

Landslide susceptibility modelling in a part of Himachal Pradesh, India: An integrated method based on machine learning and geospatial techniques

Rudraksh MOHAPATRA

DPS, VK, Delhi 110070, India; rudrakshmohapatra@gmail.com

Abstract

Landslides are one of the most destructive natural hazards in the mountainous regions across the globe including the western Himalayas of India. Hence, it is essential to implement mitigation plans, evacuation measures, and an infrastructure plan based on precise, efficient landslide susceptibility models. Current methods of landslide susceptibility mapping are improving constantly, using geospatial techniques to incorporate visual representation of the environment. However, these current methods are often opinion driven, due to lack of consensus on which factors take precedence over others. This study aims to provide a different approach namely a machine learning based approach towards Landslide Susceptibility Mapping, integrating GIS to give an accurate visual representation of the surrounding areas ranked by order of susceptibility in/and around Kullu Valley of western Himalaya, India. The landslide conditioning factors used in the study involve both static and dynamic data such as slope, land use, land cover, and rainfall variables. The research found that although the Extremely Randomised Trees provide a considerably more accurate assessment of the study area's vulnerability, the Random Forest Regressor has greater overall accuracy. There is a significant relationship between the model's outputs and past landslides. According to the study, there would be significantly more regions with high susceptibility to the effects of climate change on landslides by 2030. The application can identify the geographical distribution of landscape risk and is significantly less time-consuming than current methods of susceptibility analysis. Machine learning models could be crucial in evacuation efforts and in preventing damage to life and property.

Keywords: extremely randomised trees model; geo-informatics; landslide susceptibility; machine learning; rainfall variability; Western Himalaya

Introduction

Landslides are movement of enormous amounts of rock, debris or earth down an incline and are some of the most common natural hazards in mountainous areas (Formetta *et al.*, 2014; Lee *et al.*, 2017). These cause severe damage to life and property when they occur near human settlements. Sometimes landslides even have great influence on human society, In addition to the economic impact. It has caused hundreds of billions of loss of property and loss of life (Aleotti and Chowdhury, 1999). According to a statistic by the World Health

Organisation, approximately 4.8 million people have been affected by landslides in the last decade alone and over 18,000 lives have been lost (Sun *et al.*, 2022; <https://www.who.int/health-topics/landslides#tab=tab>). To categorise the causes of downslope movement of debris, rocks, or earth material under the influence of the force of gravity is a complex process. However, the underlying machinations of landslides is intricate, with a complex combination of various factors such as slopes, geological conditions, soil types, vegetation covers, rainfall and even anthropogenic factors contributing to the occurrence of a landslide (Nakileza and Nedala, 2020; Jung, 2021). The devastation to infrastructure and human life during landslides can be prevented with the effective implementation of mitigation strategies and well-informed disaster preparedness measures (Wickramasinghe (2021).

The young fold mountains of the Himalayas are one of the most active mountainous regions, which makes them particularly vulnerable to landslides which have significant impact on human social sustainability (Kuniyal *et al.*, 2019). The most frequent means of transportation in hilly areas is road transit, which are at risk in the event of a landslide. Landslides are increasingly a common incident in mountain region leading to loss of life and properties in many parts of the developing world particularly in Himalayas (Turner 2018). Landslides in the Himalayan region mostly obstruct contact on transportation routes, a problem further enhanced by the lack of proper shelter and alternate route plans for landslide hazards in India (Uniyal, 2017; NLRMS, 2019). Early warning tends to reduce risk by enhancing public awareness and response to the warning when hazards occur (Sufri *et al.*, 2020). A vast majority of the damage to human life and property can be minimised by the implementation of suitable and effective mitigation strategies at the right time (Sufri *et al.*, 2020). Unfortunately, by the time proper mitigation measures have been implemented in the landslide affected areas, precious time has been lost. Landslide risk prediction can be a time-consuming and costly procedure, although extremely complex risk map analysis at various sizes has aided developments in computer modelling of landslide risks (Wilkinson *et al.*, 2002; Chae *et al.*, 2017; Korup and Stolle 2014; Merghadi *et al.*, 2020).

There are two major approaches for assessing landslide susceptibility, qualitative or opinion driven models and quantitative studies. In opinion driven models, the susceptibility assessment depends exclusively on the judgement of the investigators/expert (Van Westen *et al.*, 1999). These methods are applied to small areas as they lack physical concept of slope failure (Xie *et al.*, 2004). As for quantitative studies, the most prevalent ones are statistical models, which are data driven models. In the data driven method the data on past landslides is used for the evaluation and is based on the logic that the same conditions that caused landslides in the past would cause them again in the future (Nadim *et al.*, 2011). The three most prevalent data-driven approaches are multivariate statistical methods, bivariate statistical methods, and artificial neural network analysis (Lu and Rosenbaum 2003). In bivariate statistical analysis individual conditioning factors such as geology, geomorphology, slope, etc. are combined with locations of land slide and each parameter's weighted values are calculated. Whereas, the relationship between independent variables like conditioning factors, such as slope, geology etc. and a dependent variable e.g., landslide occurrence is combined and evaluated in multivariate statistical methods. Based on all of this, a guideline has been provided to prepare an overall map with different zones classified on their degree of susceptibility to landslides, such as low, moderate, and high landslides zones (Fell *et al.*, 2008).

In recent years various quantitative techniques and approaches have been developed for landslides susceptibility modelling (LSM). In the recent years, machine learning has been used with the statistical models, physical based models, and opinion-driven (i.e., heuristic) models are considered the four main types of Landslide Susceptible Modelling approaches for the recent landslides' studies (Pham *et al.*, 2016; Chang *et al.*, 2019; Tien Bui *et al.*, 2019; Yunus *et al.*, 2019). At the other side, the opinion-driven models are based on structuring a model with minimum basic information which is considering that information by ranking the landslide influencing factors, which is mainly by the prior existing knowledge-based opinion. In the last few years, the statistical models, got advantage due to the advancements in GIS. In a result several number of quantitative method as well as various techniques have been introduced and implemented successfully for landslides modelling that help in acquiring information of landslide patterns and their triggering mechanisms

(Kalsnes and Nadim, 2012; Pham *et al.*, 2016; Dou *et al.*, 2020). For the primary concern of statistical models, the machine learning stresses optimization and performance rather than the inference (Camilo *et al.*, 2017; Liu *et al.*, 2019; Tien Bui *et al.*, 2019).

There are three main objectives of this study, all of which work towards the general aim of finding a machine learning integrated GIS application that can quickly generate a susceptibility map of the area, given certain inputs. The three objectives of this study:

1. To assess the viability of a dynamic model which generates landslide susceptibility of the terrain and identifies the vulnerable villages and roads of the landslide-prone area integrating machine learning models and GIS.
2. To generate an instantaneous prediction of landslide-prone areas, affected places, and roads using real-time inputs whenever required by the disaster mitigation team as well for the locals and tourists.
3. To predict future landslides using projected rainfall and other static inputs for policy making and further development of the susceptible area.

