1039–1082. https://doi.org/10.1007/s10994-017-5633-9

Bhardwaj BK, Pal S (2012) Data Mining: A prediction for performance improvement using classification. Int J Comput Sci Inf Secur 9:. https://doi.org/10.48550/ARXIV.1201.3418

Brenning A (2005) Spatial prediction models for landslide hazards: review, comparison and evaluation. Nat Hazards Earth Syst Sci 5:853–862. https://doi.org/10.5194/nhess-5-853-2005

Bröcker J, Smith LA (2007) Increasing the Reliability of Reliability Diagrams. Weather Forecast 22:651–661. https://doi.org/10.1175/WAF993.1

Budimir MEA, Atkinson PM, Lewis HG (2015) A systematic review of landslide probability mapping using logistic regression. Landslides 12:419–436. https://doi.org/10.1007/s10346-014-0550-5

Bui DT, Lofman O, Revhaug I, Dick O (2011) Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. Nat Hazards 59:1413–1444. https://doi.org/10.1007/s11069-011-9844-2

Carrara A, Pike RJ (2008) GIS technology and models for assessing landslide hazard and risk. Geomorphology 94:257–260. https://doi.org/10.1016/j.geomorph.2006.07.042

Chen W, Pourghasemi HR, Panahi M, et al (2017a) Spatial prediction of landslide susceptibility using an adaptive neuro-fuzzy inference system combined with frequency ratio, generalized additive model, and support vector machine techniques. Geomorphology 297:69–85. https://doi.org/10.1016/j.geomorph.2017.09.007

Chen W, Xie X, Wang J, et al (2017b) A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. CATENA 151:147–160. https://doi.org/10.1016/j.catena.2016.11.032

Cohen J (1960) A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 20:37–46. https://doi.org/10.1177/001316446002000104

Costanzo D, Chacón J, Conoscenti C, et al (2014) Forward logistic regression for earth-flow landslide susceptibility assessment in the Platani river basin (southern Sicily, Italy). Landslides 11:639–653. https://doi.org/10.1007/s10346-013-0415-3

Dinov ID (2018) Decision Tree Divide and Conquer Classification. In: Data Science and Predictive Analytics. Springer International Publishing, Cham, pp 307–343

Esteves JT, de Souza Rolim G, Ferraudo AS (2019) Rainfall prediction methodology with binary multilayer perceptron neural networks. Clim Dyn 52:2319–2331. https://doi.org/10.1007/s00382-018-4252-x

Felicísimo ÁM, Cuartero A, Remondo J, Quirós E (2013) Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. Landslides 10:175–189. https://doi.org/10.1007/s10346-012-0320-1

Friedman M (1937) The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. J Am Stat Assoc 32:675–701. https://doi.org/10.1080/01621459.1937.10503522

Goyal S, Maheshwar (2019) Naïve Bayes Model Based Improved K-Nearest Neighbor Classifier for Breast Cancer Prediction. pp 3–11

Guzzetti F, Galli M, Reichenbach P, et al (2006) Landslide hazard assessment in the Collazzone area, Umbria, Central Italy. Nat Hazards Earth Syst Sci 6:115–131. https://doi.org/10.5194/nhess-6-115-2006

Hong H, Pradhan B, Xu C, Tien Bui D (2015) Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines. CATENA 133:266–281. https://doi.org/10.1016/j.catena.2015.05.019

Jahed Armaghani D, Shoib RSNSBR, Faizi K, Rashid ASA (2017) Developing a hybrid PSO–ANN model for estimating the ultimate bearing capacity of rock-socketed piles. Neural Comput Appl 28:391–405. https://doi.org/10.1007/s00521-015-2072-z

Kalantar B, Pradhan B, Naghibi SA, et al (2018) Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). Geomatics, Nat Hazards Risk 9:49–69. https://doi.org/10.1080/19475705.2017.1407368

Kavzoglu T, Mather PM (2003) The use of backpropagating artificial neural networks in land cover classification. Int J Remote Sens 24:4907–4938. https://doi.org/10.1080/0143116031000114851

Lee S (2005) Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data. Int J Remote Sens 26:1477–1491. https://doi.org/10.1080/01431160412331331012

