

Artificial intelligence approaches for spatial prediction of landslides in mountainous regions of western India

Pankaj Prasad^{1*}, Victor Joseph Loveson¹, Sumit Das², Priyankar Chandra³

¹Geological Oceanography Division, CSIR- National Institute of Oceanography,
Dona Paula-403004, Goa, India

²Department of Geography, Savitribai Phule Pune University, Pune 411007, Maharashtra, India

³Department of Geography, Institute of Science, Banaras Hindu University,
Varanasi-221005, Uttar Pradesh, India

*Corresponding Author :Pankaj Prasad

Email: ppankaj.earthscience@gmail.com; ppankaj@nio.org

Mobile No.: +919883329432

Abstract

The prediction of landslide is a complex task but preparing the landslide susceptibility map through artificial intelligence approaches can reduce life loss and damages resulting from landslides. The purpose of this study is to evaluate and compare the landslide susceptibility mapping (LSM) using six machine learning models, including random forest (RF), deep boost (DB), stochastic gradient boosting (SGB), rotation forest (RoF), boosted regression tree (BRT), and logit boost (LB) in the mountainous regions of western India. The landslide inventory map consisting of 184 landslide locations has been divided into two groups for training (70% dataset) and validation (30% dataset) purposes. Fourteen landslide triggering factors including slope, topographical roughness index, road density, topographical wetness index, elevation, slope length, drainage density, stream power index, geomorphology, rainfall, soil, lithology, lineament density, and normalized difference vegetation index have been considered using the boruta approach for the LSM. The results reveal that the RF model has the highest precision in terms of AUC (0.88; 0.89), kappa (0.62; 0.50), accuracy (0.81; 0.77), and specificity (0.86; 0.86) both in the study region and secondary region, respectively. Hence, it can be concluded that the RF is an effective and promising technique as compared to DB, SGB, RoF, BRT, and LB for landslide susceptibility assessment in the research area as well as in regions having similar geo-environmental configuration.

Keywords: Landslide susceptibility; Boruta; GIS; Machine learning; Western India

1. Introduction

Globally, the frequency of natural disasters has been dramatically increased over the last few decades (UNISDR, 2015). Each year, a loss of billions of dollars and thousands of lives due to natural disasters, such as earthquakes, tsunamis, cyclones, floods, and landslides are recorded (Yesilnacar and Topal, 2005; Yilmaz, 2009). Almost there is no region in this earth which is free from the impact of such devastating natural hazards (Suzen and Doyuran, 2003). The occurrence of landslides constitutes about 17% of the natural disasters worldwide as reported by the Center for Research on the Epidemiology of Disasters (Pourghasemi et al., 2012; Nohani et al., 2019). Yilmaz (2009) and Pourghasemi et al. (2012) believe that landslide events will tend to rise in the next decades owing to deforestation, land-use change, and changing climatic condition. Many scientists have assessed landslide hazards and delineated the high-risk zones in past few decades. In general, landslide mapping techniques can be divided into two major methods, i.e. direct and indirect. In case of direct mapping, geomorphologists determine the degree of susceptibility based on the experience and knowledge of terrain characteristics, hillslopes etc. (Van et al., 2003). However, indirect mapping utilizes different statistical or high-performance deterministic machine learning or hybrid models to predict the landslide probable areas on the basis of previous landslide distribution. In the preceding two decades, multitudes of LSM studies have been undertaken with increasing popularity of Geographic Information System (Aleotti and Chowdury, 1999; Guzzetti et al., 2006).

The pioneering works of the landslide mapping were done in 1970s (Carrara et al., 1977; Brabb et al., 1978). Since then, various approaches and methods have been put forwarded and employed for LSM, including heuristic terrain and susceptibility zoning, geomorphological mapping, statistical modelling, physically-based numerical modelling etc. (Montgomery and Dietrich, 1994; Hansen et al., 1995; Aleotti and Chowdhury, 1999; Guzzetti et al., 1999; Lee et al., 2001; Van et al., 2008; Reichenbach et al., 2018). However, the above approaches are no longer capable because they were restricted to large amounts of field data in a smaller area. From the beginning of the current century, statistical modelling with high-resolution GIS data has been in use in different geographical regions for LSM, including frequency ratio (Yilmaz, 2009; Nohani et al., 2019), logistic regression (Pourghasemi et al., 2013; Zhao and Chen, 2020), weight of evidence (Lee et al., 2004a; Regmi et al., 2010). The statistical methods are less accurate compared to the machine learning methods because it can overcome the statistical limitations and predict non-linear structures (Tehrany et al., 2013; Costache and Bui, 2019; Prasad et al., 2020). Hence, in recent years, several machine learning models have been developed and applied successfully for LSM, including random forest (Kim et al., 2018; Arabameri et al., 2019), rotation forest (Hong et al., 2018; Pham et al., 2020a), support vector machine (Chen et al., 2017; Phong et al., 2019), boosted regression tree (Youssef et al., 2016; Park and Kim, 2019), naive bayes

(Bui et al., 2012; Pham et al., 2020a), adaboost (Hong et al., 2018; Wu et al., 2020), and artificial neural network (Lee et al., 2004b; Yilmaz, 2009). Such methods have also been employed in various spatial prediction mapping, viz. flood susceptibility (Khosravi et al., 2018; Pham et al., 2020b; Shahabi et al., 2020), land subsidence (Bui et al., 2018; Oh et al., 2019), gully erosion (Amiri et al., 2019; Pourghasemi et al., 2020; Chen et al., 2021b), piping erosion (Hosseinalizadeh et al., 2019; Chen et al., 2021), land-use change (Aburas et al., 2019; Jamali, 2020), and groundwater potentiality (Rahmati et al., 2016; Naghibi et al., 2019). Besides statistical and machine learning models, novel models with high accuracy have been also used in landslide mapping (Bui et al., 2017; Nguyen et al., 2017; Abedini et al., 2019; Chen et al., 2020; Chen and Li., 2020; Pham et al., 2020a).

In the aforementioned studies, the performance of the applied models has been assessed with respect to particular study regions. It is imperative to show the model applicability in another region for general findings (Prasad et al., 2020). To the best of our knowledge, in landslide susceptibility mapping, the models have been validated by applying them in more than one geographic region. Thus, the present study aims to map the landslide-prone zones in the western India through different approaches of artificial intelligence and to validate the models in another region. In spite of being a high landslide prone area, the present area of investigation has not received much attention from the researchers. The current study is the first attempt to prepare the landslide susceptibility map hereof using machine learning approaches. The LSM is foremost mitigative step to reduce the havoc fatalities and infrastructural damages as a dire consequence of frequent severe landslides along the precarious Ghats escarpment triggered by torrential rainfall during the Indian Summer Monsoon. In addition to the comparison of different machine learning approaches, the scientific delineation of landslide-prone regions will be helpful for proper planning and management of the study area.

2. Study area and GIS database

2.1 Study area

The study area lying approximately between 17°50 to 18°19 North latitudes and 73°18 to 73°40 East longitudes, comprises an approximate area of 1180 km² (Fig. 1). The study region represents a hilly topography ranging from 4 m to 1404 m above the mean sea level. The average slope of the area is about 15°, though some areas are close to the vertical slope. The study area shows a general tilting facing towards the Arabian Sea.

Geologically, the area spans from upper Cretaceous to lower Eocene, which comes under the Sahyadri group of formation. From the geomorphological perspective, more than fifty percent of the area belongs to structural origin highly and moderately dissected plateau. The climatic condition of the study region is subtropical monsoon type. The minimum and maximum average temperatures of the area are 25°C and 30°C, respectively. The Arabian branch of south-west monsoon brings huge amount

of cloud that leads to intense orographic rainfall along the hilly slopes. The area receives about 3000 mm annual mean rainfall during the monsoon seasons (June-September) (Department of Agriculture, Maharashtra State). The principal river Savitri, descends from the Mahabaleshwar peak of Western Ghats and flows for about 110 km before it empties into the Arabian sea.

2.2 GIS database

2.2.1 Landslide inventory map

Inventory data is crucial for any natural hazard mapping especially to understand the landslide distribution and effective variables (Kadavi et al., 2018; Panahi et al., 2020). In the present study, the landslide locations were collected from the Public Works Department (PWD) office, Google Earth historical images and intensive field survey (Fig. 2). A total of 184 landslide locations were taken into account for this study. The dataset was divided into two groups 70% (129 locations) as training data for model fitting and the rest 30% (55 locations) as testing data for model predictive capability assessment (Fig. 1). Simultaneously, an equal number of non-landslide localities were recognized from the topographical map, field survey, and satellite images for the purpose of model establishment. Likewise, the inventory map of another region was generated (Fig. 1).

2.2.1 Landslide triggering factors

The important step for the LSM is to prepare a spatial database of landslide conditioning factors (Reichaenbach et al., 2018; Nohani et al., 2019). The predisposing variables of the landslide were selected after extensive literature survey and field investigation (Hong et al., 2018; Wang et al., 2019). These factors are: slope, elevation, slope length (LS), topographical wetness index (TRI), stream power index (SPI), topographical wetness index (TWI), drainage density (DD), road density (RD), lineament density (LD), normalized difference vegetation index (NDVI), rainfall, lithology, geomorphology, and soil (Fig. 3a-3n). The terrain variables (elevation, slope, LS, SPI, TWI, and TRI) were extracted from the SRTM digital elevation model (DEM) with 30 m spatial resolution. Each conditioning variable was transformed into a spatial database of 30× 30 m size and the grid of the study region was generated by 1337 columns and 1740 rows (1312960 pixels; 1180 km²).

The slope is one of the most important triggering variables since it directly influences the landslide occurrences (Yalchin, 2008). The slope map of the study region was categorized into five classes: 0-7.19, 7.19-15.25, 15.25-24.18, 24.18-35.98, and 35.98-73.40 degrees. The altitude of the area is classified into five classes. LS represents slope steepness and slope length of the topography, which affects the soil loss potential (Naghibi et al., 2019). The LS values are extracted by the given mathematical formula (Moore and Burch, 1986).

$$LS = (A \times \text{cellsize}/22.13)^{0.4} \times (\sin\theta/0.0896)^{1.3} \quad (1)$$

here A and θ indicate the flow accumulation and slope values in degree, respectively.

The ratio between the land surface and flat surface in a specific area is defined as TRI (Stambaugh and Guyette, 2008). TRI map are derived from the following formula

$$TRI = Abs(max^2 - min^2) \quad (2)$$

where maximum and minimum altitude are written as max and min, respectively.

SPI and TWI are vital hydrological factors for LSM. The erosive power and discharge in a specific area are expressed by SPI (Moore and Grayson, 1991; Chapi et al., 2017). TWI is defined as the ratio between the catchment area and slope (Gallant, 2000; Chapi et al., 2017). Both variables were calculated from the catchment area (A) and slope (β) by the given formulae (Moore et al., 1991).

$$SPI = A \times \tan\beta \quad (3)$$

$$TWI = \ln(A/\tan\beta) \quad (4)$$

The stability of slope is controlled by the saturation level of material, which is indirectly related to the drainage density (Yalchin, 2008). The construction of roads beside high slope area reduces burden on both the heel of slope and topography that causes landslide (Yalchin, 2008). Lineament is the linear or curvilinear feature on the earth surface (Das et al., 2018). It represents weak zones where the slope instability leads to the landslide. Lineament map of the study area was prepared from the shaded relief maps. The density maps of drainage, road, and lineament were prepared in Arc GIS 10 environment using spatial analyst tools.

To prepare the NDVI thematic layer, band 4 (red band) and band 5 (infrared) of Landsat 8 OLI (<https://earthexplorer.usgs.gov/>) image (date: 29.10.2018) were used. The NDVI value was extracted as follows:

$$NDVI = (NIR - R)/(NIR + R) \quad (5)$$

where NIR and R denote the infrared and red bands of the electromagnetic spectra, respectively.

The lithology map obtained from the Geological Society of India (GSI) was reclassified into four broad categories: Karla, Indrayani, Diveghat, and Purandgarh formation on 1:250000 scale. The geomorphology map of the study area was reclassified into five major groups namely, structural origin moderately dissected plateau (SoMDp), structural origin highly dissected plateau (SoHDp), structural origin low dissected plateau (SoLDp), denudational origin moderately dissected lower plateau (DoMDLp), denudational origin-pediment pediplain complex (DoPPc) on 1:50000 based on Landsat 8 image, NRSC base map, topographical maps and field survey scale. The thematic map of soil was modified using the base map of the National Bureau of Soil Survey and Land Use planning (NBSSLUP) on 1: 500,000 scale. The soil map was categorized into three broad classes: Typic Ustorthents, Lithic Ustorthents, and Udic Rhodustalfs . Rainfall has been the most influential factor in landslide occurrences especially in the Indian sub-continent due to arrival of the southwest monsoon.

The rainfall data of the last five years (2013-2018) was collected from the rain gauge stations of the department of agriculture, Maharashtra state. Based on this data, rainfall map was classified into four classes using the inverse distance weighted method. Among the classification methods, natural break method was extensively applied for classification of landslide variables (Phong et al., 2019; Chen and Chen, 2021). In the present study, natural break method was used for the classification of the causative factors of landslide in ArcGIS platform.

3 Methodology

The methodology flow-chart of the present study is presented in Figure 4.

3.1 Boruta and multicollinearity feature selection methods

Variable selection is the essential function in natural hazard mapping. The factor determination method improves the model performance by increasing the data quality and reducing noise, overfitting, dimensionality of feature space and preventing redundancy problems (Bui et al., 2019). There are two popular features choosing methods for modeling including wrapper and filter accessible. In the present study, the boruta algorithm was used for feature selection based on wrapper method. The classification in boruta approach is implemented by voting of multiple unbiased weak classifier decision trees. The scale of importance of the variable is obtained as the loss of classification accuracy due to the random alternate of factor values between aims. The algorithm works as follows: (i) At the first step, it adds randomness to the dataframe by shuffling each feature which are called shadow feature (ii) Then, it trains the RF classifier to measure the Z score values and checks whether the attribute has a higher Z score than the maximum Z score among shadow attributes (MZSA). (iii) Finally, the features having value less than MZSA are rejected for being irrelevant and vice versa (Kursa and Rudnicki, 2010; Amiri et al., 2019; Prasad et al., 2021). Shadow minimum, mean, and maximum values represent the Z score of a shadow attribute. On the other hand, tentative variables lying close to the maximum shadow (maximum Z score), are confirmed or rejected based on the mean shadow (mean Z score).

