



# Integration of GIS and Machine Learning Techniques for Mapping the Landslide-Prone Areas in the State of Goa, India

Babitha Ganesh<sup>1</sup> · Shweta Vincent<sup>1</sup>  · Sameena Pathan<sup>2</sup> · Silvia Raquel Garcia Benitez<sup>3</sup>

Received: 28 November 2022 / Accepted: 18 April 2023 / Published online: 6 June 2023  
© The Author(s) 2023

## Abstract

A landslide susceptibility map (LSM) assists in reducing the danger of landslides by locating the landslide-prone locations within the designated area. One of the locations that are prone to landslides in India's Western Ghats of which Goa is a part. This article presents the LSMs prepared for the state of Goa using four standard machine learning algorithms, namely Logistic Regression (LR), Support Vector Machine (SVM), *K*-Nearest Neighbour (KNN), and Random Forest (RF). In order to create LSMs, a 78-point landslide inventory, as well as 14 landslide conditioning factors, has been used, including slope, elevation, aspect, total curvature, plan curvature, profile curvature, yearly rainfall, Stream Power Index, Topographic Wetness Index, distance to road, depth to bedrock/soil depth, soil type, lithology, and land use land cover. The most pertinent features for the models' construction have been chosen using the Pearson correlation coefficient test and the Random Forest method. The presence of landslides is shown to be strongly influenced by the distance to road, slope of the terrain, and the annual rainfall. The LSMs generated were classified into five levels ranging from very low susceptibility level to very high susceptible. The prediction accuracy, precision, recall, F1-score, area under the ROC (AUC-ROC), and True Skill Statistics (TSS) have been used to analyse and compare the LSMs created using various methodologies. All of these algorithms perform pretty well, as evidenced by the overall accuracy scores of 81.90% for LR, 83.33% for SVM, 81.94% for KNN, and 86.11% for RF. SVM and RF are the better approaches for forecasting landslide vulnerability in the research area, according to TSS data. The maximum AUC-ROC of 86% was achieved by the RF algorithm. The results of performance metrics lead to the conclusion that the tree-based RF approach is most appropriate for producing LSM for the state of Goa. The results of this study indicate that more landslide-prone areas can be found in the Sattari, Dharbandora, Sanguem, and Canacona regions of Goa.

**Keywords** Landslide conditioning factor (LCF) · Landslide Inventory Mapping (LIM) · Landslide susceptibility mapping (LSM) · Western Ghats

✉ Shweta Vincent  
shweta.vincent@manipal.edu

Babitha Ganesh  
babitha.mit@learner.manipal.edu

Sameena Pathan  
sameena.bp@manipal.edu

Silvia Raquel Garcia Benitez  
sgab@pumas.iingen.unam.mx

<sup>1</sup> Department of Mechatronics, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka 576104, India

<sup>2</sup> Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka 576104, India

<sup>3</sup> Electrical and Computing Coordination, Instituto De Ingenieria, UNAM, Mexico, Mexico

## Introduction

A 'Landslide' is the term used to depict the mass developments happening over a sliding surface of the land. The population, properties, and economic activities, including public services, are all severely damaged by this potentially harmful phenomena (Van Westen, 1993). Nearly 66 million people reside in areas that are severely prone to landslides, and 17% of casualties in these areas are attributed to landslides, which are recurrent calamities in mountainous areas (Achu et al., 2022a, 2022b).

An area of biodiversity, the Western Ghats (WG) mountain range in the Indian sub-mainland, is currently under grave risk for the valley's residents and environment

due to altering precipitation trends and changes in land use. WG is a mountain range crossing the states of Gujarat, Maharashtra, Goa, Karnataka, Kerala, and Tamil Nadu (Yunus et al., 2021). Goa is a tiny state located between the Arabian Sea to the west and the WG's eastern slopes (Mascarenhas et al., 2009). In the majority of locations in Goa, lateral soils have overflowed, making them one of the most perilous kinds of soil. Similar to how the precipitation seems to be quite intense along the WG, the soundness of lateritic slants has emerged as a crucial factor in the occurrence of landslides in the state of Goa (Achu et al., 2020).