Lian C, Zeng Z, Yao W, Tang H (2013) Displacement prediction model of landslide based on a modified ensemble empirical mode decomposition and extreme learning machine. Nat Hazards 66:759–771. https://doi.org/10.1007/s11069-012-0517-6
Maheshwar, Kaushik K, Arora V (2015) A Hybrid Data Clustering Using Firefly Algorithm Based Improved Genetic Algorithm. Procedia Comput Sci 58:249–256. https://doi.org/10.1016/j.procs.2015.08.018
Maheshwar, Kumar G (2019) Breast Cancer Detection Using Decision Tree, Naïve Bayes, KNN and SVM Classifiers: A Comparative Study. In: 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, pp 683–686
Marrapu BM, Jakka RS (2013) Landslide Hazard Zonation Methods : A Critical Review
Mohamad ET, Armaghani DJ, Momeni E, et al (2018) Rock strength estimation: a PSO-based BP approach. Neural Comput Appl 30:1635–1646. https://doi.org/10.1007/s00521-016-2728-3
Nefeslioglu HA, Sezer E, Gokceoglu C, et al (2010) Assessment of Landslide Susceptibility by Decision Trees in the Metropolitan Area of Istanbul, Turkey. Math Probl Eng 2010:1–15. https://doi.org/10.1155/2010/901095
O'brien RM (2007) A Caution Regarding Rules of Thumb for Variance Inflation Factors. Qual Quant 41:673–690. https://doi.org/10.1007/s11135-006-9018-6
Pandey VK, Tripathi AK, Sharma KK (2022) Implications of landslide inventory in susceptibility modeling along a Himalayan highway corridor, India. Phys Geogr 43:440–462. https://doi.org/10.1080/02723646.2021.1872857
Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830
Pham BT, Pradhan B, Tien Bui D, et al (2016) A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). Environ Model Softw 84:240–250. https://doi.org/10.1016/j.envsoft.2016.07.005
Pham BT, Tien Bui D, Pourghasemi HR, et al (2017) Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. Theor Appl Climatol 128:255–273. https://doi.org/10.1007/s00704-015-1702-9
Pourghasemi hamid reza, Jirandeh abbas goli, Pradhan B, et al (2013) Landslide susceptibility mapping using support vector machine and GIS at the Golestan Province, Iran. J Earth Syst Sci 122:349–369. https://doi.org/10.1007/s12040-013-0282-2
Pourghasemi HR, Rahmati O (2018) Prediction of the landslide susceptibility: Which algorithm, which precision? CATENA 162:177–192. https://doi.org/10.1016/j.catena.2017.11.022
Pradhan B (2013) A 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. https://doi.org/10.1016/j.cageo.2012.08.023
Quinlan JR (1996) Improved Use of Continuous Attributes in C4.5. J Artif Intell Res 4:77–90
Rokach L (2016) Decision forest: Twenty years of research. Inf Fusion 27:111–125. https://doi.org/10.1016/j.inffus.2015.06.005
Saaty TL (1990) How to make a decision: The analytic hierarchy process. Eur J Oper Res 48:9–26. https://doi.org/10.1016/0377-2217(90)90057-I
Saha AK, Gupta RP, Sarkar I, et al (2005) An approach for GIS-based statistical landslide susceptibility zonation?with a case study in the Himalayas. Landslides 2:61–69. https://doi.org/10.1007/s10346-004-0039-8
Saito H, Nakayama D, Matsuyama H (2009) Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: The Akaishi Mountains, Japan. Geomorphology 109:108–121. https://doi.org/10.1016/j.geomorph.2009.02.026
Tien Bui D, Pradhan B, Lofman O, Revhaug I (2012) Landslide Susceptibility Assessment in Vietnam Using Support Vector Machines, Decision Tree, and Naïve Bayes Models. Math Probl Eng 2012:1–26. https://doi.org/10.1155/2012/974638
Tien Bui D, Tuan TA, Klempe H, et al (2016) Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides 13:361–378. https://doi.org/10.1007/s10346-015-0557-6
Trigila A, Iadanza C, Esposito C, Scarascia-Mugnozza G (2015) Comparison of Logistic Regression and Random Forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). Geomorphology 249:119–136. https://doi.org/10.1016/j.geomorph.2015.06.001
Wang L-J, Guo M, Sawada K, et al (2016) A comparative study of landslide susceptibility maps using logistic regression, frequency ratio, decision tree, weights of evidence and artificial neural network. Geosci J 20:117–136. https://doi.org/10.1007/s12303-015-0026-1
Yao X, Tham LG, Dai FC (2008) Landslide susceptibility mapping based on Support Vector Machine: A case

study on natural slopes of Hong Kong, China. Geomorphology 101:572–582. https://doi.org/10.1016/j.geomorph.2008.02.011

