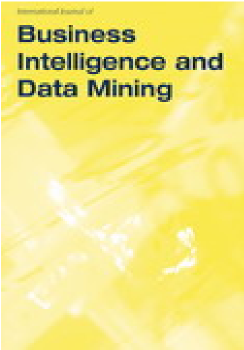


International Journal of Business Intelligence and Data Mining

ISSN online: 1743-8195 - ISSN print: 1743-8187
https://www.inderscience.com/ijbidm

Landslide susceptibility assessment along the major transport corridor using decision tree model: a case study of Kullu-Rohtang Pass

Nirbhav, Anand Malik, Maheshwar, Mukesh Prasad

DOI: 10.1504/IJBIDM.2024.10054983

Article History:

Received:	31 August 2021
Last revised:	08 March 2022
Accepted:	14 April 2022
Published online:	01 December 2023

Landslide susceptibility assessment along the major transport corridor using decision tree model: a case study of Kullu-Rohtang Pass

Nirbhav*

Department of Geography,
Delhi School of Economics,
University of Delhi,
Delhi, India
Email: captainnirbhav@gmail.com
*Corresponding author

Anand Malik

Swami Shraddhanand College,
University of Delhi,
Delhi, India
Email: anandmalik1111@gmail.com

Maheshwar

TGT Computer Science,
Directorate of Education,
Delhi Government,
Delhi, India
Email: maheshwar1524@gmail.com

Mukesh Prasad

Faculty of Engineering and Information Technology,
School of Computer Science,
Australian Artificial Intelligence Institute,
University of Technology Sydney,
Sydney, Australia
Email: mukesh.prasad@uts.edu.au

Abstract: To lessen damages from landslides, the key challenge is to predict the events precisely and accurately. The objective of this study is to assess landslide susceptibility in the study area. To achieve this objective, a detailed landslide inventory has been prepared based on imagery data and frequent field visits of 153 rock slides and 44 debris slides. Nine landslide factors were prepared initially and their relationships with each other and with the type of landslide was analysed. Information gain ratio measure is used to eliminate triggering factors with least score. Train_test_split method was used to classify the dataset into training and testing groups. Decision tree classification model of machine learning was applied for landslide susceptibility model (LSM). The performance was evaluated using classification report and receiver operating characteristic (ROC) curve. Results obtained have proven that the decision tree classification model performed well with good accuracy in forecasting landslide susceptibility.

Keywords: landslide susceptibility modelling; LSM; machine learning; decision tree classification.

Reference to this paper should be made as follows: Nirbhav, Malik, A., Maheshwar and Prasad, M. (2024) 'Landslide susceptibility assessment along the major transport corridor using decision tree model: a case study of Kullu-Rohtang Pass', *Int. J. Business Intelligence and Data Mining*, Vol. 24, No. 1, pp.1–24.

Biographical notes: Nirbhav has received his PhD degree from Department of Geography, Delhi School of Economics, University of Delhi (2020). His research interests include landslide susceptibility modelling, disaster risk reduction (DRR), climate change, remote sensing and GIS applications, machine learning, spatio-temporal database, data mining, image processing and analysis. He is currently teaching in Swami Shraddhanand College, University of Delhi. He has authored and co-authored several research papers in national and international journals. He has attended several conferences, workshops and research projects both at national and international level.

Anand Malik is an Associate Professor of Geography at SSNC (University of Delhi), since 1998. He received his Master's in Geoinformatics, ITC (University of Twente), The Netherlands, Certificate Course in Geosciences from Indian Institute of Remote Sensing (National Remote Sensing Centre), India; Certificate Course in Risk Management from University of Geneva (CERG 2013), Geneva, Switzerland. He has authored seven books and presented several papers in national and international seminars. His research interests include geo-spatial data analysis and process based dynamic modelling (slope hazards): run out modelling of snow avalanche, debris flows and landslides.

Maheshwar has received Master's in Information Technology from Delhi Technological University, Delhi in 2013. He has published several research papers in different reputed journals. He has presented his research work in different national and international conferences. His research interests include data mining, machine learning, artificial intelligence, pattern recognition and text analysis. Currently, he is working in the capacity of a teacher in Directorate of Education, Delhi.

Mukesh Prasad is a Senior Lecturer in the School of Computer Science (SoCS), Faculty of Engineering and Information Technology (FEIT), University of Technology Sydney (UTS), Australia. His research expertise lies in the development of new methods in artificial intelligence, machine learning and data analytics approached within the domain of computer vision, healthcare, biomedical, internet of things and brain computer interface and marketing research.