In addition, multicollinearity analysis was performed to verify the results of boruta method. Multicollinearity is the popular statistical method to find out whether any positive correlation between the conditioning factors exists or not. Researchers like Chen et al. (2018a), Pourghasemi et al. (2020) showed that the influencing variables containing the variance inflation factors (VIF) value more than 5 and tolerance (TOL) value less than 0.2 were subjected to multicollinearity while in the studies of Nguyen et al. (2017), Wang et al. (2019) the VIF value >10 or TOL value <0.1 of the controlling factors were rejected.

3.2 Landslide events and variable inter-relationship by Weight of evidence (WoE) method

The WoE is a quantitative approach used to forecast the occurrence of hazard events (Armas, 2012). Based on Bayes rule, the approach was primarily propounded for mineral potential mapping (Bonham-carter, 1994; Chen et al., 2018a). With the march of time, WoE approach was popularized in other spatial prediction mapping (Rahmati et al., 2015; Aghdam et al., 2017; Chen et al., 2018a; Arabameri et al., 2018). In this study, WoE method was used to determine the correlation between the landslide events and effective parameters. The primary purpose of this model is to measure the positive (W_i^+) and negative weights (W_i^-).

$$W_i^+ = \ln \left[\frac{P(B|A)}{P(B|\bar{A})} \right] \quad (6)$$

$$W_i^- = \ln \left[\frac{P(\bar{B}|A)}{P(\bar{B}|\bar{A})} \right] \quad (7)$$

where P and ln stand for the probability and natural logarithm, respectively. B and \bar{B} indicate the presence and absence of desired sub-class of landslide predisposing factors, respectively whereas A and \bar{A} are the presence and absence of landslide events, respectively.

3.3 Application of machine learning models

Random forest introduced by Breiman (2001), is a non-parametric method for classification and regression analysis, The RF algorithm creates many decision trees to enhance the performance of the model (Breimen, 2001; Rahmati et al., 2016). Each decision tree is originated by bootstrap samples and leaves around one-third of samples for validation purpose using the out of bag (OOB) error. The advantages of the model are: i) large dataset handling, ii) no overfitting, iii) OOB error estimation, iv) low bias and low variance, v) no requirement of prior data transformation and rescaling, vi) higher prediction performance. To tune the model, the algorithm requires two important criteria: number of trees (n_{tree}) and number of features (m_{try}).

Deep boost is a new ensemble learning algorithm formulated by Cortes, Mohri, and Syed (2014). The method is updated and some new solutions that are left about the theory underlying AdaBoost. It consists of deep decision trees or members of complex families. The method is a capacity-conscious criterion to achieve high accuracy without overfitting the data. The advantage of the algorithm is to minimize the corresponding learning bound and the mixture weight assigned to each sub-family (Cortes et al., 2014; Chen et al., 2021). The algorithm can extend by examining non-differentiable convex surrogate losses. As far as we are concerned, the DB model has been used for the first time in natural hazard mapping.

Stochastic gradient boosting is an ensemble machine learning technique that merges both the advantages of bagging and boosting methods and is used for classification and regression purpose (Friedman, 2001, 2002; Zhou et al., 2016). SGB is the modification of adaboost and adabtive bagging

method. SGB has many advantages such as i) it uses random sub-sample (selected without replacement) of the dataset at each stage of boosting process instead of the whole dataset, ii) at each step, small trees are generated rather than large entire classification trees, iii) since small trees are developed, SGB method is highly resistant to overfitting iv) the steepest gradient algorithm of the model is resistant to outliers (Friedman, 2002; Lawrence et al., 2004; Moisen et al., 2006).

Rotation forest is an ensemble learning method devised by Rodriguez et al. (2006) to enhance the predictive performance of individual classifiers. In this method, every single classifier is merged with one another and each classifier may be trained using different i) training datasets; ii) subsets of features; iii) parameters of the classifier; or iv) classifier models (Polikar et al., 2010; Nguyen et al., 2017). The model is a combination of the random subspace and bagging method with principal component analysis (PCA) for the ensemble classifier. By applying PCA technique, the training dataset was randomly classified into K subsets (Rodriguez et al., 2006; Nguyen et al., 2017; Naghibi et al., 2019).

Boosted regression tree utilizes both machine learning and statistical techniques for classification and regression problems (Youssef et al., 2016; Naghibi et al., 2016). BRT algorithm uses boosting technique to combine a large number of simple tree for maximizing the predictive capacity (Elith et al., 2008; Naghibi et al., 2019). The advantages of this model are: i) it replaces the missing data using surrogates, ii) it can deal with any type of predictive variable (numeric, categorical, binary), iii) the predictive value get affected by monotone transformation, and different scales of measurement. To run the BRT model, three parameters are required to optimize it, namely interaction depth, shrinkage or learning rate, and number of trees. Interaction depth represents the number of nodes and the shrinkage indicates the contribution of trees.

Logit Boost developed by Friedman and Tibshirani (2000), uses boosting method to reduce the bias and variance (Oh et al., 2019). The LB method is a minor variation of popular boosting technique (Adaboost). In this method, the additive logistic regression function is used to classify the dataset (Tehrany et al., 2018). The main advantages of the LB technique are: i) it can handle noisy data, ii) it has the ability to reduce the training errors, iii) it is used for the multiclass tasks (Oh et al., 2019). The vector values of the model define the input variables and two output groups (landslide and non-landslide events).

3.4 Validation of the models

Validation is an imperative stage in machine learning modelling for the evaluation of the study, without which the models have not any scientific significance (Rahmati et al., 2016; Naghibi et al., 2016; Prasad et al., 2020). In the validation dataset, dependent variables are not considered for verifying the model efficiency to predict the location of landslides. In the present study, the models were validated based on a set of quantitative parameters such as accuracy, kappa index, sensitivity, specificity, and

area under the receiver operating characteristics curve (AUROC). These quantitative criteria were assessed using the possible consequences (PC), namely true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The detailed information of these quantitative parameters is available in Chapi et al. (2017).

The ROC curve is a scientific method to examine the overall performance of the models (Rahmati et al., 2016; Chapi et al., 2017; Bui et al., 2019). The AUROC is a quantitative index for model capability to predict correctly the occurrence or non-occurrence of specific events. AUC value lying in between 0 to 1, was classified as poor (0.5–0.6), average (0.6–0.7), good (0.7–0.8), very good (0.8–0.9), and excellent (0.9–1) (Chen et al., 2018b). The higher AUC value, more will be the model perfection (Chapi et al., 2017; Naghibi et al., 2017; Das 2019, 2020). AUROC value is calculated as follows:

$$AUROC = \frac{\sum TP + \sum TN}{P + N} \quad (8)$$

here the P and N refer to the total numbers of landslide occurrences and non-landslide occurrences in the study area, respectively.

4. Results

4.1 Feature selection and important landslide triggering variables

The selection of significant conditioning factors in machine learning techniques is of vital importance to improve model accuracy by eliminating impertinent factors (Khosravi et al., 2018; Bui et al., 2019). In the current research, the boruta technique was used for the relevant variable selection. From the nineteen landslide predisposing factors, boruta evaluation elected fourteen suitable variables (slope, TRI, RD, TWI, elevation, LS, DD, SPI, geomorphology, rainfall, soil, lithology, LD, and NDVI) whereas five factors including standard curvature, plan curvature, profile curvature, land-use, and aspect were removed (Fig 5a). These rejected variables were also excluded in the study of Chen et al., 2018b. The boruta approach is capable of determining the mean, median, minimum, and maximum importance of the factors displayed in Table 1. In congruence with the result of boruta algorithm, the result of multicollinearity analysis presented in Table 2, showed that the standard, plan and profile curvature were disqualified since there VIF value exceeded 10 and TOL fell below 0.1.

After choosing the significant factors that are crucial for the modelling, the variable importance function of RF method was used to measure the feature importance for LSM in the study area. The measured importance values of the landslide conditioning variables were 100 (slope), 94.49 (TRI), 65.99 (elevation), 56.98 (TWI), 56.34 (LS), 46.26 (DD), 44.56 (RD), 41.99 (rainfall), 34.84 (SPI), 33.47 (NDVI), 27.54 (LD), 22.04 (geomorphology), 8.44 (soil), and 0 (lithology). These results of the variable importance analysis revealed that the slope, TRI, elevation, and TWI had more influence on landslides, followed by LS, DD, RD, rainfall, SPI, NDVI, LD, geomorphology, soil, and lithology (Fig

5b). From the literature, it was noticed that the contribution of the conditioning factors in a particular region varies with the use of different measuring methods (Chen et al., 2018b). In the case of present study, variable importance results of both boruta and RF methods were almost same that confirmed the predictive ability of the algorithms.

4.2 Spatial interaction of landslides and conditioning factors using WoE model

The spatial interconnection between landslide events and affecting variables applying WoE method is presented in Table 3. In this table, C/SC value is the final weight used for the above relationship. The C/SC value of the present study was lying between -9.60 and 9.11. The higher positive value shows strong correlation between landslide events and conditioning factors, whereas negative weight indicates poor relation. Regarding slope, increasing slope angle is directly related to the landslide events in a certain range (Bui et al., 2017; Chen and Chen, 2021). This fact was reflected from the current research where most of the landslide happened in the higher slope angle classes viz. 4° - 34° and 35° - 73° with 8.87 and 7.69 C/SC values, respectively whereas lower slope angle held negative weights. The elevation is another significant factor of landslide events as shown in the study of Wang et al. (2019) and Arabameri et al., (2019). The analysis of WoE for landslide events and elevation expressed that the moderate elevation (430-713 m) recorded the maximum C/SC value (9.11). On the other side, lower (4-154 m) and higher (920-1404 m) elevation classes had negative weights. The increasing slope length is favorable for water infiltration into sub-soil that enhances the possibility of slope failure (Bui et al., 2017). The LS value was highest (7.61) in the class of 0.42-2.10, indicating strong relation with landslide occurrences. The weight of the C/SC went upward with increasing topographic roughness. The C/SC values of sub classes 0.13-0.58, 0.58-0.85, and 0.85-1.23 of TRI were -9.60, 7.41, and 7.98, respectively. The zero value of all the classes of SPI made it clear that SPI had no relation with landslide events. The first class (2.23-5.68) of TWI contented the highest C/SC value (6.50) implying the presence of high soil moisture content rendered landslide probability high. The maximum weight (7.81) for DD was found in the second class (0.51-0.75), followed by 0.13-0.51 (-1.08), 0.75-0.96 (-2.73), and 0.96-1.23 (-4.08). The region having positive DD value denoted greater chances of landslide occurrence. With respect to RD, the C/SC value (4.93) was observed as the highest in class 1.01-4.11 where more than forty five percent of landslides occurred. In the case of LD, results showed that the class of 0-0.28 was the greatest weight of 2.50. A high vegetation density area minimizes the chance of landslide by reducing soil erosion and its transmission. As per NDVI results the 0.21-0.35 class contained the highest C/SC value (2.57) suggesting high susceptibility to landslides followed by -0.03-0.21 (1.32), 0.35-0.41 (-0.06), and 0.41-0.58 (-2.12). The longer period of monsoon rainfall in the study area had direct impact on landslide occurrence because of more runoff caused soil loosening and sediment flow. The maximal rainfall class (619-656 mm) in the area under study indicates that around

fifty percent of landslides took place in this group. Regarding lithology, the highest weights were obtained from the bdd (3.31) and bdp (2.47) groups. Since, the rock of these groups belongs to 350-450 m elevation (landslide occurrence zone), it can be inferred that they are more weathered. For geomorphology, structural origin highly dissected plateau has strongly correlated with the landslide events of C/SC value 6.97. This sub class also obtained maximum score by AHP method in the research of Bera et al., 2018. In the case of soil, Udic Rhodustalfs soil was only positive weight (6.21) among the soil groups.

4.3. Model construction

In machine learning models, researchers applied 10 cross-validation method in order to enhance the precision of classification and to reduce the noise (Wang et al., 2020). Therefore, the present work adopted this method with five repetitions using 70% of the landslide and non-landslide points. These points along with their extracted fourteen conditioning factor values comprised the training dataset whereas the remaining points were used for validation purpose. The exact pixel values of the derived parameters influenced the validation result due to the higher range of the values. To solve this problem, values were transformed into 0 to 1 range with the help of min-max normalization. After the model establishment with training dataset, testing dataset was utilized to verify the performance and predictive capability of each model. This complete procedure was done using the R and Arc GIS software.

4.4 Performance and comparison of the models

In this study, AUC and selected statistical measurements were used for evaluation of the models. These criteria considered both training and testing results for model comparison and application. The training dataset defines the model fitness whereas the testing dataset represents the predictive capability of the model (Bui et al., 2018; Prasad et al., 2021). In the case of training dataset, among the six models, the RF performed excellently with the highest value of 1 quantified by accuracy, Kappa, sensitivity, specificity, and AUC in both the primary and secondary study area (Table 4). Landis and Koch, (1977) determined that the kappa value > 0.8 means the perfect match of measured value with observed value. The value of 1 for sensitivity and specificity represent that all landslide and non-landslide points were correctly classified. The AUC values of the remaining models for primary and secondary regions were 0.89, 0.99 (DB); 0.87, 1 (SBM); 0.95, 0.97 (RoF); 0.82, 0.89 (BRT); and 0.86, 1 (LB), respectively. The highest AUC value of RF method in both the region manifested the better goodness of fit compared to the rest.

The established models were corroborated with the testing dataset to check the predictive ability of the models. The results of validation dataset revealed that RF method held the highest as measured by accuracy (0.81), kappa (0.62), specificity (0.86), and AUC (0.88) compared to other models (Fig. 6a,

Table 5). It is also observed that the RF model appeared to be ideal in the secondary region based on accuracy (0.77), kappa (0.50), specificity (0.86), and AUC (0.89) (Fig. 6b, Table 5). For primary and secondary regions, AUC values of the other models were 0.87; 0.82 (DB), 0.84; 0.85 (SBM), 0.83; 0.82 (RoF), 0.86; 0.81 (BRT), and 0.77; 0.76 (LB) respectively. The comparative results from Tables 4 and 5 exhibited suitability of the models in the following descending manner: RF DB, SBM, RoF, BRT, and LB. The model's AUC values which were more than 0.8 (except LB) confirmed the high precision landslide susceptibility mapping in the areas of investigation. In a nutshell, RF model outperformed other models in both the regions.

