Journal of the Indian Society of Remote Sensing
https://doi.org/10.1007/s12524-025-02190-3

RESEARCH ARTICLE

Effect of Various Feature Engineering Techniques and Hyper-Parameters Tuning of Machine Learning Approaches for Landslide Susceptibility Map Generation

Vivek Saxena[1] · **Upasna Singh**[2] · **L. K. Sinha**[1]

Received: 6 September 2023 / Accepted: 3 March 2025

Abstract
Landslide Susceptibility Map (LSM) is one of the essential tools for town planner's/ government institutions to make the decision about land use planning. Utilisation of LSM map, will prevent the anthropogenic activities to be executed at essential level at high or very high landslide susceptible locations and thus mitigate the risk due to such natural hazards. The present study integrates the two strategies for LSM modelling in which the first one is to evaluate the effect of various feature engineering techniques applied on causative factors measured at ratio scale while the another one is hyper-parameters tuning of implemented Machine Learning (ML) algorithm i.e. decision tree. All the models were created using statistical/ML techniques based on fourteen pre-processed causative factors which are acquired or derived primarily using remote sensing. These factors includes topographic, anthropogenic, and environmental features while geological factor is acquired from the open domain. The Area under Receiver Operating Characteristics and overall accuracy measures were used for assessing the performance of various implemented approaches for LSM. The study area considered for the present study is Rishikesh to Gangotri axis situated in the Uttarakhand province of India. Comparative study of such statistical and ML models using different types of feature engineering techniques as well as hyper-parameters tuning is the first time attempted and can be used in futuristic analysis of LSM generation for other geographic areas too.

Keywords Frequency ratio · J48 Decision tree · JENKS natural break · Landslide susceptibility map · Logistic regression · Naivebayes

Introduction

A natural hazard arises due to various environmental changes, and brings devastating risk including loss of property or lives at varying intervals and differing duration across the earth. Landslide natural hazard is described as the movement of rock mass, debris, or earth surface down a slope caused by gravitational force. Landslides are frequently triggered by a variety of events including heavy or prolonged rainfall, earthquakes, quick snow melting, volcanic activity, fluctuating reservoir levels, and a wide range of anthropogenic activities. World over landslide has caused huge economic losses and human causalities year on year. According to the "Center for Research on the Epidemiology of Disaster (CRED)" a minimum of seventeen percent of all fatalities from natural hazard are due to landslides (Lacasse & Nadim, 2009).

Landslide susceptibility maps (LSM) demarcate the earth's surface into various zones and depicts the susceptibility of that area, towards a landslide in terms of probability. Landslide susceptibility is sometimes got confused with landslide hazard zonation which seems to be feasible for a small area with excellent data coverage and involves the time dimension for mass movement (Carrara & Pike, 2008; Fell et al., 2008). Overall, it can be concluded that LSM have paramount importance in land use planning

✉ Vivek Saxena
vivek.dgre@gov.in

Upasna Singh
upasnasingh@diat.ac.in

L. K. Sinha
Sinha.lokesh61@gmail.com

[1] DGRE, DRDO, Chandigarh, India

[2] Deptt. of Computer Science & Engineering, DIAT, Pune, India

Published online: 29 April 2025

and hence good governance by district/town administrators and decision makers.

During the transition period of the early 90 s to late 90 s, Brabb in California; Carrara and Guzzetti in Italy have done pioneer work in this subject (Brabb, 1991; Carrara et al., 1991; Guzzetti et al., 1999). The field based method require the demarcation of area in various zones by the domain expert through extensive field study. Expert may have their inherent biases, which will further hamper the quality outcome of the model. This type of method come under qualitative methods of investigation.

The advancement in information and space technology, lead researchers towards the exploitation of Remote Sensing (RS) data for generating LSM. A shift was observed with researchers using the quantitative method over RS data. Further, since the study area for the generation of LSM is huge, so purely field-based methods are ruled out. The first popular method extensively used was the sum of weighted overlays i.e. Bureau of Indian Standards BIS (Anbalagan, 1992; Ghosh et al., 2009). BIS method is a field intensive method where field observations are taken for the causative factors and the weighted sum of each thematic features is calculated in the raster domain. Further, this sum goes for various threshold checks and accordingly, the area pixels are demarcated in various zones of LSM. These weights and thresholds have been defined by the group of experts but may not yield a high accuracy due to very dynamic terrain topography and geology. Another method which is also an expert-driven heuristic method, is called the Analytical Hierarchy Process (AHP) and given by SAATY (Saaty, 2002). Based on the geographic area being investigated, expert assign the weights and carry out pairwise comparisons for all causative factors (one vs all). A weight matrix is created by the domain expert followed by the computation of eigenvalues, and eigenvectors of this weight matrix. The AHP methods are more accurate and less time consuming than other expert driven methods as experts assign the weight for a particular geographic area that will be best suited for that area only. Various researchers have applied AHP methods for different terrain topography and reported an AUROC score of around 80% (Chen et al., 2016; Kumar & Anbalagan, 2016; Yalcin et al., 2011).

In the last two decades, various researchers have further advanced the field by investigating the LSM generation using statistical and machine learning (ML) methods. These methods are also known as data-driven techniques and establish the relationship between dependent attribute/ variable to various causative factors. The main strength of these methods is lower bias and better accuracy in comparison to expert driven methods. Marjanovic et al. (2019) have proposed various concepts, starting from sampling strategies, cross scaling, landslide inventories, choices of causative factors, and ML techniques; for improving the LSM accuracy. The various popular data-driven statistical and ML methods which has been applied for LSM gene or LSM generation includes weight of evidence (Regmi et al., 2010; Reichenbach et al., 2018), frencyratio (Park et al., 2013; Yalcin et al., 2011), naive bayes (Pham et al., 2016, 2018), information value (Chen et al.,2020), Artificial Neural Network (ANN) (Chen et al., 2017; Kalantar et al., 2018), logistic regression (Kalantar et al., 2018; Kavzoglu et al., 2014; Marjanović et al., 2011), decision tree (Saravanan & Gayathri, 2017; Zhang et al., 2017), and support vector machine (Hong et al., 2017; Pham et al., 2018).

The main contributions of the present study are summarised as.

a) The application of three feature engineering techniques (JENKS natural break, equal interval, quantile) on causative factors measured at ratio scale.
b) Hyper-parameters tuning of ML algorithms by keeping at par in-sample and out-sample error while model complexity is kept at optimum level.
c) Compare the predictive performance of two most popular ML algorithms i.e. logistic regression, J48 decision tree with conventional methods like frequency ratio, naïve bayes, and information value for the study area.

The point mentioned at ‘a’ above, is not very commonly attempted for generating LSM of any geographical area and almost all research studies till now attempted evaluating LSM via equal interval grouping or an arbitrary method based on expert judgment (Hussain et al., 2022; Park et al., 2013; Pham et al., 2015; Chen et al., 2020). Further, the ML algorithms are implemented with default hyper-parameters values available in the library without discussing the effect of these hyper-parameters on model generalization/ overfitting scenario. This issue is addressed in contribution mentioned at point ‘b’ above. David has given the ‘no free lunch’ theorem which states that ‘there is no single technique or algorithm that is best for all situations and data sets (Wolpert & Macready, 1997). Hence, as mentioned in point c above, various popular ML/statistical algorithms were implemented and their comparative performance is evaluated for the study area.

The remainder of the paper is organized as follows: Sect. "Study Area" gives an overview of the high landslide prone study area. Sect. "Methodology" describes the dataset creation, data cleaning, proposed data pre-processing and the various models implemented for creating LSM. In Sect. "Results and Discussion", all experimental results are discussed and comparatively evaluated using various performance assessment measures. Finally, Sect. "Conclusion" concludes the study.

Study Area

The study area is located in Uttarakhand province of India and is part of the great Himalaya starting from Rishikesh district to Gangotri glacier along the road axis (NH34) with a buffer area of 3 km on both sides. The study area polygon is bounded by a latitude of $31^0 07' 45''$N to $30^0 05' 47''$N and a longitude of $78^0 20' 23''$E to $78^0 96' 77''$E. Further, the study area is located in north western part of Uttarakhand with crossing/adjacent to the district of Uttarkashi from the north side; Tehri Garhwal and Uttarkashi from the east and west sides; and Dehradun and Pauri Garhwal from the south. Total study area covers 1096.68 km^2. Figure 1 shows the study area along with the road network and marked landslide inventories distributed in that area. The study area has a good network of metaled roads and the tallest dam in Asia known as "Tehri Dam" (covering an area of 42 km^2) which is also the fourth tallest earth and rock fill dam in the world. Topography for most of the study area places is hilly and has the highest altitude of 4840 m while the lowest altitude is 320 m.

Methodology

Pre-Processing

Pre-processing is carried out by following steps in sequence starting with (1) dataset creation (both dependent attributes i.e. landslide inventory and causative factors) (2) multicollinearity test and correlation coefficient test for checking redundancy of causative factors (3) converting the ratio scale based causative factors to interval scale using three different strategies, (4) preparing training and test dataset with 70:30 ratio respectively.

Dataset Creation

Since LSM is generated for strategic decision making rather tactical one, hence time dependent input data like rainfall/precipitation is not considered (Chen et al., 2016; Hong et al., 2017). Similarly, the dataset which requires extensive field study have also not been considered in the present study. These variables are NDVI (Chen et al., 2017; Hong et al., 2017; Marjanovic & Caha, 2011), texture, soil thickness, tree type, age and diameter (Park et al., 2013; Sarkar & Kanungo, 2004). However, above mentioned variables have been considered for evaluating LSM by some researchers referred above.

Landslide Inventory Mapping A total of 154 landslides are identified and marked as polygon using google earth high-resolution natural color imageries in QGIS software. The smallest marked landslide site cover an area of 970 m^2 while the largest landslide cover an area of 544,785 m^2. Figure 1 shows all the overlaid landslides polygon over classified elevation map in the study area. Figure 2 shows the prominent landslides observed and validated along the road axis in the study area and marked sequentially in Fig. 1.

Preparation of Causative Factors A total of fourteen causative factors are created based on the data available in the open domain which includes the Shuttle Radar Topography Machine (SRTM) Digital Elevation Model (DEM), Google Earth (GE) high resolution satellite imageries and Geological Survey of India (GSI) maps. All these causative factors can be categorised into four subgroups mentioned below.

A. **Topographic factors** which includes DEM [V1] and its derivatives like slope [V2], aspect [V3], slope length [V4], plan curvature [V5], profile curvature [V6], Stream Power Index (SPI)[V7], Sediment Transport Index (STI)[V8] and Topographic Wetness Index (TWI) [V9].