The major aim for this study was to utilise human inputs, and a machine learning model to quickly put forth a proper susceptibility map of the area, which can be worked as an early warning allowing for effective mitigation resources allocation. The model would then be used to give the susceptibility prediction for 2030. The machine learning model, which is an Extremely Randomised Trees Classifier Model, uses both static inputs (landforms, aspect, and slope) and dynamic inputs (rainfall) to output a susceptibility map, which allows us to understand which area or villages would be the most severely affected. This would not only aid in timely evacuation measures but also impose measures to minimise the damage caused to these villages and human life by having a better understanding of the villages that are prone to landslide. This is especially useful as the remoteness of landslide prone areas are a hindrance to proper implementation of mitigation strategies. This would aid in early warning and mitigation strategies as well as resources allocation for further development, improving policy making and infrastructure development. The model can take the latest inputs, which further increases the viability of the project. Thus, the aim of this project is to create an integrated model using machine learning and Geographic Information System (GIS) to improve upon existing methods of susceptibility mapping to show which villages would be the most adversely affected allowing for better mitigation strategies.

Materials and Methods

Methods and their proposed use in creating a landslide prediction system using machine learning and GIS, based on open-source software and publicly available data

There are a few problems associated with the methods that conventional approaches towards Landslide Susceptibility Mapping take, including both quantitative and qualitative approaches. The majority of present modelling programmes are script-based, which adds a level of complexity that necessitates the use of specialists or trained individuals. Industry-standard GIS software is also expensive, which adds to the impediment to its adoption in these sectors. As a result, critical landslide prediction methods are becoming increasingly difficult to deploy in many high-risk areas where people reside. These issues also highlight not only the accessibility but also limitations of current methods (Merghadi *et al.*, 2020).

The lack of data in many high-altitude areas is also a problem which limits any attempt at making successful Landslide Susceptibility Maps (Du *et al.*, 2020). The major thought behind this research was to overcome some of the problems presented above by creating a landslide prediction system using machine learning and GIS that will be accessible as well economical, based on open-source software and publicly available data.

Study area

The study area considered in the research is the Kullu valley region of the Himalayas, where landslides occur frequently. The data collected on the Kullu region from the Geological Survey of India (GSI) and the Indian Water Resources Information System (WRIS) was used to create the landslides conditioning factors or input features and past data on landslides was used to create the output features. Its mountainous geography and severe seasonal monsoon rains make the area frequently prone to landslides. Because of the availability of large-scale high-quality data, advances in both hardware and software to efficiently process the machine learning techniques progress has taken place. Such progress helps to integrate these data to generate a model which is capable of generating very precise predictions. It also provides an alternative solution to conduct proper susceptibility analysis is through machine learning. However, there is scepticism towards unconventional methods in this particular application of Landslide Susceptibility Mapping. But the results of this study show that machine learning techniques are a viable option as well.

The majority part of the study area is a part of populated Kullu district and forms a part of the Beas and Parvati River basin system in Himachal Pradesh, Western Himalaya. At a height of 1279 metres, it falls in the Middle Inner Himalayan Range, which typically has an average altitude varying between 1,300 m and 3,000 m. The physiography of the region entails high rugged topography lined with narrow valleys, incised meanders, river terraces, fan terraces and fans. Some common features of the study area include ridges, hills, gorges, spurs, precipitous cliffs and escarpments alongside several mountain peaks with heights extending over 5500 metres.

The Kullu district in particular is home to 4,37,903 people according to the 2011 census, and is an important hub of tourism in the area, owing to its rich cultural history. The landscape of this valley has changed significantly since 1961, especially population alone has increased 28 to 69 persons/ km² in this mountain region from 1961 to 2001 (census 2011; India, <https://www.census2011.co.in>).

The climate of the region is cold and dry throughout the year, and experiences three seasons annually. The cold season lasts from October and well into February. This is followed by a hot season from March to June. The rainy season lasts from July to September, during which the district experiences moderate rainfall of 1405.7 mm annually, where 57% occurs during the months from July to September, and the rest from January to February. The region contains mainly acidic soils varying 50-100 cm depth in the district. Sandy loam and Clay loam soils also dominate the majority of soil types in the district, along with quartzite intercalated with phyllite, schist and gneiss.

The study area (Figure 1) was chosen due to the fragility of its mountainous ecosystem, which makes it a hotbed for landslides. The lack of properly documented detailed meteorological data over the region makes the study of its climatic shifts and other events a challenge. However, the change in these climate patterns brought about after the industrial revolution has witnessed an overall increase in the concentrations of greenhouse gases. The introduction of these new components into the climate of the region has brought some significant changes, and it is imperative that the effects of this on natural hazards be studied, which the study aims to do.

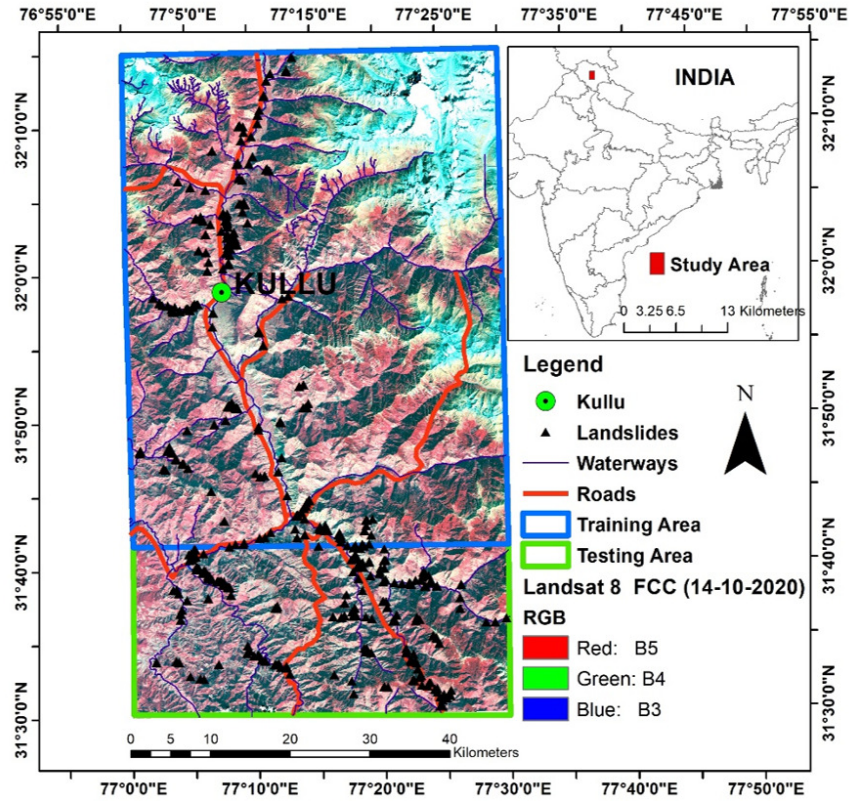


Figure 1. Study area
 The False Colour Composite (FCC) image shows the study area, with demarcation of testing and training sites shown using the coloured borders.

