AN EXPLORATORY ATTEMPT TO MAP LANDSLIDE SUSCEPTIBILITY OF ARUNACHAL PRADESH USING FREQUENCY-RATIO AND RANDOM FOREST BASED MODELLING

Shubham Maurya PwC India, Mumbai, India, shubh07maurya@gmail.com
Manohara K N Department of Civil Engineering, IIT Guwahati, n.manohara@iitg.ac.in
Arindam Dey Center for Disaster Management and Research, IIT Guwahati, arindam.dey@iitg.ac.in
Rishikesh Bharti Center for Disaster Management and Research, IIT Guwahati, r.bharti@iitg.ac.in

ABSTRACT: Landslides are among the most destructive geohazard all over the world. For systematic landslide evaluation and management, a reliable hazard zonation is required. In this study, the rainfall-induced landslide susceptibility of the entire state of Arunachal Pradesh is investigated through the implication of Random Forest (RF), a machine learning (ML) model. In this regard, various landslide conditioning factor and the importance of considering multiple conditioning factors in the mapping process is highlighted. These conditioning factors for landslides, i.e., topographic factors (slope, aspect, curvature, elevation), geologic factors (structural geology, soil type), climatic factors (precipitation) and land use land cover factors (vegetation cover, land use, human activity) are prepared from multiple sources. The RF models were trained and validated with the help of 212 historical landslide events, along with an equal number of non-landslide events, which were divided into training (70%) and validation (30%) sets. The study revealed that factors such as elevation, rainfall, slope gradient, stream power index (SPI) and land use have a significant influence on the spatial distribution of the landslides. The results of the analysis have been validated by various statistical indices such as area under curve (AUC), root mean square error (RMSE) and Kappa coefficient. The results of this study culminated in landslide susceptibility mapping (LSM) that are produced for the peak monsoon month i.e. July. The findings offer valuable insights into the development of more efficient landslide predictive models that can be used by decision-makers and land-use managers to mitigate landslide hazards.

Keywords: Landslide susceptibility mapping (LSM), Landslide conditioning factors, Frequency-ratio method, Random Forest model

INTRODUCTION

Landslides pose a severe threat to human life, infrastructure, and the environment, causing numerous casualties and financial losses every year [1]. The costs associated with landslides can be significant, as they may require extensive recovery efforts and rebuilding. Furthermore, the effects of landslides can extend beyond immediate economic losses, such as environmental degradation, loss of biodiversity, and disruption of ecosystems. The recognition of landslides as a major hazard underscores the need for effective prevention and mitigation strategies to minimize their impact on society and the environment [2,3]. Although it is impossible to completely prevent landslides, it is possible to limit their effects through prediction and the implementation of appropriate measures. This can be achieved through reliable monitoring and mapping of landslide-prone areas. Therefore, it is crucial to use reliable methods and analyses to identify and mitigate the risks associated with landslides [4].

With the help of scientific analysis, it is now possible to map areas that are susceptible to landslides and predict when a landslide might occur [5]. Landslide susceptibility mapping (LSM) is a method that has gained increasing attention in recent years as a means of assessing landslide risk and hazard. LSM allows the mapping of areas that are susceptible to landslides, which is crucial for predicting the spatial and temporal occurrence of landslides. LSM has now become a popular land use management technique due to its ability to provide valuable support for decision-making by land managers. The objective of conducting landslide analysis in Arunachal Pradesh is to develop a better understanding of the factors that contribute to landslide occurrence and to identify areas that are at high risk of landslides. In Arunachal, each year, the number of landslides getting recorded is increasing rapidly, and so is the loss of lives. The region's humid and wet tropical and subtropical climate, along with steep slopes, rugged terrain, and diverse geology, create the ideal environment for the top layer of soil to disintegrate. Additionally, the soil degradation caused by Jhum cultivation is a significant environmental concern in Arunachal Pradesh [6]. The region's frequent landslides have emphasized the importance of accurately identifying the areas that are prone to such incidents, leading to a demand for planners, policymakers, and land managers to create precise maps of these locations. Nowadays, data-driven techniques, such as machine learning, are being increasingly used in the scientific and engineering fields to address a variety of problems. One area where these methods have recently shown notable success is in generating maps that indicate the likelihood of landslides occurring in a given area. These techniques are particularly useful in situations where there is limited geotechnical data available and can help to map large regions that are at risk of landslides. Comparatively less research has been done on landslide assessment in the Arunachal Pradesh region, despite the fact that this region is known for its complex geological setting and high landslide susceptibility. In this study, the use of random forest, a machine learning model, in landslide assessment over the Arunachal Pradesh region is explored.

STUDY AREA AND DATA USED

Study Area

Arunachal Pradesh is a north eastern state of India, located in the foothills of the eastern Himalayas. The state extends from 26.28 ° N and 91.20 °E to 29.30 °N and 97.30 °E, with an area of approximately 83,743 km^2. It is located in the easternmost part of the country and shares its borders with Bhutan, China, and Myanmar. The elevation range in Arunachal Pradesh varies significantly, from approximately 100 m above sea level in valleys to over 7,000 m above sea level in its towering mountain peaks as illustrated in Fig. 1. The state experiences heavy rainfall throughout the year, with the monsoon season spanning from June to September. The average annual rainfall experienced is around 3,000 mm, with some areas receiving as much as 4,000 mm of rainfall. Maximum rainfall occurs in July, while December sees the least rainfall. Arunachal Pradesh is characterized by rugged terrain and abundant natural resources. Unfortunately, this challenging landscape increases the vulnerability to natural disasters, specifically landslides.

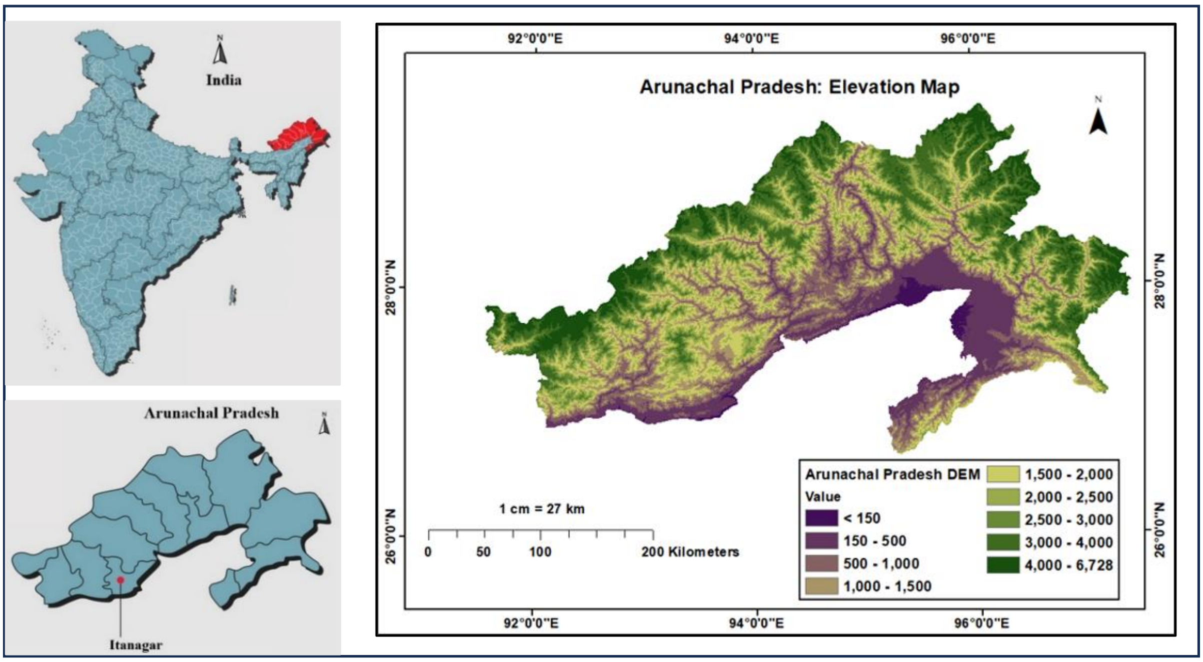

Fig. 1 Location of the study area.