Yeon Y-K, Han J-G, Ryu KH (2010) Landslide susceptibility mapping in Injae, Korea, using a decision tree. Eng Geol 116:274–283. https://doi.org/10.1016/j.enggeo.2010.09.009

Yilmaz I (2010) Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine. Environ Earth Sci 61:821–836. https://doi.org/10.1007/s12665-009-0394-9

Zare M, Pourghasemi HR, Vafakhah M, Pradhan B (2013) Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: a comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms. Arab J Geosci 6:2873–2888. https://doi.org/10.1007/s12517-012-0610-x

Table 1 Dataset types and their sources. 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)

Data Type	**Data Base**	**Resolution and Scale**	**Data Derivative**
Topographic map	Survey of India (SoI)	RF 1: 50, 000	The boundary of the study area, transport route
Imagery data	Google Earth, Landsat 8	30 meters	Land use and land cover
ASTER DEM	USGS	30 meters	Slope, distance to Road, distance to Drainage, Elevation
Landslide data	BRO. Manali PWD, Kullu and NDMA govt. reports		Landslide locations, frequency, types of landslides, 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

Table 2 Landslide triggering features

S. No	Landslide triggering factors	Class
1	Slope (degree)	(1) very gentle (0-15º); (2) gentle (15º -30º); (3) moderate (30º -45º);(4) steep (45º -60º); (5) very steep (>60º)
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) fluvglacial deposits and quaternary alluvium; (6) quartzite schist; (7) carbonaceous slate and limestone; (8) Biotite schist and kyanite 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
8	Aspect	(1) Flat (2) North (3) Northwest (4) East (5) Southeast (6) South (7) Southwest (8) Northeast (9) West
9	Relative Relief (m)	(1) <100 (2) 101-184 (3) >184

Table 3 Multicollinearity analysis of landslide triggering features

Features	VIF	Tolerance
Slope	1.193	0.839
Elevation	1.592	0.628
LULC	1.188	0.841
Distance to road	1.485	0.673
Distance to drainage	1.260	0.794
Geology & geomorphology	1.543	0.648
Lineament density	1.340	0.746
Aspect	1.193	0.839
Relative Relief	1.592	0.628

Table 4 Pearson's Correlation between landslide triggering features

	Slope	Elevation	LULC	Distance to road	Distance to drainage	Geology & Geomorphology	Lineament Density	Aspect	Relative Relief
Slope	1								
Elevation	0.39	1							
LULC	0.2	0.097	1						
Distance to road	-0.24	-0.55	-0.16	1					
Distance to drainage	0.14	0.32	0.25	-0.56	1				
Geology & Geomorphology	0.24	0.68	0.23	-0.55	0.37	1			
Lineament Density	0.3	0.7	-0.046	-0.29	0.18	0.69	1		
Aspect	-0.12	0.24	0.037	0.03	-0.11	0.098	0.45	1	
Relative Relief	0.098	0.044	0.167	0.249	-0.012	0.66	-0.020	0.127	1

Table 5 IGR values for landslide triggering features

Features	Gain ratio	Standard Deviation
Elevation	0.603995	±0.026
Distance to drainage	0.135158	±0.024
Lineament density	0.094374	±0.027
Distance to road	0.073846	±0.025
LULC	0.046241	±0.143
Slope	0.031531	±0.124
Geology & geomorphology	0.024267	±0.086
Aspect	0.001268	±0.073
Relative Relief	0.001179	±0.012

Table 6 Machine learning models performance

Parameters	Decision Tree	KNN	Gaussian Naïve Bayes	SVM	MLP
True positive	34	31	29	33	34
True negative	34	31	30	32	31
False positive	0	3	4	2	3
False negative	0	3	5	1	0
Positive predicted value (%)	100	91.18	87.88	94.29	91.89
Negative predicted value (%)	100	91.18	85.71	96.96	100
Sensitivity	100	91.18	85.29	94.18	100
Specificity	100	91.18	88.24	93.94	91.18
Accuracy	100	91.18	86.76	94.12	95.59

Table 7 AUROC and Kappa index values of landslide models on the training dataset

Number	Landslide Model	AUROC	Kappa Index
1	Decision Tree	1.000	1.0
2	KNN	0.966	0.824
3	Gaussian Naïve Bays	0.933	0.735
4	SVM	0.977	0.912
5	MLP	0.894	0.912

Table 8 Machine learning models validation

Parameters	Decision Tree	KNN	Gaussian Naïve Bayes	SVM	MLP
True positive	20	20	20	20	20
True negative	20	19	20	20	20
False positive	0	1	0	0	0
False negative	0	0	0	0	0
Positive predicted value (%)	100	95.24	100	100	100
Negative predicted value (%)	100	100	100	100	100
Sensitivity	100	100	100	100	100
Specificity	100	95	100	100	100
Accuracy	100	97.5	100	100	100

.