Database and methodology

Various thematic data have been used to fulfil the objectives of the study. The Geological Survey of India (GSI) provided the data on the parameters that have been determined crucial on predicting landslides. These include land-use land-cover data, slope, lithology, aspect, geomorphological landforms, roads, waterways etc. data and point data on past landslide locations along with villages. The data is open sourced and is freely available from GSI, India web portal (<https://bhukosh.gsi.gov.in/Bhukosh/Public>). As for the climatic data like rainfall, which constitutes the major precipitation form in the study area, was obtained from the Water Resources Information Portal. The rainfall data was taken for the past 120 years (1901-2020) (Table 1) (<https://indiawris.gov.in/wris/#/rainfall>). Geographically, the rainfall data was collected on the neighbouring five districts.

Therefore, a moving average window of 10 years was used to estimate the rainfall patterns and scenarios in the districts. Inverse Distance Weightage Interpolation (IDW) is a function in ArcGIS which relies on the inverse relation of a variable with the distance from a certain number of input points. This means that the further a given point is from the input point, the less impact the variable has. Usually, the exact magnitude of the inverse relation is controlled through the Power value. For the purposes of this study, it was determined that a Power value of 2 would be an accurate measure. One limitation of this method is that the results may be inaccurate if the distance between the given point and input points is too big. However, a work around for this solution was to use the data from the neighbouring five districts (names of districts in Table 1).

Table 1. Moving average Time Series Rainfall Data

Annual Average Rainfall per decade (10 years), 120 years and prediction for 2030	Locations and regions: State of Himachal Pradesh (HP), a large part of Western Himalaya Region					
	Mandi District	Kangra District	Shimla District	Lahul and Spiti District	Kullu District	Kinnaur District
1901-1910	158.68	325.77	183.90	92.01	110.53	69.41
1911-1920	153.61	344.51	178.66	145.65	120.32	64.21
1921-1930	159.97	372.75	198.21	176.99	135.51	106.63
1931-1940	153.69	362.74	188.19	100.73	115.84	78.41
1941-1950	177.52	401.56	212.18	131.23	138.40	116.64
1951-1960	241.48	397.83	150.60	111.97	162.73	51.23
1961-1970	227.14	362.28	160.47	85.45	135.66	75.27
1971-1980	175.50	298.23	162.60	90.39	156.41	104.33
1981-1990	188.33	318.32	136.77	85.15	120.68	67.48
1991-2000	204.42	386.24	153.66	125.20	138.65	88.53
2001-2010	168.99	275.15	141.29	103.28	122.99	92.96
2011-2020	198.07	272.75	148.22	101.76	145.76	96.33
Annual Avg. Rainfall (in mm) of 120 Yrs.	183.95	343.18	167.90	112.48	133.62	84.29
Prediction for Year 2030	207.49	304.98	135.25	92.86	144.64	92.48

A variable search radius was used, as the distance between each individual district to the study area in Kullu district may vary. The rainfall data for Kullu district was also used, however, since data for only one meteorological station was available, it could not be applied ubiquitously over the whole district. Hence the need for Inverse Distance Weightage Interpolation Method to calculate the average annual distribution of rainfall over the whole region. The data obtained both in raster and vector format which was later on used in GIS analysis. The Survey of India Topographic sheets (SOI) have been used for the ground detail.

Methodology

1. Identifying prospective Input and Output Factors

There are several intricate factors to consider while selecting factors which would be of importance in their influence on landslides. There were nine landslides conditioning factors which were selected for analysis in the study and fed into the model. The nine input factors are aspect, landforms, land-use-land-cover, landslide polygons, lithology, roads, rainfall, slope (degree) and waterways.

a) Aspect (Map)

The orientation of the land's incline is known as aspect. This feature is important in determining the cause of landslides as it measures the orientation towards which the landslide may occur.

b) Geomorphological landforms (Map)

The topography of the land with geographical features of the areas are displayed in the given map.

c) Land-use-land-cover (Map)

The physical coverage and usage of the land as represented by spatial information constitute the Land-cover map. These include areas of forests, grasslands, croplands, etc. Land use maps on the other hand involve the anthropogenic activities that work towards changing or maintaining the land (Watson and Philip 1985).

d) Landslide polygons (Map)

This map included the point data on locations where landslides had occurred. The reason behind selecting this feature was the fact that areas to the periphery where landslides had occurred before are prone to landslides, and hence would be an important source of information in Landslide Susceptibility Mapping.

e) Lithology (Map)

The lithology of the area has been subdivided into the following classes (Classes)

f) Roads (Map)

The roadways of the study area, which form the major means of transportation, play a crucial role in understanding the escape routes during evacuation measures and are of utmost importance.

g) Rainfall (Map)

The rainfall distribution of the study area, calculated using the IDW method taking in inputs from the five surrounding districts and the Kullu district itself was converted as follows.

h) Slope degree (Map)

The incline of the land plays an intuitive role in the occurrence of landslides, and is one of the most influential factors. The slope degree distribution of the area has been shown in the map.

i) Waterways (Map)

The water ways formed by the Beas River forms the main drainage system of the study area.

2. Compiling the data in a processable format

All nine input features were thus obtained. These data were imported in ArcGIS (10.5) and converted to point shapefile. Then by using Inverse Distance Weighting (IDW) interpolation has performed to this shapefile to create raster file for entire region having pixel size 29 m.

For machine generated model nine sets of input data and one set of output data was put to use here (Figure 2). The next step was to make the data suitable for analysis, this included slicing the arrays to get the desired values, and finally compiling them. To convert the data, which was in Tagged Image File Format (TIF), it was converted in Comma Separated Value Files using the Geographic Data Abstraction Library (GDAL) in python.

This conversion was then read by python using the Numpy library and represented as an array. The total number of data points was 4435251, split according to their latitude and longitude. There were nine such arrays representing the nine input features considered in the study.

3. Reconfiguring the data into training and testing datasets

One of the best ways to determine a machine learning model's accuracy is to create a testing dataset. An approximate ratio deemed suitable for dividing the dataset into training and testing was 3:1. On the basis of this the array was divided into the training dataset, which was fed into the two models, Random Forest and Extremely Randomised Trees. The output dataset was also subsequently divided into Training and Testing. It is important to note that the output was the susceptibility maps generated using conventional methods, and the loss was computed comparing the models output with it. However, the model's accuracy was determined on the number of past landslide locations which fell in the regions the model deemed as highly susceptible.