Data Used

In this comprehensive study, the SRTM DEM data with a resolution of 30 m was derived from the USGS Earth Explorer website. Landslide data, crucial for the analysis, was obtained from two distinct sources: the NASA Global Precipitation Measurement Mission and Bhukosh Geological Survey of India (GSI). Soil type information, a key component in understanding susceptibility, was gathered from the FAO/UNESCO Soil Map of the World. To accurately delineate the study area, the state boundary shape file was sourced from DIVA GIS, complemented by data from Indian Remote Sensing and GIS sources. Precise rainfall data, fundamental for landslide risk assessment, was acquired from the India Water Resources Information System (WRIS). Furthermore, geological insights vital for the study were derived from Bhukosh GSI, contributing to a comprehensive understanding of the region's terrain. The land use and land cover data, crucial for contextualizing environmental factors, were acquired with a resolution of 30 m from the Bhuvan NRSC website.

Landslide Inventory

In landslide susceptibility mapping (LSM) studies, the development of a comprehensive landslide inventory, encompassing spatial parameters, geometry, material composition, and movement characteristics, is a fundamental and essential task. This process operates under the assumption that the conditions leading to past landslides mirror those conducive to future occurrences. The study area encompasses a total of 212 landslide points, along with an equal number of non-landslide points that were meticulously delineated from a preliminary landslide susceptibility map of the designated study area generated using the Frequency Ratio method. This entire dataset was systematically partitioned into two subsets: a training set comprising 70% of the data and a validation set with the remaining 30%. Various landslide conditioning factors were judiciously incorporated, encompassing topographic elements like slope, aspect, curvature, and elevation. Geologic factors, including structural geology and soil type, were integrated, alongside climatic considerations such as precipitation. Furthermore, land cover factors, comprising vegetation cover, land use, and human activity, were derived from diverse and reliable sources.

Landslide Conditioning Factors

Landslide conditioning factors (LCFs) constitute the array of physical and environmental conditions influencing the occurrence of landslides within a given area. This study employs 12 distinct conditioning factors to comprehensively assess the potential for landslides. The integration of Geographic Information System (GIS) technology facilitated the seamless amalgamation and conversion of diverse data types at a pixel size of 30 m x 30 m. Ensuring standardization and analytical consistency, the data was georeferenced in the UTM coordinate system Zone 46 with a datum of WGS 84. Within landslide susceptibility studies, Landslide Conditioning Factors (LCFs) are systematically categorized as either nominal or ordinal, each offering distinct insights. Nominal LCFs encompass categorical data lacking inherent order or ranking, such as geology, soil type, slope aspect, and land use/land cover. Conversely, ordinal LCFs involve continuous data, such as altitude, slope angle, plan curvature, and profile curvature, where the order holds significance. Integrating both types of LCFs enhances the study's ability to furnish a comprehensive understanding of landslide potential in a specific area. The utilization of ordinal LCFs proves particularly valuable, providing detailed information that identifies areas more susceptible to landslides. This information can be useful in developing appropriate mitigation measures to reduce the risk of landslides in the future. In this study, landslide condition factors are classified into five types based on their impact and origin.

Topographic Factors

Digital Elevation Models (DEM): Digital Elevation Models (DEMs) is an imperative tool in landslide studies due to their ability to provide a precise digital representation of terrain elevation. These digital models provide three-dimensional visualizations of terrain features that are essential inputs for landslide susceptibility and hazard assessments. Digital Elevation Models (DEMs) are a critical component of landslide susceptibility mapping because they provide information about the elevation and slope of the terrain [7,8].

Slope: Slope is a critical topographic attribute that plays a significant role in landslide occurrence and susceptibility [9]. Slope refers to the angle of inclination of a terrain surface relative to the horizontal plane. The slope of a terrain surface can influence various physical processes that can lead to slope failure, including soil erosion, rock weathering, and landslides [10]. The slope gradient ranges from 0° in flat areas to 84.5° in nearly vertical cliffs, indicating a wide variability across different regions of Arunachal Pradesh.

Aspect: Aspect refers to the orientation or direction that a slope is facing, measured in degrees clockwise starting from the north. It is an important factor in landslide studies as it affects the amount of solar radiation and moisture that a slope receives, which in turn can affect the stability of the slope and its susceptibility to landslides. The aspect of a slope can influence the distribution of vegetation [11,12]. In this study, the slope aspect was determined using digital elevation model (DEM) data and was classified into ten different categories based on their orientation.

Curvature: Curvature is an important factor in landslide assessment as it can affect the stability of a slope [13]. Curvature refers to the degree of change in slope angle or shape along a surface, and it can be measured in both the plan and profile directions [14,15]. The profile & plan curvature map in the study was divided into three classes: concave (< - 0.5), flat (- 0.5 – 0.5), and convex (> .05).

Topographic Position Index: The Topographic Position Index (TPI) is a statistical measure used in geographic information systems (GIS) and remote sensing to quantify the topographic position of a point relative to its surrounding landscape. TPI is calculated by subtracting the average elevation of a surrounding area from the elevation of a central point, resulting in a relative measure of topographic position [16,17].

Stream Power Index: Stream Power Index (SPI) can be used to assess the potential for landslides in a particular area. SPI can be used to quantify the erosive power of a stream or river, which can help predict the likelihood of landslides occurring in adjacent slopes [18,19]. Stream power index is calculated using formula given by Moore [20]**.** The calculation involved employing ArcGIS software, with a digital elevation model serving as the input data. The calculated SPI values in the study area ranges from -13 to 16.4, with more negative value indicating areas prone to flooding & higher positive value indicating greater chance to erosion.

Topographic wetness index (TWI): Topographic Wetness Index (TWI) is a terrain analysis parameter that is commonly used in hydrology and land surface studies to identify and quantify the spatial distribution of wetland areas. TWI is a dimensionless parameter that combines topographic and hydrological factors to describe the relative wetness or dryness of a landscape [21,22].

Geological Factors

Soil Type: The type of soil present in a particular area is a crucial element in assessing the probability and intensity of landslides. Different soil types have varying physical and mechanical properties, such as porosity, permeability, shear strength, and cohesion, which, in turn, affect the stability of slopes and the probability of landslides [23,24]. In Arunachal Pradesh, a soil map delineates six major types: 'Sandy clay loam', 'Loam', 'Clay', 'Sandy-Loam', 'Clay-Loam' and 'Glacier-mix'. Predominantly, the region is covered by Loamy Soil, succeeded by Clay Loam and Sandy Clay Loam.

Distance to Fault: The proximity of a fault can significantly impact the risk of landslides, as faults can create areas of weakness in surrounding rock or soil, alter drainage patterns, and increase erosion [14,25]. However, the importance of distance to fault in landslide risk assessment depends on various factors, such as the type of fault, slope characteristics, and surrounding terrain.

Geology: Geological factors, such as the rock type, orientation and structure, individually or in combination, can influence the susceptibility of a slope to landslides. Fractures and faults in rocks can also provide planes of weakness along which landslides can occur [26].

Environmental Factors

Land use and land cover: Land use and land cover play a crucial role in affecting the potential occurrence of landslides. Land use and land cover changes can alter the stability of slopes and increase the susceptibility to landslides [27]. Land use refers to the human activities that take place on a particular piece of land, such as agriculture, forestry, urbanization [28]. Land cover, on the other hand, refers to the physical characteristics of the land, such as vegetation, soil type, and topography.

Triggering Factors

Rainfall: Rainfall is a primary element that influences the susceptibility and occurrence of landslides [29]. Rainfall triggers moisture percolation contributing to the saturation of shallower layers, thereby reducing the shear strength of soil or rock masses and subsequently leading to the initiation and propagation of landslides. Rainfall-triggered landslides are prevalent in regions with steep slopes, loose soil, and heavy rainfall, and can cause extensive damage to property and infrastructure and even result in loss of life [30]. Rainfall intensity is another critical factor in landslide occurrence [31]. The current study utilizes rainfall data from the IWRIS (India Water Resources Information System) website, covering a period of 30 years (i.e. 1992-2022_. The study focused on the maximum cumulative rainfall that occurred in July as this is typically the wettest month in the region.