4.5 Landslide susceptibility map construction

The preparation of the landslide susceptibility maps is essential to know the possibility of future landslide occurrences in an area. After the training processes, each pixel of the study area was computed for the prediction of landslide. These values ranging from 0 to 1 indicate low to high chances of landslide occurrences. In general, the classification of the derived probability values includes automatic and user-defined methods (Chen et al., 2021a). However, there are several standard techniques in GIS to reclassify the probability maps, namely natural breaks, min-max normalizations, standard deviations, equal intervals, geometrical, and quantiles. A specific method is selected based on the research objective and distribution of data (Chapi et al., 2017). It was learned from the previous studies that many researchers had employed the natural break method for generating the susceptibility maps (Hong et al., 2018; Park and Kim, 2019; Panahi et al., 2020; Chen et al., 2021a). This classification method has advantage of no class bias as well as minimum intra-class and maximum inter-class deviations (Arabameri et al., 2020). In the present work, the distribution of probability values of the models abruptly changed. For this reason, the landslide susceptibility maps were classified using the natural break method in ArcGIS environment into four groups, including low, moderate, high, and very high (Fig. 7). The low and moderate prone zones of the entire study area account for 22.74%, 40.06 (RF); 30.08%, 40.94% (DB); 38.99%, 30.37% (SBM); 54.59%, 28.75% (RoF); 57.13%, 15.93% (BRT); and 9.02%, 28.8 (LB), respectively. The high susceptible classes of RF, DB, SBM, RoF, BRT, and LB models constitute 26.06%, 20.13%, 19.37%, 13.02%, 13.51%, and 44.3% of the study region, respectively. The share of 10.58% (RF), 8.83% (DB), 11.25% (SBM), 3.61% (RoF), 13.42% (BRT), and 17.85% (LB) belong to very high vulnerable zones (Fig. 8). The lowest minimal variations of high and very high susceptibility classes indicate the high model precision (Naghibi et al., 2017; Prasad et al., 2020). In the present study, minuscule difference between RF, DB, and SBM, models with regard to high and very high landslide prone zones ensured the models high accuracy (Fig. 8). The low susceptibility classes of the RoF and BRT models occupied much larger area than the moderate susceptibility classes in compared to other models (except LB model). This is

mainly on account of the fact that the degree of prediction varies model to model for their different algorithm. Regarding LB model, susceptible classes did not correspond with the classes of other models because of less precision in prediction. The study pointed out that the north-eastern, eastern, and south-eastern parts of the research area are more vulnerable to landslides owing to the steepness, roughness, and wetness of these mountainous regions.

5. Discussion

5.1 Variable Contribution Analysis

Although many studies have been accomplished worldwide on LSM, still there is no specific set of variables (Van et al., 2006; Chen and Li, 2020). The influencing factors vary from one study region to another due to heterogeneity of the earth. In the current study, a total of nineteen factors were chosen consulting recent studies carried out in the different regions with more or less same elevation as of the study area under consideration and subsequently fourteen relevant factors were selected based on boruta algorithm for LSM (Hong et al., 2018; Kadavi et al., 2018; Nohani et al., 2019; Arabameri et al., 2019; Wang et al., 2019). It is worth mentioning that the result of boruta approach almost matched with the result of multicollinearity which reflects the robustness of boruta algorithm in feature selection. It can be seen that the topographical factors such as slope, TRI, and elevation exerted more influence on landslide occurrences compared to the categorical factors (lithology, soil, and geomorphology) in the study area. Researcher inferred that topographical variables are the most responsible for landslide incidents (Nakileza and Nedala, 2020; Arabameri et al., 2020). This is because of the topographical factors mainly slope angle which plays a key role in slope stability and controls the shear and normal strength on the plane of weakness. The study area is basically more or less homogeneous bedrock region (Sahyadri group) with a thin veneer of top soil. Therefore, the factors lithology and soil did not have much influence on landslide occurrence. Most of the slides in this area that occurred within accessible locations are rock falls or debris flow. On the other side, WoE method expressed the influence variability of sub classes of each landslide governing factor. Based on the result, the highest C/SC values noted from the slope and elevation sub classes revealed that the landslide occurrences mostly took place into moderate elevation (430-713 m) with higher slope angle (24-73°) probably due to more weathering at this height of the study area.

5.2 Previous works and future work proposal

Mathematically, landslide susceptibility is the likelihood of the occurrence of slope failures in location while the geo-environmental condition is fixed (Guzzeti et al., 2005). LSM is generally done to predict the potential geographic locations of future landslides (Guzzeti et al. 1999; 2006). In recent years, machine learning approaches have become more popular and outstripped other conventional methods in landslide susceptibility and hazard mapping (Bui et al., 2019; Arabameri et al., 2020). For the

mapping of potential landslide locations, a large number of machine learning methods have been developed, artificial neural network (Lee et al., 2001), support vector machine (Dou et al., 2015), random forest (Miner et al., 2010), neuro-fuzzy (Pradhan, 2013), naïve bayes (Bui et al., 2012), quadratic discriminant (Rossi et al., 2010), adaptive neuro-fuzzy inference system-satin bowerbird optimizer (Chen et al., 2021a) just to name a few. However, there is no universal model for spatial mapping of landslide but the applied model can be appropriate for similar geo-environmental settings (Chen et al., 2020; Prasad et al., 2020). Therefore, the present study aims to select suitable method for LSM in the mountainous regions of western India. In this research, six machine learning approaches, including RF, DB, SBM, RoF, BRT, and LB were compared in the study area as well as in the secondary region to choose the significant model. Among these models, RF model showed the highest precision, consistent performance, and minimal difference between training and validation results, followed by DB, SBM, RoF, BRT, and LB for LSM in both the regions. The robustness of the RF method is also manifested in other spatial prediction mappings with very high precision (Catani et al., 2013; Naghibi et al., 2019; Costache et al., 2019; Amiri et al., 2019; Prasad et al., 2020). A study by Chen et al. (2018b) in Chongren County, China, located at about an elevation (1-1230 m) as of the current study area, RF model proved to be the most suitable with respect to the performance of other applied models in landslide prediction. Arabameri et al. (2018) found that the RF model was the best fit in the Gallicash River Watershed, Iran. Bera et al. (2018) worked on LSM in mountainous region of western India. In this study, accuracy of fuzzy gamma operator (81%) and analytical hierarchy process (AHP) (69%) models were lesser than the present research. In the above research works, the researchers employed several methods including novel model only to one respective study area, which restricted the applicability of the models. To widen applicability of the models, it is important to apply the selected models in alike geographic environment (Prasad et al., 2020). In the present study, the outcomes of the models were almost same for both the primary and secondary regions. Therefore, the comparability of the models remained consistent with the either regions under investigation. From the results, it can be stated that the RF model is appropriate for LSM in the western India and also in regions with similar geo-environmental conditions. Likely, in future, more researches are required to conduct in different geographical areas around the world for the assessment of applicability of machine learning models. Furthermore, it can be suggested that the integration of SAR and optical datasets may improve the precision of the models and help to identify the landslide points in a larger area.

5.3 Advantages and disadvantages of the models

The machine learning method is applied in natural hazard and resource mapping because of its unique advantageous capability of automatically performing the task of searching several datasets aimed at obtaining necessary information (Chen et al., 2020). It is also noted that the predictive accuracy of the

model, to some extent, depends on the quality of data and the establishment of the model from training dataset (Nohani et al., 2019). In present study, an attempt was made to find out the probable reasons behind the better performance of RF method and comparatively poor achievement of LB model. RF method suitably fitted resulting in reduction of the error for validation dataset in the primary and secondary regions. The reason for better performance of the RF method may be due to composition of multiple decision trees and interconnection capability of conditioning variables and non-linearity (Catani et al., 2013; Prasad et al., 2020). Besides, RF method escapes the overfitting of dataset which is common problem in artificial intelligence approaches. In the case of LB method, the performance is relatively less in either region. The underperformance of LB model may be because of i) steady property indicated by logistic loss, ii) any change of output value does not affect the loss value, iii) oversimplification in model establishment can reduce the validation accuracy, iv) problem in node split gain and value fitting (Sun et al., 2014).

6. Conclusions

The Landslide susceptibility mapping through high performing artificial intelligence models is an indispensable step for downscaling future loss of properties and human lives. Therefore, in this study, six machine learning approaches (RF, DB, SGB, RoF, BRT, and LB) were employed for effective LSM. To confirm the model applicability, the selected methods were tested in another region as well. The major findings of the study are as follows:

- i) Among the nineteen landslide causing factors, fourteen (slope, elevation, LS, SPI, TRI, TWI, RD, LD, DD, NDVI, rainfall, lithology, soil, and geomorphology) were selected and remaining five (aspect, land-use, plan curvature, profile curvature and standard curvature) were discarded with the help of boruta approach for LSM.
- ii) From the attribute evaluation of WoE method, it was found that the landslide events mostly occurred in the moderately elevated area within higher slope angle where the topographical roughness is maximum. As per the variable importance of RF method, topographical factors were most important compared to the other factors for the LSM.
- iii) Out of six methods, RF model is the most desirable for its superlative performance in model fitness and validation in both the primary and secondary regions. Besides, novel application of the second most suitable model DB in LSM has been successful which widens its scope in natural hazard mapping.
- iv) According to landslide susceptibility classes of RF method, about 11% and 26% of the study area come under the very high and high landslide vulnerable zones, respectively.

The outcomes of this study can be useful to the planners for planning and management of future landslides events. To meet the purpose, RF method can be opted for preparing LSM in alike geographic environment.

Acknowledgement

We acknowledge the support from the University Grant Commission (3160 NET-June 2015) to the first author. The authors are thankful to the Director CSIR-NIO for constant encouragement. The authors are also grateful to the reviewer for critical comments and constructive suggestions that improved the manuscript significantly. Field support from the survey team members is thankfully acknowledged. The NIO contribution number is.....

Declaration of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article.

References

- Abedini, M., Ghasemian, B., Shirzadi, A., Shahabi, H., Chapi, K., Pham, B.T., Bin Ahmad, B., Tien Bui, D., 2019. A novel hybrid approach of bayesian logistic regression and its ensembles for landslide susceptibility assessment. *Geocarto International* 34, 1427–1457. <https://doi.org/10.1080/10106049.2018.1499820>
- Aburas, M.M., Ahamad, M.S.S. and Omar, N.Q., 2019. Spatio-temporal simulation and prediction of land-use change using conventional and machine learning models: a review. *Environmental Monitoring and Assessment*, 191(4), 205. <https://doi.org/10.1007/s10661-019-7330-6>
- Aghdam, I.N., Pradhan, B. and Panahi, M., 2017. Landslide susceptibility assessment using a novel hybrid model of statistical bivariate methods (FR and WOE) and adaptive neuro-fuzzy inference system (ANFIS) at southern Zagros Mountains in Iran. *Environmental Earth Sciences*, 76(6), 237. DOI: [10.1007/s12665-017-6558-0](https://doi.org/10.1007/s12665-017-6558-0)
- Aleotti, P. and Chowdhury, R., 1999. Landslide hazard assessment: summary review and new perspectives. *Bulletin of Engineering Geology and the Environment*, 58(1), 21-44. <https://doi.org/10.1007/s100640050066>
- Amiri, M., Pourghasemi, H.R., Ghanbarian, G.A. and Afzali, S.F., 2019. Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. *Geoderma*, 340, 55-69. <https://doi.org/10.1016/j.geoderma.2018.12.042>
- Arabameri, A., Rezaei, K., Pourghasemi, H.R., Lee, S. and Yamani, M., 2018. GIS-based gully erosion susceptibility mapping: a comparison among three data-driven models and AHP knowledge-based technique. *Environmental Earth Sciences*, 77(17), 628. <https://doi.org/10.1007/s12665-018-7808-5>
- Arabameri, A., Pradhan, B., Rezaei, K. and Lee, C.W., 2019. Assessment of landslide susceptibility using statistical-and artificial intelligence-based FR–RF integrated model and multiresolution DEMs. *Remote Sensing*, 11(9), 999. <https://doi.org/10.3390/rs11090999>

- Arabameri, A., Saha, S., Roy, J., Chen, W., Blaschke, T. and Tien Bui, D., 2020. Landslide susceptibility evaluation and management using different machine learning methods in the Gallicash River Watershed, Iran. *Remote Sensing*, 12(3), 475. [doi:10.3390/rs12030475](https://doi.org/10.3390/rs12030475)
- Armaş, I., 2012. Weights of evidence method for landslide susceptibility mapping. Prahova Subcarpathians, Romania. *Natural Hazards*, 60(3), 937-950. <https://doi.org/10.1007/s11069-011-9879-4>
- Bera, S., Guru, B. and Ramesh, V., 2019. Evaluation of landslide susceptibility models: A comparative study on the part of Western Ghat Region, India. *Remote Sensing Applications: Society and Environment*, 13, 39-52. <https://doi.org/10.1016/j.rsase.2018.10.010>
- Bonham-Carter, G.F., 1994. *Geographic Information Systems for Geoscientists: Modelling with GIS*, Elsevier, UK.
- Brabb, E.E., Pampeyan, E.H., and Bonilla, M.G., 1978. Landslide susceptibility in San Mateo County, California. US Geological Survey Miscellaneous Field Studies Map, MF-360, Map at 1: 62,500 scale.
- Breiman, L. 2001. "Random Forests." *Machine Learning*. 45 (1), 5–32.
- Budimir, M.E.A., Atkinson, P.M., Lewis, and H.G., 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, D.T., Pradhan, B., Lofman, O. and Revhaug, I., 2012. Landslide susceptibility assessment in vietnam using support vector machines, decision tree, and Naive Bayes Models. *Mathematical Problems in Engineering*, <https://doi.org/10.1155/2012/974638>
- Bui, D.T., Nguyen, Q.P., Hoang, N.D. and Klempe, H., 2017. A novel fuzzy K-nearest neighbor inference model with differential evolution for spatial prediction of rainfall-induced shallow landslides in a tropical hilly area using GIS. *Landslides*, 14(1), 1-17. [DOI 10.1007/s10346-016-0708-4](https://doi.org/10.1007/s10346-016-0708-4)
- Bui, D.T., Shahabi, H., Shirzadi, A., Chapi, K., Pradhan, B., Chen, W., Khosravi, K., Panahi, M., Bin Ahmad, B. and Saro, L., 2018. Land subsidence susceptibility mapping in south korea using machine learning algorithms. *Sensors*, 18(8), 2464. <https://doi.org/10.3390/s18082464>
- Bui, D.T., Shirzadi, A., Shahabi, H., Geertsema, M., Omidvar, E., Clague, J.J., Thai Pham, B., Dou, J., Talebpour Asl, D., Bin Ahmad, B. and Lee, S., 2019. New ensemble models for shallow landslide susceptibility modeling in a semi-arid watershed. *Forests*, 10(9), 743. <https://doi.org/10.3390/f10090743>
- Carrara, A., Carratelli, E.P., and Merenda, L., 1977. Computer-based data bank and statistical analysis of slope instability phenomena. *Zeitschrift für Geomorphologie*, 21, 187–222.
- Catani, F., Lagomarsino, D., Segoni, S., and Tofani, V., 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. *Natural Hazards and Earth System Sciences*, 13(11), 2815-2831. <https://doi.org/10.5194/nhess-13-2815-2013>