4. Model training

The next step was to train the model on various criteria. The tree-based models required significant changes in their criteria. For the random forest, a number of factors were tweaked, and it was determined that 50 estimators which calculate the loss of the model up till a depth of 100 yielded the best results for the Random Forest application.

As for the Extremely Randomised Trees model, 300 estimators which operated on the Gini criterion at a maximum depth of 20 fit the best for that particular application.

5. Producing the Landslide Susceptibility Maps

Finally, the output was exported as a TIF file and used the projection system EPSG:32643 (commonly used for Northern latitudes). The model generated susceptibility map was then compared with the available susceptibility map of the area with having same three landslide susceptibility classes as low, moderate and high. The village points were overlaid on the model generated susceptibility map and were classified as low moderate and high based on the contents of susceptibility classes. A buffer of 1 km of the active landslide points were created in Arc GIS and the same is overlaid with the model generated susceptible map. The classified villages were confirmed once again based on their proximity to the active landslide buffer zone. The three prominent roads of the area were also overlaid on the model generated susceptibility map and buffers of 100 m, 200 m, 500 m and 1000 m were created respectively. On the basis of the occurrences of landslides within the buffer zones of roads the roads were also classified based on vulnerability to effects of landslides. As mentioned in objectives section the rainfall data were used to forecast for 2030 based on the value of the rainfall a predicted susceptibility map has also been prepared using static and dynamic data, which shows the Early Warning Capabilities of the model.

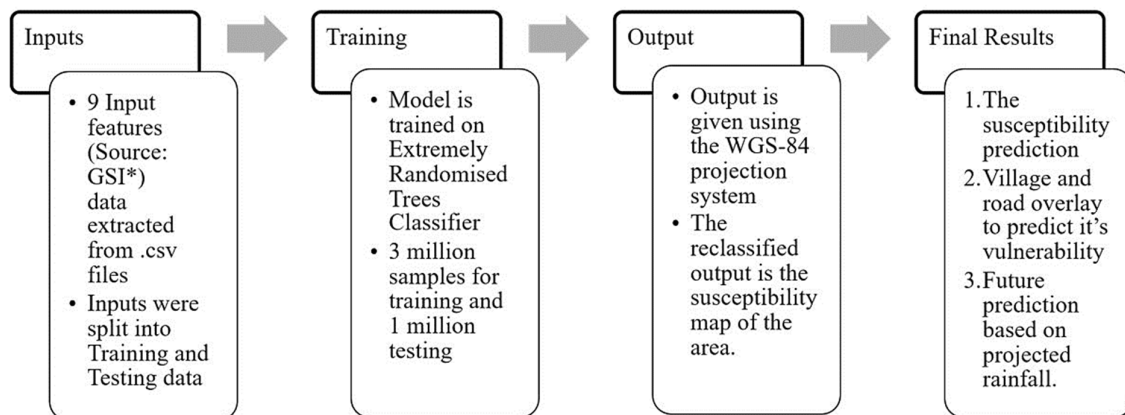


Figure 2. Models used in the study. Steps used in the present study (Extremely Randomised Trees; Overall steps remain same for all models)

1. Random Forest

Random Forest model is a combination of several decision tree architectures so that an individual tree depends on the values of a random vector independently sampled with the same distribution applied across all the trees in the forest. The overall error of the entire model usually converges as the number of trees in the forest increases. A Random Forest also has the variance reduced as compared to a decision tree due to multiple decision tree structures being present. The overall accuracy rate of a Random Forest Model is high due to the random splitting of nodes in each tree on the basis of proximities and Out of Bag (OoB) (Breiman, 2001). However, a Random Forest model sometimes suffers from the problem of bias, as the factor which results in the least error is often advanced with in progressing the tree node.

2. Extremely Randomised Trees

A single decision tree in an extremely randomised trees model works under the following framework (Merghadi *et al.*, 2020):

- i) random selection of m features as a candidate set of splitting features;
- ii) within each of features F_i with $i \in 1, 2, \dots, m$ a single random cut-point is drawn, and the performance of this feature with this cut-point is evaluated based on specific criteria;
- iii) the m features are then paired with their randomly selected cut-points. The pair with the 'best' performing splitting feature and cut-point is chosen which adheres to the selected criteria.

There are two main criteria which are considered, namely the Gini criteria and the Entropy criteria. For the purposes of this study, the Entropy criteria yielded better results.

Extremely Randomised Trees model is an ensemble-based learning method which work on the same principles as a decision tree (Geurts *et al.*, 2006). However, on top of a regular decision tree, an Extremely Randomised Tree adds an additional layer of randomness into the model as compared to a Random Forest model, and the individual nodes of each tree are split randomly, instead of quality, as is done in a Random Forest model. It is composed of several different decision trees, whose final results are then averaged out to reduce the bias. The problem of bias is a common occurrence in decision trees, however that issue is in large part reduced in an Extremely Randomised Trees model. The randomness in the selection of the node features and the averaged-out value of each decision tree works towards reduction in bias and variance. Furthermore, bootstrap value is false, making use of the entire dataset to build each tree. Alongside a proper trade-off between variance and bias, there is also computational efficiency to be considered in the context of this study on natural disasters. The selection of each node was based on the criteria of entropy, which the study notes yielded much lower loss than other criteria.

Rainfall Data for year 2030 is predicted by fitting a linear trend using FORECAST (x, known_ys, known_xs) formula in excel.

- x- The x value data point to use to calculate a prediction.
- known_ys - The dependent array or range of data (y values). Here Annual Avg. Rainfall data in ten years up to 2020 is considered.
- known_xs - The independent array or range of data (x values). Here tenth year from 1910 to 2020.

Algorithm of the model

The algorithm of the extremely randomised trees model is as follows:

1. Data Processing: The TIFF files of the thematic layers are converted into CSV files using the GDAL library. The data is then converted to arrays containing their values.

2. Input Creation: The combined arrays are then arranged in a combined array. This is then transposed in order to fit the model. The data then is split into training and testing.

3. Model Training:

The model is then trained on 3 million samples. The testing data consists of 1 million samples.

4. Parameter Tuning: The parameters were experimented on.

The optimal parameters were found to be:

estimators: 300

max_depth: 20

criterion: entropy

5. Obtaining Results:

The AUC, model score, feature importance and Mean Absolute Error values were then obtained. The testing data was predicted and the future prediction was also done.