Other Possible Influential Factors

Distance to a Road: The distance to a road is a crucial factor in assessing the susceptibility and risk of landslides. Roads can have significant impacts on the stability of slopes, especially when they are constructed on steep terrain or cut into hillsides [14]. Roads can have both positive and negative impacts on landslide susceptibility and hazard. On one hand, roads can contribute to landslide occurrence by altering the natural drainage patterns and increasing the weight of the slope. Conversely, it can provide access to landslides for monitoring, mitigation, and emergency response activities.

Distance to Stream: Streams can contribute to landslide risk in multiple ways, making the distance to a stream an important factor in determining landslide susceptibility [14]. Generally, slopes located closer to streams have a higher risk of landslides, as streams can erode and saturate the slope, making it more susceptible to failure.

METHODOLOGY

This study employs a systematic and comprehensive methodology to assess landslide susceptibility, illustrated in the flowchart shown in Fig. 2. It begins by collecting diverse data, including Digital Elevation Model (DEM), satellite images, meteorological, lithological, and historical landslide records. This forms a comprehensive dataset for subsequent analyses. Next, the study integrates DEM data with various datasets to develop conditioning factors, enhancing understanding of terrain vulnerabilities. Using these factors and incorporating landslide point data, a preliminary Landslide Susceptibility Map (LSM) is constructed via the Frequency Ratio (FR) method. To improve accuracy, non-landslide points are strategically extracted from regions with low vulnerability. Further probability modelling through Random Forest (RF) utilizes both landslide and non-landslide points, contributing to nuanced susceptibility assessment. Lastly rigorous evaluation of the model's performance and prediction accuracy is undertaken, ensuring the reliability and validity of the final Landslide Susceptibility Mapping. This comprehensive and systematic approach positions the study as a valuable contribution to understanding and mitigating landslide risks in the considered study area. Random Forest approach offer improved accuracy and efficiency in landslide assessment by automating the analysis of large datasets of landslide conditioning factors. Unlike time-consuming statistical models that necessitate manual input and spatial analysis, such machine learning (ML) techniques automate the identification of relationships between dependent and independent variables [32]. This automation not only enhances the efficiency, but also saves time and effort in the process of assessment.

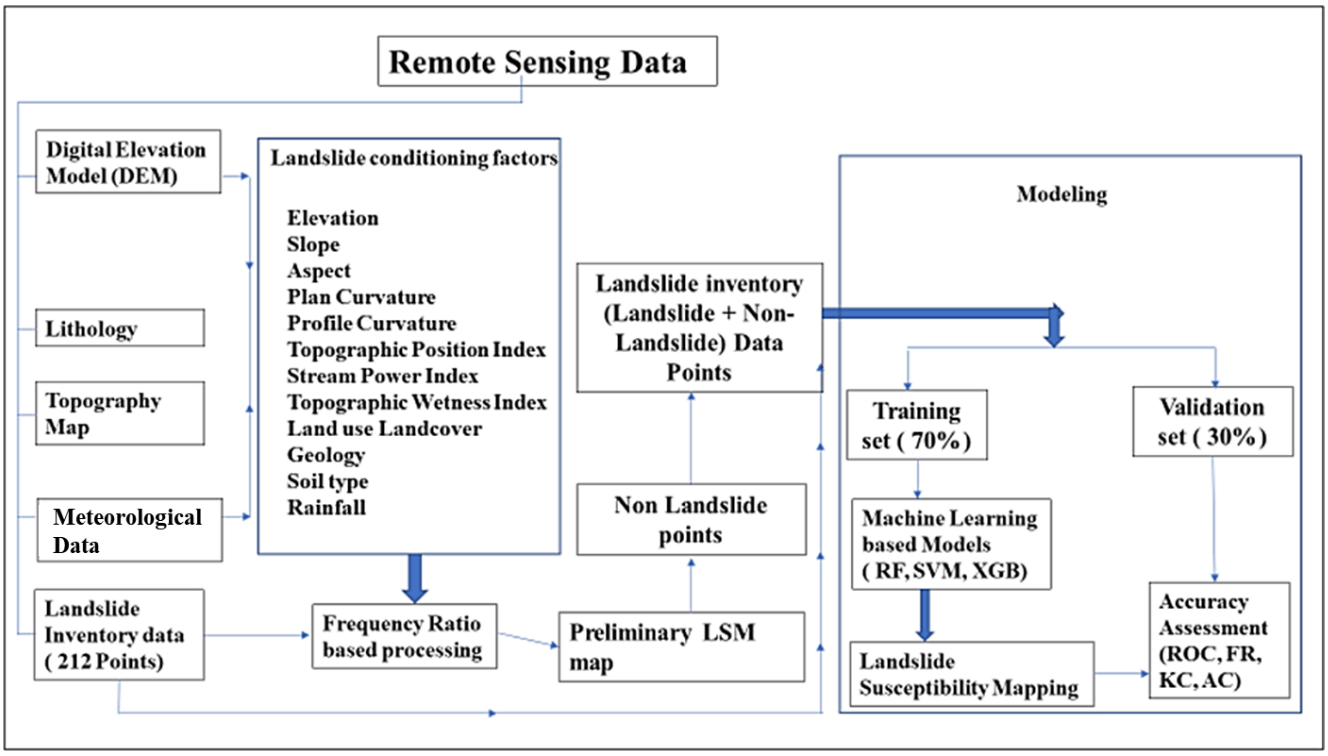
Remote Sensing Data
Digital Elevation Model (DEM)
Lithology
Topography Map
Meteorological Data
Landslide Inventory data (212 Points)
Landslide conditioning factors
Elevation
Slope
Aspect
Plan Curvature
Profile Curvature
Topographic Position Index
Stream Power Index
Topographic Wetness Index
Land use Landcover
Geology
Soil type
Rainfall
Frequency Ratio based processing
Preliminary LSM map
Non Landslide points
Landslide inventory (Landslide + Non-Landslide) Data Points
Modeling
Training set (70%)
Validation set (30%)
Machine Learning based Models (RF, SVM, XGB)
Landslide Susceptibility Mapping
Accuracy Assessment (ROC, FR, KC, AC)

Fig. 2 Work flow of the adopted research methodology

It is worth mentioning that the selection of non-landslide points is a crucial step in landslide susceptibility mapping, and various approaches have been used to address this issue. One commonly used approach is to select non-landslide points randomly throughout the study area, ensuring that they are distributed evenly across the landscape [33-35]. Researchers have also used a buffer zone approach [36-38], where a buffer zone of 100 to 500 m is created around the landslide points. The non-landslide points are then randomly generated by excluding this buffer zone, ensuring that they are away from the landslide points. Some researchers have also considered masking water bodies in the study area along with creating a buffer zone of 100 to 500 m around the landslide points and then randomly selecting the non-landslide points from the remaining area [39,40]. If the selection of non-landslide points is not made correctly, the resulting susceptibility map may not accurately reflect the areas that are truly susceptible to landslides. This can lead to inaccurate risk assessments and ineffective land management strategies. While the random selection of non-landslide points can be a good approach for small study areas, it may not be appropriate for larger study areas, as it may result in a biased representation of the landscape. The study area considered herein is quite large; the state of Arunachal Pradesh spans over 83743 km^2. For a larger study area, it may be necessary to use more sophisticated techniques for selecting non-landslide points, such as ‘stratified sampling’ or ‘systematic sampling’; however, they are out of context of the present study.

A unique approach has been followed in the present study. Firstly, various available and relevant techniques [41-43] are reviewed that can be used to construct a simple LSM map by using values corresponding to landslide points only. After conducting a thorough literature review, the Frequency Ratio (FR) method was chosen as it was found to perform well in similar studies [44]. This approach allowed for the identification of best possible non-landslide locations for a large study area without the need for additional data or complex techniques.