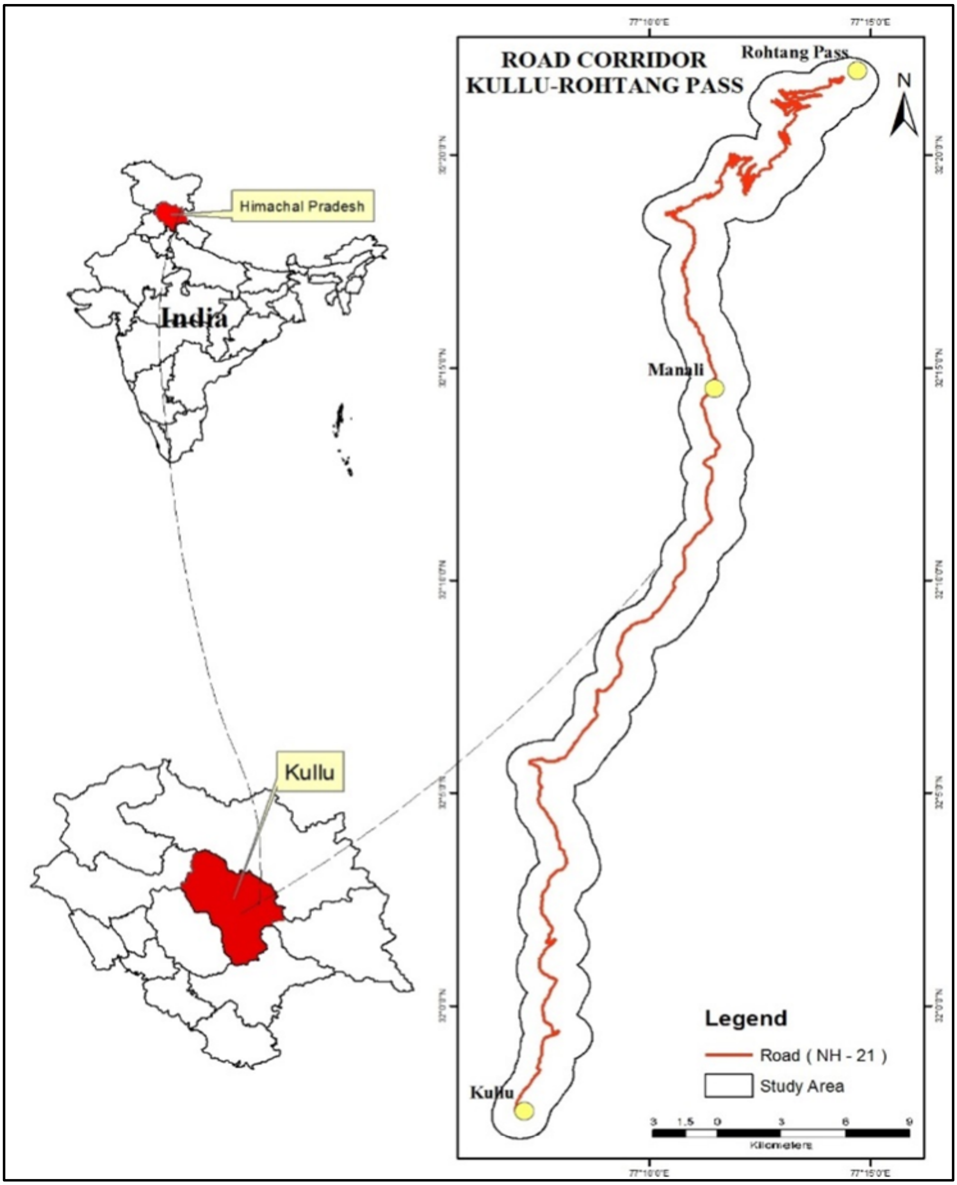
1 Introduction

Landslide is a hazard in form of mass movement mainly occurred in hilly or mountainous terrain causing an extensive loss to the economy and human settlements (Loi et al., 2017). The landslide events are quite frequent during monsoon season (July to September) in India induced by episodic rainfall (Bhambri et al., 2016; Ambrosi et al., 2018). The Kullu-Rohtang Pass transport corridor is highly susceptible to landslides and the construction activities along the road corridor have put an additional threat inducing these events. Rapid urbanisation has increased the demand for more land and thus increasing the frequency of landslides incidents (Singh and Pandey, 1996). LSM is very useful in order to assess landslide hazard and risk, also it is considered very helpful for land use planning and environmental impact assessment. This method has emerged very convenient and beneficial for policy makers and engineers for making and executing appropriate policies and strategies to lessen landslides risk (Sassa, 2017). Landslides can be classified into many categories based on various deformation patterns (Hunger et al., 2014). The conditions influencing the occurrence of any landslide event are different for each landslide. For example, the rock slide occurs majorly on steep mountains while debris slide takes place on gentle slope. Therefore, it becomes crucial to perform the landslide susceptibility assessment by considering the distinctions between these types of landslides. Landslide susceptibility model (LSM) can be grouped under two major categories: qualitative and quantitative assessment. Qualitative assessment comprises of methods which are primarily based on inventories, comprehension and knowledge, while quantitative assessment includes physically-based techniques and data induced models (Hussin et al., 2016). The data-based models include decision tree, Naïve Bayes, k-nearest neighbour (KNN) and so forth. In data-based model, the machine learning models are observed more efficient and have performed better than traditional models such as analytic, experience and opinion-based models.

Machine learning models have been practiced in many research areas like data mining (Maheshwar et al., 2015), pattern recognition (Narayanan et al., 2016), medical diagnosis (Goyal and Maheshwar, 2019; Maheshwar and Kumar, 2019) and artificial intelligence (Ghahramani, 2015) and have shown better results. Different machine learning models have been used for landslide identification and susceptibility modelling (Wang et al., 2021; Arabameri et al., 2021). Before executing LSM, it is better to understand the landslide mechanism and different triggering factors causing the landslides. Feature selection methods can be used to analyse the correlation among the triggering factors and occurrence of landslide events. These methods provide powerful techniques to select the major triggering factors for LSM.

The primary purpose of this study is to recognise the different types of landslides and their triggering factors along the transport corridor from Kullu to Rohtang Pass. This transport corridor suffers from massive landslides during monsoon season due to the presence of unstable slope in the area. So, an accurate LSM is required which can be helpful in taking suitable and appropriate landslide risk mitigation measures. This study focuses on developing a machine learning-based LSM model with better spatial agreement and good accuracy for the research area. In this study, information gain is used as attribute selection measure and then the decision tree machine learning model is applied on the reduced dataset. The results showed that the decision tree model achieve a quite satisfactory predictive accuracy. Anaconda tool with Python programming language has been used to carry out the programming and experimental related work. The reason for using Python programming language is its simplicity and enriched libraries for carrying out machine learning related research.

Figure 1 Demarcation of the study area (see online version for colours)



2 Study area

2.1 General features

In the present study, the transport corridor (NH-21) from Kullu-Rohtang Pass having a length of 90 km has been selected for LSM. Geographically, this area lies between the latitudes 32°0′0″N to 32°20′0″N and longitudes 77°5′0″E to 77°15′0″E (Figure 1). In this study, a buffer of one kilometre on both sides of the road has been taken for LSM.

The selected area comes under lesser Himalaya mountain ranges having an uneven geomorphology with average to high elevation ranging from 1,279 m to 3,979 m from mean sea level (MSL). The average rainfall in this area is about 1,363 mm. Maximum number of landslides from Kullu to Rohtang Pass are mainly recorded in the months of July to September as the rainfall is maximum in these months. The maximum and minimum temperature varies from 25° Celsius to 4° Celsius.

There are different varieties of soil found in the study area such as brown hill soil, red loamy soil and mountain meadow soil. Agricultural practices are done predominantly in mountain cut terraces and river terraces. Thrusts like Vaikrita, Jutogh and Kullu are also found in the study area. These thrusts are dynamic in nature and play a significant role in neo-tectonics of this area. The location of the study area is along the banks of the river Beas which is the main source of drainage. The drainage pattern of the river Beas in the study area reflects primary stage of dendritic pattern with visible sign of parallel dendritic and trellis patterns in between. The urbanisation, mining, road construction and deforestation are major activities which increase the vulnerability of landslide occurrences (Saha et al., 2005). The occurrence of landslide events along the transport corridor affects the transportation and sometimes even completely cut off the supply lines affecting the economic activities adversely.

2.2 Landslide type

The type of a landslide depends on various environmental, local and regional terrain conditions. In the study area, two landslide types have been identified:

1 Rock slide: Rock slide (Hunger et al., 2014) is a major type of landslide that mostly occurs in multistage patterns [Figure 2(a)]. Most of the rock slides occurrences are due to the gravitational pressure and erosional influences. On steep mountainous ridges due to the development of large structural joints, the occurrence of large scale rock slide is quite often. Constructional activities in gentle slope terrain are the main reason behind happenings of these events as the slope may lose the equilibrium state under the influence of artificial cutting.

2 Debris slide: Debris slide (Iverson, 2015) is the downward movement of the combination of rocks material, organic matters, loose soil, water in the form of slurry with size of sand particles of at least 50% of flowing material, flowing down the slope [Figure 2(b)]. Debris slide mainly occurs because of intense water flow, speedy snowmelt which is eroding and mobilising loose rocks and soil particles.

Figure 2 Landslide types, (a) rock slide (b) debris slide (see online version for colours)

(a) (b)

3 Methodology

3.1 Landslide triggering factors analysis

In this study, the methods used for landslide triggering factors analysis include information gain ratio (IGR) and decision tree classifier.

3.1.1 Information gain ratio

IGR is well known and widely used attribute method of selection (Quinlan, 1996; Tien Bui et al., 2016). Attributes having higher value of the IGR have higher ability of prediction for the model. Assuming, the training data D comprises of n number of samples. The information required to categories a sample in D is obtained by

$$Info(D) = -\sum_{i=1}^{m} \log_2 \left(p_i \right) \tag{1}$$

Here m is used to denote the number of various classes and p_i is the probability that a sample in D associates to class c_i and is computed by $|C_{i,D}|/|D|$. For each triggering factor A, which divides the training data D into v partitions (D_1, D_2, …, D_v) the information gain is calculated by using

$$Info_A(D) = \sum_{j=1}^{v} \frac{\left|D_j\right|}{|D|} \times Info\left(D_j\right) \tag{2}$$

The IGR for factor A is calculated as:

$$GainRatio(A) = \frac{Info(D) - Info_A(D)}{SplitInfo(A)} \tag{3}$$

Here *SplitInfoA* is calculated by using the following formula:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \tag{4}$$

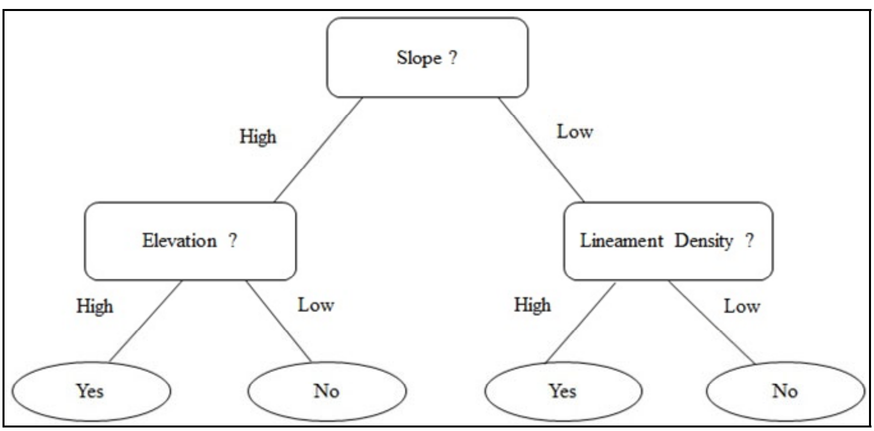
3.2 Decision tree classification

Classification (Bertsimas and Dunn, 2017; Dinov, 2018) is a supervised learning technique of machine learning. Decision tree learning is one among many classification models which uses the decision tree as a predictive model that maps a given testing sample to one of the predefined class of the data. In decision tree classification model, all internal nodes indicate a test on an attribute while the leaf nodes show the class label (Figure 3). The branch shows the results of the test conducted on the attribute at internal node.