Feature importance is calculated as follows:

$$f_i = \frac{\sum_{j: \text{node } j \text{ splits on feature } i} n_j}{\sum_{k \in \text{all nodes}} n_k}$$

- f_i = the importance of feature i
- n_j = the importance of node j

Normalized Feature Importance:

$$\text{norm}f_{ij} = \frac{f_j}{\sum_{j \in \text{all features}} f_j}$$

Final computation of feature importance:

$$RFfi_i = \frac{\sum_{j \text{ all features}} normfi_{ij}}{T}$$

- $RFfi_i$ the importance of feature i calculated from all trees in Random Forest Model
- $normfi_{ij}$ = the normalised feature importance for i tree j item
- T = Total number of trees

6. Visualising the Results:

The results of the testing data and prediction was visualized using Rasterio module in python. The final reclassification however was done in GIS (Figure 3).

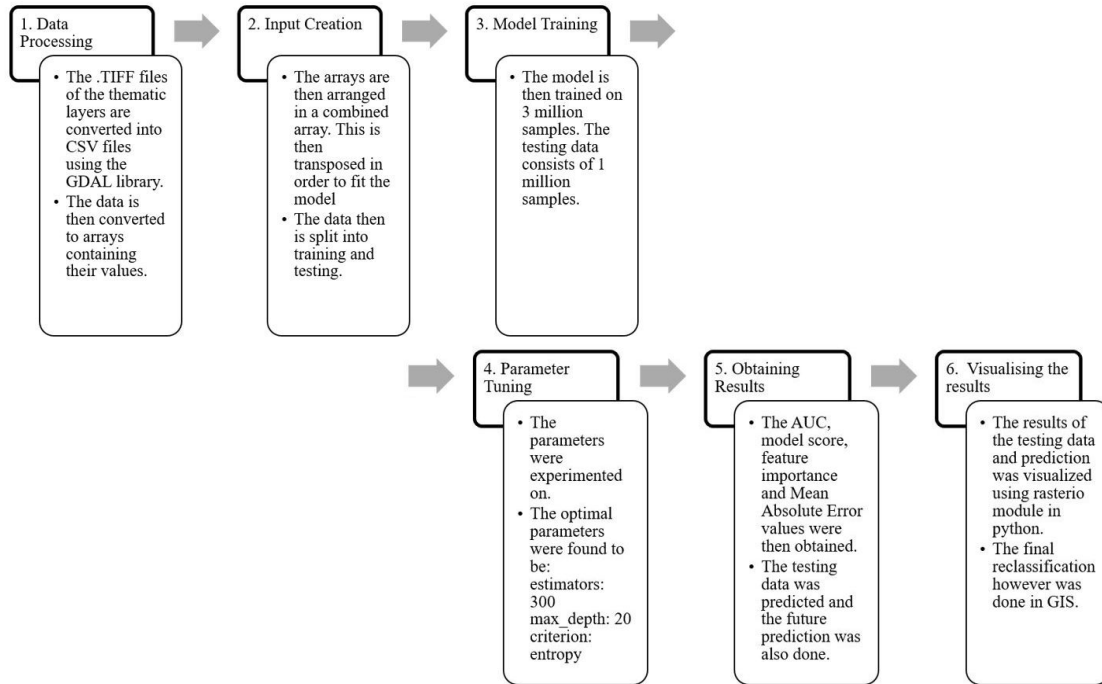


Figure 3. Algorithm of the model (for Extremely Randomised Trees)

Results and Discussion

There were three main results: model generated susceptibility of Random Forest, Extremely Randomised Trees, and the features importance as ascertained by the models in python. The statistical learning susceptibility map of the area has been shown in Figure 4A. The areas marked in red are zones of high susceptibility to landslides; those in yellow are moderately susceptible; and those in green are areas of low susceptibility. The results of the Extremely Randomised Trees Classifier are shown in Figure 4B. In contrast to the Random Forest Classifier Figure 4C, this model designates more areas to the highly susceptible group. The Random Forest Classifier results are shown in Figure 4C. The study notes that the model deems a major portion of the study area as moderately susceptible. This is a diversion from the statistical learning model, where the moderate and highly susceptible areas are limited

The model worked with 300 estimators at a depth of 20 with nine input features (Figure 4B). The feature importance as calculated by the model is shown in the Figure 5.