In the present study, several categorical variables such as land use and land cover, soil type, and geology are considered as landslide conditioning factors. However, since most random forest models work with numerical values, many of the qualitative categorical variables were transformed into numerical ones through a process called ‘encoding’. In this study, certain features, such as ‘Distance to Road’, ‘Distance to Stream’, and ‘Distance to Fault’, were excluded from the analysis. This decision was made due to anomalies in the data and limitations in the inventory of these features. For instance, it was found that if various faults were mislocated, it would generate notable inaccuracy in the analysis. Hence, even though these parameters have their own influence on the landslide potential, yet it was decided to drop these parameters from the analysis for generating preliminary LSM.

Preparation of Non-Landslide Points

Frequency Ratio Method

The frequency ratio (FR) method is a statistical approach that uses the frequency of landslides in a given area to assess the susceptibility of that area to future landslides. The method assumes that areas with high landslide frequency are more susceptible to landslides in the future [45]. The FR method involves dividing a study area into different cells and then calculating the ratio of the number of landslides that occurred in each cell to the total number of landslides in the study area. This ratio is then compared to the frequency of the specific factor or condition that is known to trigger landslides in that area [46]. The (FR) values are calculated as follows:

$$FR = (\mathbf{a} / A) / (\mathbf{b} / B) \qquad (1)$$

where FR is the frequency ratio, **a** is the total number of landslide pixels in a specific class of a factor, A is the total number of landslide occurrences in all classes of that factor, **b** is the total number of pixels containing landslide, and B is the total number of pixels in the study area.

Application of Frequency Ratio Method

In order to use the FR method, each factor is first converted into different classes. The ArcGIS software was then utilized to determine the number of cells (corresponding to both landslide and non-landslide areas). The FR for each factor's class or type was then obtained by dividing the landslide occurrence ratio by the area ratio, as shown in the above equation. This procedure allowed to assess the relative importance of each class of the factor in terms of its association with landslides. The resulting FR values for all landslide influence factors were calculated. Further, as proposed by Lee and Talib [47], the Landslide Susceptibility Index (LSI) was obtained by summing the factor ratings using Eq. (1).

$$LSI = \Sigma\ (FR_i) \qquad (2)$$

where $\mathbf{\Sigma}(FR_i)$ represents the sum of the FR values for each class of the various landslide influencing factors, and i represents the class number.

LSM Developed by the FR Method in the Study Area

The LSM map is generated using the LSI values calculated for each cell in the study area. This map can help identify areas with varying degrees of susceptibility to landslides. Non-landslide points were extracted from this map, assuming that they are located in areas with low susceptibility. This operation is performed in ARCGIS environment. Figure 3 displays the LSM map of Arunachal Pradesh created using the Frequency Ratio method.

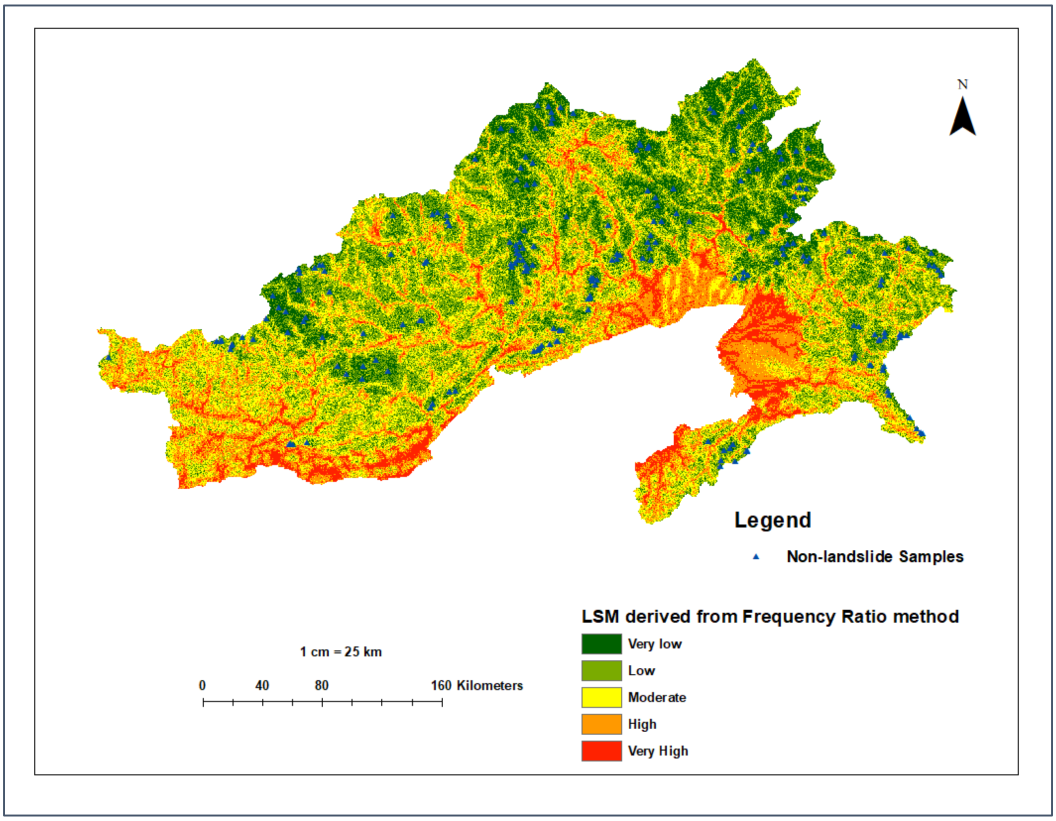

Fig. 3 Preliminary landslide susceptibility map of the study area obtained from the frequency ratio method

Random Forest (RF) Modelling

Random Forest (RF) is a powerful machine learning (ML) algorithm used for supervised classification and regression tasks. The algorithm is based on decision trees, which are a collection of if-then statements [48]. The main idea behind RF is to use the predictive power of different decision trees. Each decision tree is built independently of the others, i.e., with different subsets of features. The final prediction is based on the majority vote of all the decision trees [49]. Random forest can handle non-linear relationships between the features and the landslide occurrence. Overall, RF has proven to be a valuable tool in landslide assessment [50,51].

In this study, raster maps are georeferenced to WGS84 46N and converted into vector points. A table is created with X and Y coordinates, corresponding to the pixel values containing information on 12 landslide conditioning factors. The 'Extract All Multi-Points' function in ARCGIS is employed to generate a table with 424 rows and 15 columns. Each column represents a specific factor, such as such as landslide labels, coordinate information, each of the landslide conditioning factors, and thus streamlining the overall data for further analysis. The study transitions to a Jupyter notebook for data exploration, focusing on identifying correlations and addressing missing values. Within the notebook, the dataset is thoroughly explored to discern correlations among factors and handle missing values effectively. The dataset is then divided into two sets, with 70% of the data used for training and the remaining 30% used for validating and testing the model. The research employs random forest modelling, allowing for different complexity levels to determine effective landslide susceptibility mapping. Figure 5 shows the schematic working of the RF Algorithm.