- **DEM [V1]** is downloaded from USGS Earth Explorer at 30 m spatial resolution and re-projected to the UTM 44N projection system. In mountain areas, various factors like weather, vegetation, and anthropogenic activity are closely related to elevation (Ngo et al., 2021).
- **Slope [V2]** map is derived from the DEM and each pixel represent the "steepest flow direction using eight neighbourhood pixels". It is a crucial parameter which affect the stability of landmass and hence landslide (Yuan et al., 2019).
- **Aspect [V3]** define the direction of the steepest down slope, which is again derived from DEM. This is the main factor behind rain water seepage, runoff and absorption of solar radiation (Wang et al., 2019).
- **Plan curvature [V4]** is the curvature of the surface perpendicular to the slope direction. Plan curvature relates to the convergence and divergence of flow across a surface. The derived plan curvature has a value between -0.027 to 0.026, and the variable is measured at ratio scale. A positive value indicates the surface is sideward convex at that cell while a negative plan indicates the surface is sideward concave at that cell
- **Profile curvature [V5]** is the curvature of the surface in the direction of slope. Profile curvature affects the acceleration or deceleration of flow across the surface and hence influences erosion and deposition. The derived profile curvature has a value between -0.035

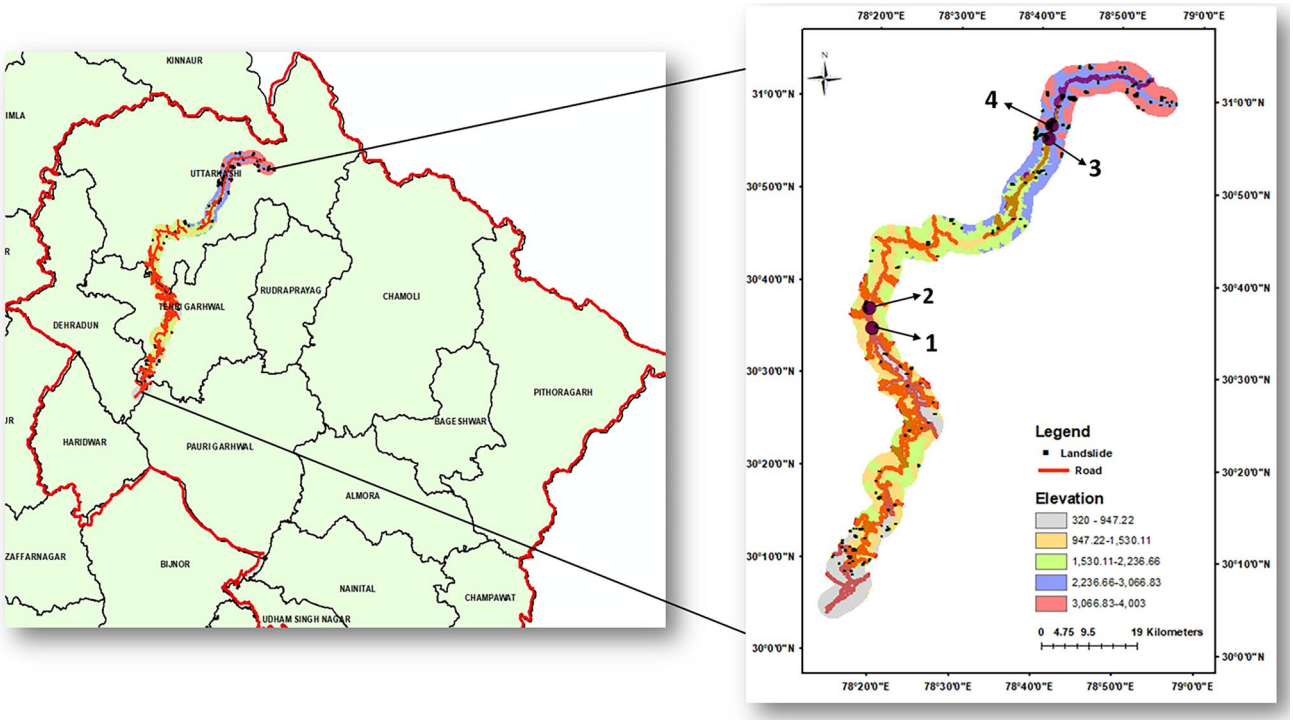


Fig. 1 Study area covering locations of landslides and road network

to 0.034 and the Variable is measured at a ratio scale. A negative value indicates that the surface is upwardly convex at that cell. A positive value indicates that the surface is upwardly concave at that cell. A value of zero indicates that the surface is linear.

Both plan and profile curvature are second-order derivatives of DEM or first order derivatives of slope along and perpendicular to direction of slope respectively. They define the convexity or concavity of the slope shape and hence surface runoff and erosion process which contribute towards landmass movement (Chen et al., 2021).

Fig. 2 Prominent landslides validated along the road axis in the study area depicted in Fig. 1

- **Slope length [V6]** is defined as the distance from the point of origin of overland flow to the point where either the slope gradient decreases enough that deposition begins, or runoff water enters a well-defined channel (wet/dry).
- **Topographic Wetness Index [V7]** explains the potential of soil moisture present at that cell or steady state wetness of soil or quantifies topographic control on hydrological processes. It is calculated using a raster calculator with the formula given in Eq. 1.

$$TWI = \ln[(Flowaccumulation)/\{cellsize * \tan(slope\_in\_radian)\}] \quad (1)$$

A higher value of TWI indicates that soil is more likely to reach saturation and cause slippage (Sun et al., 2020).

- **Stream Power Index [V8]** measures the erosive power of flowing water/stream on that cell. As catchment area and slope gradient increase, the amount of water contributed by upslope areas and the velocity of water flow increase. Hence, stream power index and erosion risk also increases. It is computed using Eq. 2.

$$SPI = (Flowaccumulation) * \tan(slope\_(in\_radian))/(cell\_size) \quad (2)$$

- **Sediment Transport Index [V9]** defines whether erosion or deposition will occur at that cell. Sediment transport will be due to gravity and the movement of fluid/water.

$$STI = (As/22.1 * 30))^{0.6}(Sin(slope\_in\_radian)/0.0896)^{1.3} \quad (3)$$

As = upstream/upslope area or catchment area or flow accumulation.

Larger the value of SPI or STI means stronger erosion and transportation capacity of water flow. This will lead to less stability of the slope adjacent to the river channel (Pradhan & Kim, 2014).

All eight derivatives [V2 to V9] are derived using direct function call or raster calculator of QGIS software using DEM. Figure 3 shows the categorised DEM (based on Jenks criteria) and Figure 4 shows the aspect map.

B. **Geological factors** taken into consideration are fault line [V10] and lithology [V11]. Fault line is digitised using high resolution GE imageries, while the lithology map is downloaded from the GSI website. Figure 5 shows the lithology map of the study area. Lithology defines the physical and chemical properties of the rock which governs the slope stability. Further, location close to fault line with discontinuous rock type have a high probability of slope failure.

C. **Environmental factors** taken into consideration are landcover [V12] and river [V13]. Both landcover and river feature are digitised using GE imageries. Figure 6 shows the landcover map. Landcover and river factors have been used in most of the studies for LSM (Park et al., 2013; Pham et al., 2016).

D. **Anthropogenic factor** like the road [V14] feature is again digitised using GE imageries. Figure 7 represents the road and fault line marked for the study area. Road cutting disturb the natural equilibrium of any slope and the locations close to the road have a high susceptibility to landslides. This feature has been used in numerous studies for LSM.

Features at sl. no. V1 to V9 are generated in the raster domain, while features like O & V10 to V14 are generated in the vector domain. All vector features are then rasterised at the same spatial resolution of 30 m and having the same coordinate reference system i.e. UTM 44N. Further, all variables are clipped to the study area polygon and got same no. of study area pixels while non-study area pixels got assigned NULL / no data values. All variables are then converted into an ASCII file having 8,812,680 samples, and after removing NULL values i.e. outside study area pixels; a truncated dataset ASCII file is created having all 15 variables in which there were 5928 samples were representative of landslide pixels and the rest are of non-landslides (1,212,624). The total number of samples (pixels) in the population are 1,218,552.

Multicollinearity Test and Correlation Coefficient Test

When two or more independent factors are highly correlated with each other, multicollinearity might be present in the data. Although multicollinearity is not an issue for non-parametric algorithms but as a pre-processing measure, a check for multicollinearity is performed to drop any correlated attributes which will reduce the computational time for algorithm implementation and hence its complexity also. Multicollinearity can be checked by computing variation inflation factor (VIF), pairwise scatter plot, and finding the Eigen values in a correlation matrix. If VIF is 1, features are not correlated. If VIF is less than 5, there is little correlation. If VIF is between 5–10, there is a mild correlation between the variables. If VIF is more than 10, there is a high correlation between the variables and issue needs to handled (Lee et al., 2018; Zhou et al., 2021).

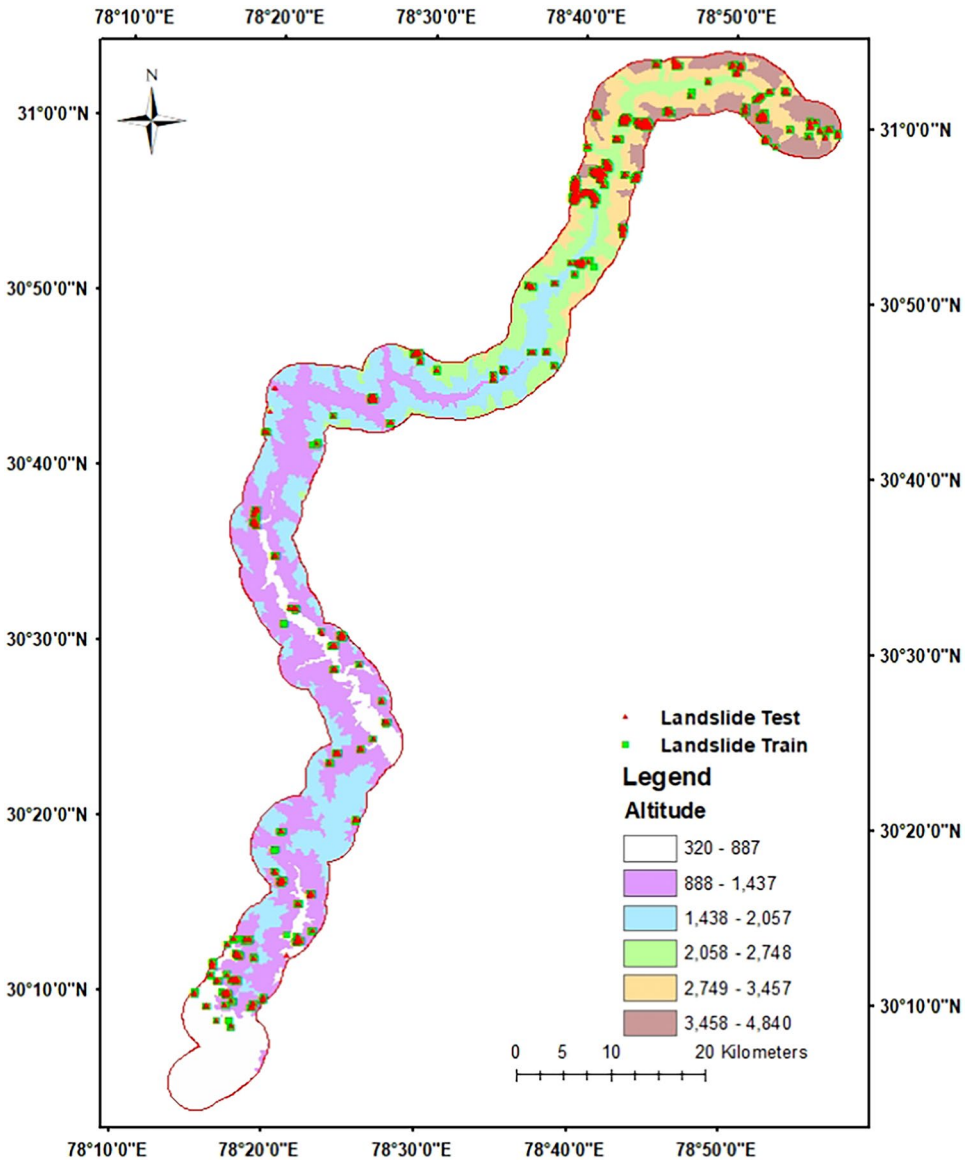


Fig. 3 Categorised DEM based on JENKS natural break

A multi-collinearity test is performed on all the causative factors measured at the ratio scale (DEM and its derivatives i.e. V1 to V9) after removing the no data values or NULL values. Using the Variance Inflation Factor (VIF) function of the 'statsmodels' python package, the VIF of each variable is computed against all others. Result is summarised in Table 1. A VIF value of less than 10 signifies little correlation, so we have not dropped any of the causative factors. The maximum VIF is found to be 7.16 and it becomes prudent to assume that all variables are significant. Further, the Pearson Correlation Coefficient (PCC) signifies the correlation between any two variables. Hence, PCC is computed for all DEM, and its derivative factors. The result is summarised in Table 2.