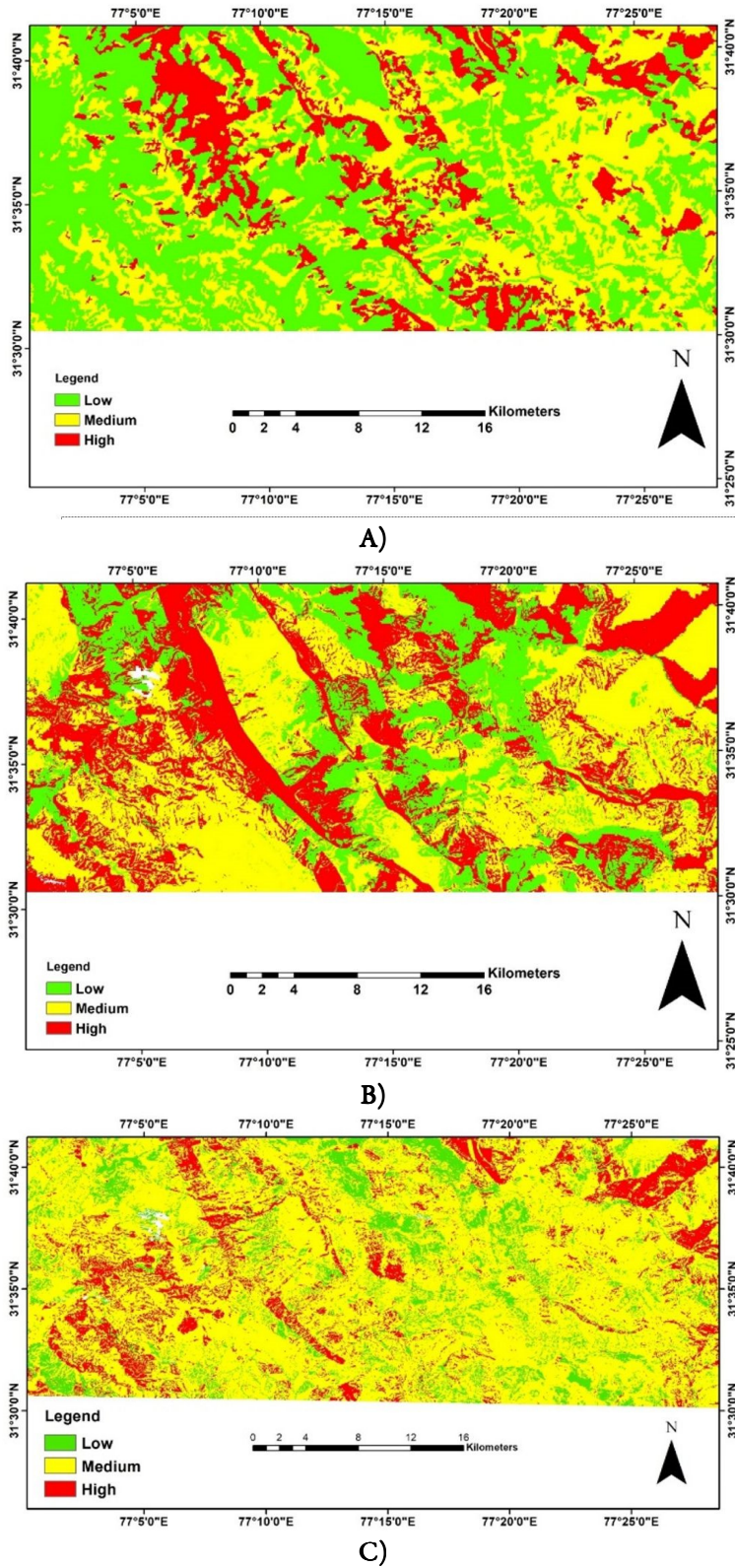


Figure 4. Landslide susceptibility assessment of the study area given by (A) Statistical Models; (B) Extremely Randomised Trees; (C) Random Forest Model

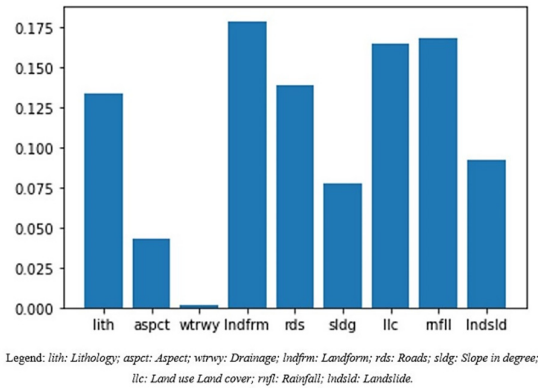


Figure 5. Feature importance as calculated by the model. The machine learning model designates landforms to be the most influential factor followed by rainfall and land use land cover

Feature importance is the decrease in node impurity weighted by the probability of reaching that node. The node probability is calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature. Node Importance for each tree is calculated using Gini Importance. The accuracy of the statistical learning model is about 72%, this means that the areas it designated as susceptible contained about 145 out of 202 previous landslide locations in the area (Figure 4A). The accuracy of the Random Forest Regressor model was about 85.14%, this means that the areas it designated as susceptible included about 187 out of 202 previous landslide locations in the area (Figure 4C).

The Extremely Randomised Trees model has an accuracy of 79.78%. This means that the areas it designated as susceptible contained about 161 out of 202 (79.78%) of active landslides in the area. This accuracy metric was deemed appropriate as the landslide points which occurred in the past are established ground truths, thus the model’s validity is assured. It performed better than conventional learning methods as the areas marked as susceptible by conventional learning contained only 71.78% of active landslides as shown in Table 2. Table 2 shows the full breakdown of all 3 models.

Table 2. Comparison between conventional learning and model prediction

Susceptibility Category	No. of active landslides in conventional learning	No of active landslides in Random Forest Regressor Model	No of active landslides in Extremely Randomised Trees Classifier
Low	57	15	41
Moderate	43	157	44
High	102	30	117
Total Landslides	202	202	202
Accuracy	71.78%	85.14%	79.70%

The study notes that while the Random Forest Regressor has a higher accuracy overall, the Extremely Randomised Trees is a much better evaluation of the study area’s susceptibility. This is because while the Random Forest model may have more previous landslide points under its susceptible classes, the majority of them lie in the moderately susceptible zone. The Extremely Randomised Trees model on the other hand contains most of the previous landslide under the high susceptibility areas.

The villages, active landslides, and classified roads as per vulnerability to landslides projected on the Random Forest model generated susceptibility map has been shown in Figure 6A. The villages, active landslides and classified roads as per vulnerability to landslides projected on the Extremely Randomised trees model generated susceptibility map has been shown in Figure 6B. The landslide susceptibility map predicted for 2030 is shown in Figure 6C. The study notes that the prediction for 2030 has significantly more areas designated to high susceptibility. The areal statistics comparison of different classes is tabulated in Table 3.

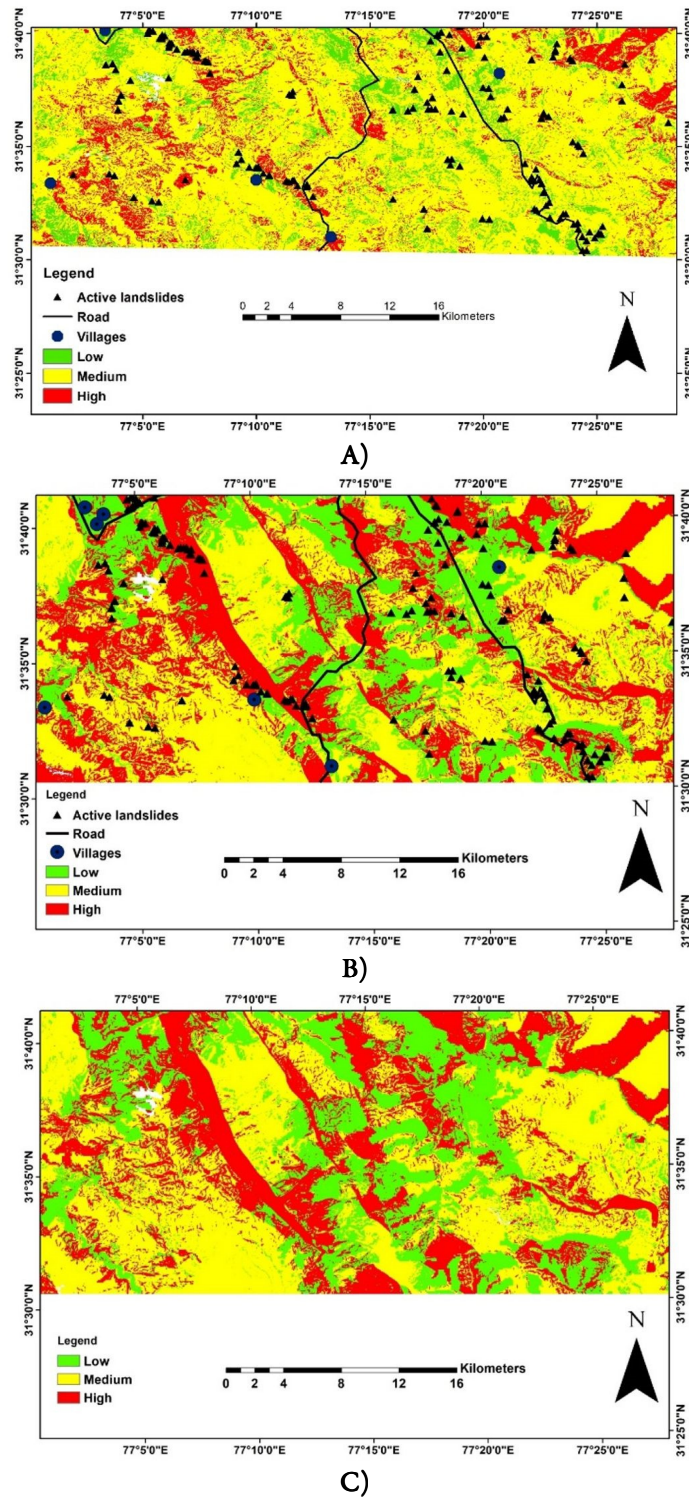


Figure 6. Active landslide and village projection, and predicted susceptibility analysis using extremely randomised trees: (A) Active landslide and village projection on Random Forest Analysis; (B) Active landslide and village projection on Extremely Randomised Trees Analysis; the study notes that at least four villages remain in high susceptible areas to landslides; (C) Predicted Susceptibility Analysis for 2030 using Extremely Randomised Trees