At each level of the decision tree, the triggering factors are decided by the splitting principle. The splitting principle uses the attribute selection method to find the best triggering factor to be used as split point. A triggering factor with maximum IGR value is used as the split-point.

For this study, initially nine triggering factors are considered. Two factors with least IGR values are eliminated. So rest seven factors with higher IGR values are used as split points. The splitting principal is applied for deciding the triggering factors at each level.

Figure 3 Decision tree for landslide prediction



So, if a sample, *X*, is given for which the landslide type class is unknown, the triggering factors values of the sample are tested for the decision tree. An approach from the root node to the leaf node is traced. The leaf node tells the class of landslide type to which the given sample belongs.

4 Data preparation and analysis

In this study, ASTER DEM with 30m spatial resolution has been used for topographic analysis. Benchmarks have been digitised from the survey of India (SoI) topographic sheet nos. 52 H/3 and 52 H/4 on the scale 1:50, 000. Analysis of geographic and topographic variables such as aspect, slope, elevation, relative relief, distance to road and distance to drainage is done by using ASTER DEM, USGS (Table 1).

Google Earth and Landsat 8 OLI, USGS are used for preparing land use and land cover (LULC) data and landslide inventory map. Geological and geomorphological data is prepared by using geological quadrangle maps, GSI and ground water prospects maps by NRSA for preparing lineament density data. Landslide locations, types, frequency and year of occurrence have been collected from BRO, Manali and PWD, Kullu.

4.1 Landslide inventory

The basic principle of LSM is based on detailed landslide inventory. A definitive, dependable and precise landslide inventory data is very important for predicting accuracy of landslide (Pandey et al., 2021). Frequent field visits for inspection and exploration were performed to examine the factors influencing the landslide in the selected area. Fifty four landslides had been recorded by BRO along the transport corridor in 2011. This inventory was further updated and the number of landslides increased to 96 in 2012. It was updated to 143 in 2015. Detailed field investigation of 54 landslides occurred along the transport corridor was carried out using Global Position System (GPS) and Google earth images in 2018 (Table 2). Spatial temporal map has been prepared using 18 years (2000 to 2018) landslide data collected from BRO, Manali and PWD, Kullu. A total of 197 landslides were identified with 153 of rock slides and 44 of debris slides.

Table 1 Types and sources of dataset

Data type	*Database*	*Resolution and scale*	*Data derivative*
Topographic map	Survey of India (SoI)	RF 1: 50, 000	Boundary of the study area, transport route
Imagery data	Google Earth, Landsat 8	30 meter	Land use and land cover, landslide inventory map
ASTER DEM	USGS	30 meter	Slope, aspect, road, drainage, elevation and relative relief map
Landslide data	BRO, Manali, PWD, Kullu and NDMA govt. reports		Landslide locations, frequency, types of landslide, year of occurrence, road damage and cost
Ancillary data	Geological Quadrangle Map, GSI	1:250,000	Geology and geomorphology map
	GSI and Ground Water Prospects Map by NRSA	1:250,000 and 1:50,000	Lineament density map

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

Table 2 Characteristics of the landslides occurred along the Kullu-Rohtang Pass transport corridor (2018)

Landslide ID	*Landslide location (km)*	*Landslide type*	*Slope (degree)*	*LULC*	*Elevation (m)*	*Lineament density*	*Distance to road (m)*	*Geology and geomorphology*	*Distance to drainage (m)*
L 1	0–2.35	Rock slide	45–60	Settlement	<1,000	Medium	200–400	Schist and quartzite	100–200
L 2	2.35–2.86	Rock slide	45–60	Settlement	<1,000	Medium	200–400	Slate, limestone phylite	100–200
L3	2.86–3.78	Rock slide	45–60	Settlement	<1,000	Medium	200–400	Slate, limestone phylite	100–200
L 4	3.78–4.61	Rock slide	0–15	Settlement	<1,000	Medium	200–400	Slate, limestone phylite	100–200
L 5	4.61–4.92	Rock slide	0–15	Settlement	<1,000	Medium	200–400	Quartzite schist	100–200
L 6	4.92–13.85	Rock slide	45–60	Settlement	<1,000	Medium	200–400	Quartzite schist	100–200
L 7	13.85–14.25	Debris slide	45–60	Dense forest	<1,000	Medium	200–400	Quartzite schist	100–200
L 8	14.25–15.58	Debris slide	45–60	Dense forest	<1,000	High	<200	Quartzite schist	<100
L 9	15.58–16.23	Rock slide	15–30	Dense forest	<1,000	High	<200	Quartzite schist	<100
L 10	16.23–17.39	Rock slide	15–30	Dense forest	<1,000	High	<200	Quartzite schist	<100
L 11	17.39–19.08	Rock slide	45–60	Settlement	<1,000	High	<200	Quartzite schist	<100
L 12	19.08–19.76	Rock slide	45–60	Settlement	<1,000	High	<200	GGG	<100
L 13	19.76–22.64	Rock slide	45–60	Settlement	1,000–2,000	High	<200	GGG	<100
L 14	22.64–23.79	Rock slide	45–60	Settlement	1,000–2,000	Medium	200–400	GFD	100–200
L 15	23.79–27.83	Debris slide	45–60	Agriculture	1,000–2,000	Medium	200–400	GFD	100–200
L 16	27.83–28.64	Rock slide	45–60	Agriculture	1,000–2,000	Medium	200–400	GFD	100–200
L 17	28.64–30.42	Rock slide	45–60	Agriculture	1,000–2,000	Medium	200–400	GFD	100–200
L 18	30.42–31.66	Rock slide	45–60	Agriculture	1,000–2,000	Low	>400	GFD	>200
L 19	31.66–33.72	Debris slide	0–15	Settlement	1,000–2,000	Low	>400	GFD	>200
L 20	33.72–35.2	Rock slide	45–60	Settlement	1,000–2,000	Low	>400	GFD	>200

Notes: Granite gneiss and granitoid (GGG), glacio-fluvial deposits (GFD), highly dissected hill and valley (HDHV) and biotitie schist, kynite gneiss (BSKG).

Table 2 Characteristics of the landslides occurred along the Kullu-Rohtang Pass transport corridor (2018) (continued)

Landslide ID	*Landslide location (km)*	*Landslide type*	*Slope (degree)*	*LULC*	*Elevation (m)*	*Lineament density*	*Distance to road (m)*	*Geology and geomorphology*	*Distance to drainage (m)*
L 21	35.2–36.28	Debris slide	15–30	Settlement	1,000–2,000	Medium	200–400	GFD	100–200
L 22	36.28–36.86	Debris slide	15–30	Settlement	1,000–2,000	Medium	200–400	GFD	100–200
L 23	36.86–37.08	Rock slide	45–60	Settlement	1,000–2,000	Medium	200–400	GFD	100–200
L 24	37.08–38.86	Rock slide	45–60	Settlement	1,000–2,000	Medium	200–400	GFD	100–200
L 25	38.86–40.7	Rock slide	>60	Settlement	2,000–3,000	Medium	200–400	GFD	100–200
L 26	40.7–42.68	Rock slide	>60	Settlement	2,000–3,000	Medium	200–400	BSKG	100–200
L 27	42.68–43.1	Rock slide	>60	Settlement	2,000–3,000	Low	>400	BSKG	>200
L 28	43.1–44.02	Rock slide	>60	Settlement	2,000–3,000	Low	>400	BSKG	>200
L 29	44.02–44.39	Rock slide	>60	Settlement	2,000–3,000	High	<200	BSKG	<100
L 30	44.39–45.56	Rock slide	>60	Settlement	2,000–3,000	High	<200	HDHV	<100
L 31	45.56–46.3	Rock slide	>60	Settlement	2,000–3,000	High	<200	HDHV	<100
L 32	46.3–47.02	Rock slide	>60	Settlement	2,000–3,000	High	<200	HDHV	<100
L 33	47.02–60.77	Rock slide	>60	Settlement	>3,000	High	<200	HDHV	<100
L 34	60.77–61.84	Debris slide	>60	Settlement	>3,000	High	<200	HDHV	<100
L 35	61.84–61.72	Rock slide	15–30	Sparse forest	>3,000	High	<200	HDHV	<100
L 36	61.72–62.5	Rock slide	15–30	Sparse forest	>3,000	High	<200	HDHV	<100
L 37	62.5–63.08	Rock slide	15–30	Sparse forest	>3,000	High	<200	HDHV	<100
L 38	63.08–63.96	Rock slide	15–30	Sparse forest	>3,000	High	<200	HDHV	<100
L 39	63.96–64.48	Debris slide	15–30	Sparse forest	>3,000	High	<200	HDHV	<100
L 40	64.48–67.94	Debris slide	>60	Sparse forest	>3,000	High	<200	HDHV	<100

Notes: Granite gneiss and granitoid (GGG), glacio-fluvial deposits (GFD), highly dissected hill and valley (HDHV) and biotitie schist, kynite gneiss (BSKG).