As observed from Table 2, SPI and STI are highly correlated. The rest of the variables do not signify any substantial correlation value, which may lead to truncation of that feature for further processing. So, one of the features among STI and SPI may be omitted/dropped, and the performance of the LSM can be gauged.

Converting the Ratio Scale Factors (DEM and its Derivatives Except Aspect) to Interval Scale

DEM being a numeric attribute hence all its derivatives are also measured at a ratio scale. All these variables are converted into an interval scale. The intervals are defined using three available techniques in ARCGIS and implemented using PYTHON which are as follows:

A. JENKS natural break criteria which is similar to OTSU method used in image processing and Fisher's Linear

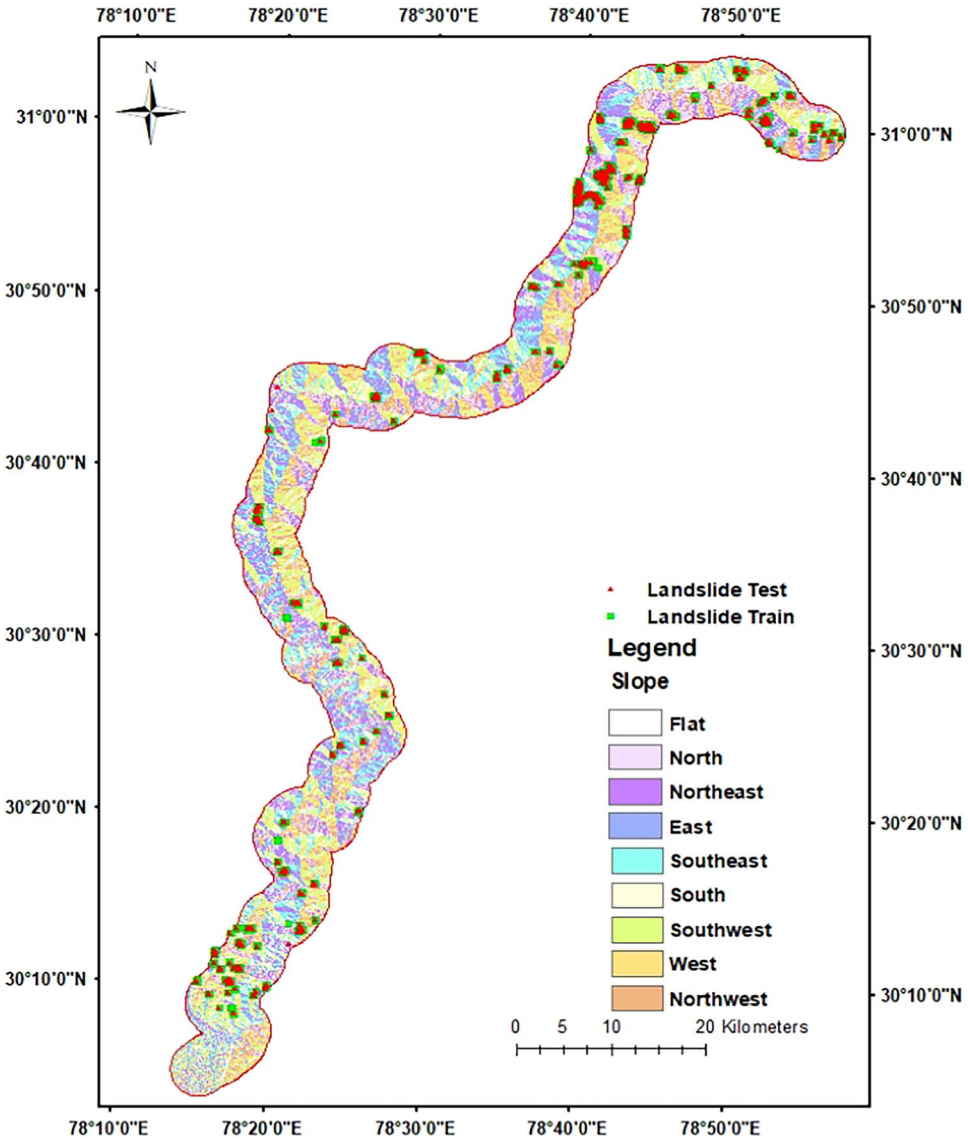


Fig. 4 Aspect map of study area

Discriminant Analysis (LDA) method which is very popularly found applications in classical statistics, machine learning and pattern recognition field for data clustering (Anowar et al., 2021; Huang et al., 2012). This approach maximise the inter class variance and minimize the intra class variance. Using this method, data points of all generated factors except aspect, are divided into six classes as same no. of classes are created in other causative factors like distance from road, river and fault line.

B. Second grouping method used is equal interval which has been used by various authors/researchers in available LSM literature till date (Chen et al., 2020; Pham et al., 2016). Here also six classes are created for all data points of each causative factor as mentioned in point A above.

C. Third method is the quantile method which puts the same no. of samples in each class making the histogram of each bin equal. Total six classes are created for all the data points of each causative factor. Only exception to this equal frequency is one class in slope length, STI and SPI factor, for which the total no. of samples for '0' values were higher than the expected no. of samples in each class which is '203,092' in our case.

Considering the above three techniques, we got three groups of data generated i.e. group 1 as per natural break, group 2 as per equal interval, and group 3 as per quantile criteria.

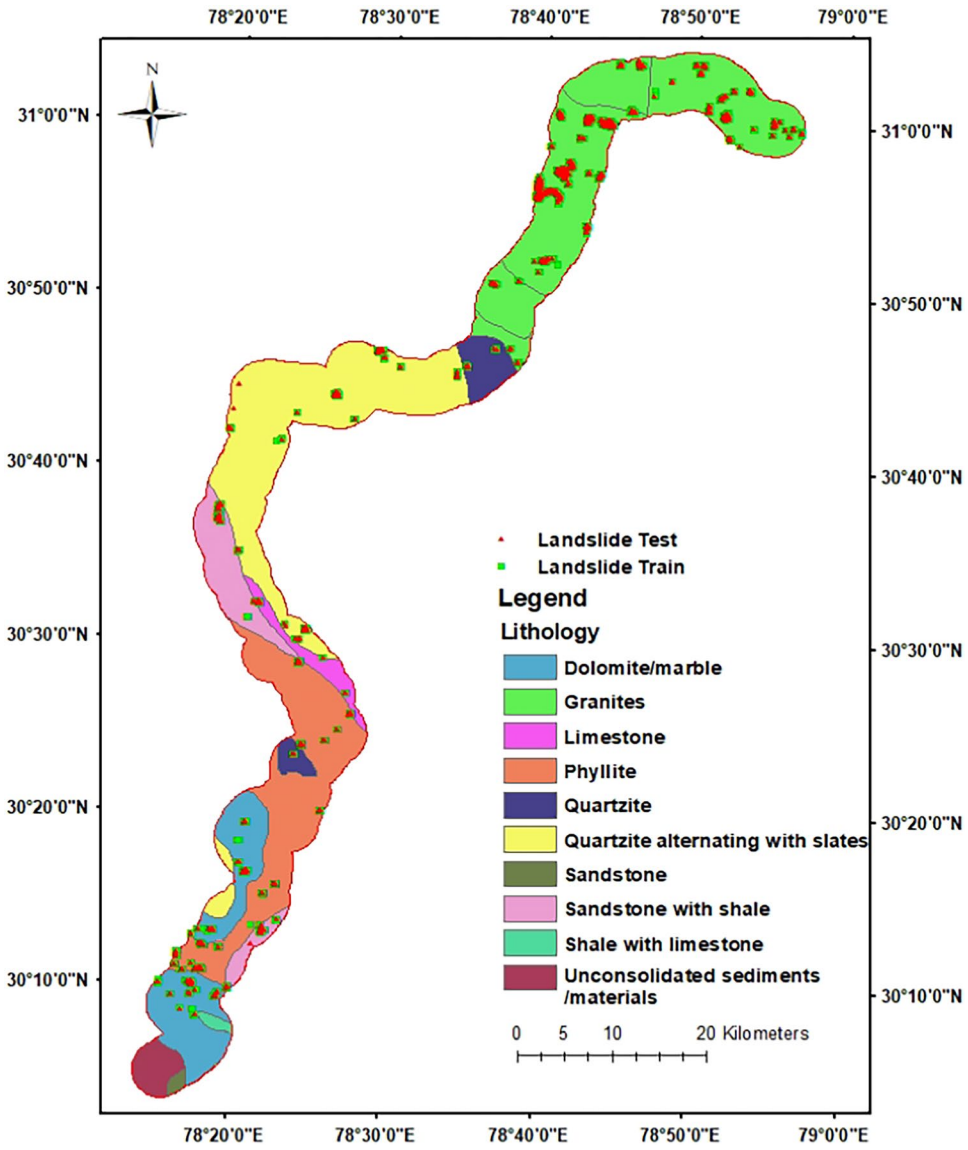


Fig. 5 Lithology map

Feature Importance Analysis

A total of fourteen causative factors are considered for predicting the landslide susceptibility at a particular point. But the contribution made by each of these factors in various susceptibility model is different. So, the information gain ratio (IGR) technique is used to quantitatively evaluate the importance of each causative factor using their average merit value (Zhou et al., 2018). The greater the average merit of a factor, the more important role it plays in the prediction of landslide. For the computation of information gain, we are assuming that the training data X consists of n samples, and belongs to the class C_i (landslide, non-landslide). Then, the information entropy is:

$$Info(X) = -\sum_{i=1}^{2} \frac{n(C_i, X)}{|X|} log_2 \frac{n(C_i, X)}{|X|} \tag{4}$$

The amount of information $(X_1, X_2, X_3, \ldots\ldots\ldots, X_{14})$ split of X regarding the causal factor Y is given as:

$$Info(X, Y) = -\sum_{j=1}^{m} \frac{X_i}{|X|} log_2 Info(X) \tag{5}$$

The Information Gain Ratio (IGR) of the landslide causal factor Y is calculated as:

$$IGR(X, Y) = Info(X) - Info(X, Y) \tag{6}$$

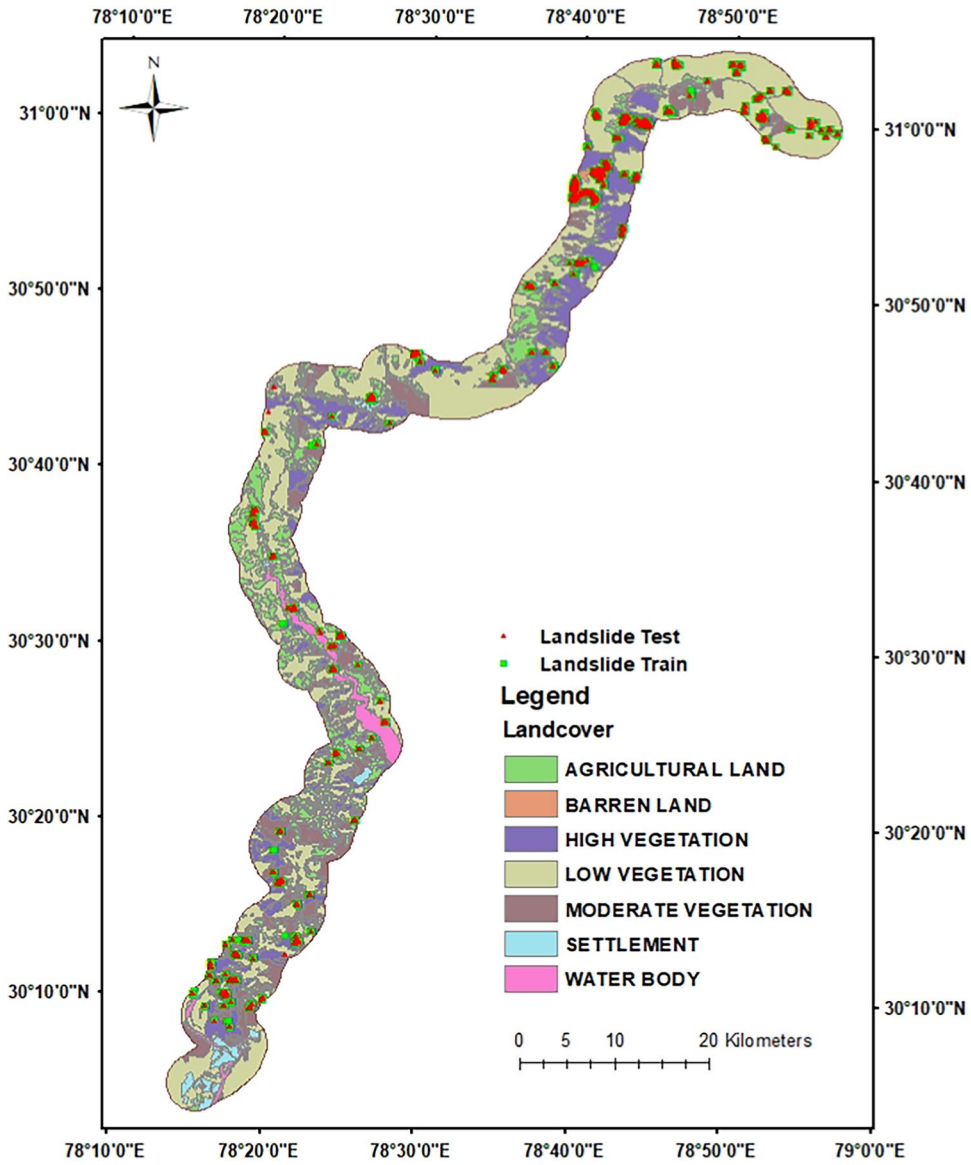


Fig. 6 Landcover map

Preparing Training and Test Dataset

The generated dataset in all three grouping techniques are divided into train and test sub groups with 70:30 ratios respectively with stratification on which ensures the landslide samples got divided proportionately in train and test data. Further, samples are drawn randomly for each group. Entire pre-processing steps, along with model evaluation steps are summarized in the flow chart shown in Fig. 8.

Landslide Susceptibility Assessment Models

Frequency Ratio

Frequency Ratio (FR) technique is a statistical approach to simulate environmental conditions (Hidayat et al., 2019; Javad et al., 2014). This method is based on the relationship between the spread of the landslide point with reference to each of the independent variable's classes. This relationship explains the contribution of each variable's classes to the occurrence of the landslides. Once computed the FR (as per Eq. 7), one can calculate the Landslide Susceptibility Index (LSI) as mentioned in Eq. 8. LSI predict each sample's probable susceptibility w.r.t. landslide and can be divided into two class/multi

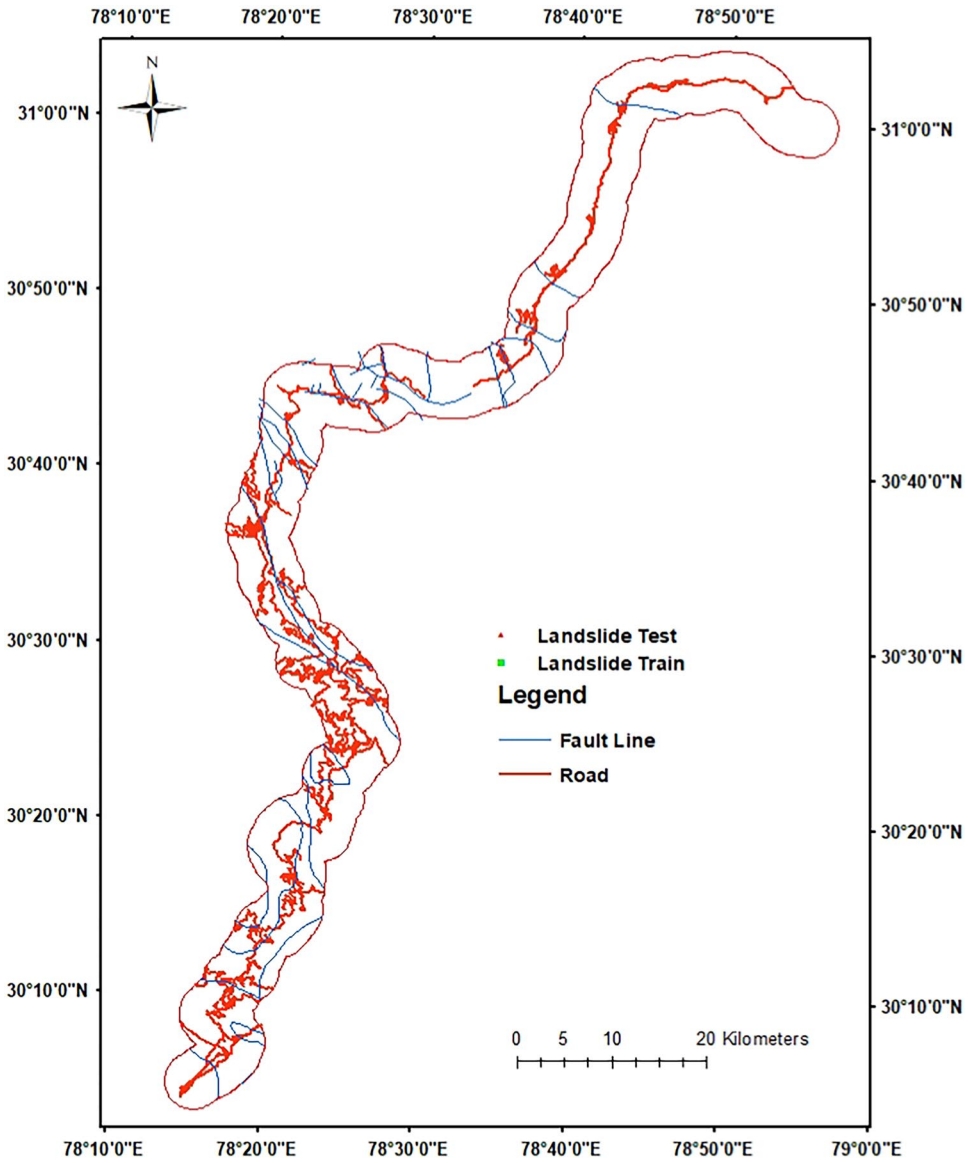


Fig. 7 Fault line and Road map

class output as per the need based on JENKS natural break criteria.

$$\text{Frequency Ratio (FR)} = \frac{\text{Percentage of grids showing landslide occurrence}}{\text{Percentage of grids in domain}} \quad (7)$$

$$\text{LSI (P)} = \sum_{n=1}^{14} \text{FRi} \quad (8)$$

Table 3 depicts the FR values for each causative factor's sub class for group 1 data.

Table 1 VIF value for all causative factors derived from DEM

FEATURES	VIF FACTOR
DEM	6.1574
ASPECT	3.8829
PLAN_CURVATURE	1.5912
PROFILE_CURVATURE	1.3756
SLOPE_LENGTH	1.9892
SLOPE	7.1666
SPI	3.7922
STI	5.2652
TWI	5.8921

Table 2 PCC value for DEM and its derived feature

	Dem	Aspect	Plan Curvature	Profile Curvature	Slope Length	Slope	SPI	STI	TWI
Dem	1.0000	0.0072	0.0137	−0.0253	0.0920	0.4137	0.0365	0.0962	0.0149
Aspect	0.0072	1.0000	0.0011	0.0001	0.0073	0.0386	0.0017	0.0104	0.0005
Plan Curvature	0.0137	0.0011	1.0000	−0.5193	−0.2491	0.0131	−0.1316	−0.3138	−0.1439
Profile Curvature	−0.0253	0.0001	−0.5193	1.0000	0.0866	−0.0050	0.0701	0.1609	0.0567
Slope Length	0.0920	0.0073	−0.2491	0.0866	1.0000	0.0208	0.2904	0.4407	0.3092
Slope	0.4137	0.0386	0.0131	−0.0050	0.0208	1.0000	−0.0022	0.1206	−0.0093
SPI	0.0365	0.0017	−0.1316	0.0701	0.2904	−0.0022	1.0000	0.8340	0.0957
STI	0.0962	0.0104	−0.3138	0.1609	0.4407	0.1206	0.8340	1.0000	0.1610
TWI	0.0149	0.0005	−0.1439	0.0567	0.3092	−0.0093	0.0957	0.1610	1.0000

Information Value

The Information Value (IV) model is a bivariate statistical method which is used to predict the spatial relationship between landslides and landslide factor's classes and similar to FR in nature (Chen et al., 2020; Wubalem, 2021). In simple words, the information value is used to determine the degree of influence of individual causative/independent factor's classes over the landslide occurrence. The formula for the calculation of the information value is given in.

equation no. 9.

$$\mathrm{IV}(\mathrm{Y},\mathrm{Xi}) = \log_2 \frac{\mathrm{S}_0^{\mathrm{i}}/\mathrm{S}}{\mathrm{A}_0^{\mathrm{i}}/\mathrm{A}} \tag{9}$$

IV(Y, Xi) is the information value under the causal factor x_i; S is the total area of the landslide; S_0^i is the landslide area under the factorx_i; A is the total area of the study area; A_0^i is the area under the factorx_i.

Further, the information value is normalized (IV_{norm}) between 0.1 to 0.99 as the calculated information value has a value from -1 to 1 and the LSI value may go negative also due to summation. The same is also summarized in Table 3 for group 1 data.

LSI can be computed using the formula given in Eq. 10.

$$\mathrm{LSI}\ (\mathrm{P}) = \sum\nolimits_{\mathrm{n}=1}^{14} \mathrm{IVnorm} \tag{10}$$

LSI predict each sample's probable susceptibility w.r.t. landslide and can be divided into two class/multi class output as per the need based on JENKS natural break criteria.

Further, both FR and IV can be computed using Eq. 7 and 9 for group 2 and 3 data, and additional tables akin to table no. 3 are created for these group data.

Naïve Bayes

A well-known and classic algorithm i.e. Naïve Bayes (NB) which still works well in so many scenarios for classification problem (Hong et al., 2018; Pham et al., 2017). It has got the underlying assumption of each feature is independent and contribute to the outcome equally. Although practically independent assumption does not hold true but still this give far superior results with linear training time.

Let X = (X1,X2,…..,X14) be a vector of fourteen causative factors and y has two classes i.e. landslide and non-landslide. The Naïve Bayes classifier can be summarised using Eq. 11.

$$\mathrm{YNB} = \operatorname{argmax} \mathrm{P}(\mathrm{Yi}) \prod_{i=1}^{14} P(\mathrm{Xi}/\mathrm{Yi}) \tag{11}$$

where $P(Y_i)$ is the prior probabilities of event Y_i (landslide/non landslide) and $P(X_i/Y_i)$ is the conditional probability which can be calculated using Eq. 12.

$$P\left(\frac{X_i}{Y_i}\right) = \frac{1}{\sqrt{2\pi\sigma}} e \frac{-\left(X_i - \mu\right)}{2\sigma^2} \tag{12}$$

where μ is the mean and σ is the standard deviation of Xi.