Depending on the geophysical, meteorological, and topographical variables, landslide susceptibility mapping (LSM) estimates the likelihood of a landslide event occurring in the area of interest (Chen et al., 2021). Information on earlier landslide catastrophes and all of the factors causing failures in the study area must be used in the LSM preparation phase. Geographic information systems (GIS) and satellite remote sensing (RS) have received widespread use and are regarded as valuable tools for observing the changes in the land. In order to comprehend and track changes in the topography of the research region, RS provides economic data expressed in several electromagnetic bands and acquired at various time intervals (Lee, 2019). The postulated LSM techniques can be broadly divided into qualitative and quantitative techniques. In the qualitative approach, it is the field expert's task to identify the slope failure attributes through field investigation and location mapping. The use of statistical or machine learning (ML) algorithms on data that can be derived by superimposing a map of the study area with thematic variables pertaining to structural, geological, hydrological, and geotechnical characteristics that impact the probability of the occurrence of landslides at the defined study area is an integral part of quantitative methods (Dias et al., 2021). The main objective of Tsangaratos et al. (2017), is to produce the LSM using Logistic Regression (LR), Weight of Evidence (WoE), and Random Forest (RF) in Nancheng County, China. The mapping approach made use of a database with information on 112 prior landslides. Lithology, altitude, slope, aspect, Topographic Wetness Index (TWI), sediment transport index (STI), profile curvature, plan curvature, distance to rivers, distance to faults, and distance to roads are the thematic variables considered landslide conditioning factors (LCF). The statistical methods FR, information value, and Certainty Factor (CF) have been applied in Wubalem (2021) to develop LSM at Uatzau catchment area, in north-western Ethiopia. Six LCFs and 514 slope failure points have been used in this model. There have been quite a few activities carried out to get LSM ready for the Western Ghats regions. The study conducted by Vijith et al. (2014) aims to form LSM with

the help of GIS-based WoE method for the upland catchment of river Meenachil, Kottayam district, Kerala. The LCFs such as geomorphology, drainage density, soil type, soil thickness, land use, Normalized Difference Vegetation Index (NDVI), slope, aspect, relative relief, slope length, curvatures, flow path length, and TWI have been generated, and its weights were determined. An attempt is made to map the landslides in the Tevankarai Ar sub-watershed, Kodaikanal, India, utilizing binary Logistic Regression analysis (Ramani et al., 2011). This model depicts the link between the independent factors chosen for research (predictor variables) via the best-fitting function and the dependent variable (landslide presence or absence). The regression analysis employs a forward stepwise Logistic Regression model with maximum likelihood estimation. The traditional statistical algorithms such as FR, WoE, Statistical Index, and Shannon entropy are easy to implement but they are less accurate compared to the more advanced statistical methods such as ML and deep learning (DL). ML techniques have been drawing good attention as they are producing LSMs with better accuracy. The effectiveness of deep and machine learning approaches, such as deep neural networks (DNN) and Random Forests (RF), has been tested against the outcomes of conventional statistical techniques in the study carried out by Achu et al., 2022a, 2022b, in portions of the Kozhikode and Wayanad districts of Kerala state. According to the study, DL and ML approaches work best with datasets that have fewer landslide points. Six distinct ML models, including Adaptive Boosting (AdaBoost), Naive Bayes (NB), Neural Network (NNET), Random Forest (RF), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost), were employed in another study (Jennifer, 2022) to create the LSM of Kerala's Idukki district. Comparing the AdaBoost model to the other five models, it was the most accurate and consistently useful.