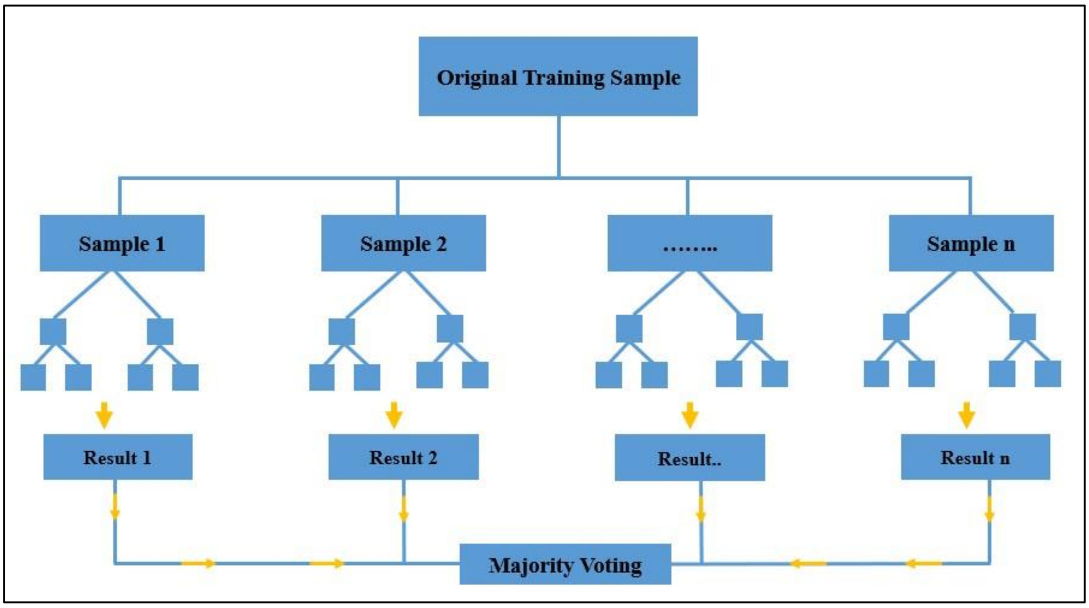

Fig. 4 Schematic working diagram of the RF algorithm

Modelling Prediction and Performance

The landslide susceptibility index (LSI) was calculated for every pixel in the study area based on the trained RF model. The LSI values were then classified into four susceptibility categories, namely 'low', 'moderate', 'high', and 'very high' susceptible zones. The performance of the prediction models can be evaluated using various metrics such as accuracy, sensitivity, specificity, precision, and area under the curve (AUC) [52]. The results obtained from landslide susceptibility modelling through Random Forest can provide valuable practical information for the selection of mitigation and remediation strategies. One of the key benefits is that the prediction capability of the model can be quantitatively assessed, which helps in understanding the accuracy of the model's predictions. Another significant advantage is the ability to assess the model's capability to differentiate the most landslide prone zones, and importantly, the less susceptible areas or safest zones. This knowledge can be useful for identifying areas that are at high risk of landslides and for developing effective mitigation and remediation strategies to prevent or minimize damage. In order to assess the predictive performance of the RF model, several metrics were utilized in the current study. Specifically, the receiver-operator characteristic-AUC (i.e. ROC-AUC), kappa coefficient, root mean square error (RMSE), mean absolute error (MAE), and test accuracy were applied. To ensure the reliability of the results, validation datasets (that were not used during the model training process) were employed to evaluate the predictive performance of the model.

RESULTS AND DISCUSSIONS

Variable Contribution Analysis

'Variable contribution analysis' or 'feature importance' is a key aspect of landslide assessment and modelling. Feature importance refers to the degree to which each input variable or feature contributes to the prediction of landslide susceptibility. The identification of important features is crucial for understanding the factors that contribute to landslide occurrence and for developing effective landslide risk management strategies. In this study, the inbuilt 'Feature importance' function of the Random Forest algorithm is used to find the importance of the landslide variables. The result shows that 'elevation' is the most dominant contributing factor, followed by SPI, slope, rainfall, aspect, built-up area, plan curvature, topographic position index (TPI), and followed by others. However, water (consisting of rivers and stream bodies), few soil types (clay loam soil and sandy clay loam soil), crops and glacier have less feature importance score, and thus are of lower importance in landslide susceptibility modelling of Arunachal Pradesh. Crops have very low feature importance due to the fact that the crops are mostly planted on relatively flat terrain where the soil is more stable and less prone to erosion as compared to other land use types.

Additionally, crops can help absorb excess water and reduce surface runoff through its root systems, which can reduce potentiality of landslides during heavy rainfall events. Sandy loam soil also have a very low feature importance score. This is attributed to that fact that sandy loam soil possesses the unique characteristics that allows it to quickly drain out the infiltrated rainwater, thereby reducing the scenarios of waterlogging and flooding and reducing the potential to landsliding as well. However, it is to be noted that these attributes are a part of feature importance because of encoding and are not explicitly separate conditioning factors such as 'slope' or 'elevation'.

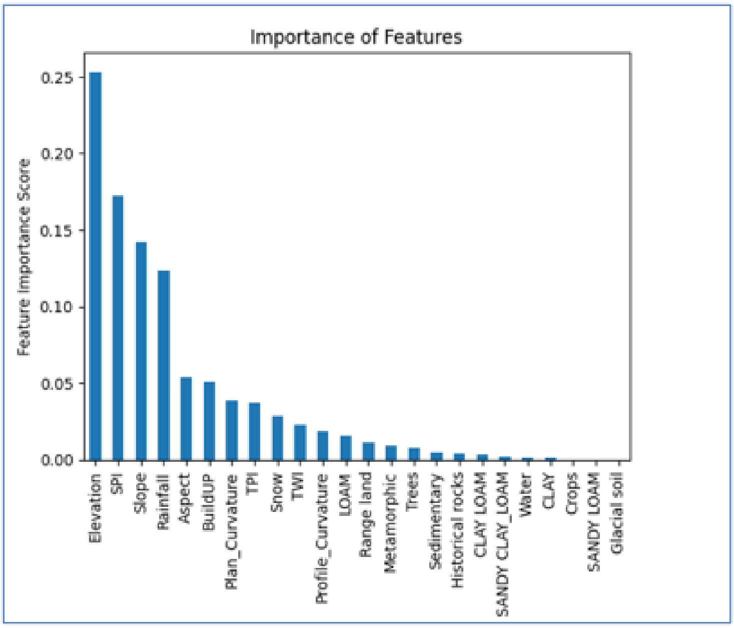

Fig 5. Feature importance based on Random Forest modeling

Accuracy Assessment and Comparison

The AUC curve is a widely used metric in model assessment and evaluates the performance of the model in terms of sensitivity and specificity. In addition to the AUC curve, the MSE (Mean Squared Error) and MAE (Mean Absolute Error) are used to assess the accuracy of the model's predictions. Lower values of MSE and MAE indicate better performance, while higher MSE & MAE indicates a worse fit of the model. A kappa value of 0.8 or higher indicates strong agreement between the predicted and actual values, while a value between 0.6 and 0.8 indicates moderate agreement, and a value less than 0.6 indicates poor agreement.

Accuracy Assessment of Frequency Ratio Method

The AUC curve was prepared in the ArcGIS environment where true positive and false positive points are used as input. True positive points represented the landslide points and false positive points represented the other remaining areas in LSM map. The AUC obtained by the FR method was 0.740. This value is in the reliable range and, hence, can be used for deriving out the non-landslide points from the constructed preliminary LSM map.

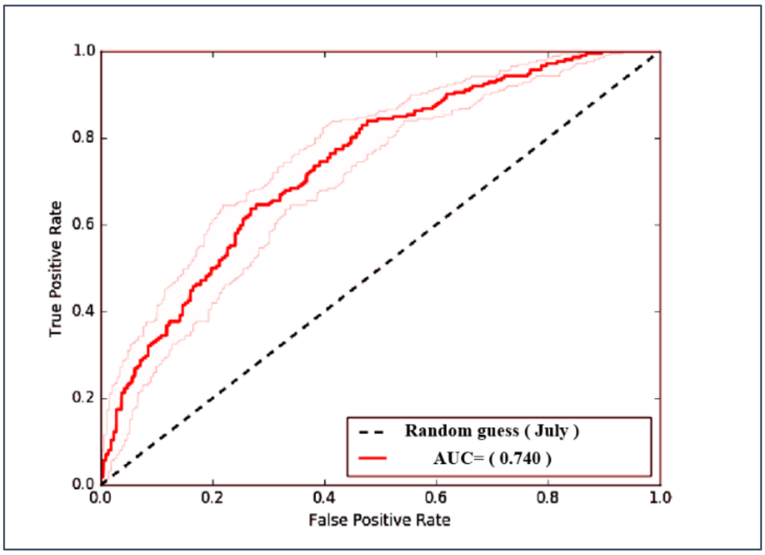

Fig. 6 Prediction rate curves for the susceptibility maps produced using frequency ratio method

Accuracy Assessment of Machine Learning Model

Table 1 shows the performance of the RF model evaluated using statistical methods such as Area Under the Curve (AUC), Root mean Square Error (RMSE), Mean Absolute Error (MAE) & Kappa Coefficients. The value of 0.98, as obtained for AUC, signifies the model's high success in predicting potential future landslide occurrences in the region. Additionally, the accuracy assessment from the test dataset yielded a commendable accuracy rate of 93%. These results collectively affirm the effectiveness and reliability of the model in its predictive capabilities for landslide events in the specified area.…

Table 1 Area under the curve and RMSE values of the models for July month.