Logistic Regression

Logistic Regression (LR) model predicts the occurrence of an event as a linear set of predictor / causative factors (Akgun, 2012; Kavzoglu et al., 2014, 2019). The probability of a landslide event can be determined from the following equation:

$$\mathrm{P} = \mathrm{P}(\mathrm{Y}/\mathrm{X}) = \frac{1}{1+e^{-z}} \tag{13}$$

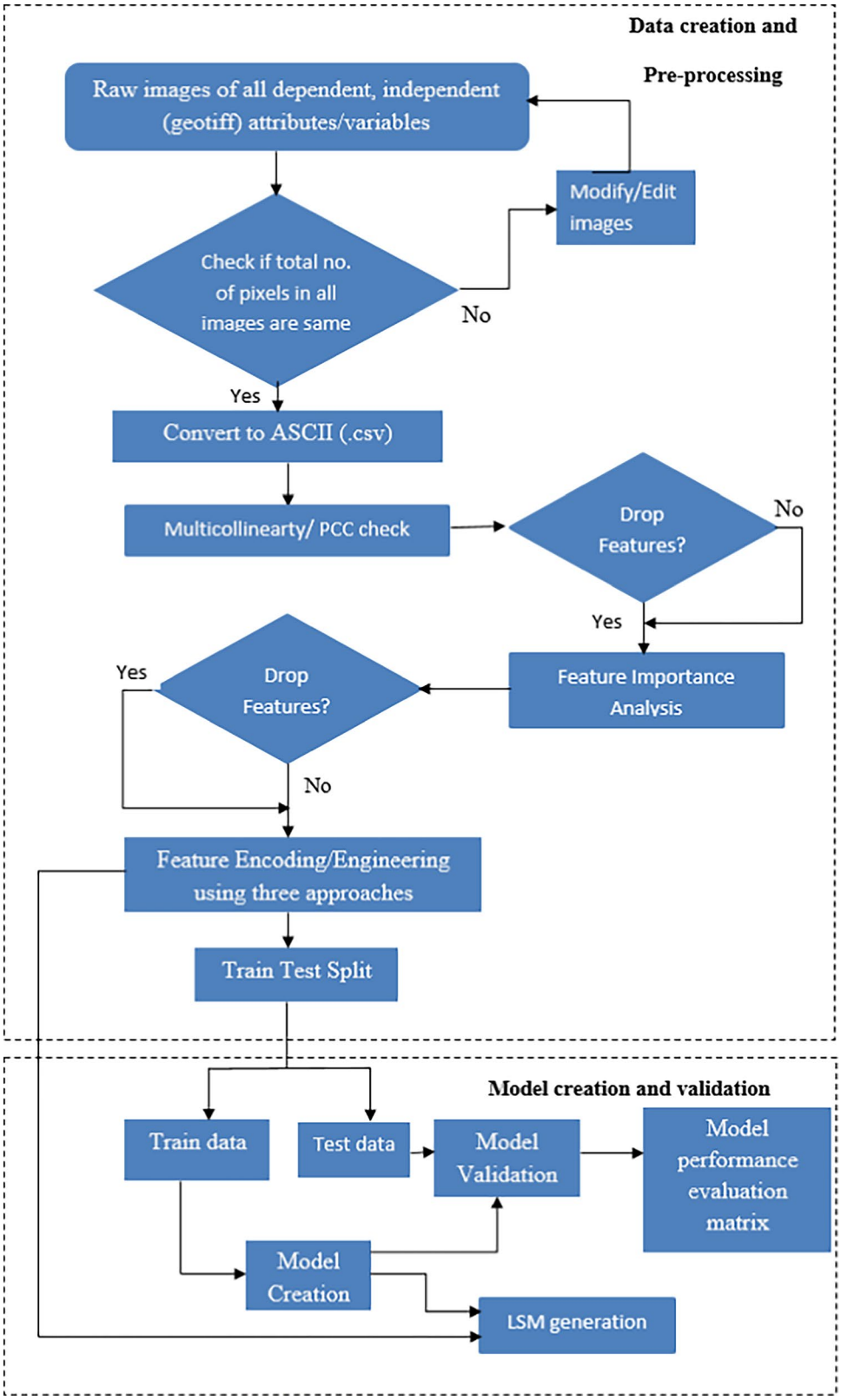


Fig. 8 Flow chart for pre-processing and model evaluation adopted in present study

Table 3 Results of FR & normalised IV computed for each causative factor's subclass for group 1 data

Factor	Class	Grids in domain *A*	No. of grids showing landslide occurrence *B*	% of grids showing landslide occurrence $C=(0.5+\text{Bi})/\sum\text{B}$	% of grids in domain $D=(0.5+\text{Ai})/\sum\text{A}$	Frequency ratio (*FR*) = C/D	$IV=\log_2$ (C/D)	*Normalized IV*
Plan Curvature	< = 2.543456388 (0)	15,784	190	0.05	0.02	2.48	1.31	0.99
	−2.543456387 – −0.957240183 (1)	102,526	731	0.18	0.12	1.47	0.55	0.54
	−0.957240182 – −0.092031344 (2)	254,934	1077	0.26	0.30	0.87	−0.20	0.10
	−0.092031343 – 0.773177495 (3)	322,112	1229	0.30	0.38	0.78	−0.35	0.01
	0.773177496 – 2.215192226 (4)	133,224	685	0.17	0.16	1.06	0.08	0.26
	Greater than 2.215192226 (5)	24,406	238	0.06	0.03	2.01	1.01	0.81
Profile Curvature	< = 3.46856041 (0)	19,593	239	0.06	0.02	2.51	1.33	0.99
	−3.46856040 – −1.346149886 (1)	100,398	556	0.13	0.12	1.14	0.19	0.33
	−1.346149885 – −0.133343871 (2)	241,301	991	0.24	0.28	0.84	−0.24	0.08
	−0.133343870 – 1.079462141 (3)	334,473	1260	0.30	0.39	0.77	−0.37	0.01
	1.079462142 – 3.050271913 (4)	128,432	791	0.19	0.15	1.27	0.34	0.42
	Greater than 3.050271913 (5)	28,789	313	0.08	0.03	2.24	1.16	0.89
Aspect	< = 0 (0) i.e. flat	825	0	0.00	0.00	0.12	−3.01	0.01
	0 – 22.5 or greater than equal to 337.5 (1)	108,754	134	0.03	0.13	0.25	−1.98	0.27
	22.5 – 67.5 (2)	100,135	242	0.06	0.12	0.50	−1.01	0.51
	67.5 – 112.5 (3)	112,810	464	0.11	0.13	0.85	−0.24	0.71
	112.5 – 157.5 (4)	100,800	901	0.22	0.12	1.84	0.88	0.99
	157.5 – 202.5 (5)	113,422	739	0.18	0.13	1.34	0.42	0.87
	202.5 – 247.5 (6)	110,504	819	0.20	0.13	1.52	0.61	0.92
	247.5 – 292.5 (7)	113,488	669	0.16	0.13	1.21	0.28	0.84
	292.5 – 337.5 (8)	92,248	182	0.04	0.11	0.41	−1.30	0.44
Fault	50 (0)	21,958	37	0.01	0.03	0.35	−1.51	0.01
	100 (1)	21,793	43	0.01	0.03	0.41	−1.29	0.15
	150 (2)	21,835	38	0.01	0.03	0.36	−1.46	0.04
	200 (3)	21,457	44	0.01	0.03	0.43	−1.23	0.18
	250 (4)	21,266	47	0.01	0.02	0.46	−1.12	0.24
	300 (5)	744,677	3941	0.95	0.87	1.09	0.12	0.99

Table 3 (continued)

Factor	Class	Grids in domain *A*	No. of grids showing landslide occurrence *B*	% of grids showing landslide occurrence $C=(0.5+\mathrm{Bi})/\sum \mathrm{B}$	% of grids in domain $D=(0.5+\mathrm{Ai})/\sum \mathrm{A}$	Frequency ratio $(FR)=\mathrm{C/D}$	$IV=\log_2$ (C/D)	*Normalized IV*
River	40 (0)	49,530	121	0.03	0.06	0.50	−0.99	0.01
	80 (1)	16,361	133	0.03	0.02	1.68	0.75	0.97
	120 (2)	15,714	131	0.03	0.02	1.72	0.78	0.99
	160 (3)	15,385	104	0.03	0.02	1.40	0.48	0.82
	200 (4)	15,057	101	0.02	0.02	1.39	0.47	0.82
	240 (5)	740,939	3560	0.86	0.87	0.99	−0.02	0.55
Road	40 (0)	44,701	154	0.04	0.05	0.71	−0.49	0.01
	80 (1)	37,214	149	0.04	0.04	0.83	−0.28	0.40
	120 (2)	32,516	156	0.04	0.04	0.99	−0.02	0.87
	160 (3)	29,103	135	0.03	0.03	0.96	−0.06	0.79
	200 (4)	26,101	117	0.03	0.03	0.93	−0.11	0.70
	240 (5)	683,351	3439	0.83	0.80	1.03	0.05	0.99
DEM	< = 887 (0)	114,459	198	0.05	0.13	0.36	−1.49	0.01
	888 –1437 (1)	267,850	742	0.18	0.31	0.57	−0.81	0.26
	1438 –2057 (2)	206,998	398	0.10	0.24	0.40	−1.34	0.07
	2058 –2748 (3)	102,454	1089	0.26	0.12	2.19	1.13	0.97
	2749 –3457 (4)	98,758	1089	0.26	0.12	2.27	1.18	0.99
	Greater than 3457 (5)	62,467	634	0.15	0.07	2.09	1.06	0.95
Slope	< = 10.87785184 (0)	98,977	60	0.01	0.12	0.13	−2.99	0.01
	10.87785185 –21.16771168 (1)	159,845	395	0.10	0.19	0.51	−0.98	0.43
	21.16771169 –29.69359556 (2)	200,048	624	0.15	0.23	0.64	−0.64	0.50
	29.69359557 –37.92548343 (3)	190,260	986	0.24	0.22	1.07	0.09	0.66
	37.92548344 –47.33335529 (4)	140,835	1100	0.27	0.17	1.61	0.68	0.78
	Greater than 47.33335529 (5)	63,021	985	0.24	0.07	3.21	1.68	0.99

Table 3 (continued)

Factor	Class	Grids in domain *A*	No. of grids showing landslide occurrence *B*	% of grids showing landslide occurrence $C=(0.5+\mathrm{Bi})/\sum \mathrm{B}$	% of grids in domain $D=(0.5+\mathrm{Ai})/\sum \mathrm{A}$	Frequency ratio (*FR*) = C/D	$IV=\log_2$ (C/D)	*Normalized IV*
TWI	< = 3.375663230 (0)	86,791	323	0.08	0.10	0.77	−0.38	0.01
	3.375663231 –5.585415369 (1)	181,784	838	0.20	0.21	0.95	−0.08	0.42
	5.585415370 –8.679068363 (2)	132,180	778	0.19	0.15	1.21	0.28	0.89
	8.679068364 –11.44125854 (3)	215,240	836	0.20	0.25	0.80	−0.32	0.09
	11.44125855 –13.65101068 (4)	186,190	1061	0.26	0.22	1.17	0.23	0.83
	Greater than 13.65101068 (5)	50,801	314	0.08	0.06	1.27	0.35	0.99
STI	< = 1.129179711 (0)	738,255	3197	0.77	0.87	0.89	−0.17	0.01
	1.129179712 –4.516718846 (1)	95,301	730	0.18	0.11	1.58	0.66	0.36
	4.516718847 –10.72720726 (2)	15,452	181	0.04	0.02	2.41	1.27	0.62
	10.72720727 –20.88982466 (3)	3216	29	0.01	0.00	1.89	0.91	0.47
	20.88982467 –40.65046961 (4)	692	12	0.00	0.00	3.71	1.89	0.89
	Greater than 40.65046961 (5)	70	1	0.00	0.00	4.37	2.13	0.99
SPI	< = 7.843282782 (0)	841,212	4051	0.98	0.99	0.99	−0.01	0.01
	7.843282783 –39.21641391 (1)	10,371	80	0.02	0.01	1.60	0.67	0.14
	39.21641392 –113.7276003 (2)	1247	14	0.00	0.00	2.39	1.26	0.24
	113.7276004 –282.3581801 (3)	136	5	0.00	0.00	8.28	3.05	0.57
	282.3581802 –615.6976984 (4)	18	0	0.00	0.00	5.56	2.47	0.46
	Greater than 615.6976984 (5)	2	0	0.00	0.00	41.11	5.36	0.99
Slope	< = 108.0233130 (0)	455,076	1690	0.41	0.53	0.76	−0.39	0.01