- Chapi, K., Singh, V.P., Shirzadi, A., Shahabi, H., Bui, D.T., Pham, B.T. and Khosravi, K., 2017. A novel hybrid artificial intelligence approach for flood susceptibility assessment. *Environmental Modelling & Software*, 95, 229-245. <https://doi.org/10.1016/j.envsoft.2017.06.012>
- Chen, W., Pourghasemi, H.R., Kornejady, A. and Zhang, N., 2017. Landslide spatial modeling: Introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques. *Geoderma*, 305, 314-327. <https://doi.org/10.1016/j.geoderma.2017.06.020>
- Chen, W., Chen, W., Li, H., Hou, E., Wang, S., Wang, G., and Peng, T., 2018a. GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models. *Science of the Total Environment*, 634, 853–867. <https://doi.org/10.1016/j.scitotenv.2018.04.055>
- Chen, W., Peng, J., Hong, H., Shahabi, H., Pradhan, B., Liu, J., Zhu, A.X., Pei, X. and Duan, Z., 2018b. Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. *Science of the total environment*, 626, 1121-1135.
- Chen, W., Chen, Y., Tsangaratos, P., Ilia, I. and Wang, X., 2020. Combining evolutionary algorithms and machine learning models in landslide susceptibility assessments. *Remote Sensing*, 12(23), 3854. <https://doi.org/10.3390/rs12233854>
- Chen, W. and Li, Y., 2020. GIS-based evaluation of landslide susceptibility using hybrid computational intelligence models. *Catena*, 195, 104777. <https://doi.org/10.1016/j.catena.2020.104777>
- Chen, X. and Chen, W., 2021. GIS-based landslide susceptibility assessment using optimized hybrid machine learning methods. *Catena*, 196, 104833. <https://doi.org/10.1016/j.catena.2020.104833>
- Chen, Y., Chen, W., Janizadeh, S., Bhunia, G.S., Bera, A., Pham, Q.B., Linh, N.T.T., Balogun, A.L. and Wang, X., 2021. Deep learning and boosting framework for piping erosion susceptibility modeling: spatial evaluation of agricultural areas in the semi-arid region. *Geocarto International*, 1-27. <https://doi.org/10.1080/10106049.2021.1892212>
- Chen, W., Chen, X., Peng, J., Panahi, M. and Lee, S., 2021a. Landslide susceptibility modeling based on ANFIS with teaching-learning-based optimization and Satin bowerbird optimizer. *Geoscience Frontiers*, 12(1), 93-107. <https://doi.org/10.1016/j.gsf.2020.07.012>
- Chen, W., Lei, X., Chakraborty, R., Pal, S.C., Sahana, M. and Janizadeh, S., 2021b. Evaluation of different boosting ensemble machine learning models and novel deep learning and boosting framework for head-cut gully erosion susceptibility. *Journal of Environmental Management*, 284, 112015. <https://doi.org/10.1016/j.jenvman.2021.112015>
- Cortes, C., Mohri, M. and Syed, U., 2014. Deep boosting. *Proceedings of the 31 st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.*
- Costache, R. and Bui, D.T., 2019. Spatial prediction of flood potential using new ensembles of bivariate statistics and artificial intelligence: A case study at the Putna river catchment of Romania. *Science of the Total Environment*, 691, 1098-1118. <https://doi.org/10.1016/j.scitotenv.2019.07.197>

- Das, S., Pardeshi, S.D., Kulkarni, P.P., and Doke, A., 2018. Extraction of lineaments from different azimuth angles using geospatial techniques: a case study of Pravara basin, Maharashtra, India. *Arabian Journal of Geosciences*, 11 (8), 160.
- Das, S., 2019. Geospatial mapping of flood susceptibility and hydro-geomorphic response to the floods in Ulhas basin, India. *Remote Sensing Applications Society and Environment*, 14, 60-74.
- Das, S., 2020. Flood susceptibility mapping of the Western Ghat coastal belt using multi-source geospatial data and analytical hierarchy process (AHP). *Remote Sensing Applications Society and Environment* 20, 100379.
- Dou, J., Paudel, U., Oguchi, T., Uchiyama, S., and Hayakawa, Y.S., 2015. Shallow and DeepSeated Landslide Differentiation using support Vector Machines: a Case Study of the Chuetsu Area, Japan. *Terrestrial Atmospheric and Oceanic Sciences*, 26, 227. [https://doi.org/10.3319/TAO.2014.12.02.07\(EOSI\)](https://doi.org/10.3319/TAO.2014.12.02.07(EOSI))
- Elith, J., Leathwick, J.R. and Hastie, T., 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802-813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>
- Friedman, J.H., and Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2), 337-407.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.
- Friedman, J.H., 2002. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4), 367-378. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2)
- Gallant, J.C., 2000. *Terrain analysis: principles and applications*. John Wiley & Sons.
- Guzzetti, F., Carrara, A., Cardinali, M. and Reichenbach, P., 1999. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. *Geomorphology*, 31, 181-216. [https://doi.org/10.1016/S0169-555X\(99\)00078-1](https://doi.org/10.1016/S0169-555X(99)00078-1)
- Guzzetti, F., Reichenbach, P., Cardinali, M., Galli, M., and Ardizzone, F., 2005. Probabilistic landslide hazard assessment at the basin scale. *Geomorphology*, 72, 272–299. <http://dx.doi.org/10.1016/j.geomorph.2005.06.002>
- Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., and Galli, M., 2006. Estimating the quality of landslide susceptibility models. *Geomorphology*, 81, 166-184.
- Hansen, A., 1984. Landslide hazard analysis. In: Brunsden, D., Prior, D.B. (Eds.), *Slope Instability*. Wiley & Sons, New York, 523–602
- Hong, H., Liu, J., Bui, D.T., Pradhan, B., Acharya, T.D., Pham, B.T., Zhu, A.X., Chen, W. and Ahmad, B.B., 2018. Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). *Catena*, 163, 399-413. <https://doi.org/10.1016/j.catena.2018.01.005>

- Hosseinalizadeh, M., Kariminejad, N., Rahmati, O., Keesstra, S., Alinejad, M. and Behbahani, A.M., 2019. How can statistical and artificial intelligence approaches predict piping erosion susceptibility?. *Science of the Total Environment*, 646, 1554-1566. <https://doi.org/10.1016/j.scitotenv.2018.07.396>
- Jamali, A., 2020. Land use land cover modeling using optimized machine learning classifiers: a case study of Shiraz, Iran. *Modeling Earth Systems and Environment*, 1-12. <https://doi.org/10.1007/s40808-020-00859-x>
- Kadavi, P.R., Lee, C.W. and Lee, S., 2018. Application of ensemble-based machine learning models to landslide susceptibility mapping. *Remote Sensing*, 10(8), 1252. <https://doi.org/10.3390/rs10081252>
- Khosravi, K., Pham, B.T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I. and Bui, D.T., 2018. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Science of the Total Environment*, 627, 744-755. <https://doi.org/10.1016/j.scitotenv.2018.01.266>
- Kim, J.C., Lee, S., Jung, H.S., and Lee, S., 2018. Landslide susceptibility mapping using random forest and boosted tree models in Pyeong-Chang, Korea. *Geocarto International*, 33(9), 1000-1015. <https://doi.org/10.1080/10106049.2017.1323964>
- Kursa, M.B. and Rudnicki, W.R., 2010. Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11), 1-13.
- Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. *biometrics* 159e174.
- Lawrence, R., Bunn, A., Powell, S. and Zambon, M., 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, 90(3), 331-336. <https://doi.org/10.1016/j.rse.2004.01.007>
- Lee, S., Ryu, J., Min, K., and Won, J., 2001. Development of two artificial neural network methods for landslide susceptibility analysis. In: *IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium, (Cat.No.01CH37217).5.2364-2366*. <https://doi.org/10.1109/IGARSS.2001.978003>
- Lee, S. and Choi, J., 2004a. Landslide susceptibility mapping using GIS and the weight-of-evidence model. *International Journal of Geographical Information Science*, 18(8), 789-814. <https://doi.org/10.1080/13658810410001702003>
- Lee, S., Ryu, J.H., Won, J.S. and Park, H.J., 2004b. Determination and application of the weights for landslide susceptibility mapping using an artificial neural network. *Engineering Geology*, 71, 289-302. [https://doi.org/10.1016/S0013-7952\(03\)00142-X](https://doi.org/10.1016/S0013-7952(03)00142-X)
- Miner, A., Vamplew, P., Windle, D.J., Flentje, P., and Warner, P., 2010. A Comparative Study of Various Data Mining Techniques as Applied to the Modeling of Landslide Susceptibility on the Bellarine Peninsula. *Geologically Active, Proceedings of the 11th IAEG Congress of the International Association of Engineering Geology and the Environment, Auckland, New Zealand*.

- Moisen, G.G., Freeman, E.A., Blackard, J.A., Frescino, T.S., Zimmermann, N.E. and Edwards Jr, T.C., 2006. Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological Modelling*, 199(2), 176-187. <https://doi.org/10.1016/j.ecolmodel.2006.05.021>
- Montgomery, D.R., and Dietrich, W.E., 1994. A physically based model for the topographic control of shallow landsliding. *Water Resources Research*, 30 (4), 1153–1171. <http://dx.doi.org/10.1029/93wr02979>
- Moore, I.D., and Burch, G.J., 1986. Physical basis of the length-slope factor in the universal soil loss equation 1. *Soil Science Society of America Journal*, 50(5), 1294-1298. [doi:10.2136/sssaj1986.03615995005000050042x](https://doi.org/10.2136/sssaj1986.03615995005000050042x)
- Moore, I.D. and Grayson, R.B., 1991. Terrain-based catchment partitioning and runoff prediction using vector elevation data. *Water Resources Research*, 27(6), 1177-1191. <https://doi.org/10.1029/91WR00090>
- Moore, I.D., Grayson, R.B., and Ladson, A.R., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, 5(1), 3-30. <https://doi.org/10.1002/hyp.3360050103>
- Naghibi, S.A., Pourghasemi, H.R., and Dixon, B., 2016. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental Monitoring and Assessment*, 188, 44. <https://doi.org/10.1007/s10661-015-5049-6>
- Naghibi, S.A., Ahmadi, K., and Daneshi, A., 2017. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. *Water Resources Management*, 31(9), 2761-2775. <https://doi.org/10.1007/s11269-017-1660-3>
- Naghibi, S.A., Dolatkordestani, M., Rezaei, A., Amouzegari, P., Heravi, M.T., Kalantar, B. and Pradhan, B., 2019. Application of rotation forest with decision trees as base classifier and a novel ensemble model in spatial modeling of groundwater potential. *Environmental Monitoring and Assessment*, 191(4), 248. <https://doi.org/10.1007/s10661-019-7362-y>
- Nakileza, B.R. and Nedala, S., 2020. Topographic influence on landslides characteristics and implication for risk management in upper Manafwa catchment, Mt Elgon Uganda. *Geoenvironmental Disasters*, 7(1), 1-13. <https://doi.org/10.1186/s40677-020-00160-0>
- Nguyen, Q.K., Tien Bui, D., Hoang, N.D., Trinh, P.T., Nguyen, V.H. and Yilmaz, I., 2017. A novel hybrid approach based on instance based learning classifier and rotation forest ensemble for spatial prediction of rainfall-induced shallow landslides using GIS. *Sustainability*, 9(5), 813. <https://doi.org/10.3390/su9050813>
- Nohani, E., Moharrami, M., Sharafi, S., Khosravi, K., Pradhan, B., Pham, B.T., Lee, S. and Melesse, A., 2019. Landslide susceptibility mapping using different GIS-based bivariate models. *Water*, 11(7), 1402. <https://doi.org/10.3390/w11071402fgull>
- Oh, H.J., Syifa, M., Lee, C.W. and Lee, S., 2019. Land subsidence susceptibility mapping using bayesian, functional, and meta-ensemble machine learning models. *Applied Sciences*, 9(6), 1248. <https://doi.org/10.3390/app9061248>