Table 2 Characteristics of the landslides occurred along the Kullu-Rohtang Pass transport corridor (2018) (continued)

Landslide ID	*Landslide location (km)*	*Landslide type*	*Slope (degree)*	*LULC*	*Elevation (m)*	*Lineament density*	*Distance to road (m)*	*Geology and geomorphology*	*Distance to drainage (m)*
L 41	67.94–68.45	Debris slide	>60	Sparse forest	>3,000	High	<200	HDHV	<100
L 42	68.45–68.79	Rock slide	>60	Sparse forest	>3,000	High	<200	HDHV	<100
L 43	68.79–69.8	Debris slide	>60	Sparse forest	>3,000	High	<200	HDHV	<100
L 44	69.8–70.26	Rock slide	45–60	Sparse forest	>3,000	High	<200	HDHV	<100
L 45	70.26–81.74	Rock slide	45–60	Barren land	>3,000	High	<200	HDHV	<100
L 46	81.74–82.8	Debris slide	>60	Barren land	>3,000	High	<200	HDHV	<100
L 47	82.8–83.64	Rock slide	>60	Barren land	>3,000	High	<200	HDHV	<100
L 48	83.64–83.7	Rock slide	>60	Barren land	>3,000	High	<200	HDHV	<100
L 49	83.7–84.1	Rock slide	>60	Barren land	>3,000	High	<200	HDHV	<100
L 50	84.1–84.56	Debris slide	>60	Barren land	>3,000	High	<200	HDHV	<100
L 51	84.56–85.08	Rock slide	>60	Snow cover	>3,000	High	<200	Snow cover	<100
L 52	85.08–85.62	Debris slide	>60	Snow cover	>3,000	High	<200	Snow cover	<100
L 53	85.62–86.2	Debris slide	>60	Snow cover	>3,000	High	<200	Snow cover	<100
L 54	86.2–86.92	Rock slide	>60	Snow cover	>3,000	High	<200	Snow cover	<100

Notes: Granite gneiss and granitoid (GGG), glacio-fluvial deposits (GFD), highly dissected hill and valley (HDHV) and biotitie schist, kynite gneiss (BSKG).

The types of landslide, i.e., rock slide and debris slide are according to the records of BRO, Manali and PWD, Kullu. The landslide inventory is prepared using Geographic Information System (GIS) software (ArcGIS 9.3) with the help of satellite imageries and GPS way points [Figure 4(a)]. The largest and the smallest landslides mapped along the transport corridor were 4,000 m^3 and 120 m^3 respectively. The highest number of landslides occurred on the terrain or steep slopes are due to geological conditions, continuous deformation of structures in the topography and enormous amount of road cuts in cracked rocks. The maximum landslide (rock slide and debris slide) prone area lies between 45° to 60° and 60° to 80° slope classes.

4.2 Landslide triggering factors

As per the field investigations and evaluation of data, initially nine causing factors were arranged for the assessment of landslide susceptibility: lineament density, slope, elevation, relative relief, aspect, LULC, geology and geomorphology, distance to road and distance to drainage (Champati Ray et al., 2007; Cao et al., 2021). Continuous factors (slope, elevation, lineament density and so on) had been discretised using their normalised values which are calculated by using analytical hierarchical process (AHP) model (Saaty, 1990).

4.2.1 Slope

The slope was categorised in to five categories: very gentle (0–15°), gentle (15°–30°), moderate (30°–45°), steep (45°–60°), very steep (>60°) [Figure 4(b)]. The rock slides are mainly found in steep and very steep slopes while debris occurs in moderate and steep slopes. The normalised values of gentle, very gentle, moderate, steep and very steep slopes were 0.0335, 0.0580, 0.1118, 0.2523, and 0.5443 respectively.

4.2.2 Elevation

The elevation in the study area was divided into four categories: (<1,000 m), (1,000–2,000 m), (2,000–3,000 m), (>3,000 m) [Figure 4(c)]. The normalised values of these categories were 0.0903, 0.0461, 0.1807 and 0.6827 respectively. The landslide had highest normalised value of 0.6827 and occurred frequently in range of (>3,000 m).

4.2.3 Land use and land cover

LULC is an important triggering factor causing the landslide. LULC was classified into six categories: dense forest, agriculture, sparse forest, settlement, barren land and snow cover [Figure 4(d)]. The normalised values for these categories were 0.0373, 0.1854, 0.1438, 0.2035, 0.3981 and 0.0316 respectively. From the normalised value mentioned above, it can be inferred that landslide occurs quite frequently in barren land areas.

Figure 4 (a) Landslide inventory map (b) Slope (c) Elevation (d) LULC (e) Geology and geomorphology (f) Lineament density (g) Aspect (h) Distance to road (i) Distance to drainage (j) Relative relief (see online version for colours)