Many of the locations undergoing landslides have not yet been investigated by the researchers, despite the fact that a few LSMs have been developed for the various WG of India. Goa is one of the states in WG, witnessing several landslides during monsoon season with not much attention towards developing LSMs. The study carried out by Nagarajan et al. (2000) created a technique and tested its efficacy in identifying landslide-prone locations in the state of Goa. A small portion of the WG known as the Kumbharli Ghat has been chosen as the study area. As there was no inventory available, they have identified landslide conditioning factors by analysing the information of previous landslides occurred at different areas such as Kumbharli Ghat and Varandh Ghat. Landslide-prone locations have been determined based on the domain specialists' expertise of the area. Without using any statistical algorithms, this investigation was solely qualitative. The

accuracy of the solution is subjected to the knowledge of the experts. In one more study (Mhaddolkar, 2017), researchers have presented the study of landslides that have occurred in the past. They considered only two landslides that occurred at the locations Ponda and Neura and investigated to understand the conditions for the slope failure. Through this, it was inferred that inappropriate practices of hill cutting are one of the reasons for the landslides at these places. The solution provided by their study is limited to those two landslides and does not provide any general solution to locate landslide-prone areas in other places of Goa. A methodology has been proposed in Kessarkar et al. (2011) to create an LSM for the Canacona region of Goa state. The method entails computing the Landslide Potential Index (LPI) and categorizing this LPI to prepare various susceptible zones, such as low, medium, high, and extremely high. The slope and height of the landscape, elevation, vegetation, soil type, and drainage pattern are only a few of the variables taken into account while determining LPI. The resolution of the maps used in deriving the conditioning factors was 90 m. It was difficult to obtain the LSM with high accuracy because of the lack of sufficiently resolution satellite data for establishing the triggering components.

It is obvious that no LSMs have been produced using ML algorithms for the full state of Goa, based on our understanding of the literature that is currently available. Traditional statistical techniques are used to construct the LSMs for some Goa regions. In our article, LSM has been created for the entire state of Goa using different ML algorithms such as LR, SVM, *K*-Nearest Neighbours (KNN), and Random Forest (RF) with an ensemble approach. The use of recently captured hyperspectral good resolution satellite data, GIS software, and ML algorithms resulted in accurate LSMs for the state of Goa. The study area has a wide variety of landslide conditioning elements that have been created, and the LSM of the study area has been prepared using the most significant triggering factors selected after applying feature selection methods. The work adds to strengthening our understanding of the relationship between landslides and potential causal factors by using established methodologies in the region of Goa that has not previously been extensively researched. The results will provide an important basis for future efforts to reduce disaster-related risk and plan for land use in the entire state of Goa. While ArcGIS 10.8.2 by Environmental Systems Research Institute (ESRI) was used to compile the data and create the LSMs, the computational process was carried out using Python Language in Google Colaboratory.

## Description of Study Area

The study area, located between N 14° 53' and N 15° 48', is the part of Western Ghats. Figure 1 displays the map of the study area. With a beachfront that runs its whole length of more than 100 kms, the state is wider to the north. E 73° 39' and E 74° 20', respectively, are the approximate western and eastern extents. The state can be split into three types of topography based on its physiography, ranging from low-lying coastal–estuarine plains in the west to an undulating region in the centre to the steep slopes of the Western Ghats on the state's eastern border. The steeper and higher ranges that make up the Western Ghats are found in Goa's eastern and southern regions. With the exception of the ranges in South Goa, the Western Ghats generally slope from north to south. They are more than 40 km from the sea in North Goa. However, towards the south, the trend of the hills is virtually east–west, which is connected to the underlying rock structure. Here, the lowest portions of the Karmal Ghat, a western arm of the Ghats, literally touch the ocean (Mascarenhas et al., 2009). According to the reports of Climate Research and Services, India Meteorological Department (IMD), with an average of over 2900 mm from June to September, the Southwest Monsoon is when Goa has the heaviest rainfall. Most of the geological formations are covered in laterite, alluvium, and sand, as one could anticipate from the humid tropical climate. The distinctive property of laterite is that it is soft underneath and hard on top (Mascarenhas et al., 2009).

## Methodology and Experiments

The methodology used for this study involves five steps as depicted in Fig. 2, (a) construction of landslide inventory, (b) preparation of LCF maps, (c) selection of most suitable LCFs, (d) ML classification model building, and (e) assessment of model performance.