- Panahi, M., Gayen, A., Pourghasemi, H.R., Rezaie, F. and Lee, S., 2020. Spatial prediction of landslide susceptibility using hybrid support vector regression (SVR) and the adaptive neuro-fuzzy inference system (ANFIS) with various metaheuristic algorithms. *Science of the Total Environment*, 139937. <https://doi.org/10.1016/j.scitotenv.2020.139937>
- Park, S. and Kim, J., 2019. Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance. *Applied Sciences*, 9(5), 942. <https://doi.org/10.3390/app9050942>
- Pham, B.T., Prakash, I., Dou, J., Singh, S.K., Trinh, P.T., Tran, H.T., Le, T.M., Van Phong, T., Khoi, D.K., Shirzadi, A. and Bui, D.T., 2020a. A novel hybrid approach of landslide susceptibility modelling using rotation forest ensemble and different base classifiers. *Geocarto International*, 35(12), 1267-1292. <https://doi.org/10.1080/10106049.2018.1559885>
- Pham, B.T., Avand, M., Janizadeh, S., Phong, T.V., Al-Ansari, N., Ho, L.S., Das, S., Le, H.V., Amini, A., Bozchaloei, S.K., Jafari, F., Prakash, I., 2020b. GIS based hybrid computational approaches for flash flood susceptibility assessment. *Water*, 12 (3), 686. <https://doi.org/10.3390/w12030683>
- Phong, T.V., Phan, T.T., Prakash, I., Singh, S.K., Shirzadi, A., Chapi, K., Ly, H.B., Ho, L.S., Quoc, N.K. and Pham, B.T., 2019. Landslide susceptibility modeling using different artificial intelligence methods: A case study at Muong Lay district, Vietnam. *Geocarto International*, 1-24. <https://doi.org/10.1080/10106049.2019.1665715>
- Polikar, R., DePasquale, J., Mohammed, H.S., Brown, G. and Kuncheva, L.I., 2010. Learn++. MF: A random subspace approach for the missing feature problem. *Pattern Recognition*, 43(11), 3817-3832. <https://doi.org/10.1016/j.patcog.2010.05.028>
- Pourghasemi, H.R., Mohammady, M. and Pradhan, B., 2012. Landslide susceptibility mapping using index of entropy and conditional probability models in GIS: Safarood Basin, Iran. *Catena*, 97, 71-84. <https://doi.org/10.1016/j.catena.2012.05.005>
- Pourghasemi, H.R., Moradi, H.R., and Aghda, S.F., 2013. Landslide susceptibility mapping by binary logistic regression, analytical hierarchy process, and statistical index models and assessment of their performances. *Natural Hazards*, 69(1), 749-779. <https://doi.org/10.1007/s11069-013-0728-5>
- Pourghasemi, H.R., Sadhasivam, N., Kariminejad, N. and Collins, A.L., 2020. Gully erosion spatial modelling: Role of machine learning algorithms in selection of the best controlling factors and modelling process. *Geoscience Frontiers*, <https://doi.org/10.1016/j.gsf.2020.03.005>
- 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. *Computer & Geoscience*, 51, 350–365. <https://doi.org/10.1016/j.cageo.2012.08.023>
- Prasad, P., Loveson, V.J., Kotha, M., and Yadav, R., 2020. Application of machine learning techniques in groundwater potential mapping along the west coast of India. *GIScience & Remote Sensing*, 57(6), 735-752. <https://doi.org/10.1080/15481603.2020.1794104>
- Prasad, P., Loveson, V.J., Das, B. and Kotha, M., 2021. Novel ensemble machine learning models in flood susceptibility mapping. *Geocarto International*, 1-23. <https://doi.org/10.1080/10106049.2021.1892209>

- Rahmati, O., Pourghasemi, H.R. and Zeinivand, H., 2015. Flood susceptibility mapping using frequency ratio and weights-of-evidence models in the Golastan Province, Iran. *Geocarto International*, 31(1), 42-70. <https://doi.org/10.1080/10106049.2015.1041559>
- Rahmati, O., Pourghasemi, H.R., and Melesse, A.M., 2016. Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran Region, Iran. *Catena*, 137, 360-372. <https://doi.org/10.1016/j.catena.2015.10.010>
- Regmi, N.R., Giardino, J.R. and Vitek, J.D., 2010. Modeling susceptibility to landslides using the weight of evidence approach: Western Colorado, USA. *Geomorphology*, 115(1-2), 172-187. <https://doi.org/10.1016/j.geomorph.2009.10.002>
- Reichenbach, P., Rossi, M., Malamud, B.D., Mihir, M. and Guzzetti, F., 2018. A review of statistically-based landslide susceptibility models. *Earth Science Reviews*, 180, 60-91. <https://doi.org/10.1016/j.earscirev.2018.03.001>
- Rodriguez, J.J., Kuncheva, L.I. and Alonso, C.J., 2006. Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28(10), 1619-1630. [doi: 10.1109/TPAMI.2006.211](https://doi.org/10.1109/TPAMI.2006.211)
- Rossi, M., Guzzetti, F., Reichenbach, P., Mondini, A.C., and Peruccacci, S., 2010. Optimal landslide susceptibility zonation based on multiple forecasts. *Geomorphology*, 114, 129–142. <https://doi.org/10.1016/j.geomorph.2009.06.020>
- Shahabi, H., Shirzadi, A., Ghaderi, K., Omidvar, E., Al-Ansari, N., Clague, J.J., Geertsema, M., Khosravi, K., Amini, A., Bahrami, S. and Rahmati, O., 2020. Flood Detection and Susceptibility Mapping Using Sentinel-1 Remote Sensing Data and a Machine Learning Approach: Hybrid Intelligence of Bagging Ensemble Based on K-Nearest Neighbor Classifier. *Remote Sensing*, 12(2), 266. <https://doi.org/10.3390/rs12020266>
- Stambaugh, M.C. and Guyette, R.P., 2008. Predicting spatio-temporal variability in fire return intervals using a topographic roughness index. *Forest Ecology and Management*, 254(3), 463-473. <https://doi.org/10.1016/j.foreco.2007.08.029>
- Sun, P., Reid, M.D. and Zhou, J., 2014. An improved multiclass LogitBoost using adaptive-one-vs-one. *Machine learning*, 97(3), 295-326. <https://doi.org/10.1007/s10994-014-5434-3>
- Suzen, M.L., and Doyuran, V., 2003. A comparison of the GIS based landslide susceptibility assessment methods: multivariate versus bivariate. *Environmental Geology*, 45, 665-679.
- Tehrany, M.S., Pradhan, B. and Jebur, M.N., 2013. Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. *Journal of Hydrology*, 504, 69-79. <https://doi.org/10.1016/j.jhydrol.2013.09.034>
- Tehrany, M.S., Jones, S., Shabani, F., Martínez-Álvarez, F. and Bui, D.T., 2018. A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using logitboost machine learning classifier and multi-source geospatial data. *Theoretical and Applied Climatology*, 137, 637-653. <https://doi.org/10.1007/s10661-019-7362-y>
- UNISDR, C., 2015. The human cost of natural disasters: A global perspective.

- Van W.C.J., Rengers, N., and Soeters, R., 2003. Use of Geomorphological information in indirect landslide assessment. *Natural Hazards*, 30, 399-419.
- Van W.C.J., Van Asch, T.W., and Soeters, R., 2006. Landslide hazard and risk zonation-Why is it still so difficult? *Bull. Eng. Geol. Environ.* 65, 167–184.
- Van W.C.J., Castellanos, E., and Kuriakose, S.L., 2008. Spatial data for landslide susceptibility, hazard, and vulnerability assessment: an overview. *Engineering Geology*, 102, 112–131. <http://dx.doi.org/10.1016/j.enggeo.2008.03.010>
- Wang, Y., Fang, Z. and Hong, H., 2019. Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China. *Science of the Total Environment*, 666, 975-993. <https://doi.org/10.1016/j.scitotenv.2019.02.263>
- Wang, Y., Sun, D., Wen, H., Zhang, H. and Zhang, F., 2020. Comparison of random forest model and frequency ratio model for landslide susceptibility mapping (LSM) in Yunyang County (Chongqing, China). *International journal of environmental research and public health*, 17(12), 4206. <https://doi.org/10.3390/ijerph17124206>
- Wu, Y., Ke, Y., Chen, Z., Liang, S., Zhao, H. and Hong, H., 2020. Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. *Catena*, 187, 104396. <https://doi.org/10.1016/j.catena.2019.104396>
- Yalcin, A., 2008. GIS-based landslide susceptibility mapping using analytical hierarchy process and bivariate statistics in Ardesen (Turkey): comparisons of results and confirmations. *Catena*, 72(1), 1-12.
- Yesilnacar, E. and Topal, T., 2005. Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). *Engineering Geology*, 79, 251-266. <https://doi.org/10.1016/j.enggeo.2005.02.002>
- Yilmaz, I., 2009. Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: a case study from Kat landslides (Tokat—Turkey). *Computers & Geosciences*, 35(6), 1125-1138. <https://doi.org/10.1016/j.cageo.2008.08.007>
- Youssef, A.M., Pourghasemi, H.R., Pourtaghi, Z.S., and Al-Katheeri, M.M., 2016. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides*, 13(5), 839-856. <https://doi.org/10.1007/s10346-015-0614>
- Zhao, X. and Chen, W., 2020. Optimization of computational intelligence models for landslide susceptibility evaluation. *Remote Sensing*, 12(14), 2180. <https://doi.org/10.3390/rs12142180>
- Zhou, J., Shi, X.Z., Huang, R.D., Qiu, X.Y. and Chen, C., 2016. Feasibility of stochastic gradient boosting approach for predicting rockburst damage in burst-prone mines. *Transactions of Nonferrous Metals Society of China*, 26(7), 1938-1

List of caption for figures

- Figure 1** Location maps of the study area (a), and secondary region (b)
- Figure 2** Recent field photos of occurred landslides in the study area
- Figure 3** Thematic layers of the Landslide triggering factors (a) slope, (b) elevation, (c) LS, (d) TRI, (e) SPI, (f) TWI, (g) drainage density, (h) road density, (i) lineament density, (j) NDVI, (k) rainfall, (l) lithology, (m) geomorphology, and (n) soil. The current methodology of the research
- Figure 4** The current methodology of the research
- Figure 5** Feature selection by boruta approach (a), variables importance in landslide susceptibility mapping (b)
- Figure 6** ROC curve of the models in study area (a), secondary area (b)
- Figure 7** Landslide susceptibility mapping derived from (a) RF, (b) DB, (c) SGB, (d) RoF, (e) BRT, and (f) LB models
- Figure 8** Area of different landslide susceptibility classes of the models

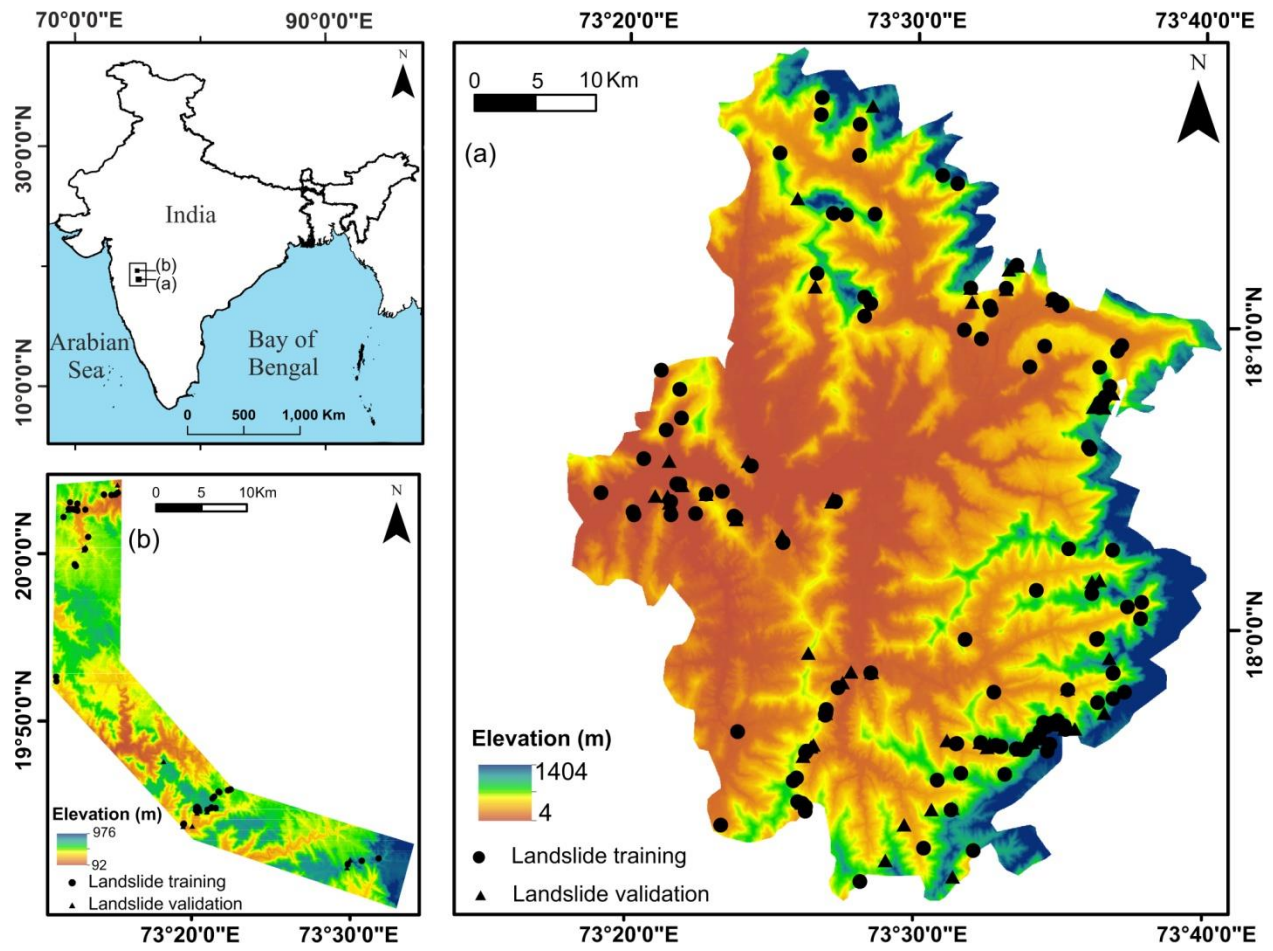
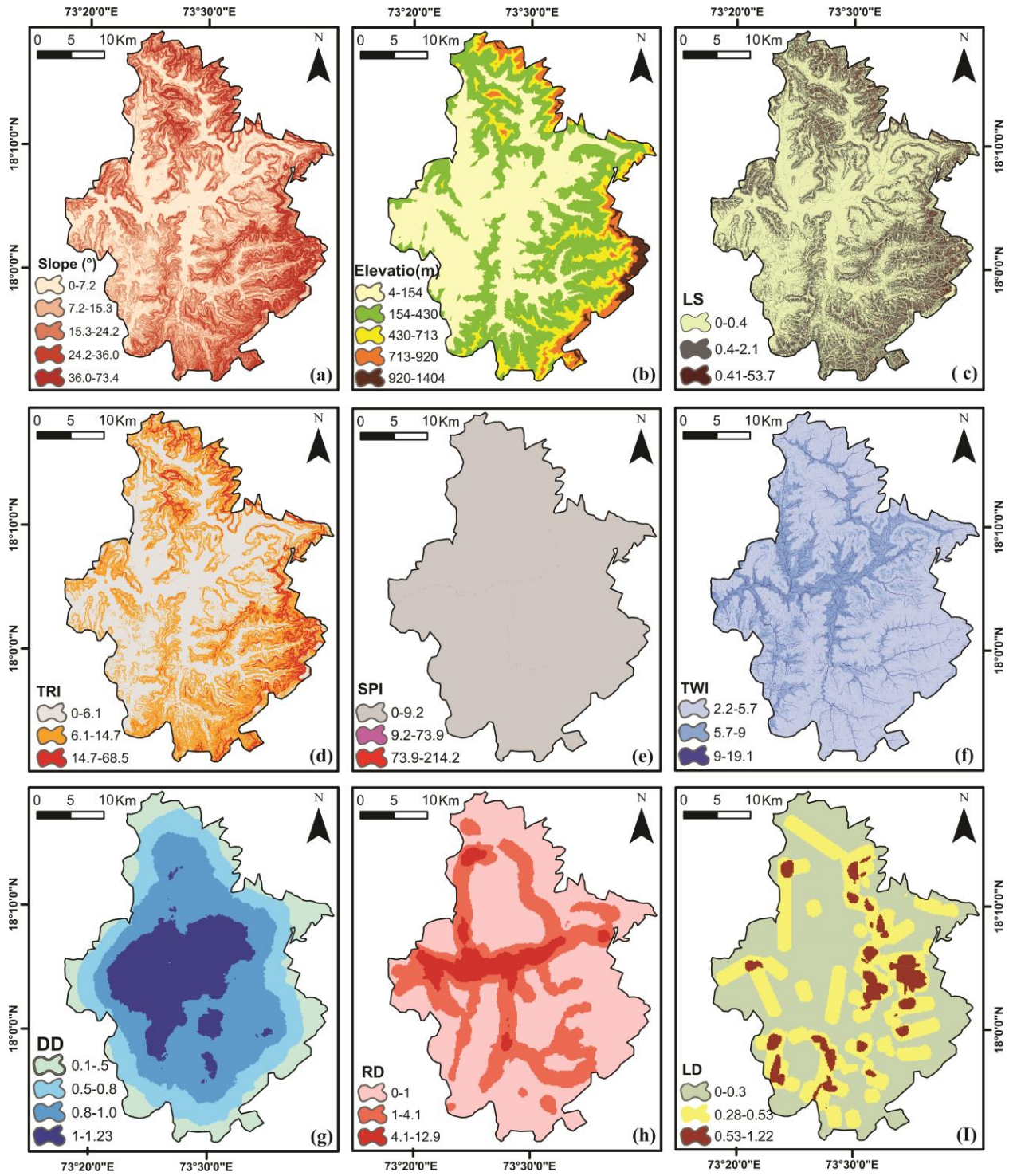