Model	Test Accuracy	AUC (%)	RMSE	MAE	Kappa Coefficient
Random Forest	0.930	98.0	0.265	0.07	0.859

Landslide Susceptibility Mapping

Based on the framework reported earlier, For the purpose of comparative assessment, a landslide susceptibility map was developed for each district of Arunachal Pradesh. Random forest model was used to construct the LSM map. The RF classification produces decision values for each pixel, which indicate the likelihood of that pixel belonging to each class. These decision values are then converted into probability values, which ranges from 0 to 1 and are summed to 1 for each pixel. In the ArcGIS software, these probability values are represented as a rule image, where each pixel's probability value represents the 'true' probability of that pixel belonging to each class. To further classify the landslide susceptibility, the index file is reclassified into four susceptibility classes: 'Very High', 'High', 'Moderate', and 'Low'. This enables the identification of areas with a high likelihood of landslides, as well as areas with lower susceptibility to landslides [53].

Figures 7-10 depict the landslide susceptibility maps of the Upper Subansiri district, Upper Dibang Valley district, Tawang district & West Siang district of Arunachal Pradesh, with green color indicating low susceptibility while red indicating very high susceptibility. The portrayed maps were developed for the monsoon season in the area, so that the widespread extent of the landslide during the monsoon season can be identified. Similar mapping was conducted for other districts of Arunachal Pradesh; however, has not been portrayed here for the sake of brevity. The adopted colour-coding scheme on the map allows the viewers to quickly identify areas that are at higher risk of landslides within the district. By comparing the susceptibility levels, the map can provide insights into susceptibility of different parts of the region, which can be useful for disaster management and risk reduction efforts.

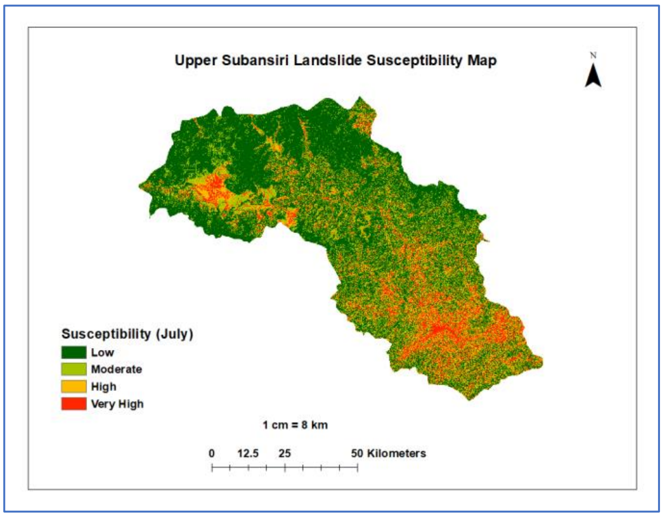

Fig. 7 LSM of Upper Subansiri district

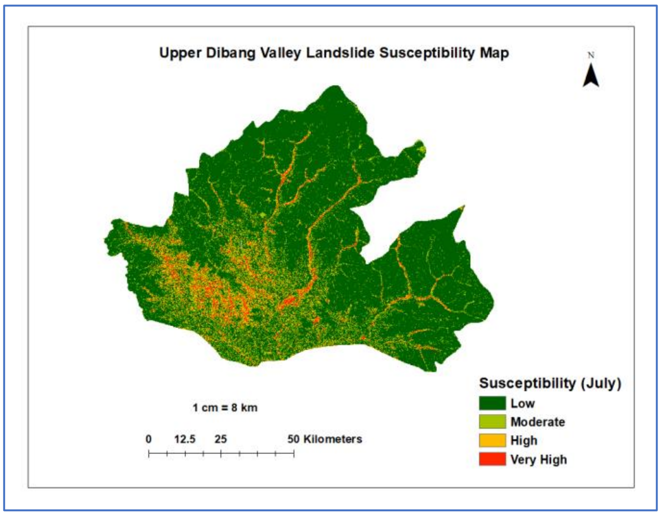

Fig. 8 LSM of Upper Dibang Valley district

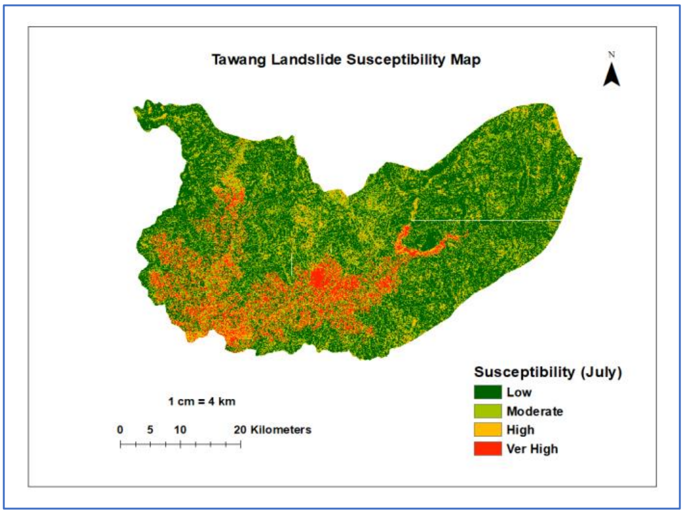

Fig. 9 LSM of Tawang district

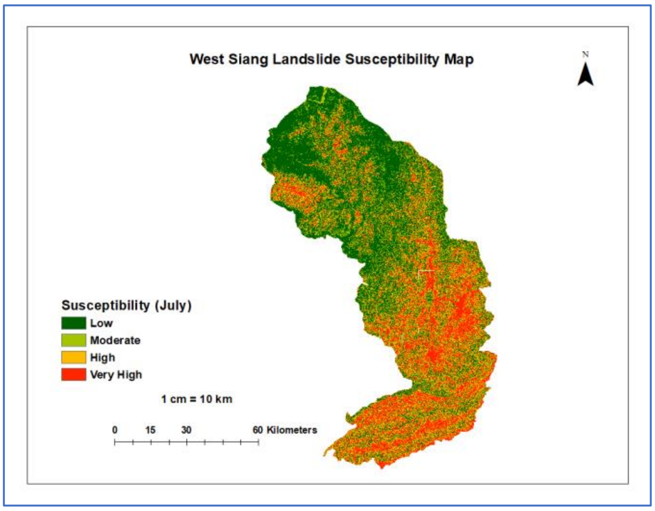

Fig. 10 LSM of West Siang district

CONCLUSIONS

Based on the present study on the development of landslide susceptibility mapping for the state of Arunachal Pradesh in India, the following conclusions are drawn:

- The integration of frequency ratio approach and random forest machine learning method provides an effective way to develop LSMs of an area, and such integration has the capability of providing a comprehensive and advanced approach to geohazard analysis.
- The Frequency Ratio (FR) method can be efficiently utilized for selecting the best possible non-landslide points selection and thereby constructing the preliminary LSM that can be further improved by employing the RF technique while accommodating the landslide occurrence points. The resulting FR-based LSM achieved a notable 74% AUC value, thereby showcasing its precision. In comparison to the conventionally adopted random selection of non-landslide locations, FR method resulted in a selection technique with improved accuracy and reliability.