Table 3. Comparison of areal coverage of different classes

Class	Statistical Learning Susceptibility		Extremely Randomised Trees Model Generated Susceptibility		Random Forest Regressor Susceptibility	
	Area	Percentage	Area	Percentage	Area	Percentage
Low	435.98	47.10	217.86	23.54	115.62	12.49
Moderate	332.77	35.95	409.82	44.28	679.10	73.37
High	155.52	16.80	294.66	31.84	127.45	13.76
Unclassified	1.31	0.14	3.24	0.35	3.41	0.37
Total	925.58	100.00	925.58	100.00	925.58	100.00

Area in sq.km; Predicted susceptibility is for the year 2030.

It can be seen that the model performs better as the area shown with low susceptibility in statistical learning is actually an active landslide zone. The model accurately predicts it to be a high susceptibility zone. The usage of machine learning models for susceptibility mapping is not as common as those of opinion driven or statistical learning models. However, there is also a sceptical eye cast on opinion driven models, or human made outputs. This particular project has a novel mixed method approach involving human inputs compiled in GIS and an output composed by Extremely Randomised Trees Classifier to generate a susceptibility map of the study area. This significantly reduces the time consumed compared to the conventional method of susceptibility mapping. The Extremely Randomised Trees Classifier was considered optimal for the skeleton model as it reduces the risk of overfitting as can be seen in models involving decision trees and has the best trade-off between low bias and low variance. Furthermore, the entropy criterion further helps in reducing variance. Overall, the model has shown to have better performance compared to other model architectures such as decision trees, random forests etc.

The results of the model have been used to categorise villages by levels of susceptibility to landslide (Table 4). Taking a buffer area around predicted landslide sites, settlements have been categorised into low, moderate and high levels of susceptibility (Figure 4C). Such an exercise can be further scaled-up for effective disaster preparedness and relief.

Table 4. Villages as per susceptibility type based on 2030 prediction

SL No	Name	Susceptibility type
1	Pandoh	Low
2	Uba	Moderate
3	Chachiot	Moderate
4	Thunag	High
5	Janjheli	Moderate
6	Banjar	Low
7	Pandoh	Low

It is observed that out of seven villages three villages each fall into low and moderate susceptibility where as one village Thunag falls into high category of susceptibility. The high susceptibility of Thunag village is because of its proximity to active landslides and road apart from the other factor slope and landform.

Conclusions

Although there are a variety of landslide susceptibility models available, they are rarely accessible, inexpensive, or simple to use by non-experts. As a result, the goal of this research was to create landslide susceptibility model that were both sustainable and practicable, using open-source geographic data and machine learning. The results showed that this application can forecast landslide susceptibility and that there is a significant relationship between model outputs and past landslides. A real-time evaluation of the situation, is made possible by this application's early warning system, which may help to determine the village specific severity based on susceptibility. The specificity of the location of individual villages is much more accurate in GIS which further adds to its credibility as a useful tool in mitigation strategies. Furthermore, the model can be used to obtain the future predicted landslide projections. Identification of most vulnerable settlements would further improve the capacity of the government and civil society institutions to respond to disasters. Factoring the degree of susceptibility at the micro-level would help in designing effective bottom-up plans for mitigation. It would make further developmental plans much easier, which is a crucial benefit and necessity for high altitude areas that describes the most vulnerable places and sections of roads affected. However, before using these technologies for disaster relief applications, more testing of regional compatibility and applicability is necessary. Machine Learning is ideal for landslide susceptibility studies and land slide monitoring is likely to become more widespread in the future decades as it can be applied to any landscape. The application of the model can be used across any terrain.

Declaration

The author has declared no conflict of interest for this publication. The author read and approved the final manuscript.

Acknowledgement

The paper has been presented at the National Symposium on i-GEOMATICS hosted by Indian Society of Geomatics (ISG,) Ahmedabad and Indian Society of Remote Sensing (ISRS), Indian Space Research Organisation (ISRO), India and received the second-best poster presentation award. The author greatly acknowledges the participants and jury of the conference for their valuable suggestions and comments.

References

- Aleotti P, Chowdhury R (1999). Landslide hazard assessment: summary review and new perspectives. *Bulletin of Engineering Geology and the Environment* 58(1):21-44. <https://doi.org/10.1007/s100640050066>
- Breiman L (2001). *Machine Learning* 45(1):5-32. <https://doi.org/10.1023/a:1010933404324>
- Camilo DC, Lombardo L, Mai PM, Dou J, Huser R (2017). Handling high predictor dimensionality in slope-unit-based landslide susceptibility models through LASSO-penalized Generalized Linear Model. *Environmental Modelling & Software* 97:145-156. <https://doi.org/10.1016/j.envsoft.2017.08.003>
- Chae B-G, Park H-J, Catani F, Simoni A, Berti M (2017). Landslide prediction, monitoring and early warning: a concise review of state-of-the-art. *Geosciences Journal* 21(6):1033-1070. <https://doi.org/10.1007/s12303-017-0034-4>
- Chang K-T, Merghadi A, Yunus AP, Pham BT, Dou J (2019). Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-based machine learning techniques. *Scientific Reports* 9(1):1-21. <https://doi.org/10.1038/s41598-019-48773-2>
- Dou J, Yunus AP, Bui DT, Merghadi A, Sahana M, Zhu Z, Chen C-W, Han Z, Pham BT (2020). Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan. *Landslides* 17(3):641-658. <https://doi.org/10.1007/s10346-019-01286-5>