Figure 1



Figure 2



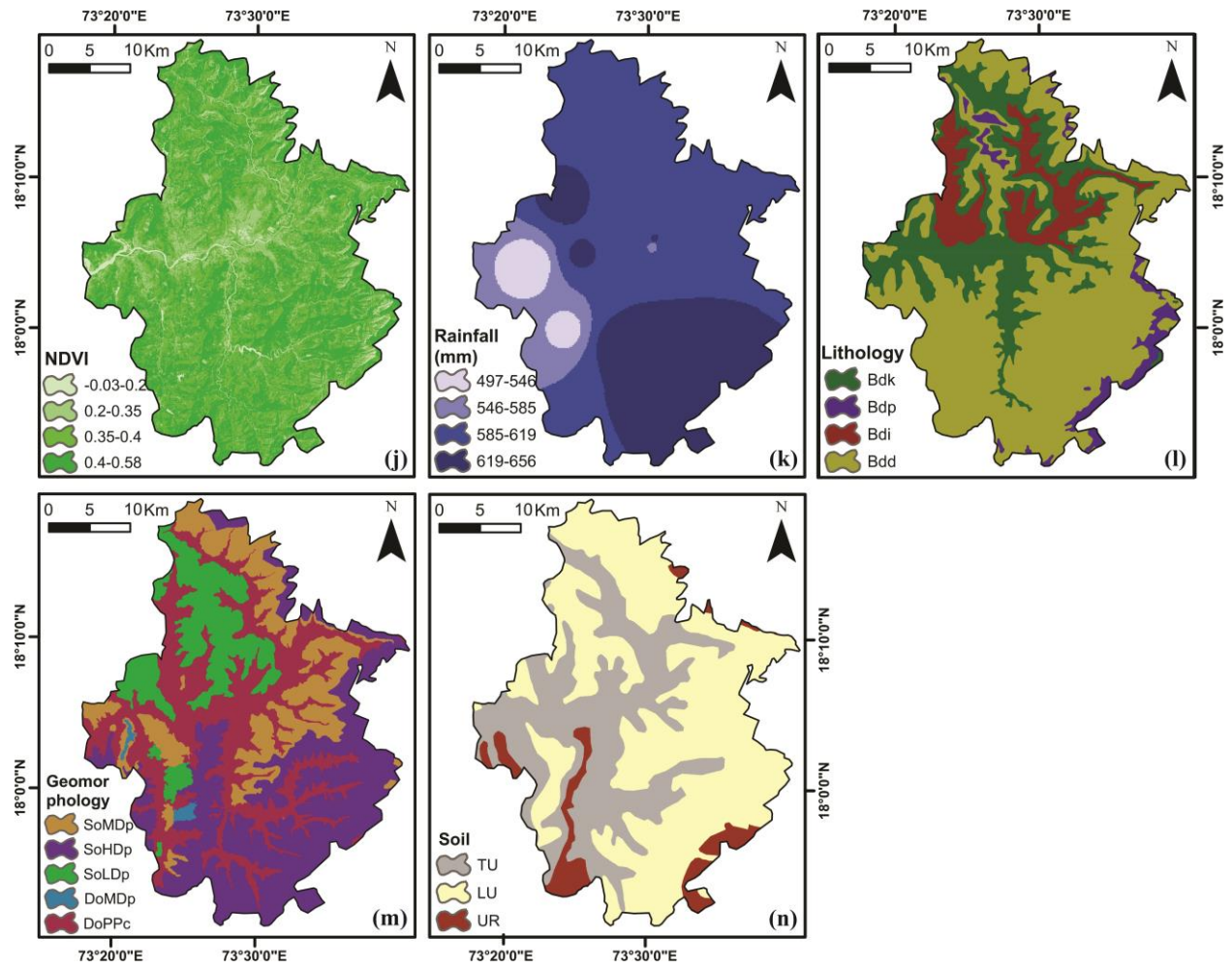


Figure 3

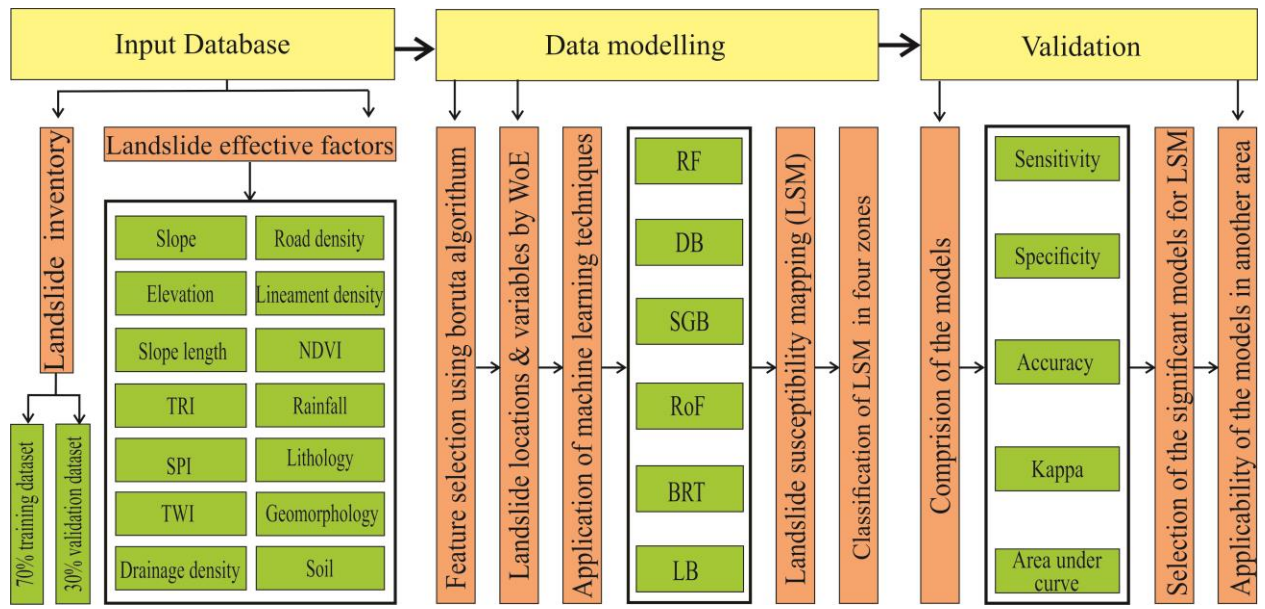


Figure 4

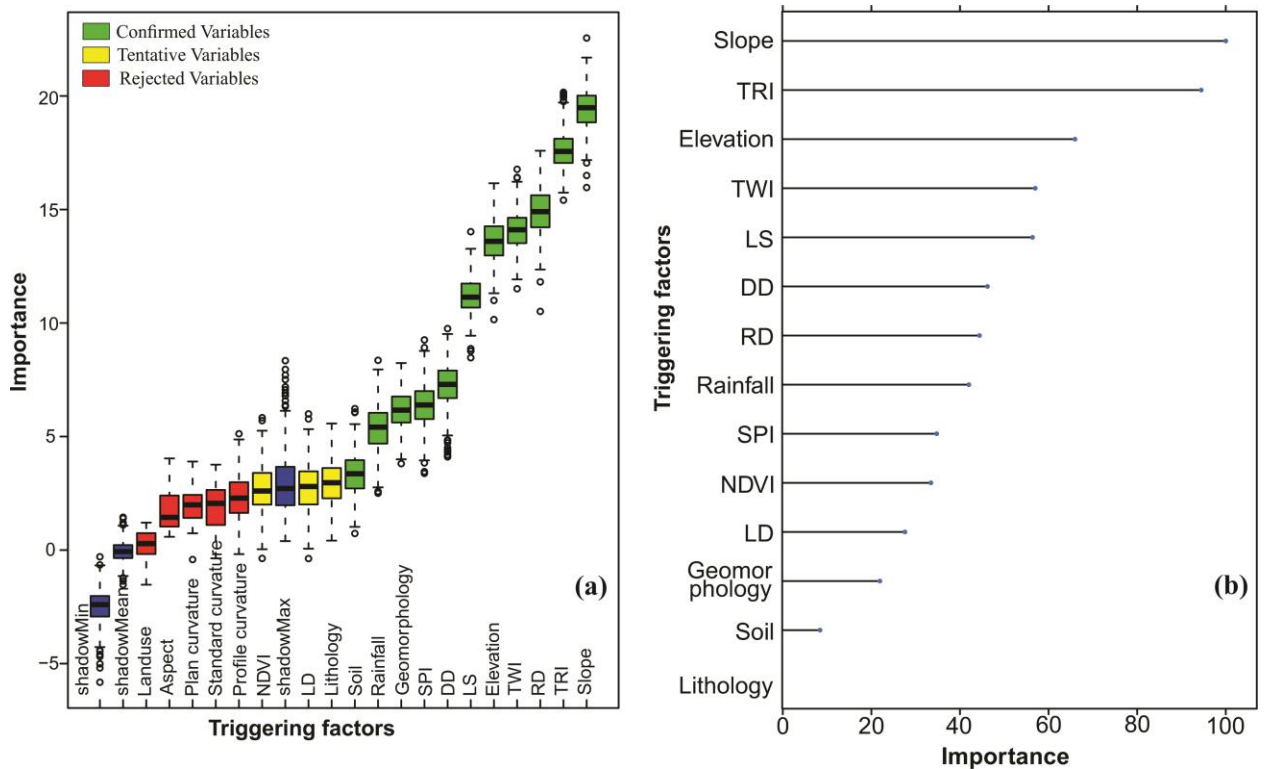


Figure 5

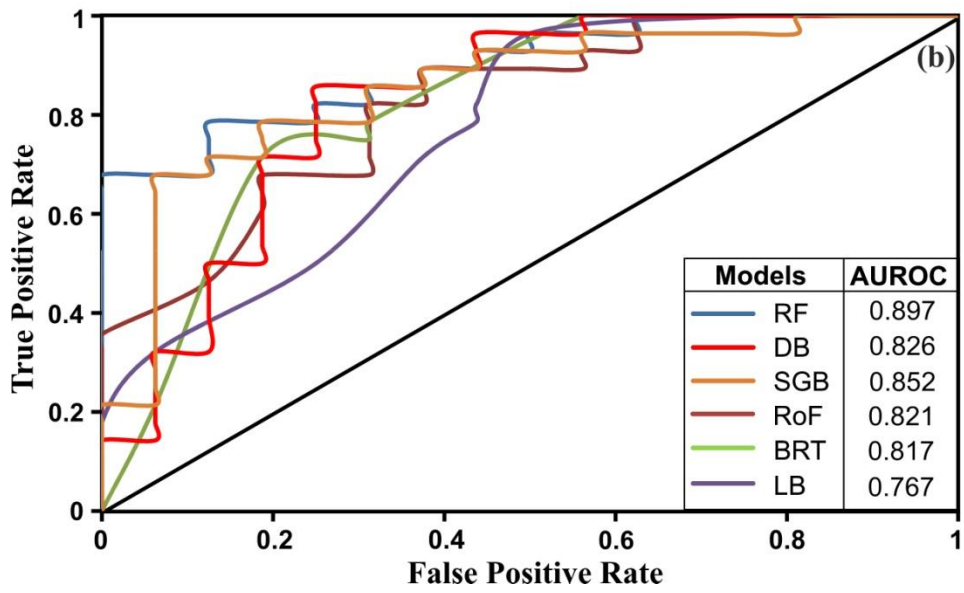
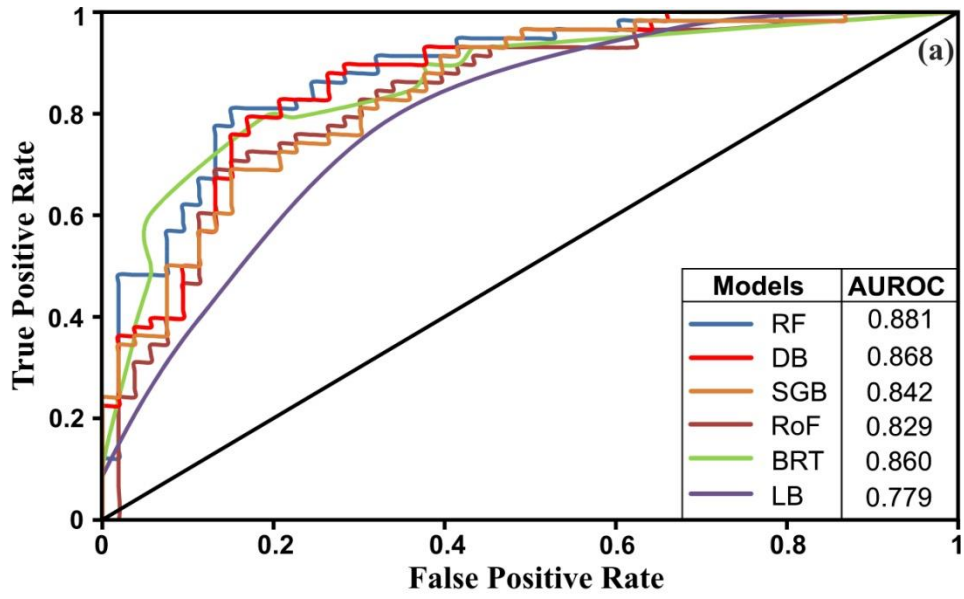


Figure 6

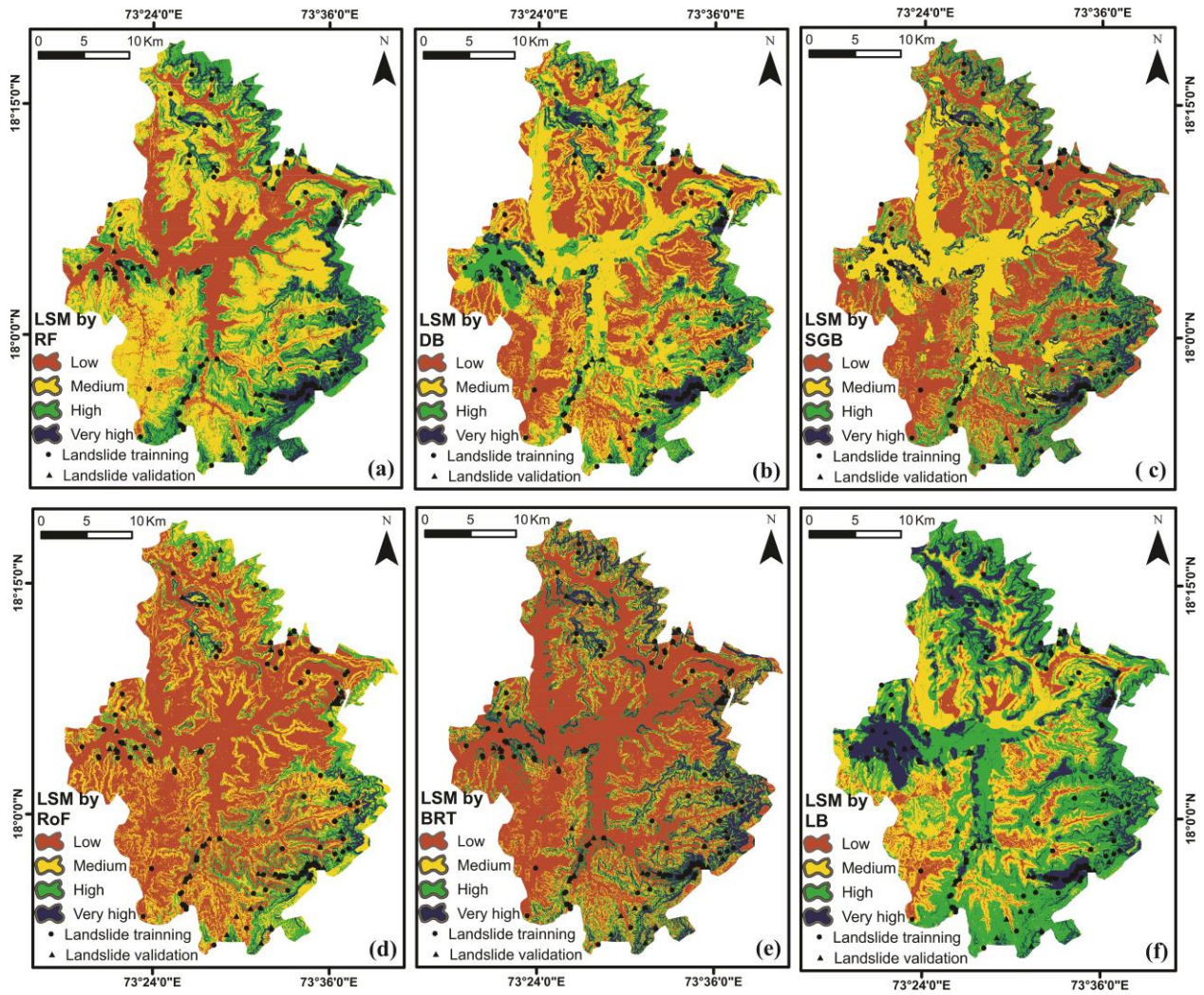


Figure 7

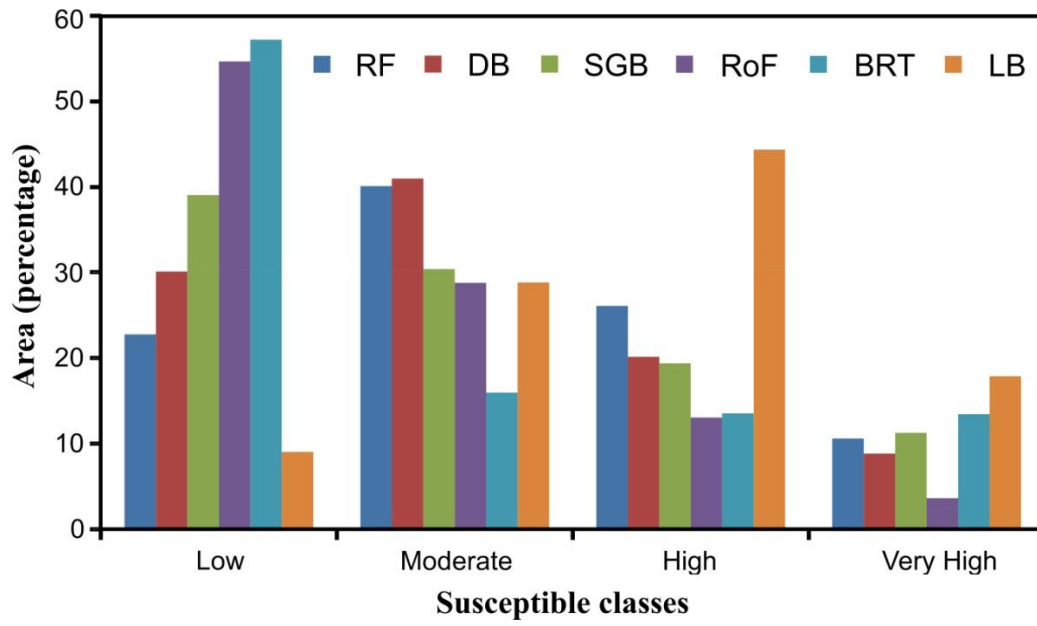


Figure 8

Table 1 Feature selection by Boruta algorithm

Conditioning factors/Importance	Mean Importance	Median Importance	Min Importance	Max Importance	Norm Hits	Decision
Slope	19.45	19.48	15.97	16.15	1	Confirmed
TRI	17.63	17.55	15.40	20.16	1	Confirmed
RD	14.92	14.90	10.51	17.58	1	Confirmed
TWI	14.09	14.10	11.50	16.77	1	Confirmed
Elevation	13.59	13.59	10.15	16.15	1	Confirmed
LS	11.17	11.13	8.47	14.02	1	Confirmed
DD	7.28	7.29	4.10	9.75	0.98	Confirmed
SPI	6.38	6.39	3.37	9.25	0.96	Confirmed
Geomorphology	6.15	6.16	3.80	8.23	0.96	Confirmed
Rainfall	5.38	5.41	2.50	8.35	0.92	Confirmed
Soil	3.36	3.36	0.73	6.22	0.63	Confirmed
Lithology	2.95	2.96	0.41	5.57	0.55	Confirmed
LD	2.77	2.80	-0.37	5.99	0.48	Tentative
NDVI	2.64	2.60	-0.36	5.83	0.45	Tentative
Profile curvature	2.32	2.28	-0.17	5.12	0.24	Rejected
Standard curvature	1.83	2.05	-0.36	3.76	0.01	Rejected
Plan curvature	1.90	1.99	-0.41	3.90	0	Rejected
Aspect	1.78	1.44	0.58	4.03	0	Rejected
Landuse	0.12	0.28	-1.52	1.20	0	Rejected

Table 2 Multicollinearity analysis.

Landslide influencing factors	Collinearity statistics	
	Tolerance (TOL)	Variance inflation factors (VIF)
Slope	0.232	4.318
TRI	0.302	3.308
RD	0.744	1.343
TWI	0.198	5.040
Elevation	0.341	2.936
LS	0.409	2.443
DD	0.520	1.922
SPI	0.663	1.509
Geomorphology	0.651	1.537
Rainfall	0.855	1.169
Soil	0.554	1.805
Lithology	0.823	1.215
LD	0.795	1.259
NDVI	0.475	2.106
Profile curvature	6.11744E-12	1.63467E+11
Standard curvature	1.99529E-12	5.0118E+11
Plan curvature	7.22278E-12	1.38451E+11
Aspect	0.815	1.226
Landuse	0.510	1.960

Table 3 Spatial relationship between landslide triggering factors and landslide events using WoE model

Parameters	Classes	% of area	% of inventory	W+	W-	C	SW+	SW-	SC	C/SC
Slope (Degree)	0-7	31.09	3.26	-2.26	0.34	-2.59	0.17	0.01	0.42	-6.25