### Landslide Inventory Mapping (LIM)

The algorithms used in this study work on the assumption that future landslides can be predicted based on the information about the previous landslides that have occurred in the study area and the conditioning factors that trigger the slope failure. Conducting field surveys and interpreting satellite images are commonly used data acquisition methods (Ganesh et al., 2023). The development of geographic information systems (GIS) in recent years has made it possible to analyse the collected data for effective landslide mapping and monitoring. Landslide Inventory Mapping (LIM) is the name given to the process of creating

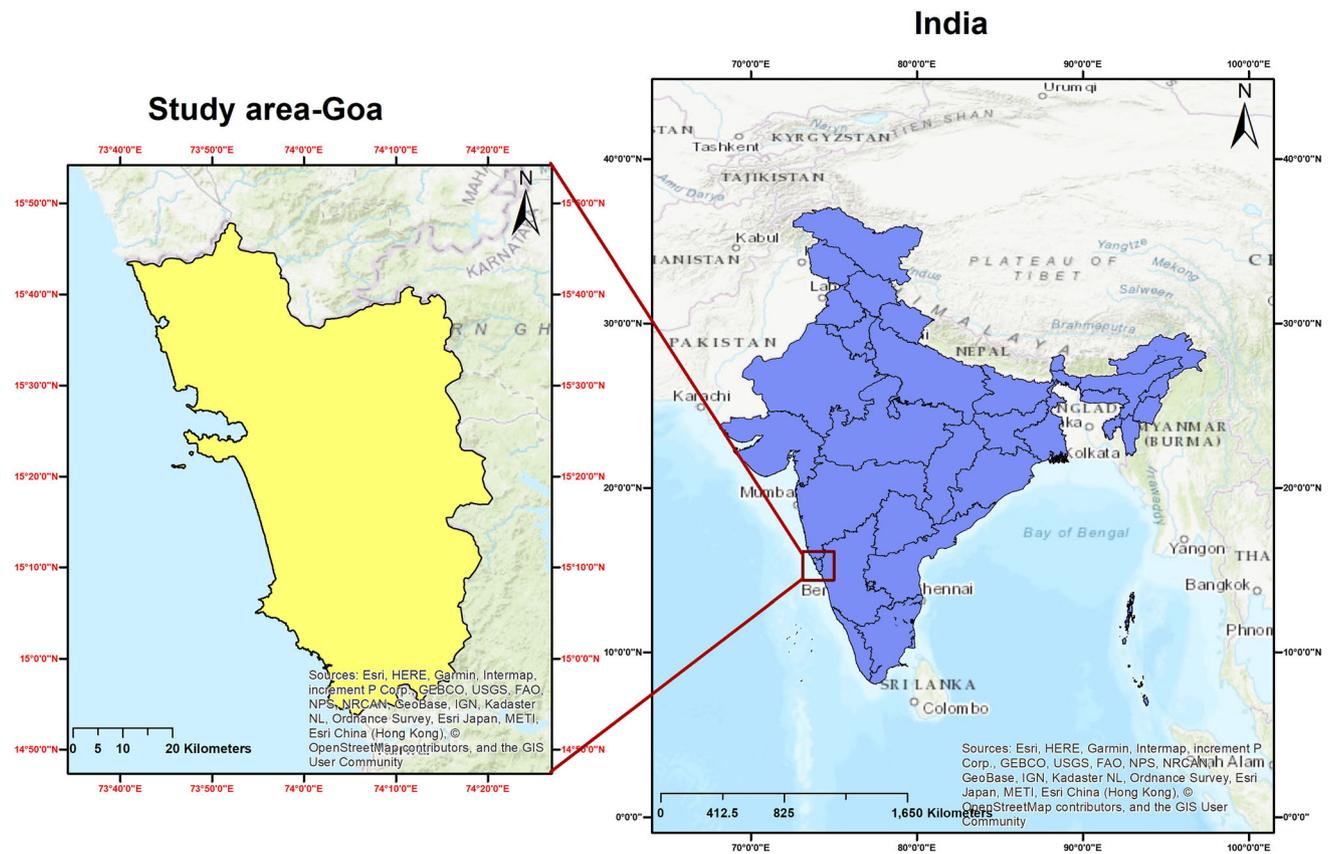
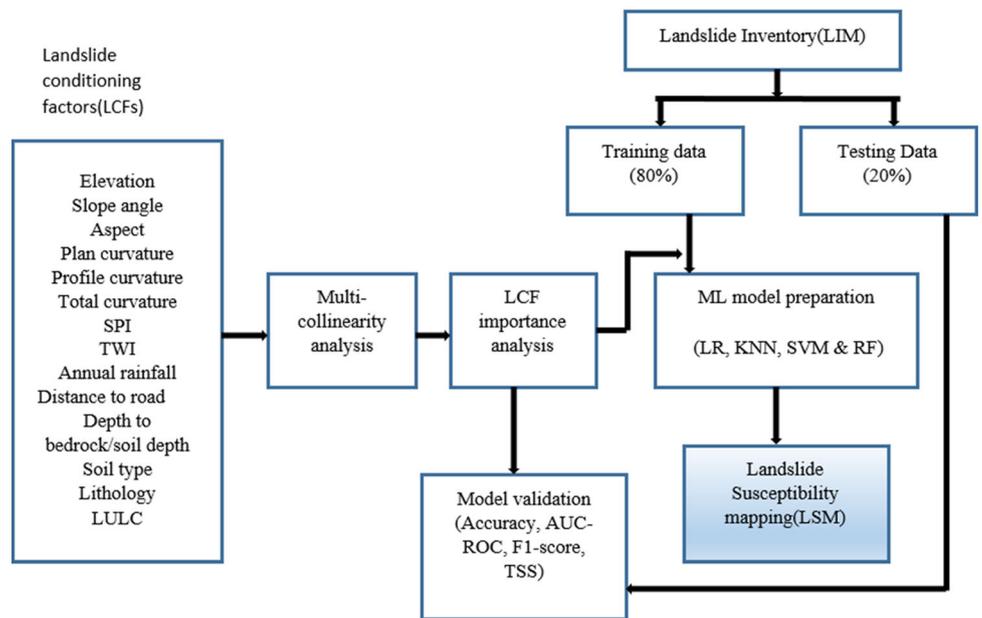