.

Table 9 AUROC and Kappa index values of the five landslide models on the test dataset

Number	Landslide Model	AUROC	Kappa Index
1	Decision Tree	1.000	1.0
2	KNN	0.999	0.95
3	Gaussian Naïve Bays	1.000	1.0
4	SVM	1.000	1.0
5	MLP	1.000	1.0

Table 10 Wilcoxon signed-rank test for pairwise comparative study of machine learning models

Number	Pairwise comparison	p-value	Significance
1	Decision tree vs. KNN	0.119	No
2	Decision tree vs. Gaussian Naïve Bays	0.021	Yes
3	Decision tree vs. SVM	0.000	Yes
4	Decision Tree vs MLP	0.741	No
5	KNN vs. Gaussian Naïve Bays	0.131	No
6	KNN vs. SVM	0.064	No
7	KNN vs. MLP	0.000	Yes
8	Gaussian Naïve Bays vs. SVM	0.696	No
9	Gaussian Naïve Bays vs. MLP	0.000	Yes
10	SVM vs. MLP	0.016	Yes

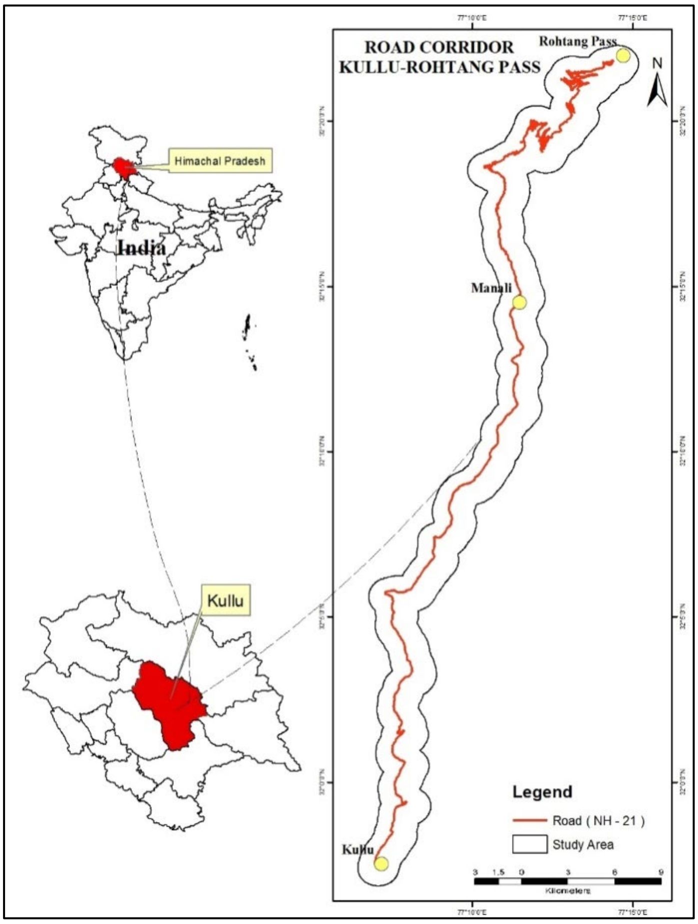