	7-15	26.69	9.23	-1.06	0.21	-1.27	0.06	0.01	0.25	-5.01
	15-24	24.19	33.15	0.32	-0.13	0.44	0.02	0.01	0.16	2.82
	24-35	13.77	38.04	1.02	-0.33	1.35	0.01	0.01	0.15	8.87
	35-73	4.26	16.84	1.37	-0.14	1.52	0.03	0.01	0.20	7.69
Elevation (m)	4-154	43.54	20.11	-0.77	0.35	-1.12	0.03	0.01	0.18	-6.09
	154-430	39.19	44.02	0.12	-0.08	0.20	0.01	0.01	0.15	1.34
	430-713	10.35	32.61	1.15	-0.29	1.43	0.02	0.01	0.16	9.12
	713-920	5.16	2.17	-0.87	0.03	-0.90	0.25	0.01	0.51	-1.77
	920-1404	1.76	1.9	-0.48	0.01	-0.49	0.50	0.01	0.71	-0.69
LS	0-0.42	60.72	28.80	-0.75	0.60	-1.34	0.02	0.01	0.16	-8.24
	0.42-2.10	35.68	64.13	0.59	-0.58	1.17	0.01	0.02	0.15	7.61
	2.10-53.72	3.58	7.06	0.68	-0.04	0.71	0.08	0.01	0.29	2.48
TRI	0.13-0.58	54.03	9.78	-1.71	0.67	-2.38	0.06	0.01	0.25	-9.60
	0.58-0.85	39.17	67.39	0.54	-0.62	1.17	0.01	0.02	0.16	7.41
	0.85-1.23	6.79	22.82	1.21	-0.19	1.40	0.02	0.01	0.18	7.98
SPI	0-9.24	99.94	100	0.00	-	-	0.01	-	-	-
	9.24-73.93	0.05	0	-	0.00	-	-	0.01	-	-
	73.93-214.22	0	0	-	0.00	-	-	0.01	-	-
TWI	2.23-5.68	67.46	94.56	0.33	-1.78	2.11	0.01	0.10	0.33	6.50
	5.68-9.06	29.49	5.44	-1.69	0.29	-1.99	0.10	0.01	0.33	-6.11
	9.06-19.12	2.75	0	-	0.03	-	-	0.01	-	-
DD	0.13-0.51	4.68	8.15	-0.20	0.03	-0.24	0.04	0.01	0.22	-1.08
	0.51-0.75	9.26	3.26	0.74	-0.41	1.15	0.01	0.01	0.15	7.81
	0.75-0.96	57.19	43.47	-0.30	0.15	-0.45	0.02	0.01	0.16	-2.73
	0.96-1.23	28.85	45.10	-0.88	0.16	-1.04	0.06	0.01	0.25	-4.08
RD	0-1.01	63.22	42.39	-0.40	0.45	-0.85	0.01	0.01	0.15	-5.70
	1.01-4.11	29.26	46.19	0.46	-0.27	0.73	0.01	0.01	0.15	4.93
	4.11-12.94	7.52	11.42	0.42	-0.04	0.46	0.05	0.01	0.23	1.99
LD	0-0.28	60.46	69.56	0.14	-0.26	0.40	0.01	0.02	0.16	2.50
	0.28-0.53	32.28	28.81	-0.11	0.05	-0.16	0.02	0.01	0.16	-1.01
	0.53-1.22	7.26	1.63	-1.50	0.06	-1.56	0.33	0.01	0.58	-2.67
NDVI	-0.032-0.21	1.91	3.26	0.53	-0.01	0.55	0.17	0.01	0.42	1.32
	0.21-0.35	13.54	20.11	0.40	-0.08	0.47	0.03	0.01	0.18	2.58
	0.35-0.41	44.23	44.02	0.00	0.00	-0.01	0.01	0.01	0.15	-0.06
	0.41-0.58	40.30	32.61	-0.21	0.12	-0.33	0.02	0.01	0.16	-2.12
Rainfall (mm)	497-545	4.69	8.15	0.55	-0.04	0.59	0.07	0.01	0.27	2.19
	545-584	9.26	3.26	-1.04	0.06	-1.11	0.17	0.01	0.42	-2.67
	584-619	57.20	43.48	-0.27	0.28	-0.55	0.01	0.01	0.15	-3.72