- Du J, Glade T, Woldai T, Chai B, Zeng B (2020). Landslide susceptibility assessment based on an incomplete landslide inventory in the Jilong Valley, Tibet, Chinese Himalayas. *Engineering Geology* 270:105572. <https://doi.org/10.1016/j.enggeo.2020.105572>
- Fell R, Corominas J, Bonnard C, Cascini L, Leroi E, Savage WZ (2008). Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. *Engineering Geology* 102(3-4):85-98. <https://doi.org/10.1016/j.enggeo.2008.03.022>
- Formetta G, Rago V, Capparelli G, Rigon R, Muto F, Versace P (2014). Integrated physically based system for modeling landslide susceptibility. *Procedia Earth and Planetary Science* 9:74-82. <https://doi.org/10.1016/j.proeps.2014.06.006>
- Geurts P, Ernst D, Wehenkel L (2006). Extremely randomized trees. *Machine Learning* 63(1):3-42. <https://doi.org/10.1007/s10994-006-6226-1>
- Jung BC (2021). *Disaster by choice. How our actions turn natural hazards into catastrophes* Ilan Kelman Oxford, UK: Oxford University Press, 2020. ISBN: 9780198841340. *World Medical & Health Policy* 14(2): 445-446. <https://doi.org/10.1002/wmh.3.452>
- Kalsnes B, Nadim F (2012). SafeLand: Changing pattern of landslides risk and strategies for its management. In: Sassa K, Rouhban B, Briceño S, McSaveney M, He B (Eds). *Landslides: Global Risk Preparedness*. Springer, Berlin, Heidelberg pp 95-114. https://doi.org/10.1007/978-3-642-22087-6_7
- Korup O, Stolle A (2014). Landslide prediction from machine learning. *Geology Today* 30(1):26-33. <https://doi.org/10.1111/gto.12034>
- Kuniyal JC, Jamwal A, Kanwar N, Chand B, Kumar K, Dhyani PP (2019). Vulnerability assessment of the Satluj catchment for sustainable development of hydroelectric projects in the northwestern Himalaya. *Journal of Mountain Science* 16(12):2714-2738. <https://doi.org/10.1007/s11629-017-4653-z>
- Lee S, Hong S-M, Jung H-S (2017). A support vector machine for landslide susceptibility mapping in Gangwon Province, Korea. *Sustainability* 9(1):48. <https://doi.org/10.3390/su9010048>
- Liu C, Fan X, Zhu C, Shi B (2019). Discrete element modeling and simulation of 3-Dimensional large-scale landslide-Taking Xinmocun landslide as an example. *Journal of Engineering Geology* 27(6):1362-1370. <https://doi.org/10.13544/j.cnki.jeg.2018-234>
- Lu P, Rosenbaum MS (2003). Artificial neural networks and grey systems for the prediction of slope stability. *Natural Hazards* 30(3):383-398. <https://doi.org/10.1023/b:nhaz.0000007168.00673.27>
- Merghadi A, Yunus AP, Dou J, Whiteley J, ThaiPham B, Bui DT, Avtar R, Abderrahmane B (2020). Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Science Reviews* 207:103225. <https://doi.org/10.1016/j.earscirev.2020.103225>
- Nakileza BR, Nedala S (2020). Topographic influence on landslides characteristics and implication for risk management in upper Manafwa catchment, Mt Elgon Uganda. *Geoenvironmental Disasters* 7(1):27. <https://doi.org/10.1186/s40677-020-00160-0>
- NLRMS (2019). *National Landslide Risk Management Strategy (2019)*. National Landslide Risk Management Strategy. A publication of the National Disaster Management Authority, Government of India, New Delhi. <https://nidm.gov.in/PDF/pubs/NDMA/26.pdf>
- Pham BT, Pradhan B, Tien Bui D, Prakash I, Dholakia MB (2016). A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). *Environmental Modelling & Software* 84:240-250. <https://doi.org/10.1016/j.envsoft.2016.07.005>
- Sufri S, Dwirahmadi F, Phung D, Rutherford S (2020). Enhancing community engagement in disaster early warning system in Aceh, Indonesia: opportunities and challenges. *Natural Hazards* 103(3):2691-2709. <https://doi.org/10.1007/s11069-020-04098-2>
- Sun D, Gu Q, Wen H, Shi S, Mi C, Zhang F (2022). A hybrid landslide warning model coupling susceptibility zoning and precipitation. *Forests* 13(6):827. <https://doi.org/10.3390/f13060827>
- Tien Bui D, Shirzadi A, Shahabi H, Geertsema M, Omidvar E, Clague J, Thai Pham B, Dou J, Talebpour Asl D, Bin Ahmad B, Lee S (2019). New ensemble models for shallow landslide susceptibility modeling in a semi-arid watershed. *Forests* 10(9):743. <https://doi.org/10.3390/f10090743>
- Turner AK (2018). Social and environmental impacts of landslides. *Innovative Infrastructure Solutions* 3:70. <https://doi.org/10.1007/s41062-018-0175-y>
- Uniyal A (2017). *National Landslide Risk Management Strategy*.

- Van Westen CJ, Seijmonsbergen AC, Mantovani F. (1999). Comparing landslide hazard maps. *Natural Hazards* 20:137-158. <https://doi.org/10.1023/A:1008036810401>
- Watson DF, Philip GM (1985). A refinement of inverse distance weighted interpolation. *Geoprocessing* 2(4):315-327.
- Wickramasinghe D (2021). Ecosystem-based disaster risk reduction. *Oxford Research Encyclopedia of Natural Hazard Science*. <https://doi.org/10.1093/acrefore/9780199389407.013.360>
- Wilkinson PL, Anderson MG, Lloyd DM (2002). An integrated hydrological model for rain-induced landslide prediction. *Earth Surface Processes and Landforms* 27(12):1285-1297. <https://doi.org/10.1002/esp.409>
- Xie M, Esaki T, Cai M (2004). A time-space based approach for mapping rainfall-induced shallow landslide hazard. *Environmental Geology* 46(6-7):840-850. <https://doi.org/10.1007/s00254-004-1069-1>
- Yunus AP, Dou J, Song X, Avtar R (2019). Improved bathymetric mapping of coastal and lake environments using Sentinel-2 and Landsat-8 Images. *Sensors* 19(12):2788. <https://doi.org/10.3390/s19122788>



The journal offers free, immediate, and unrestricted access to peer-reviewed research and scholarly work. Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the articles, or use them for any other lawful purpose, without asking prior permission from the publisher or the author.



License - Articles published in **Nova Geodesia** are Open-Access, distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) License.

© **Articles by the authors**; Licensee **SMTCT**, Cluj-Napoca, Romania. The journal allows the author(s) to hold the copyright/to retain publishing rights without restriction.

Notes:

- **Material disclaimer:** The authors are fully responsible for their work and they hold sole responsibility for the articles published in the journal.
- **Maps and affiliations:** The publisher stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.
- **Responsibilities:** The editors, editorial board and publisher do not assume any responsibility for the article's contents and for the authors' views expressed in their contributions. The statements and opinions published represent the views of the authors or persons to whom they are credited. Publication of research information does not constitute a recommendation or endorsement of products involved.