Table 3 (continued)

Factor	Class	Grids in domain *A*	No. of grids showing landslide occurrence *B*	% of grids showing landslide occurrence $C=(0.5+\text{Bi})/\sum\text{B}$	% of grids in domain $D=(0.5+\text{Ai})/\sum\text{A}$	Frequency ratio $(FR)=\text{C/D}$	$IV=\log_2$ (C/D)	*Normalized IV*
Length	108.0233131 –324.0699391 (1)	252,326	1356	0.33	0.30	1.10	0.14	0.28
	324.0699392 –669.7445408 (2)	100,315	694	0.17	0.12	1.42	0.51	0.46
	669.7445409 –1188.256443 (3)	32,400	277	0.07	0.04	1.76	0.82	0.62
	1188.25644 –2052.442948 (4)	10,250	96	0.02	0.01	1.93	0.95	0.69
	Greater than 2052.442948 (5)	2619	37	0.01	0.00	2.94	1.56	0.99
Land Cover	Agricultural Land (1)	121,385	6	0.00	0.14	0.01	−6.51	0.15
	Water Body (2)	32,950	5	0.00	0.04	0.03	−4.87	0.34
	Settlement (3)	21,884	0	0.00	0.03	0.00	−7.73	0.01
	Barren Land (4)	2235	0	0.00	0.00	0.05	−4.44	0.38
	High Vegetation (5)	133,103	94	0.02	0.16	0.15	−2.78	0.57
	Moderate Vegetation (6)	96,273	96	0.02	0.11	0.21	−2.28	0.63
	Low Vegetation (7)	445,156	3949	0.95	0.52	1.82	0.87	0.99
Lithology	Phyllite (1)	152,938	310	0.07	0.18	0.42	−1.26	0.71
	Sandstone with shale (2)	56,560	167	0.04	0.07	0.61	−0.72	0.77
	Sandstone (3)	2733	0	0.00	0.00	0.04	−4.73	0.32
	Unconsolidated sediments (4)	18,119	0	0.00	0.02	0.01	−7.46	0.01
	Shale with limestone (5)	3432	0	0.00	0.00	0.03	−5.06	0.28
	Granites (6)	257,120	2871	0.69	0.30	2.30	1.20	0.99
	Quartzite (7)	31,855	46	0.01	0.04	0.30	−1.74	0.66
	Quartzite alternating with shale (8)	216,188	410	0.10	0.25	0.39	−1.36	0.70
	Dolomite (9)	94,895	301	0.07	0.11	0.65	−0.61	0.78
	Limestone (10)	19,146	45	0.01	0.02	0.49	−1.03	0.74

The probabilities vary between 0 to 1 on a 'S' shape curve. In the above equation, Z represents the linear combination as follows:

$$Z = b_0 + b_1X_1 + b_2X_2 + \ldots\ldots + b_{14}X_{14} \quad (14)$$

LR fits the above equation to the training data with b0 as intercept and b1 to b14 as coefficient to the model specific to each causative factors Xi.

J48 Decision Tree

Decision Tree (DT) classifier is a class of inductive learning approach where a Generic model is created from a specific set of training examples (Hong et al., 2018; Zhang et al., 2017). A decision tree is a type of tree structure that resembles a flowchart, where each internal node represents a test on an attribute, each branch a test result, and each leaf node (terminal node) a class label. J48 algorithm divides the information to create minor subsets to base the decision (Bhargava et al., 2013; Tan et al., 2016). It works on the principle of greedy top-down approach and divide & conquer resulting the splitting factor of the decision tree on the basis of information gain (based on entropy/ gini index change) as an attribute. The splitting strategy stops when the leaf node belongs to the same class as the instances. Algorithm has an advantage of working on loss and missing values data also. Splitting criteria, max depth, min. samples leaves, min. samples split are the hyper-parameters which can be tuned for getting optimum performance measure.

Model Performance Evaluation Metric

Success and prediction rate processes are applied to evaluate the comparative performance of each model (Brenning, 2005; Fabbri et al., 2002). Success rate is evaluated by computing the overall accuracy over training data while the prediction rate is evaluated by computing the overall accuracy over the test dataset. Confusion matrix for both train and test datasets are created and overall accuracy is computed as a metric for model performance evaluation. Since, the classification problem attempted in this study is a true representative of highly imbalance class classification, the AUROC measure is found to be more optimistic in such a scenario (Brownlee, 2020). AUROC provide a single score to compare various model's performance and assess the prediction accuracy qualitatively by plotting the true

Table 4 Success rate, prediction rate and AUROC score for each algorithm implemented over Group 1, 2, 3 dataset

			Group 1	Group 2	Group 3
Frequency ratio (FR)	Overall accuracy	Test	0.67	0.71	0.61
		Train	0.67	0.71	0.61
	AUROC		0.73	0.74	0.73
Information value (IV)	Overall accuracy	Test	0.53	0.50	0.47
		Train	0.54	0.50	0.47
	AUROC		0.70	0.69	0.68
Decision tree (DT)	Overall Accuracy	Test	1.00	0.99	1.00
		Train	1.00	1.00	1.00
	AUROC		0.85	0.84	0.84
Naive bayes (NB)	Overall accuracy	Test	0.97	0.96	0.97
		Train	0.97	0.97	0.97
	AUROC		0.80	0.80	0.80
Logistic regression (LR)	Overall accuracy	Test	0.99	0.99	0.99
		Train	0.99	0.99	0.99
	AUROC		0.82	0.82	0.82

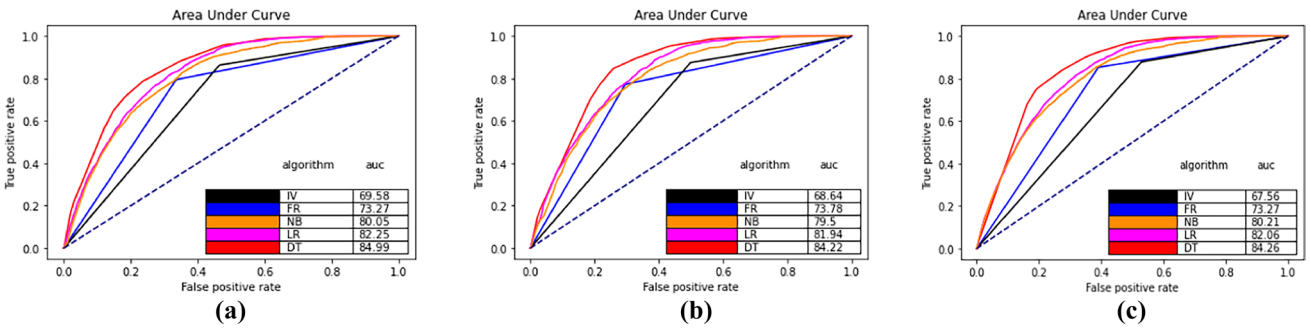


Fig. 9 **a** AUROC score for each algorithm for Group 1 dataset. **b** AUROC score for each algorithm for Group 2 dataset. **c** AUROC score for each algorithm for Group 3 dataset

positive rate at y axis and the false positive rate at x axis. A large number of researchers have used primarily the test data AUROC measure to compare the performances of various applied models (Chen et al., 2020; Hong et al., 2017; Park et al., 2013; Pham et al., 2016).

Results and Discussion

The LSM is produced by dividing the outcome into five parts using JENKS natural break. These parts are defined as very high, high, moderate, low and very low class for landslide susceptibility. The scikitlearn python package is used for implementing NB, LR, and DT algorithms over the generated datasets. Other python packages like osgeo is used to read and write the geotiff files, jenkspy is used for finding natural cluster and hence dividing the entire dataset predicted output based on statistical/ML models, into various landslide susceptibility class.

Table 4 shows the success rate, prediction rate in terms of overall accuracy over train, test data respectively and test AUROC score for each of the implemented algorithms over generated datasets. As can be seen from Table 4 and Fig. 9a, b, c; classical statistical method like FR and IV method do not exhibit good overall accuracy and AUROC score though FR outperform IV method consistently on all three datasets. Even, Naïve Bayes has shown better result in comparison to FR and IV. LR and DT have shown consistently better performance over other methods in all the three groups dataset due to their ability to classify the highly non-linear decision boundaries. FR has shown better results for the group 2 dataset i.e. equal interval while all other algorithms have shown better results on the group 1 dataset i.e. natural break. LR has shown the same result on all three groups dataset. DT has outperformed all other algorithms for all three groups dataset.

Hyper-parameters tuning gives an optimized bias-variance trade-off at the minimal model complexities (Belkin et al., 2019). Hyper-parameters tuning is not applicable to FR, IV & NB as they are simple statistical methods; however for logistic regression, hyper-parameters like penalty (with values as l2 & none) and solver (with values as newton-cg, lbfgs, liblinear, sag, saga) were evaluated using gridsearchcv (akin to brute force search) in python and best parameter found out are l2 and bilinear for penalty and solver respectively. Values for all coefficients i.e. b0 to b14 are found to be as −14.24821171, −0.04530787, 1.1252632, 0.01231773, 0.00241996, 0.07298033, 0.15599982, −0.21188946, −0.03201266, 0.2271251, 0.25997982, −0.00282946, −0.06901128, 0.09357177, and 0.10055757 respectively. Similarly, for decision tree the tuned hyper-parameters like splitting criterion, max depth, min. samples leaves, min. samples split have got the best value as 'entropy', '1', '2', and '5' respectively when optimising AUROC score using grid search.

Result demonstrates that Hyper-parameters tuning plays a very important role in the DT algorithm. If the DT model is implemented with default parameters then the AUROC score for test data is 0.56 while if implemented with tuned Hyper-parameters then it rises to 0.85. Hence, Hyper-parameters based optimum DT model is highly desirable. The results are summarised in Table 5 for group 3 data as a representative result.

DT model has given the best performance measure score among all algorithms (as per Table 4) overall groups dataset and marginally better score over group 1 dataset, hence LSM is generated for DT which is shown in Fig. 10. A total of 1272, 3247, and 1049 landslide points are lying in very high, high, and medium susceptibility demarcated area while only 259 and 101 points are lying in low and very low susceptibility demarcated area.