Lithology	619-656	28.85	45.11	0.45	-0.26	0.71	0.01	0.01	0.15	4.77
	Bdk	24.89	17.93	-0.33	0.09	-0.42	0.03	0.01	0.19	-2.17
	Bdp	4.36	8.15	0.63	-0.04	0.67	0.07	0.01	0.27	2.47
	Bdi	10.56	1.63	-1.87	0.10	-1.96	0.33	0.01	0.58	-3.37
Geomorphology	Bdd	60.17	72.28	0.18	-0.36	0.55	0.01	0.02	0.16	3.32
	SoMDp	18.07	19.02	0.05	-0.01	0.06	0.03	0.01	0.19	0.34
	SoHDp	34.24	59.78	0.56	-0.49	1.05	0.01	0.01	0.15	6.98
	SoLDp	14.52	7.07	-0.72	0.08	-0.80	0.08	0.01	0.29	-2.79
	DoMDp	0.63	0.55	-0.14	0.00	-0.14	1.00	0.01	1.00	-0.14
Soil	DoPPc	32.54	13.58	-0.87	0.25	-1.12	0.04	0.01	0.22	-5.21
	Typic Ustorthents	35.89	34.23	-0.05	0.03	-0.07	0.02	0.01	0.16	-0.47
	Lithic Ustorthents	57.59	47.28	-0.20	0.22	-0.41	0.01	0.01	0.15	-2.81
	Udic Rhodustalfs	6.51	18.47	1.04	-0.14	1.18	0.03	0.01	0.19	6.21

Table 4 Performance of the models for training dataset

Regions Factors/Models	Primary region						Secondary region					
	RF	DB	SBM	RoF	BRT	LB	RF	DB	SBM	RoF	BRT	LB
TP	131	98	100	116	76	94	57	56	57	52	50	57
TN	126	109	104	112	111	113	45	42	45	4	32	45
FP	0	17	22	14	15	13	0	3	0	5	13	0
FN	0	33	31	15	55	37	0	1	0	5	7	0
Overall	1	0.80	0.79	0.89	0.73	0.80	1	0.96	1	0.90	0.80	1
Kappa	1	0.61	0.58	0.77	0.46	0.61	1	0.92	1	0.80	0.59	1
Sensitivity	1	0.75	0.76	0.89	0.58	0.71	1	0.98	1	0.91	0.87	1
Specificity	1	0.87	0.83	0.89	0.88	0.89	1	0.93	1	0.89	0.71	1
AUC	1	0.89	0.87	0.95	0.82	0.86	1	0.99	1	0.97	0.89	1

Table 5 Performance of the models for validation dataset

Regions Factors/Models	Primary region						Secondary region					
	RF	DB	SBM	RoF	BRT	LB	RF	DB	SBM	RoF	BRT	LB
TP	40	42	39	39	34	35	10	12	11	11	11	9
TN	50	47	44	44	49	46	24	21	22	21	21	22
FP	8	11	14	14	9	12	4	7	6	7	7	6
FN	13	11	14	14	19	18	6	4	5	5	5	7
Overall	0.81	0.80	0.74	0.74	0.74	0.73	0.77	0.75	0.75	0.73	0.72	0.70
Kappa	0.62	0.60	0.49	0.49	0.49	0.45	0.50	0.48	0.46	0.42	0.42	0.35
Sensitivity	0.75	0.79	0.75	0.73	0.64	0.66	0.62	0.75	0.68	0.68	0.68	0.56
Specificity	0.86	0.81	0.75	0.75	0.84	0.79	0.86	0.75	0.78	0.75	0.75	0.78
AUC	0.88	0.87	0.84	0.83	0.86	0.77	0.89	0.82	0.85	0.82	0.81	0.76

Supplementary tables

Table: Inventory of landslides

Primary Region			Secondary Region		
SI No	Latitude	Longitude	SI No	Latitude	Longitude
1	17.9477997	73.5819016	1	19.6980991	73.4998016
2	17.9463005	73.5771027	2	19.6938	73.4973984
3	17.9480991	73.5746002	3	19.6998005	73.5297012
4	17.9370995	73.5672989	4	19.6970005	73.5121002
5	17.9337997	73.5643997	5	19.6980991	73.4998016
6	17.9381008	73.5192032	6	19.6938	73.4973984
7	17.9750996	73.4748001	7	19.6998005	73.5297012
8	18.0503998	73.4228973	8	19.6970005	73.5121002
9	18.0466003	73.4235001	9	19.7670002	73.373497
10	17.9421005	73.398201	10	19.7661991	73.3705978
11	17.8903999	73.3889008	11	19.7660999	73.3705978
12	17.9755001	73.475502	12	19.7640991	73.361702
13	17.9547005	73.4494019	13	19.7646999	73.3608017
14	17.9528008	73.4492035	14	19.7597008	73.3562012
15	17.9526997	73.4489975	15	19.7577	73.3543015
16	17.9514999	73.4487991	16	19.7488003	73.3534012
17	17.9286995	73.436203	17	19.7486992	73.3534012
18	17.9344997	73.5614014	18	19.7478008	73.3522034
19	18.0270996	73.6065979	19	19.7444992	73.3487015
20	18.2779999	73.4662018	20	19.7450008	73.3407974
21	18.2880001	73.4733963	21	19.7479992	73.3395004
22	18.0436993	73.6139984	22	19.7474003	73.3378983
23	18.1308002	73.3631973	23	19.7457008	73.3374023
24	18.1411991	73.3523026	24	19.7301006	73.3333969
25	18.1599998	73.5371017	25	19.7324009	73.324501
26	18.1648006	73.5273972	26	19.7318993	73.3239975
27	18.1882	73.530899	27	19.7315998	73.3242035
28	18.1802998	73.5317001	28	19.7297993	73.3237
29	18.1760998	73.5427017	29	19.8729	73.1889038
30	18.1781998	73.5419006	30	19.8775005	73.1884003
31	17.9857998	73.4386978	31	19.9890003	73.2070999
32	17.8775005	73.5346985	32	19.9878006	73.2080994
33	17.9162006	73.5137024	33	19.9897995	73.2067032
34	18.2926998	73.4440002	34	19.9895	73.2068024
35	18.2609005	73.4658966	35	20.0363998	73.1945038

36	18.1285992	73.6091003	36	20.0433998	73.1990967
37	18.1303997	73.6135025	37	20.0447006	73.1999969
38	18.1259003	73.607399	38	20.0443993	73.203598
39	18.1240005	73.6070023	39	20.0436993	73.206398
40	18.0993996	73.600502	40	20.0429001	73.209198
41	18.1009007	73.5996017	41	20.0440006	73.2173996
42	18.0151997	73.6308975	42	20.0594006	73.2368011
43	17.9326992	73.5770035	43	20.0590992	73.2369995
44	17.9029999	73.4360962	44	20.0589008	73.2446976

45	17.9027996	73.4358978	45	20.0589008	73.2471008
46	18.1882	73.5513992	46	20.0596008	73.2493973
47	18.1875992	73.5514984	47	20.0611992	73.2514038
48	18.1562996	73.5737	48	20.0632992	73.2522964
49	18.1814003	73.5795975	49	20.0690002	73.2509995
50	18.1821003	73.5782013	50	20.0499001	73.2089996
51	18.1823006	73.5784988	51	20.0512009	73.2012024
52	18.1791992	73.5836029	52	19.7485008	73.3375015
53	18.1802006	73.5822983	53	19.7493	73.3383026
54	18.1791	73.5822983	54	19.7460003	73.3371964
55	18.1786995	73.5821991	55	19.7462006	73.341301
56	18.1982002	73.5529022	56	19.6900997	73.4966965
57	18.1448002	73.5652008	57	19.7478008	73.3370972
58	18.0265007	73.6023026	58	20.0170002	73.2207031
59	18.0198002	73.6022034	59	20.0175991	73.2205963
60	17.9841995	73.6126022	60	19.7446995	73.3382034
61	17.9333992	73.5633011	61	19.7432003	73.3383026
62	17.9333992	73.5614014	62	19.7457008	73.3401031
63	17.9435997	73.5728989	63	19.7465	73.3491974
64	18.1954994	73.4419022	64	19.7481995	73.3579025
65	18.1881008	73.4410019	65	19.7483006	73.3584976
66	17.9648991	73.5460968	66	19.7451	73.3423004
67	17.9370995	73.5385971	67	19.7483006	73.3412018
68	17.9374008	73.5382004	68	19.7942009	73.3024979
69	17.9349995	73.5503998	69	20.0443993	73.2081985
70	17.9356995	73.5473022	70	20.0443001	73.2068024
71	17.9365005	73.5438995	71	20.0069008	73.2181015
72	17.9666996	73.5886993	72	20.0048008	73.2173996
73	17.9673004	73.5886993	73	20.0039997	73.2164993
74	18.1343994	73.6118011			
75	18.1224995	73.6085968			
76	18.1539001	73.6158981			
77	18.1569996	73.6184006			
78	18.2010002	73.5574036			
79	18.2005997	73.5579987			
80	18.2285995	73.450798			
81	18.2278996	73.4585037			
82	18.2367001	73.4304962			
83	18.2283001	73.4751968			
84	18.2831001	73.4436035			
85	18.1788998	73.4730988			
86	18.1826992	73.4694977			
87	18.1721001	73.4695969			
88	18.2502003	73.5140991			
89	18.2457008	73.5227966			
90	18.0624008	73.3729019			

91	18.0629997	73.3367004
92	18.0692997	73.4514999
93	18.0695	73.4536972
94	18.0888996	73.4048004
95	18.0911999	73.4029007
96	18.0681	73.3578033
97	18.0716	73.3565979
98	17.9447002	73.5875015
99	17.9466991	73.5867004
100	17.9449997	73.5931015
101	17.9596996	73.6059036
102	17.9619007	73.6147995
103	17.9538002	73.6099014
104	17.9655991	73.6214981
105	17.9950008	73.6055984
106	17.9948997	73.6048965
107	17.9946003	73.6052017
108	18.0062008	73.6304016
109	18.0126991	73.6227036
110	17.9761009	73.6148987
111	17.9759007	73.6147003
112	17.9200993	73.527298
113	17.8999004	73.5217972
114	17.9006996	73.5225983
115	17.9365997	73.5786972
116	17.9939003	73.5290985
117	18.0214005	73.5700989
118	18.0445995	73.5886002
119	17.8626995	73.5231018
120	17.8999996	73.5105972
121	17.9197006	73.552597
122	18.0776997	73.365097
123	18.0785007	73.3636017
124	17.9339008	73.4408035
125	17.8598995	73.4695969
126	17.8714008	73.4841003
127	17.8785	73.5062027
128	17.8915005	73.4949036
129	17.8983994	73.4376984
130	17.9009991	73.4377975
131	17.9032001	73.4356003
132	17.9036007	73.433197
133	17.9151993	73.4307022
134	17.9165001	73.4325027
135	17.9307995	73.4375
136	17.9316998	73.4381027

137	17.9333992	73.5623016
138	17.9337997	73.5593033
139	17.9344006	73.5422974
140	17.9349003	73.4420013
141	17.9349003	73.5498962
142	17.9361992	73.5242996
143	17.9365005	73.5249023
144	17.9375992	73.5683975
145	17.9391994	73.5681992
146	17.9395008	73.5686035
147	17.9396	73.5705032
148	17.9400997	73.572197
149	17.9449005	73.5870972
150	17.9473991	73.5858994
151	17.9475994	73.5818024
152	17.9477005	73.5805969
153	17.9477997	73.5789032
154	17.9479008	73.5762024
155	17.9482002	73.5799026
156	17.9482994	73.5852966
157	17.9484997	73.5748978
158	17.9496002	73.5828018
159	17.9519997	73.4487991
160	17.9668007	73.4560013
161	17.9696007	73.4587021
162	17.9755993	73.463501
163	18.0592995	73.3963013
164	18.0601997	73.3961029
165	18.0610008	73.3949966
166	18.0613995	73.3373032
167	18.0615997	73.3585968
168	18.0685997	73.3582993
169	18.0711002	73.4520035
170	18.0718002	73.3495026
171	18.073	73.3787994
172	18.0734005	73.3181
173	18.0746994	73.3881989
174	18.0786991	73.3618011
175	18.0911007	73.3572998
176	18.0923004	73.342598
177	18.1084003	73.3554993
178	18.1149998	73.3641968
179	18.1224995	73.6059036
180	18.1229	73.6022034
181	18.1254005	73.606102
182	18.1282005	73.607399

183	18.1446991	73.6056976
184	18.2618999	73.419899

Table Details of database used in landslide susceptibility mapping

Data layers	Source of data	Scale
Landslide inventory	Public work department http://cgwb.gov.in/ Google earth images Field survey	--
Rainfall and temperature	Department of agriculture Maharashtra state http://maharain.gov.in/ https://www.worldweatheronline.com/	--
Geology	Geological Survey of India (GSI) https://www.gsi.gov.in	1:250,000
Geomorphology	National remote sensing centre https://bhuvan.nrsc.gov.in SRTM https://earthexplorer.usgs.gov/ Landsat8 OLI https://earthexplorer.usgs.gov/ Field survey	1:50000 1 arc second,
Soil map	National bureau of soil survey and land use planning https://www.nbsslup.in/	1:500,000
Digital elevation model (DEM)	SRTM	1 arc second,
Satellite image	Landsat8 OLI	30 m spatial resolution