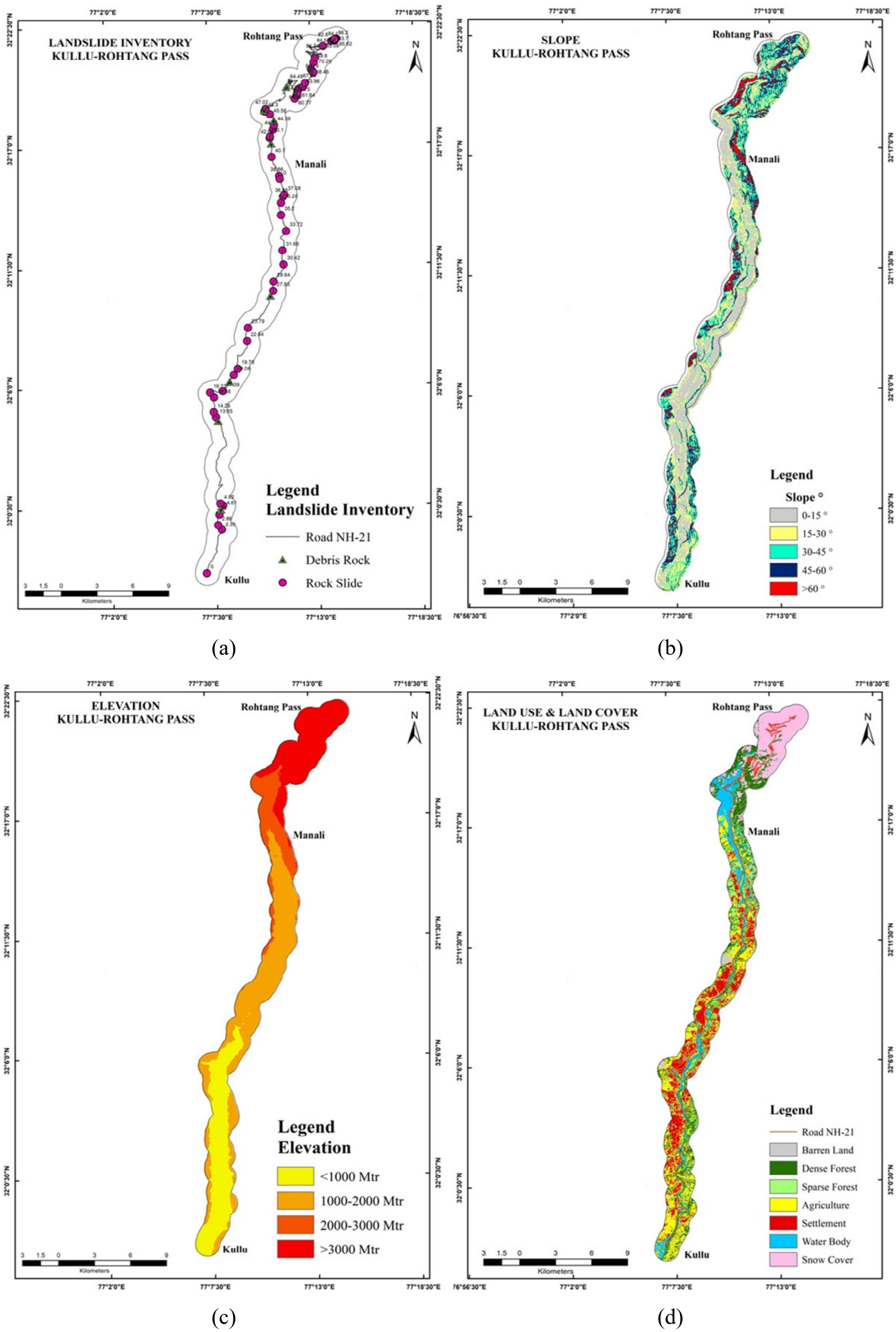


Figure 4 (a) Landslide inventory map (b) Slope (c) Elevation (d) LULC (e) Geology and geomorphology (f) Lineament density (g) Aspect (h) Distance to road (i) Distance to drainage (j) Relative relief (continued) (see online version for colours)

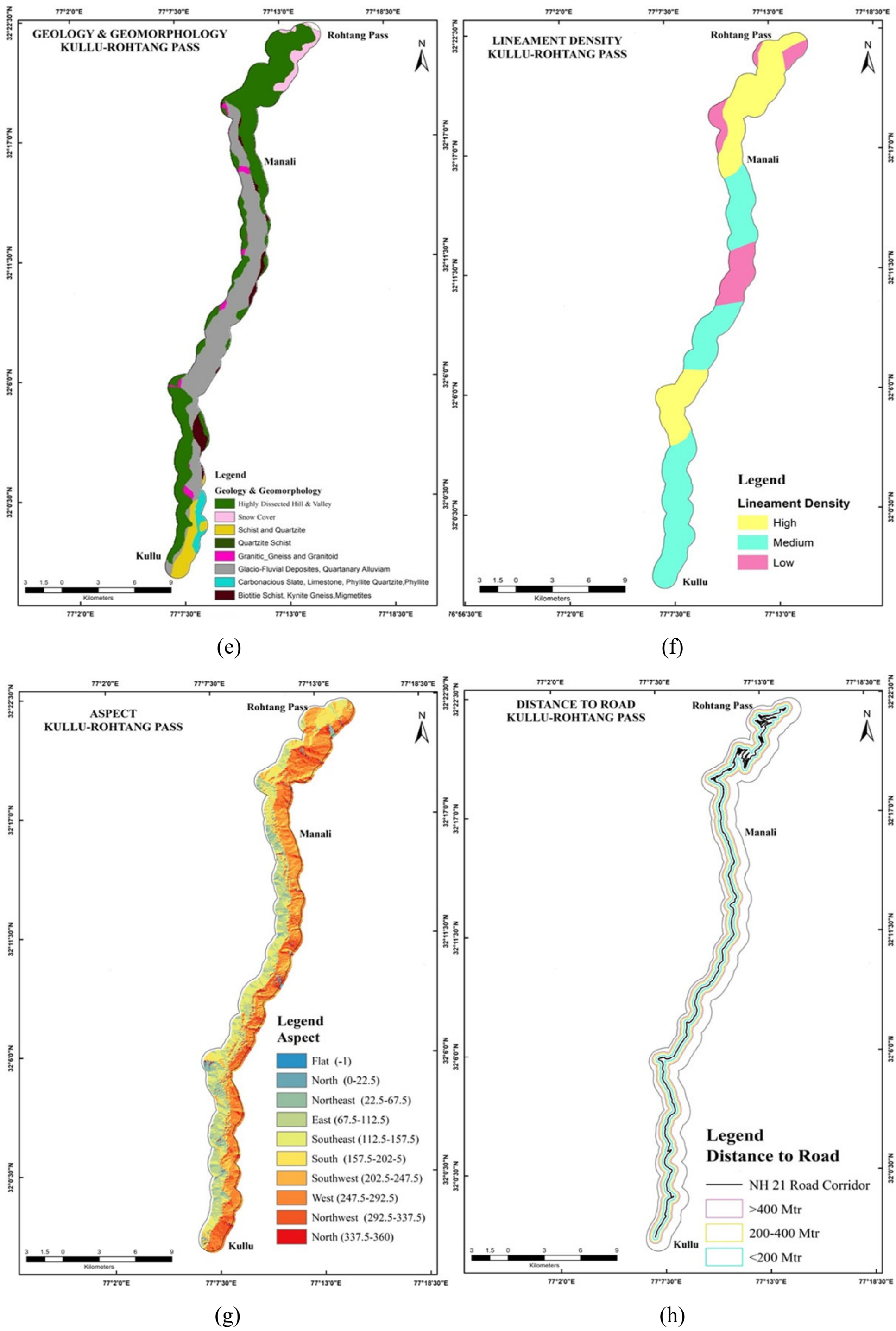
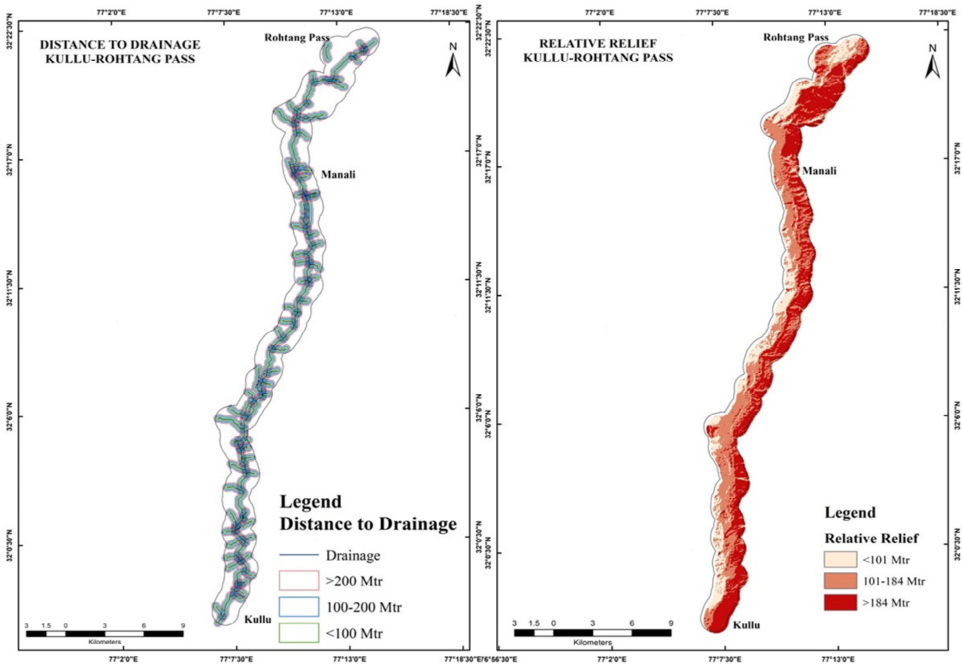


Figure 4 (a) Landslide inventory map (b) Slope (c) Elevation (d) LULC (e) Geology and geomorphology (f) Lineament density (g) Aspect (h) Distance to road (i) Distance to drainage (j) Relative relief (continued) (see online version for colours)



(i) (j)

4.2.4 Geology and geomorphology

Geology and geomorphology is an important feature in landslide study. The study area was divided into eight categories: highly dissected hill and valley, snow cover, schist and quartzite, granitic gneiss and granitoid, glacio-fluvial deposits and quaternary alluvium, quartzite schist, carbonaceous slate and limestone, biotite schist and kynite gneiss [Figure 4(e)]. The normalised values for these categories were 0.3320, 0.0744, 0.0321, 0.0369, 0.0266, 0.2113, 0.1049 and 0.1489 respectively. The highest normalised value of highly dissected hill and valley shows that it has more impact in causing the landslide comparing to other categories.