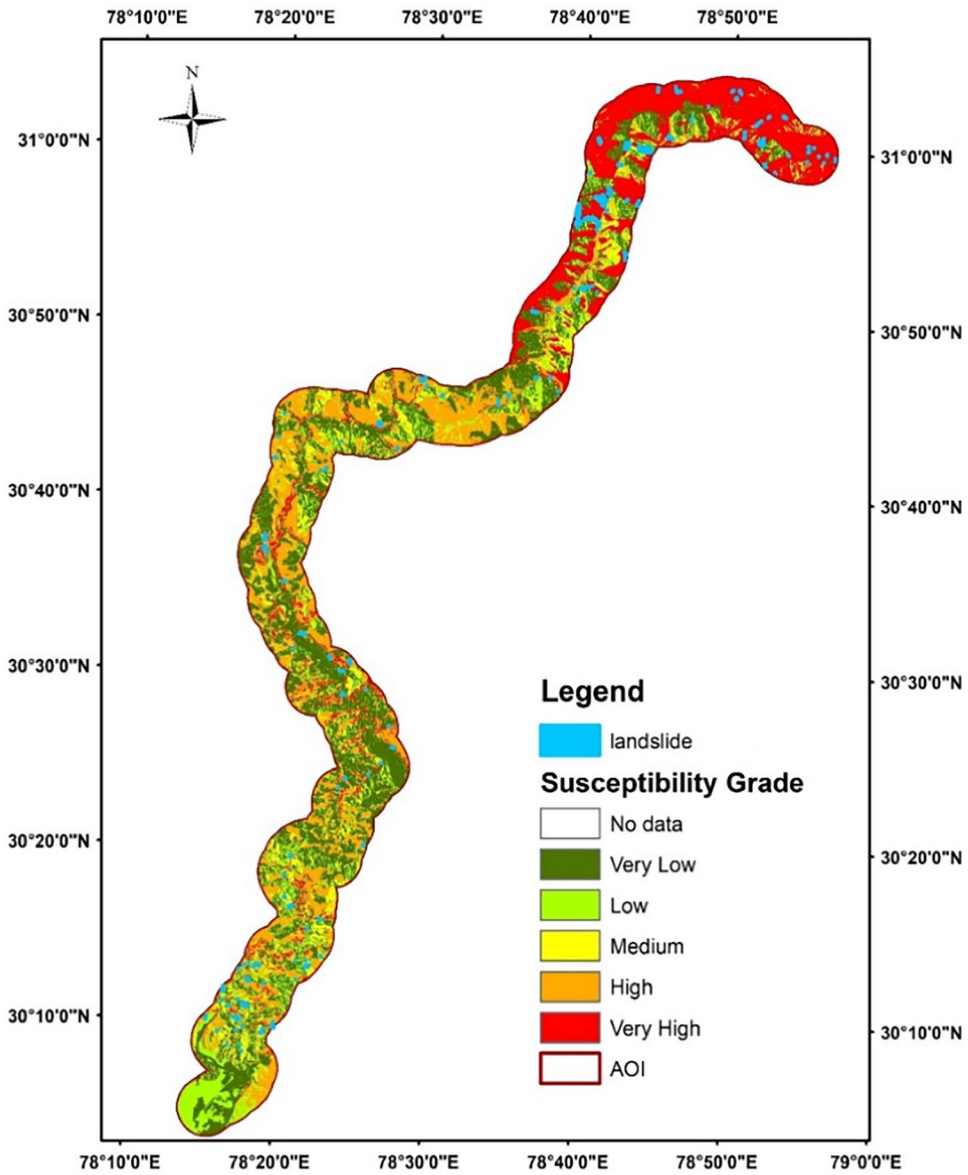
Feature importance analysis (based on information gain ratio) results are shown in Fig. 11 a, b, c for group 1,2,3 respectively. The results indicate that the distance from the river and fault line are the dominant factors in all three datasets. Although all the selected causative factors contribute more or less in LSM predictability; but in order to select the most relevant combination of causal factors, one can eliminate the factors which are less important as they may cause noise, increase computation, and reduce the accuracy of the model (Chen et al., 2020).

Further, another experiment is executed by removing the bottom four causative factors from the datasets which is based on feature importance and these factors are found to be SPI, STI, aspect, and slope length. All five algorithm models were evaluated on these three trim down datasets. It is observed that even PCC applied in Sect. "Multicollinearity test and Correlation Coefficient Test" also suggested that STI and SPI are correlated and one of the variable may be dropped at that stage itself. Figure 12 a,b,c shows the AUROC score for each models over this trim down datasets

Table 5 Results of decision tree algorithm over group 3 data using default parameters and tuned parameters

Decision tree	Train accuracy	Test accuracy	Train AUROC	Test AUROC
With default parameters	1	0.99	1	0.56
With tuned Hyper-parameters	1	1	0.86	0.85

Fig. 10 LSM for Study Area



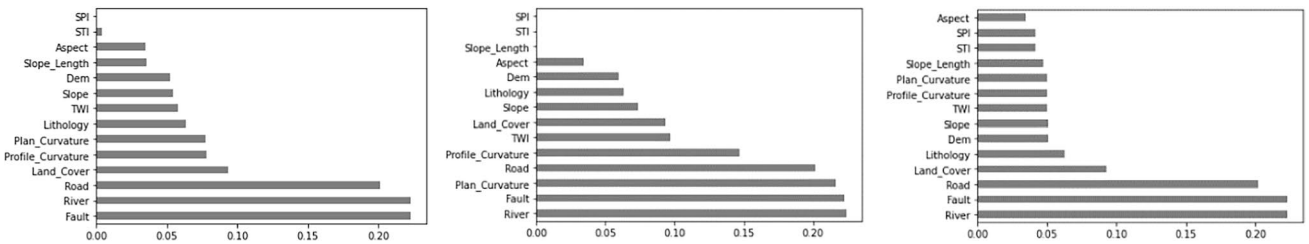


Fig. 11 **a** Feature importance analysis over Group 1 dataset **b** Feature importance analysis over Group 2 dataset **c** Feature importance analysis over Group 3 dataset

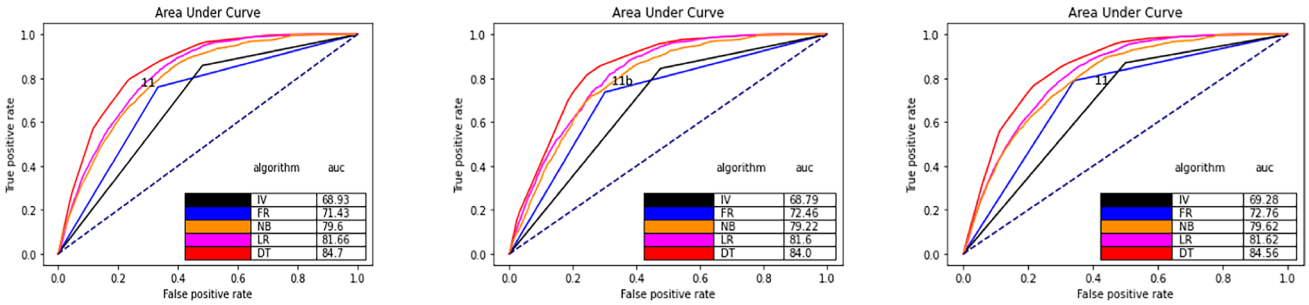


Fig. 12 **a** AUROC score for each algorithm for Group 1 trim down dataset. **b** AUROC score for each algorithm for Group 2 trim down dataset. **c** AUROC score for each algorithm for Group 3 trim down dataset

of Group 1, 2, 3. After comparing the Fig. 9 and 12, it is evident that DT, LR and NB retained the at par performances (AUROC score didn't dropped even 1%) while the model becomes less complex after removing the least important causative factors.

Conclusion

The present study has primarily considered the remote sensing based data for the generation of LSM. This study evaluates the comparative performance of five statistical/ML algorithms for LSM which were implemented over three groups of generated dataset using different feature engineering techniques available in the literature. Further, the importance of Hyper-parameters has been brought out with results obtained for the DT algorithm. It can be summarized that Hyper-parameters tuning must be an essential part for any ML algorithm for optimum model complexity while having at par in-sample and out-sample error. Landslide susceptibility class boundaries are highly non-linear due to the complex inter-play of all causative factors. The applied ML models like DT and LR have outperformed the bi-variate models like FR and IV because of their inherent characteristics in classifying the highly non-linear decision boundaries.

Wang et al. (2016a, b) have investigated the LSM using evidential belief function, and weight of evidence methods and achieved the AUROC score of 0.8 and 0.79 for the validation set, respectively. Park et al. (2013) have reported an overall accuracy of 65.27,64.35,65.51, and 68.47% for frequency ratio, AHP, logistic regression, and ANN model respectively. Hong et al. (2018) have concluded the AUROC score of 0.855, 0.85, 0.839 and 0.814 for rotation forest, AdaBoost, bagging, and J48 decision tree machine learning algorithms applied on the dataset of Gaungchang area in China. Pourghasemi and Rahmati (2017) have evaluated ten machine learning models for the Ghaemshahr region of Mazandaran province, Iran and achieved the best AUROC values of 0.84 and 0.81 for random forest and boosted regression tree, respectively. Akgun (2012) has evaluated LSM using FR and logistic regression method and achieved the AUROC score as 0.81 and 71% respectively. Wang et al. (2016a, b) have compared LR, FR, DT and other algorithms and found the AUROC score for DT as 0.70 to 0.74 with different sampling strategy while hyper-parameters tuning is not been attempted.

The AUROC score achieved in our finding is at par or better in comparison to the same model results applied over other different geographic areas by other researchers mentioned in the above paragraph's referred research findings. Even, results shows that one can drop the lowest contributing 3 to 4 causative factors without hampering the model's performance measure criteria.

In the future, other ML algorithms like ensemble methods of DT, SVM, ANN, and other advanced techniques may be evaluated over JENKS natural break based datasets rather than simply going for equal interval grouping or based on expert judgment which has been the adopted approach in most of the literature. Hyperparameters tuning plays a very important role in certain ML algorithms like DT, which have still not been investigated in detail for LSM due to lack of knowledge or unavailability of this feature in various packages like WEKA, R, SPSS, ARCGIS model builder etc. It will be meaningful research in the future for other geographic area where one should attempt to tune hyperparameters in algorithms like DT and its ensemble for a better score of performance measure while keeping the model complexities at the minimum. The generated LSM can be used as an input for generalised planning and assessment purposes, but the same LSM may not be so useful for site-specific scale assessment of landslide.

Acknowledgements Authors are thankful to the Dr. LK Sinha, Director, Defence Geo-informatics Research Establishment, Defence Research & Development Organization; Chandigarh, India, for allowing them to carry out this research work while the feedback/suggestions received from the anonymous and internal reviewer has significantly improved the quality of the research.

Author's Contribution Corresponding author (Sh. Vivek Saxena) does all the experimental work and manuscript creation. Dr. Upasna Singh in capacity of supervisor, has reviewed the paper and provided suggestion while executing this research as well as the manuscript. Dr. LK Sinha has endorsed this research to be attempted in the research lab in the capacity of Director DGRE and provided all the resources while in capacity of external supervisor, he has provided suggestion to improve the work.

Funding This study received no specific funding from public, commercial, or not-for-profit funding entities however the research is carried out to pursue doctorate degree research proposal which is cleared by Doctorate Research Management Committee (DRMC), Defence Institute of Advance Technology (DIAT) a Deemed to be University, Girinagar, Pune, India and endorsed by Director, Defence Geo-informatics Research Establishment (DGRE), Defence Research and Development Organization, Ministry of Defence, India for advancement in technology and capacity building in the research lab.

Declarations

Conflict of Interest No potential conflict of interest was reported by any of the authors for this research.

References

Akgun, A. (2012). A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio methods: a case study at İzmir, Turkey. *Landslides*, *9*(1), 93–106. https://doi.org/10.1007/s10346-011-0283-7

Anbalagan, R. (1992). Landslide hazard evaluation and zonation mapping in mountainous terrain. *Engineering Geology*. https://doi.org/10.1016/0013-7952(92)90053-2

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review, 40*, 100378. https://doi.org/10.1016/j.cosrev.2021.100378

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences, 116*(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116

Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on J48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(6).

Brabb, E. E. (1991). The world landslide problem. *Episodes Journal of International Geoscience, 14*(1), 52–61. https://doi.org/10.18814/epiiugs/1991/v14i1/008

Brenning, A. (2005). Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Natural Hazards and Earth System Sciences, 5*(6), 853–862. https://doi.org/10.5194/nhess-5-853-2005

Brownlee, J. (2020). *Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning*. Machine Learning Mastery.