Fig. 1 Study area mapping

Fig. 2 Methodology of LSM creation



a map that contains details concerning landslides. In this work, the information of historical landslides has been acquired from the Bhukosh portal of the Geological Survey of India. This inventory includes 78 landslide points. It

includes the fields such as latitude and longitude of the location, date and time of the event, type of landslide, depth of landslide, short description, link to the news if any, topo sheet, geologic, hydrologic condition of the

location, land use land cover (LULC) information, and so on. Out of 78 landslides, 13 landslides have occurred at the border of Karnataka and Goa, 3 have occurred at the border of Maharashtra and Goa, and the remaining 58 landslides have occurred inside Goa. About 66% of the total landslides that occurred were of the shallow translational type, 31% of total landslides belong to the shallow rotational type, and 2.56% of them were deep rotational slides. The majority of the landslides took place along the roads adjacent to the Ghats section. In this work, as we have created a predictive model to differentiate between landslides and non-landslides using a binary classification as the foundation, we had to integrate the landslide points with non-landslide points in order to train and test the classifiers. A total of 200 randomly selected locations were used as the non-landslide points. Therefore, the landslide inventory contains 278 data points in total as shown in Fig. 3.

### LCF Preparation

LCFs refer to the several causal factors that cause landslides to occur (Rane & Vincent, 2022). These factors

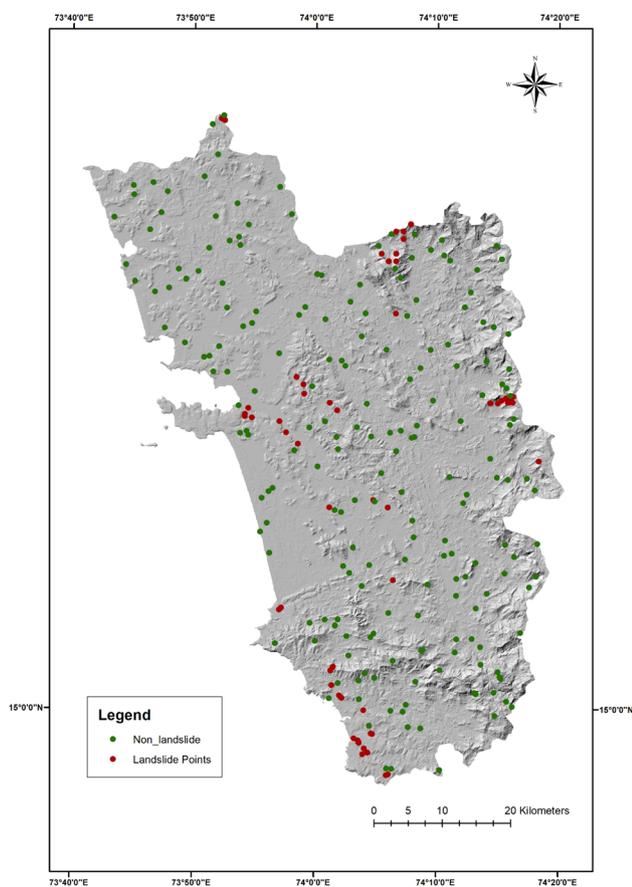


Fig. 3 Landslide inventory of Goa

directly or indirectly influence the occurrence of landslides. Based on the geological conditions of the study area, availability of data, and existing literature, we have selected 14 LCFs for this study as shown in Fig. 4. Soil type, lithology, and land use land cover (LULC) are categorical variables, and others are continuous variables. For the simplification purpose, these continuous variables have been reclassified into a set of classes. The list with description and classification methods of these LCFs is presented in Fig. 4.

Table 1 With a resolution of 30 m, all of these thematic parameters have been represented in raster format in Arc-Map. The thematic maps including the landslide inventory map have been projected onto the WGS-1984-UTM-Zone-43N coordinate system.

### LCF Selection

Multi-collinearity test using Pearson correlation coefficient: Each LCFs contribution will have an impact on the model's accuracy. As a result, prior to the mapping process, choosing the most pertinent criteria and excluding the irrelevant ones is crucial. In landslide research, the Pearson correlation method is a successful and popular feature selection technique (Feby et al., 2020).