- The relative importance of the various landslide conditioning factors, in terms of their influence on landslide occurrences, could be extracted using the ‘feature importance’ function of the Random Forest algorithm. The results indicated that for the landslide susceptibility modelling for the state of Arunachal Pradesh, landslide conditioning factors such as elevation, stream power index, slope, rainfall, aspect, built-up area, plan curvature, TPI, snow, TWI, profile curvature, loamy soil, and range lands had a significant contribution. In contrast, factors like water bodies (rivers, ponds etc.), presence of sandy loam, and glacial soils had a minimal impact on the occurrence of landslides.
- Various evaluation indicators were used to evaluate the performance of Random Forest model in generating the LSM for various districts of Arunachal Pradesh. With a test accuracy of 93%, the ROC-AUC was achieved to be 98%, while the RMSE, MAE and Kappa coefficient were obtained to be 0.265, 0.07 and 0.859, respectively. Each of them exhibited magnitudes that are indicative of a very reliable and superior prediction of the landslides in the region.

Hence, in a nutshell, a sufficiently reliable and accurate LSM is developed for various districts of Arunachal Pradesh. The application of these landslide susceptibility maps generated for the state of the Arunachal Pradesh is crucial for decision-makers, planners, and engineers. These maps can be used to identify areas that are prone to landslides and to take measures to prevent and mitigate the risks associated with them. This can ultimately lead to a reduction in the loss of life and property damage caused by landslides. The use of these maps can also help in emergency planning and response, as well as in the development of strategies for disaster management.

Developing a reliable Landslide Susceptibility Model (LSM) for Arunachal Pradesh encounters various challenges, with one major limitation being the scarcity of historical landslide data. The available data are often clustered around human settlements, offering an incomplete picture of landslide occurrences across the entire region. Moreover, the geological complexity, high relief, and steep slopes in Arunachal Pradesh contribute to the challenge. The rugged topography amplifies susceptibility, but accurately quantifying this impact is complicated by limited data on local geological conditions. Remote sensing and GIS data, fundamental for LSMs, face challenges due to factors such as cloud cover and limited accessibility. Additionally, inadequate monitoring, the absence of a robust early warning system, and limited socio-economic data contribute to the difficulties. Collaborative efforts and a multidisciplinary approach are essential to overcome these challenges and pave the way for more reliable LSMs, especially in data-scarce regions like Arunachal Pradesh. This study is a preliminary attempt in establishing a methodology in negating the uncertainties involved in the selection of the non-landslide points which greatly influence the dependability of the LSMs. This can be particularly useful in generating sufficiently reliable LSMs for the data-scarce regions.

REFERENCES

1. Kavzoglu, T., Colkesen, I., & Sahin, E. K. (2019). Machine learning techniques in landslide susceptibility mapping: a survey and a case study. Eds. S. P. Pradhan, V. Vishal, T. N. Singh, *Landslides: Theory, Practice and Modelling*, Springer Nature, Switzerland, Chapter 13, 283-301.
2. Gaidzik, K., & Ramírez-Herrera, M. T. (2021). The importance of input data on landslide susceptibility mapping. *Scientific Reports*, *11*(19334):1-14.
3. Batar, A. K., & Watanabe, T. (2021). Landslide susceptibility mapping and assessment using geospatial platforms and weights of evidence (WoE) method in the Indian Himalayan Region: Recent developments, gaps, and future directions. *ISPRS International Journal of Geo-Information*, *10*(114), 1-28.
4. Araujo, J. R., Ramos, A. M., Soares, P. M., Melo, R., Oliveira, S. C., & Trigo, R. M. (2022). Impact of extreme rainfall events on landslide activity in Portugal under climate change scenarios. *Landslides*, *19*(10), 2279-2293.
5. Getachew, N., & Meten, M. (2021). Weights of evidence modeling for landslide susceptibility mapping of Kabi-Gebro locality, Gundomeskel area, Central Ethiopia. *Geoenvironmental Disasters*, *8*(1), 1-22.
6. Paul, S., Tripathi, A. K., Burman, R. R., Panggam, M., Ray, S. K., Kalita, N., Vanlalduati, R. & Singh, A. K. (2017). Jhum cultivation and its consequences on forest and environment in Eastern Himalayan tract of India: A participatory assessment. *Range Management and Agroforestry*, *38*(1), 121-126.
7. Wu, S., Li, J., & Huang, G. H. (2008). A study on DEM-derived primary topographic attributes for hydrologic applications: Sensitivity to elevation data resolution. *Applied Geography*, *28*(3), 210-223.
8. Sarma, C. P., Dey, A., & Murali Krishna, A. (2020). Influence of digital elevation models on the simulation of rainfall-induced landslides in the hillslopes of Guwahati, India. *Engineering Geology*, 268(105523), 1-13.

9. Zhang, Y. X., Lan, H. X., Li, L. P., Wu, Y. M., Chen, J. H., & Tian, N. M. (2020). Optimizing the frequency ratio method for landslide susceptibility assessment: A case study of the Caiyuan Basin in the southeast mountainous area of China. *Journal of Mountain Science*, *17*(2), 340-357.
10. Fenton, G. A., McLean, A., Nadim, F., & Griffiths, D. V. (2013). Landslide hazard assessment using digital elevation models. *Canadian Geotechnical Journal*, *50*(6), 620-631.
11. Maren, I. E., Karki, S., Prajapati, C., Yadav, R. K., & Shrestha, B. B. (2015). Facing north or south: Does slope aspect impact forest stand characteristics and soil properties in a semiarid trans-Himalayan valley? *Journal of Arid Environments*, *121*, 112-123.
12. Zhang, J., Qiu, H., Tang, B., Yang, D., Liu, Y., Liu, Z., Ye, B., Zhou, W. & Zhu, Y. (2022). Accelerating effect of vegetation on the instability of rainfall-induced shallow landslides. *Remote Sensing*, *14*(5743), 1-18.
13. Xiao, T., Yin, K., Yao, T., & Liu, S. (2019). Spatial prediction of landslide susceptibility using GIS-based statistical and machine learning models in Wanzhou County, Three Gorges Reservoir, China. *Acta Geochimica*, *38*, 654-669.
14. Aghdam, I. N., Varzandeh, M. H. M., & Pradhan, B. (2016). Landslide susceptibility mapping using an ensemble statistical index (Wi) and adaptive neuro-fuzzy inference system (ANFIS) model at Alborz Mountains (Iran). *Environmental Earth Sciences*, *75*, 1-20.
15. Wang, Q., Li, W., Chen, W., & Bai, H. (2015). GIS-based assessment of landslide susceptibility using certainty factor and index of entropy models for the Qianyang County of Baoji city, China. *Journal of Earth System Science*, *124*, 1399-1415.
16. Pawłuszek, K., & Borkowski, A. (2016). Landslides identification using airborne laser scanning data derived topographic terrain attributes and support vector machine classification. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXIII ISPRS Congress, XLI-B8, 145-149.
17. Jebur, M. N., Pradhan, B., & Tehrany, M. S. (2014). Optimization of landslide conditioning factors using very high-resolution airborne laser scanning (LiDAR) data at catchment scale. *Remote Sensing of Environment*, *152*, 150-165.
18. Chen, C. Y., & Yu, F. C. (2011). Morphometric analysis of debris flows and their source areas using GIS. *Geomorphology*, *129*(3-4), 387-397.
19. Pourghasemi, H., Pradhan, B., Gokceoglu, C., & Moezzi, K. D. (2013). A comparative assessment of prediction capabilities of Dempster–Shafer and weights-of-evidence models in landslide susceptibility mapping using GIS. *Geomatics, Natural Hazards and Risk*, *4*(2), 93-118.
20. Moore, I. D., Gessler, P. E., Nielsen, G. A. E., & Peterson, G. A. (1993). Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, *57*(2), 443-452.
21. Selamat, S. N., Majid, N. A., Taha, M. R., & Osman, A. (2022). Landslide susceptibility model using Artificial Neural Network (ANN) approach in Langat River Basin, Selangor, Malaysia. *Land*, *11*(833), 1-21.
22. Kavzoglu, T., Sahin, E. K., & Colkesen, I. (2014). Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides*, *11*, 425-439.
23. Mezughi, T. H., Akhir, J. M., Rafek, A. G., & Abdullah, I. (2011). Landslide susceptibility assessment using frequency ratio model applied to an area along the EW highway (Gerik-Jeli). *American Journal of Environmental Sciences*, *7*(1), 43-50.
24. Sonker, I., Tripathi, J. N., & Singh, A. K. (2021). Landslide susceptibility zonation using geospatial technique and analytical hierarchy process in Sikkim Himalaya. *Quaternary Science Advances*, *4*, 100039-1-17.
25. Ramesh, V., & Anbazhagan, S. (2015). Landslide susceptibility mapping along Kolli hills Ghat road section (India) using frequency ratio, relative effect and fuzzy logic models. *Environmental Earth Sciences*, *73*, 8009-8021.
26. Regmi, A. D., Yoshida, K., Dhital, M. R., & Devkota, K. (2013). Effect of rock weathering, clay mineralogy, and geological structures in the formation of large landslide, a case study from Dumre Besei landslide, Lesser Himalaya Nepal. *Landslides*, *10*, 1-13.
27. Kumar, V., Gupta, V., & Jamir, I. (2018). Hazard evaluation of progressive Pawari landslide zone, Satluj valley, Himachal Pradesh, India. *Natural Hazards*, *93*, 1029-1047.
28. Tyagi, A., Tiwari, R. K., & James, N. (2023). Mapping the landslide susceptibility considering future land-use land-cover scenario. *Landslides*, *20*(1), 65-76.
29. Pham, B. T., Tien Bui, D., Pham, H. V., Le, H. Q., Prakash, I., & Dholakia, M. B. (2017). Landslide hazard assessment using random subspace fuzzy rules based classifier ensemble and probability analysis of rainfall data: A case study at Mu Cang Chai District, Yen Bai Province (Viet Nam). *Journal of the Indian Society of Remote Sensing*, *45*, 673-683.
30. Wei, L., Cheng, H., & Dai, Z. (2023). Propagation modeling of rainfall-induced landslides: A case study of the Shaziba Landslide in Enshi, China. *Water*, *15*(424), 1-19.
31. He, S., Wang, J., & Liu, S. (2020). Rainfall event–duration thresholds for landslide occurrences in China. *Water*, *12*(2), 494.
32. Bui, D. T., Pradhan, B., Lofman, O., Revhaug, I., & Dick, O. B. (2012). Landslide susceptibility mapping at Hoa Binh province (Vietnam) using an adaptive neuro-fuzzy inference system and GIS. *Computers & Geosciences*, *45*, 199-211.
33. Ullah, I., Aslam, B., Shah, S. H. I. A., Tariq, A., Qin, S., Majeed, M., & Havenith, H. B. (2022). An integrated approach of machine learning, remote sensing, and GIS data for the landslide susceptibility mapping. *Land*, *11*(8), 1265-1-20.