4.2.5 Lineament density

Lineament can be defined as the lines of landscape representing the geometric pattern of rocks. Lineament density is an important triggering factor that affects the phenomenon of landslides. It was categorised into three categories: low, medium and high [Figure 4(f)]. Normalised values for these categories were 0.0681, 0.1542 and 0.7775 respectively.

4.2.6 Aspect

Aspect was divided into nine categories: flat, north, northeast, east, southeast, south, southwest, west and northwest [Figure 4(g)]. The respective normalised values for these categories were 0.0432, 0.2827, 0.0182, 00289, 0.0895, 0.2423, 0.1508, 0.015 and 0.0624.

4.2.7 Distance to road

Construction of roads brings the changes in the natural topography and leads to landslides. The study area was divided into three zones classifying the impact of landslides occurred along the road corridor [Figure 4(h)]. The three zone categories were: <200 m, 200–400 m and >400 m with normalised values 0.6810, 0.0688 and 0.2500 respectively.

4.2.8 Distance to drainage

Distance to drainage is also among the main conditioning factors in landslide susceptibility analysis process. The proximity of slope to streams makes the slope more unstable and exposed to landslide events. The distance to drainage was divided into three categories: <100 m, 100–200 m and >200 m [Figure 4(i)]. The normalised values for these categories were 0.6494, 0.2941 and 0.0563 respectively.

4.2.9 Relative relief

Difference between the highest and lowest elevation of a particular area is termed as relative relief. Relative relief was also categorised into three categories: <101 m, 101–184 m, >184 m [Figure 4(j)]. The normalised values for these categories were 0.6695, 0.2668 and 0.0635 respectively.

5 Results and analysis

5.1 Landslide susceptibility analysis

To analyse landslide susceptibility, following methods have been performed:

5.1.1 Multi-collinearity problem analysis

Machine learning models respond even to the slight changes in data in their needed range. Therefore, AHP method was used to normalise each triggering factor in a range of [0.01, 0.99]. This normalised data acts as an input to the machine learning model while the landslide susceptibility index (debris slide: 0, rock slide: 1) acts as the output. train_test_split method of classification model was used to divide the data in 70% training samples and 30% testing samples to estimate the accuracy of the model.

The performance of the susceptibility model can be influenced by the multi-collinearity among the triggering factors. To evaluate the multi-collinearity among the nine triggering factors, tolerance and variance inflation factors (VIF) were applied. A tolerance of less than 0.2 or a VIF of 5 or above leads to multi-collinearity problem

(O'Brien, 2007). The smallest tolerance in the data is 0.331 and the highest VIF among these is 3.025 (Table 3).

Table 3 Multi-collinearity of factors

Factor	*VIF*	*Tolerance*
Slope	1.252	0.799
Elevation	2.961	0.338
LULC	1.266	0.790
Distance to road	2.171	0.461
Distance to drainage	1.513	0.661
Geology/geomorphology	2.005	0.499
Lineament	3.025	0.331

5.1.2 Selection and elimination of unimportant triggering factors

Initially, nine factors are arranged accordingly and observed as the landslide causing factors. IGR method has been applied to access the relevance of every individual causing factor. The gain ratio of all nine causing factor can be seen in Figure 5. The triggering factors with high gain ratio value are more important. It is evident from the outcome that lineament density has the highest gain ratio with a value of 1.77.

Figure 5 IGR for triggering factors (see online version for colours)

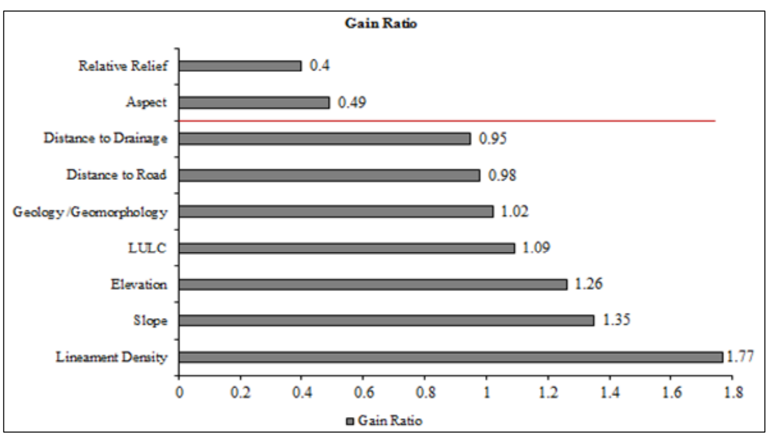


Table 4 Predictive accuracy with elimination of unimportant factors

Model	*Eliminating unimportant factors*	*AUROC*
Model-1	Without eliminating any factor	0.875
Model-2	Eliminating relative relief	0.892
Model-3	Eliminating relative relief, aspect	0.979
Model-4	Eliminating relative relief, aspect, distance to drainage	0.923

Though, all the factors are relevant for landslide prediction, but it is clear from the Figure 6 that the least important factors may reduce the predictive accuracy of the model (Pham et al., 2016). In process of finding the important triggering factors, two least important factors were eliminated at a time and the decision tree classification model was used to predict the accuracy.

As given in Table 4, the predicted accuracy of this model is improved when the irrelevant triggering factors are removed from the dataset.

The maximum accuracy is achieved when two irrelevant factors are eliminated. Thus, relative relief and aspect are eliminated from the dataset.

5.1.3 Landslide susceptibility modelling

Decision tree classification model of machine learning was executed to evaluate the susceptibility of the landslides dataset (Figure 6). As discussed earlier in Subsection 5.1.2, seven relevant triggering factors namely lineament density, slope, elevation, LULC, geology and geomorphology, distance to road and distance to drainage were determined to generate the LSM. The parameters for decision tree classifier are shown in Table 5.

Figure 6 Flowchart of LSM and prediction

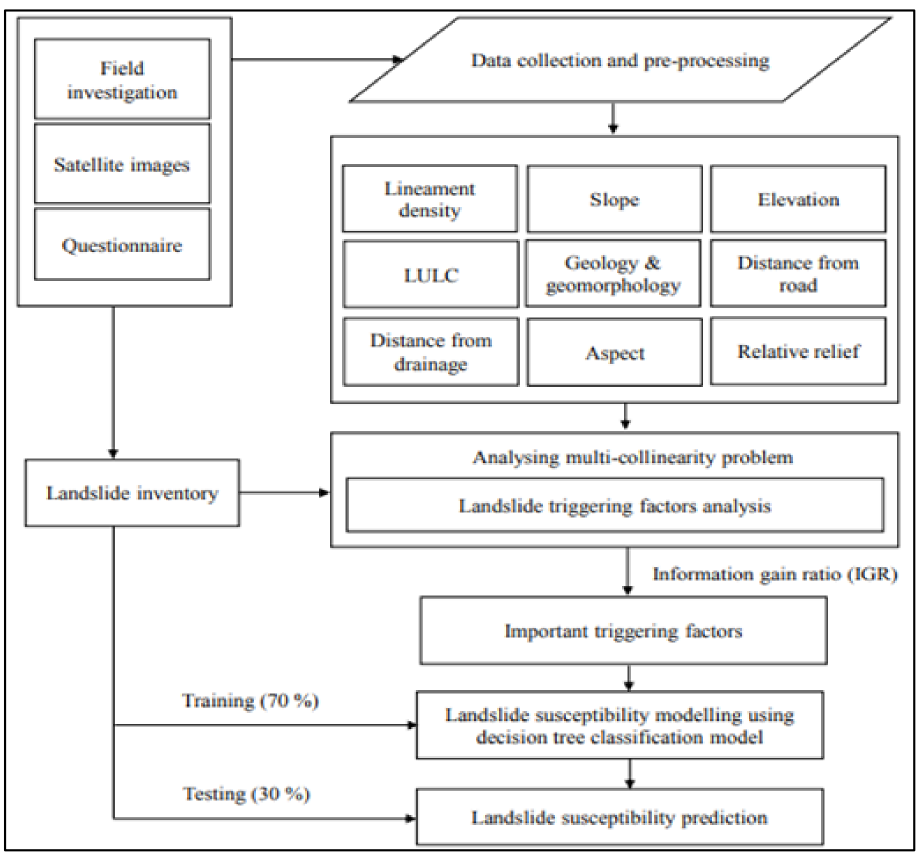


Table 5 Parameters of decision tree classifier

Factor	*Value*	*Note*
Criterion	Entropy	Measures the quality of split
min_samples_leaf	1	Minimum number of samples required to be at a leaf node
Splitter	Best	Strategy used to choose the split at each node
min_sample_split	2	Minimum number of sample required to split and internal node
min_impurity_decrease	0.0	Measure of impurity at a node split

5.2 Validation and ROC curve

5.2.1 Using classification report of the model

To evaluate the effectiveness of this LSM, validation is the most important component. With the triggering factors selected as discussed in Subsection 5.1.2, the decision tree classification model achieve predictive accuracy of 90.7% which is quite satisfactory (Figure 7).