Carrara, A., Cardinali, M., Detti, R., Guzzetti, F., Pasqui, V., & Reichenbach, P. (1991). GIS techniques and statistical models in evaluating landslide hazard. *Earth Surface Processes and Landforms, 16*(5), 427–445. https://doi.org/10.1002/esp.3290160505

Carrara, A., & Pike, R. J. (2008). GIS technology and models for assessing landslide hazard and risk. *Geomorphology, 3*(94), 257–260. https://doi.org/10.1016/j.geomorph.2006.07.042

Chen, T., Zhu, L., Niu, R. Q., Trinder, C. J., Peng, L., & Lei, T. (2020). Mapping landslide susceptibility at the Three Gorges Reservoir, China, using gradient boosting decision tree, random forest and information value models. *Journal of Mountain Science, 17*(3), 670–685. https://doi.org/10.1007/s11629-019-5839-3

Chen, W., Chen, X., Peng, J., Panahi, M., & Lee, S. (2021). 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., Li, W., Chai, H., Hou, E., Li, X., & Ding, X. (2016). *GIS-based landslide susceptibility mapping using analytical hierarchy process (AHP) and certainty factor (CF) models for the Baozhong region of Baoji City*. China: Environmental Earth Sciences. https://doi.org/10.1007/s12665-015-4795-7

Chen, W., Pourghasemi, H. R., Kornejady, A., & 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

Fabbri, A. G., Chung, C. F., Napolitano, P., Remondo, J., & Zêzere, J. L. (2002). Prediction rate functions of landslide susceptibility applied in the Iberian Peninsula. *WIT Transactions on Modelling and Simulation*, *31*.

Fell, R., Corominas, J., Bonnard, C., Cascini, L., Leroi, E., Savage, W. Z., et al. (2008). Guidelines for landslide susceptibility, hazard and risk zoning for land-use planning. *Engineering Geology, 102*(3–4), 99–111. https://doi.org/10.1016/j.enggeo.2008.03.014

Ghosh, S., Van Westen, C. J., Carranza, E. J. M., Ghoshal, T. B., Sarkar, N. K., & Surendranath, M. (2009). A quantitative approach for improving the BIS (Indian) method of medium-scale landslide susceptibility. *Journal of the Geological Society of India*. https://doi.org/10.1007/s12594-009-0167-9

Guzzetti, F., Carrara, A., Cardinali, M., & Reichenbach, P. (1999). Landslide hazard evaluation: A review of current techniques and their application in a multi-scale study. *Central Italy. Geomorphology, 31*(1–4), 181–216. https://doi.org/10.1016/S0169-555X(99)00078-1

Hidayat, S., Pachri, H., & Alimuddin, I. (2019). Analysis of Landslide Susceptibility Zone using Frequency Ratio and Logistic Regression Method in Hambalang, Citeureup District, Bogor Regency, West Java Province. *IOP Conference Series: Earth and Environmental Science*. https://doi.org/10.1088/1755-1315/280/1/012005

Hong, H., Liu, J., Bui, D. T., Pradhan, B., Acharya, T. D., Pham, B. T., & 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

Hong, H., Pradhan, B., Bui, D. T., Xu, C., Youssef, A. M., & Chen, W. (2017). Comparison of four kernel functions used in support vector machines for landslide susceptibility mapping: A case study at Suichuan area (China). *Geomatics, Natural Hazards and Risk, 8*(2), 544–569. https://doi.org/10.1080/19475705.2016.1250112

Hong, H., Shahabi, H., Shirzadi, A., Chen, W., Chapi, K., Ahmad, B. B., & Tien Bui, D. (2019). Landslide susceptibility assessment at the Wuning area, China: a comparison between multi-criteria decision making, bivariate statistical and machine learning

methods. *Natural Hazards, 96*, 173–212. https://doi.org/10.1007/s11069-018-3536-0

Huang, M., Yu, W., & Zhu, D. (2012). An improved image segmentation algorithm based on the Otsu method. *2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, 135–139. https://doi.org/10.1109/snpd.2012.26

Hussain, M. A., Chen, Z., Kalsoom, I., Asghar, A., & Shoaib, M. (2022). Landslide susceptibility mapping using machine learning algorithm: A case study along Karakoram Highway (KKH), Pakistan. *Journal of the Indian Society of Remote Sensing, 50*(5), 849–866. https://doi.org/10.1007/s12524-021-01451-1

Javad, M., Baharin, A., Barat, M., & Farshid, S. (2014). Using frequency ratio method for spatial landslide prediction. *Research. Journal of Applied Sciences Engineering and Technology*. https://doi.org/10.19026/rjaset.7.658

Kalantar, B., Pradhan, B., Naghibi, S. A., Motevalli, A., & Mansor, S. (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, Natural Hazards and Risk, 9*(1), 49–69. https://doi.org/10.1080/19475705.2017.1407368

Kavzoglu, T., Colkesen, I., & Sahin, E. K. (2019). Machine learning techniques in landslide susceptibility mapping: a survey and a case study. In *Landslides: Theory, Practice and Modelling* (pp. 283–301). Springer. https://doi.org/10.1007/978-3-319-77377-3_13

Kavzoglu, T., Sahin, E. K., & Colkesen, I. (2014). Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides, 11*(3), 425–439. https://doi.org/10.1007/s10346-013-0391-7

Kumar, R., & Anbalagan, R. (2016). Landslide susceptibility mapping using analytical hierarchy process (AHP) in Tehri reservoir rim region Uttarakhand. *Journal of the Geological Society of India*. https://doi.org/10.1007/s12594-016-0395-8

Lacasse, S., & Nadim, F. (2009). Landslide risk assessment and mitigation strategy. *Landslides - Disaster Risk Reduction*. https://doi.org/10.1007/978-3-540-69970-5_3

Lee, J.-H., Sameen, M. I., Pradhan, B., & Park, H.-J. (2018). Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. *Geomorphology, 303*, 284–298. https://doi.org/10.1016/j.geomorph.2017.12.007

Marjanovic, M., & Caha, J. (2011). Fuzzy Approach to Landslide Susceptibility Zonation. *DATESO*, 181–195.

Marjanović, M., Samardžić-Petrović, M., Abolmasov, B., & \DJurić, U. (2019). Concepts for Improving Machine Learning Based Landslide Assessment. In *Natural Hazards GIS-based Spatial Modeling Using Data Mining Techniques* (pp. 27–58). Springer. https://doi.org/10.1007/978-3-319-73383-8_2

Marjanović, M., Kovačević, M., Bajat, B., & Vožen\'\ilek, V. (2011). Landslide susceptibility assessment using SVM machine learning algorithm. *Engineering Geology, 123*(3), 225–234. https://doi.org/10.1016/j.enggeo.2011.09.006

Ngo, P. T. T., Panahi, M., Khosravi, K., Ghorbanzadeh, O., Kariminejad, N., Cerda, A., & Lee, S. (2021). Evaluation of deep learning algorithms for national scale landslide susceptibility mapping of Iran. *Geoscience Frontiers, 12*(2), 505–519. https://doi.org/10.1016/j.gsf.2020.06.013

Park, S., Choi, C., Kim, B., & Kim, J. (2013). Landslide susceptibility mapping using frequency ratio, analytic hierarchy process, logistic regression, and artificial neural network methods at the Inje area. *Korea. Environmental Earth Sciences, 68*(5), 1443–1464. https://doi.org/10.1007/s12665-012-1842-5

Pham, B. T., Pradhan, B., Tien Bui, D., Prakash, I., & Dholakia, M. B. (2016). A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). *Environmental Modelling and Software*. https://doi.org/10.1016/j.envsoft.2016.07.005

Pham, B. T., Shirzadi, A., Tien Bui, D., Prakash, I., & Dholakia, M. B. (2018). A hybrid machine learning ensemble approach based on a radial basis function neural network and Rotation Forest for landslide susceptibility modeling: A case study in the Himalayan area India. *International Journal of Sediment Research*. https://doi.org/10.1016/j.ijsrc.2017.09.008

Pham, B. T., Tien Bui, D., Indra, P., & Dholakia, M. (2015). Landslide susceptibility assessment at a part of Uttarakhand Himalaya, India using GIS–based statistical approach of frequency ratio method. *Int J Eng Res Technol, 4*(11), 338–344.

Pham, B. T., Tien Bui, D., Pourghasemi, H. R., Indra, P., & Dholakia, M. B. (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. *Theoretical and Applied Climatology*. https://doi.org/10.1007/s00704-015-1702-9

Pourghasemi, H. R., & Rahmati, O. (2017). Prediction of the landslide susceptibility: Which algorithm, which precision? *CATENA*. https://doi.org/10.1016/j.catena.2017.11.022

Pradhan, A. M. S., & Kim, Y.-T. (2014). Relative effect method of landslide susceptibility zonation in weathered granite soil: A case study in Deokjeok-ri Creek, South Korea. *Natural Hazards, 72*, 1189–1217. https://doi.org/10.1007/s11069-014-1065-z

Regmi, N. R., Giardino, J. R., & Vitek, J. D. (2010). *Modeling susceptibility to landslides using the weight of evidence approach: Western Colorado*. Geomorphology. https://doi.org/10.1016/j.geomorph.2009.10.002

Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M., & Guzzetti, F. (2018). A review of statistically-based landslide susceptibility models. *In Earth-Science Reviews*. https://doi.org/10.1016/j.earscirev.2018.03.001

Saaty, T. L. (2002). Decision making with the Analytic Hierarchy Process. *Scientia Iranica*. https://doi.org/10.1504/ijssci.2008.017590

Saravanan, N., & Gayathri, V. (2017). Classification of dengue dataset using j48 algorithm and ant colony based aj48 algorithm. *Proceedings Of The International Conference On Inventive Computing And Informatics (ICICI 2017)*, 1062–1067. https://doi.org/10.1109/icici.2017.8365302

Sarkar, S., & Kanungo, D. P. (2004). An integrated approach for landslide susceptibility mapping using remote sensing and GIS. *Photogrammetric Engineering & Remote Sensing, 70*(5), 617–625. https://doi.org/10.14358/pers.70.5.617

Sun, D., Xu, J., Wen, H., & Wang, Y. (2020). An optimized random forest model and its generalization ability in landslide susceptibility mapping: Application in two areas of Three Gorges Reservoir, China. *Journal of Earth Science, 31*, 1068–1086. https://doi.org/10.1007/s12583-020-1072-9

Tan, P.-N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

Wang, L.-J., Guo, M., Sawada, K., Lin, J., & Zhang, J. (2016a). A comparative study of landslide susceptibility maps using logistic regression, frequency ratio, decision tree, weights of evidence and artificial neural network. *Geosciences Journal, 20*, 117–136. https://doi.org/10.1007/s12303-015-0026-1

Wang, Q., Li, W., Wu, Y., Pei, Y., Xing, M., & Yang, D. (2016b). A comparative study on the landslide susceptibility mapping using evidential belief function and weights of evidence models.

Journal of Earth System Science, 125(3), 645–662. https://doi.org/10.1007/s12040-016-0686-x

Wang, Y., Fang, Z., & 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

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67–82. https://doi.org/10.3156/jfuzzy.9.5_696

Wubalem, A. (2021). Landslide susceptibility mapping using statistical methods in Uatzau catchment area, northwestern Ethiopia. *Geoenvironmental Disasters*. https://doi.org/10.1186/s40677-020-00170-y

Yalcin, A., Reis, S., Aydinoglu, A. C., & Yomralioglu, T. (2011). A GIS-based comparative study of frequency ratio, analytical hierarchy process, bivariate statistics and logistics regression methods for landslide susceptibility mapping in Trabzon. *NE Turkey Catena*. https://doi.org/10.1016/j.catena.2011.01.014

Zhang, K., Wu, X., Niu, R., Yang, K., & Zhao, L. (2017). The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area China. *Environmental Earth Sciences*. https://doi.org/10.1007/s12665-017-6731-5

Zhou, C., Yin, K., Cao, Y., Ahmed, B., Li, Y., Catani, F., & Pourghasemi, H. R. (2018). Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China. *Computers & Geosciences, 112*, 23–37. https://doi.org/10.1016/j.cageo.2017.11.019

Zhou, X., Wen, H., Zhang, Y., Xu, J., & Zhang, W. (2021). Landslide susceptibility mapping using hybrid random forest with GeoDetector and RFE for factor optimization. *Geoscience Frontiers, 12*(5), 101211. https://doi.org/10.1016/j.gsf.2021.101211