34. Bui, D. T., Shahabi, H., Shirzadi, A., Chapi, K., Alizadeh, M., Chen, W., ... & Tian, Y. (2018). Landslide detection and susceptibility mapping by AIRSAR data using support vector machine and index of entropy models in Cameron Highlands, Malaysia. *Remote Sensing*, *10*(10), 1527.
35. Zhao, F., Meng, X., Zhang, Y., Chen, G., Su, X., & Yue, D. (2019). Landslide susceptibility mapping of Karakorum Highway combined with the application of SBAS-InSAR technology. *Sensors*, *19*(12), 2685.
36. Li, X., Cheng, J., Yu, D., & Han, Y. (2021). Research on non-landslide selection method for landslide hazard mapping. *Research Square Preprint*. https://doi.org/10.21203/rs.3.rs-270737/v1
37. Nhu, V. H., Mohammadi, A., Shahabi, H., Ahmad, B. B., Al-Ansari, N., Shirzadi, A., Clague, J. J., Jaafari, A., Chen, W. & Nguyen, H. (2020). Landslide susceptibility mapping using machine learning algorithms and remote sensing data in a tropical environment. International *Journal of Environmental Research and Public Health*, *17*(14), 4933.
38. Rosi, A., Frodella, W., Nocentini, N., Caleca, F., Havenith, H. B., Strom, A., Saldov, M., Bimurzaev, G. A. & Tofani, V. (2023). Comprehensive landslide susceptibility map of Central Asia. *Natural Hazards and Earth System Sciences*, *23*(6), 2229-2250.
39. Bajni, G., Camera, C. A., & Apuani, T. (2023). A novel dynamic rockfall susceptibility model including precipitation, temperature and snowmelt predictors: a case study in Aosta Valley (Northern Italy). *Landslides*, 1-24.
40. Wang, S., Lin, X., Qi, X., Li, H., & Yang, J. (2022). Landslide susceptibility analysis based on a PSO-DBN prediction model in an earthquake-stricken area. *Frontiers in Environmental Science*, *10*, 912523.
41. Shano, L., Raghuvanshi, T. K., & Meten, M. (2020). Landslide susceptibility evaluation and hazard zonation techniques–A review. *Geoenvironmental Disasters*, *7*(1), 1-19.
42. Pardeshi, S. D., Autade, S. E., & Pardeshi, S. S. (2013). Landslide hazard assessment: Recent trends and techniques. *SpringerPlus*, *2*, 1-11.
43. Autade, S. E., Pardeshi, S. D., & Pardeshi, S. S. (2020). Advances in landslide hazard assessment in India. *Transactions of the Institution of Indian Geographers*, *42*(2), 257-271.
44. Sharma, S., & Mahajan, A. K. (2019). A comparative assessment of information value, frequency ratio and analytical hierarchy process models for landslide susceptibility mapping of a Himalayan watershed, India. *Bulletin of Engineering Geology and the Environment*, *78*, 2431-2448.
45. Choi, J., Oh, H. J., Lee, H. J., Lee, C., & Lee, S. (2012). Combining landslide susceptibility maps obtained from frequency ratio, logistic regression, and artificial neural network models using ASTER images and GIS. *Engineering Geology*, *124*, 12-23.
46. Pradhan, B., Lee, S., & Buchroithner, M. F. (2010). Remote sensing and GIS-based landslide susceptibility analysis and its cross-validation in three test areas using a frequency ratio model. *Photogrammetrie-Fernerkundung-Geoinformation*, *1*, 17-32.
47. Lee, S., & Talib, J. A. (2005). Probabilistic landslide susceptibility and factor effect analysis. *Environmental Geology*, *47*, 982-990.
48. Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.
49. Stumpf, A., & Kerle, N. (2011). Object-oriented mapping of landslides using Random Forests. *Remote Sensing of Environment*, *115*(10), 2564-2577.
50. Taalab, K., Cheng, T., & Zhang, Y. (2018). Mapping landslide susceptibility and types using Random Forest. *Big Earth Data*, *2*(2), 159-178.
51. Hong, H., Pourghasemi, H. R., & Pourtaghi, Z. S. (2016). Landslide susceptibility assessment in Lianhua County (China): A comparison between a random forest data mining technique and bivariate and multivariate statistical models. *Geomorphology*, *259*, 105-118.
52. Rahmati, O., Tahmasebipour, N., Haghizadeh, A., Pourghasemi, H. R., & Feizizadeh, B. (2017). Evaluating the influence of geo-environmental factors on gully erosion in a semi-arid region of Iran: An integrated framework. *Science of the Total Environment*, *579*, 913-927.
53. Pourghasemi, H. R., Moradi, H. R., & Fatemi Aghda, S. M. (2013). Landslide susceptibility mapping by binary logistic regression, analytical hierarchy process, and statistical index models and assessment of their performances. *Natural Hazards*, *69*, 749-779.