Figure 7 (a) Original dataset visualisation in 2D (b) Result visualisation in 2D (see online version for colours)

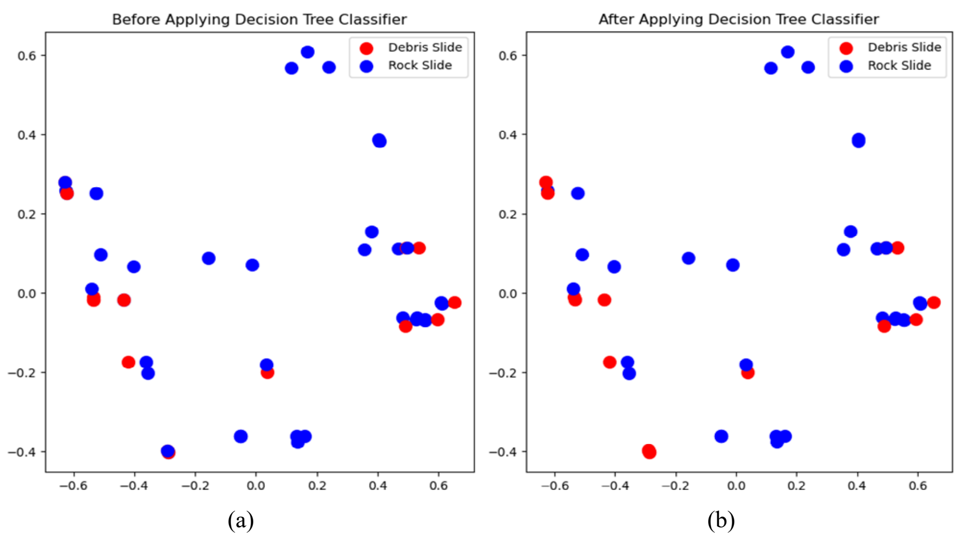


Figure 7(a) shows the plot of dataset in two-dimensions for better visualisation. Though the actual data consists of seven triggering factors, each representing the dimension of data but it is practically not possible to visualise this data with all the seven triggering factors as one can better visualise up to three-dimensions only. In this study, two-dimensions have been used to display both data and result using principal component analysis (PCA) technique. PCA (Svante et al., 1987) is mainly used to reduce the dimensionality of the data and returns principal components that can be used to display the data in lower dimensions. Figure 7(b) shows the results observed by applying the decision tree classifier on the data in two-dimensions. The classification report of the decision tree model of machine learning is shown in Table 6.

Precision and recall are two important metrics for model evaluation. Precision measures the result relevancy and is calculated by using equation (5). If the model has high precision value then it indicates that the false positive rate of the model is low. Recall measures the percentage of total relevant results that this model classifies correctly and is calculated by using equation (6). If the recall value is high, it shows the false negative rate of the model is low. For a model to be more accurate the value of precision and recall should be high.

Table 6 Classification report of model

Class	*Precision*	*Recall*	*F1-score*
0	0.75	1.00	0.86
1	1.00	0.87	0.93

So, from precision and recall values (Table 6), it can be inferred that model has good precision and recall value and hence good accuracy.

$$precision(P) = \frac{T_p}{T_p + F_p} \tag{5}$$

$$recall(R) = \frac{T_p}{T_p + F_n} \tag{6}$$

In equations above,

- T_p = total number of true positive samples
- F_p = total number of false positive samples
- F_n = total number of false negative sample
- $(T_p + F_p)$ represents the actual results and $(T_p + F_n)$ represents the predicted results.

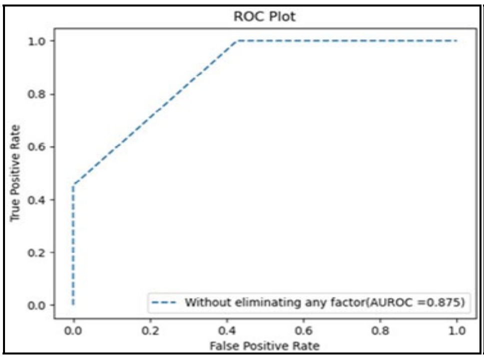
F1-score can be defined as the harmonic mean of precision and recall and can be calculated by using the equation (7). Higher the value of f1-score, better the model is.

$$f1\text{-}score = 2\frac{P \times R}{P + R} \tag{7}$$

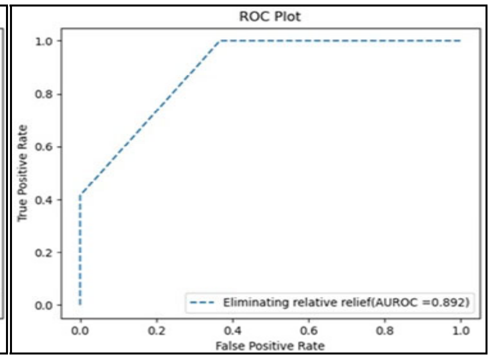
5.2.2 Using ROC curve

Receiver operating characteristic (ROC) curve is a probability curve that summarises the trade of true positive rate and false positive rate of a predictive model. It is the most widely used evaluation metric for checking the performance of a classification model. Area under the ROC curve (AUROC) is applied to evaluate the model's execution and a model with larger area under AUROC is considered as the best. Figure 8(c) shows that the AUROC of the model with the selected triggering factors is 0.979 which is greater than that of AUROCs generated with different sets of triggering factors.

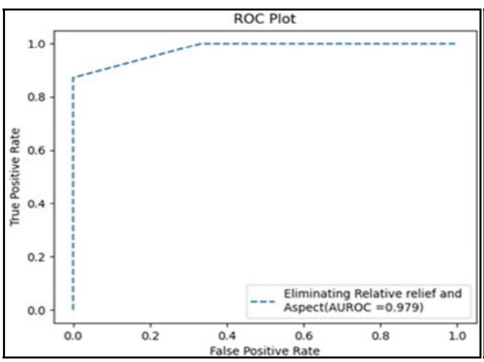
Figure 8 (a) Without eliminating any factor (b) Eliminating relative relief (c) Eliminating relative relief and aspect (d) Eliminating relative relief, aspect and distance to drainage (see online version for colours)



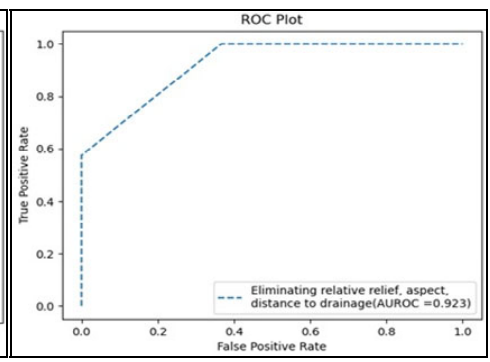
(a)



(b)



(c)



(d)

6 Discussion

In this study area, two different types of landslides were discussed: rock slide and debris slide. Each triggering factor has its different importance for a landslide occurrence. For example, lineament density had dominant effect on a landslide while relative relief had least role to play in case of landslide. Machine learning model (decision tree classifier) performed well for LSM and assessment with an AUROC of 0.979. This indicates that machine learning models can be applied to complex nonlinear problems like landslide prediction. One major concern must be notified that machine learning model performance is sensitive to data and may differ from case to case. The error in a LSM is comprised of false negative class and false positive class of data as shown in confusion matrix Table 7.