Fig. 1. The study area location

Fig. 2. (a) Rock Fall (b) Debris slide

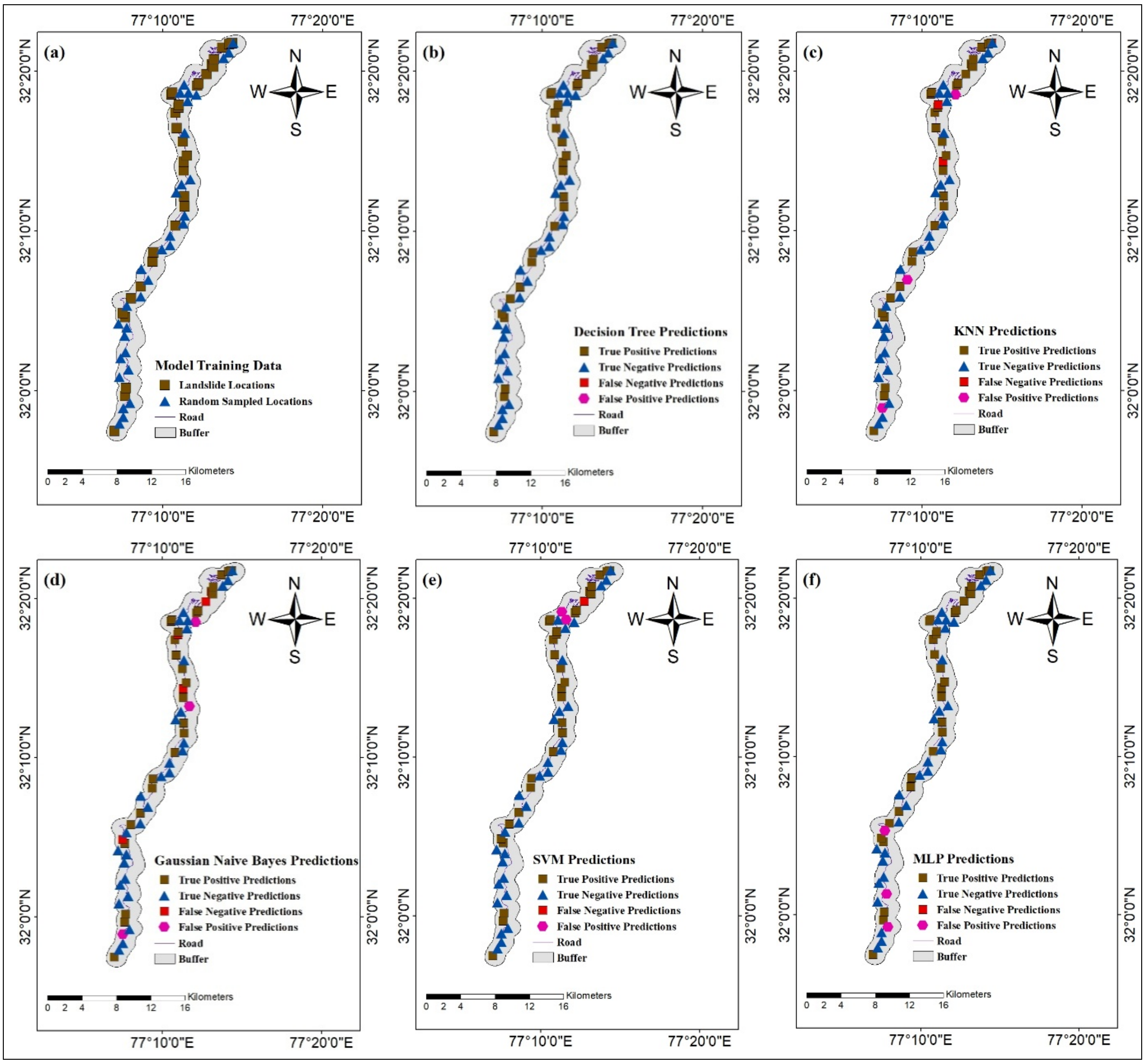

Fig. 3. Landslide susceptibility prediction on training data

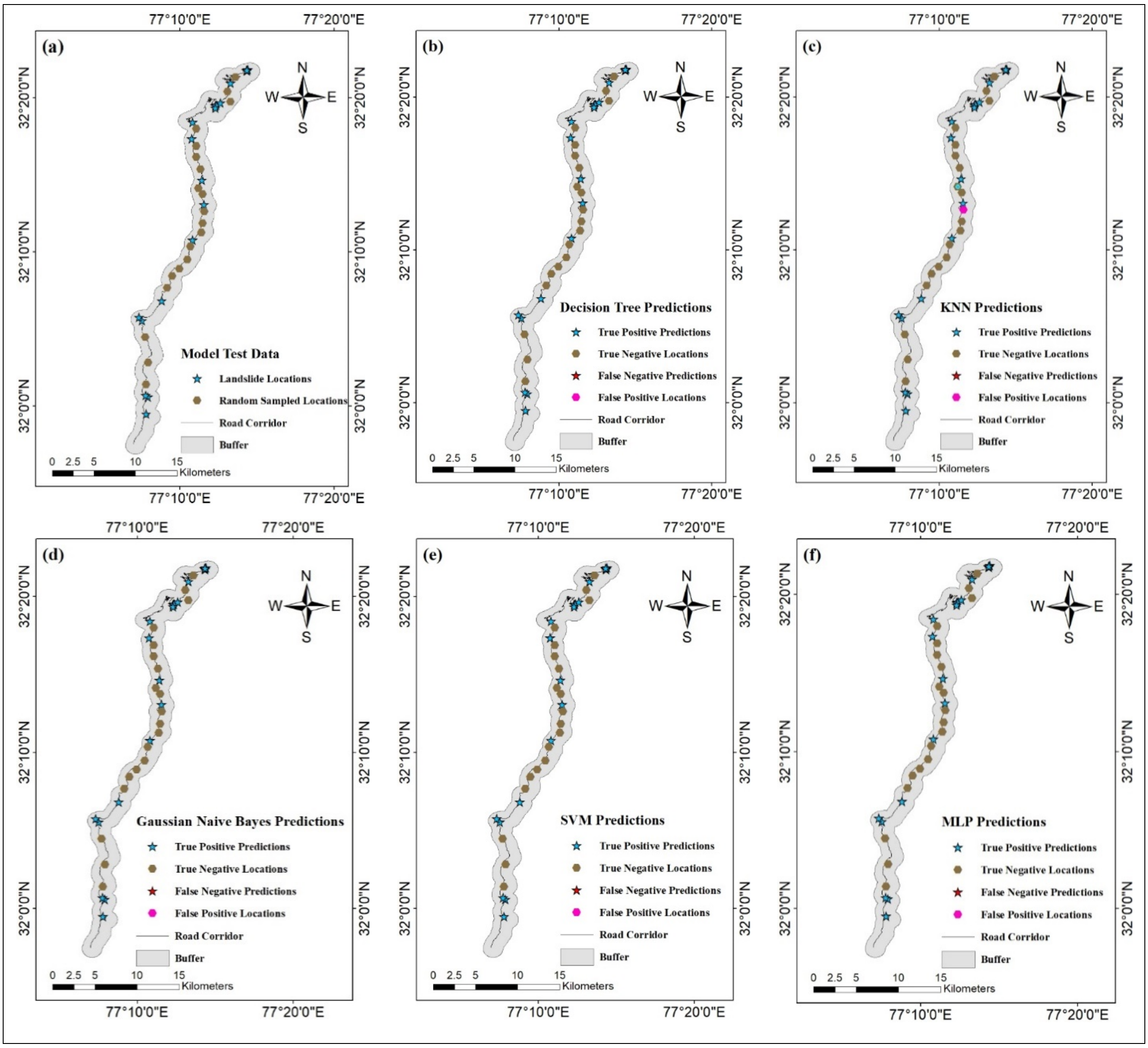

Fig. 4. Landslide susceptibility prediction on test data

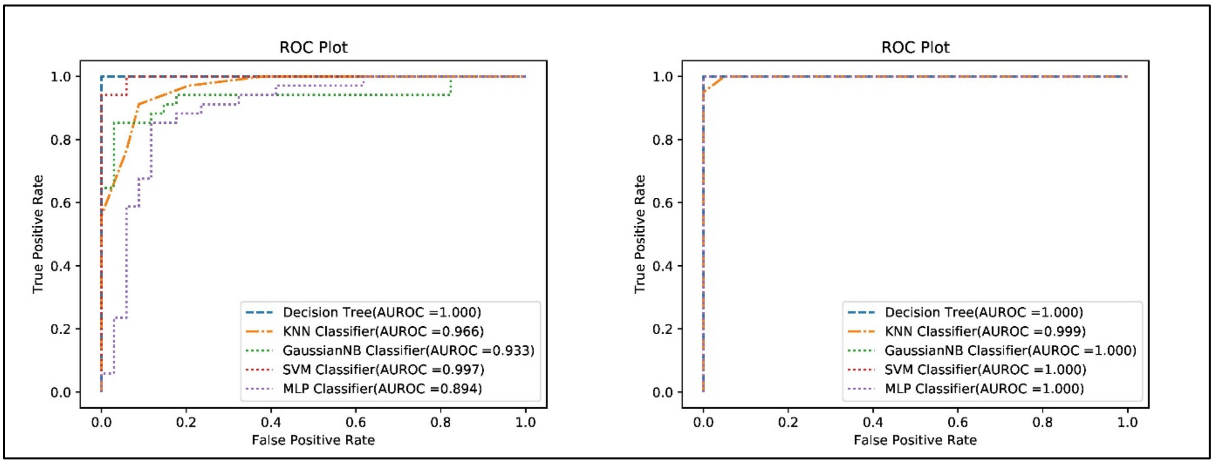

Fig. 5. AUROC curve for machine learning models on training and test data