Table 7 Confusion matrix

	Actual class	
Predicted class	True positive	False positive
	False negative	True negative

False positive class predicts areas as landslides prone that are actually not. This may restrict the land use and can lead to economy loss by not using that land for economic activities. However, if the landslide prone area is predicted as a stable slope erroneously, i.e., false negative portion increased, it may lead to serious landslide disaster and loss. Further studies of landslide susceptibility assessment, should have the prime focus on the methods to minimise the false negative error.

7 Conclusions

LSM and assessment is necessary for proper land use planning in mountain areas to reduce the disaster risks. In this study, Kullu to Rohtang Pass has been considered as the research study area and two different types of landslides are observed, the rock slide and the debris slide. Lineament density, slope, elevation, LULC, geology and geomorphology, distance to road and distance to drainage were important triggering factors for landslide prediction. IGR was the basis to evaluate the significance of each triggering factor. The four models with various eliminated factors exhibited that the irrelevant factors had negative effect on LSM and should be eliminated to achieve high predictive accuracy.

Decision tree classification model of machine learning was followed to accomplish LSM. The classification report and ROC curve were applied to assess the performance. The final outcomes show an accuracy of 90.7% and AUROC value of 0.979. Results demonstrate that decision tree classifier for landslide prediction performed well with good accuracy and can be beneficiary for decision making. This prefatory analysis would help engineers to develop suitable plans for carrying out engineering projects in the study area. The present study has many different directions for future research. Implementing different machine learning models and comparing their performances with decision tree machine learning model for landslide susceptibility in the study area will be an interesting topic of work. Further this work can be extended by using other feature selection techniques for selecting best data to fit the machine learning model.

References

Ambrosi, C., Strozzi, T., Scapozza, C. and Wegmüller, U. (2018) ‘Landslide hazard assessment in the Himalayas (Nepal and Bhutan) based on Earth observation data’, *Engineering Geology*, Vol. 237, pp.217–228.

Arabameri, A., Pal, S.C., Rezaie, F., Chakrabortty, R., Saha, A., Blaschke, T., Napoli, M.D., Ghorbanzadeh, O. and Ngo, P. (2021) ‘Decision tree based ensemble machine learning approaches for landslide susceptibility mapping’, *Geocarto International*, pp.1–35.

Bertsimas, D. and Dunn, J. (2017) ‘Optimal classification trees’, *Machine Learning*, Vol. 106, No. 7, pp.1039–1082.

Bhambri, R., Mehta, M., Dobhal, D.P., Gupta, A.K., Pratap, B., Kesarwani, K. and Verma, A. (2016) ‘Devastation in the Kedarnath (Mandakini) Valley, Garhwal Himalaya, during 16–17 June 2013: a remote sensing and ground-based assessment’, *Natural Hazards*, Vol. 80, No0. 3, pp.1801–1822.

Cao, Y., Wei, X., Fan, W., Nan, Y., Xiong, W. and Zhang, S. (2021) ‘Landslide susceptibility assessment using the weight of evidence method: a case study in Xunyang area, China’, *PLoS One*, Vol. 16, No. 1, p.e0245668.

Champati Ray, D.P., Dimri, S., Lakhera, R.C. and Sati, S. (2007) 'Fuzzy-based method for landslide hazard assessment in active seismic zone of Himalaya', *Landslides*, Vol. 4, No. 2, pp.101–111.

Dinov, I.D. (2018) 'Decision tree divide and conquer classification', in *Data Science and Predictive Analytics*, Springer, Cham, USA.

Ghahramani, Z. (2015) 'Probabilistic machine learning and artificial intelligence', *Nature*, Vol. 521, No. 7553, pp.452–459.

Goyal, S. and Maheshwar (2019) 'Naïve Bayes model based improved k-nearest neighbor classifier for breast cancer prediction', in Luhach, A., Jat, D., Hawari, K., Gao, X.Z. and Lingras, P. (Eds.): *Advanced Informatics for Computing Research, ICAICR, Communications in Computer and Information Science*, Springer, Singapore, p.1075.

Hunger, O., Leroueil, S. and Picarelli, L. (2014) 'The Varnes classification of landslide types, an update', *Landslides*, Vol. 11, No. 2, pp.167–194.

Hussin, H.Y., Zumpano, V., Reichenbach, P. et al (2016) 'Different landslide sampling strategies in a grid-based bi-variate statistical susceptibility model', *Geomorphology*, Vol. 253, pp.508–523.

Iverson, R.M. (2015) 'Scaling and design of landslide and debris-flow experiments', *Geomorphology*, Vol. 244, pp.9–20.

Loi, D.H., Quang, L.H., Sassa, K., Takara, K., Dang, K., Thanh, N.K. and van Tien, P. (2017) 'The 28 July 2015 rapid landslide at Ha Long City, Quang Ninh, Vietnam', *Landslides*, Vol. 14, No. 3, pp.1207–1215.

Maheshwar, K.G. (2019) 'Breast cancer detection using decision tree, Naïve Bayes, KNN and SVM classifiers: a comparative study', *International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, pp.683–686.

Maheshwar, K.K. and Arora, V. (2015) 'A hybrid data clustering using firefly algorithm based improved genetic algorithm', *Procedia Computer Science*, Vol. 58, pp.249–256.

Narayanan, B.N., Djaneye Boundjou, O. and Kebede, T.M. (2016) 'Performance analysis of machine learning and pattern recognition algorithms for Malware classification', *IEEE National Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS)*, Dayton, OH, pp.338–342.

O'Brien, R.M. (2007) 'A caution regarding rules of thumb for variance inflation factors', *Qual. Quant*., Vol. 41, No. 5, pp.673–690.

Pandey, V.K., Tripathi, A.K. and Sharma, K.K. (2021) 'Implications of landslide inventory in susceptibility modeling along a Himalayan highway corridor, India', *Physical Geography*, pp.1–24.

Pham, B.T., Bui, D.T., Dholakia, M., Prakash, I. and Pham, H.V. (2016) 'A comparative study of least square support vector machines and multiclass alternating decision trees for spatial prediction of rainfall-induced landslides in a tropical cyclones area', *Geotechnical and Geological Engineering*, Vol. 34, No. 6, pp.1807–1824.

Quinlan, J.R. (1996) 'Improved use of continuous attributes in C4.5', *J. Artificial Intelligence Res.*, Vol. 4, pp.77–90.

Saaty, T.L. (1990) 'How to make a decision: the analytic hierarchy process', *European Journal Operational Research*, Vol. 48, No. 1, p.926.

Saha, A.K., Gupta, R.P., Sarkar, I., Arora, M.K. and Csaplovics, E. (2005) 'An approach for GIS-based statistical landslide susceptibility zonation – with a case study in the Himalayas', *Landslides*, Vol. 2, No. 1, pp.61–69.

Sassa, K. (2017) 'The Fifth World Landslide Forum – implementing and monitoring the ISDR-ICL Sendai Partnerships', *Landslides*, Vol. 14, No. 5, pp.1857–1859.

Singh, R.B. and Pandey, B.W. (1996) 'Landslide hazard in Indian Himalaya and Canadian Rockies: a comparative analysis', in Jose, C., Clemente, I. and Tomas, F. (Eds.): *Landslides*, pp.63–69, Balkema, Rotterdam.

Svante, W., Kim, E. and Paul, G. (1987) 'Principal component analysis', *Chemometrics and Intelligent Laboratory Systems*, Vol. 2, Nos. 1–3, pp.37–52.

Tien Bui, D., Tuan, T.A., Klempe, H., Pradhan, B. and Revhaug, I. (2016) 'Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree', *Landslides*, Vol. 13, No. 2, pp.361–378.

Wang, H., Zhang, L., Yin, K., Luo, H. and Li, J. (2021) 'Landslide identification using machine learning', *Geoscience Frontires*, Vol. 12, No. 1, pp.351–364.