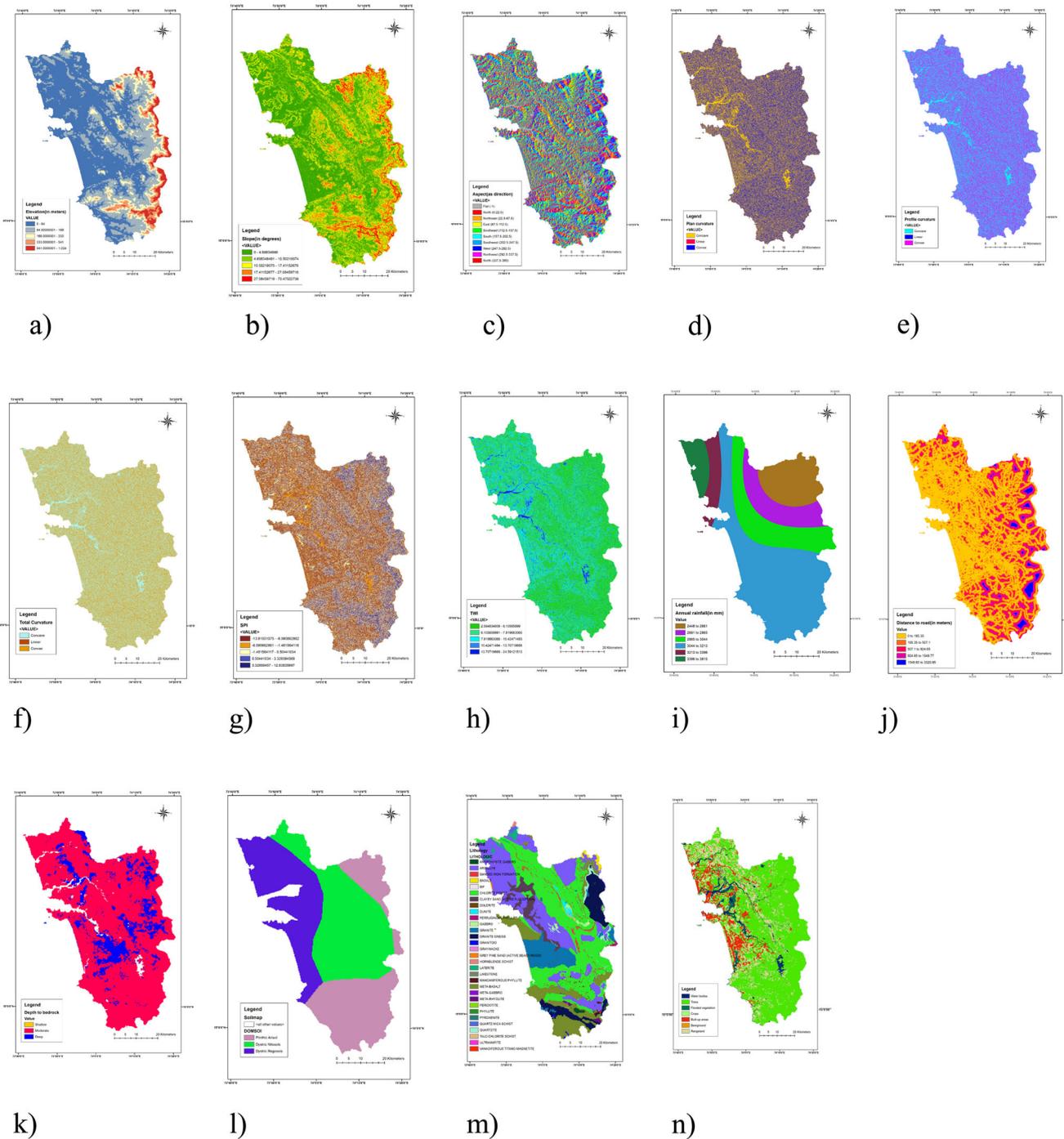
LCF importance analysis using Random Forest: It is also important to look into and get rid of any factors associated that have little to no predictive potential. One of the popular techniques for determining the significance of each LCF is to calculate the information gain ratio (IGR) (Dou et al., 2020).

### ML Model Preparation

ML classifier preparation entails splitting the entire number of data points into training data and testing data. In our study, out of the total 278 landslide and non-landslide points, 75% (222 data points) and 25% (56 data points) have been used for training and testing the classifiers, respectively. By placing the LCF map on the LIM of the research area, the values of various LCFs relevant to these training and testing data points have been derived. Then, the different classifiers have been constructed in the Google Colaboratory platform to fit the dataset. LR, SVM, KNN, and RF are the ML classifiers that were employed in this study. The description of these supervised classifiers is presented in this section.

### Logistic Regression (LR)

Based on the values of a number of predictor variables, a multimodal analysis model described as LR can be used to predict whether a feature or event will occur. The link



**Fig. 4** Landslide conditioning factors **a** Elevation, **b** slope angle, **c** aspect, **d** plan curvature, **e** profile curvature, **f** total curvature, **g** SPI, **h** TWI, **i** annual rainfall, **j** distance to road, **k** depth to bedrock, **l** soil type, **m** lithology, and **n** LULC

between the incidence and its dependencies on various variables can be quantified using Eq. (2).

$$y = \frac{\exp(z)}{1 + \exp(z)} \tag{2}$$

where  $z = b_0 + b_1X_1 + b_2X_2 + \dots + b_3X_n$ ,  $b_0$  is the intercept of the algorithm, and  $b_i$  is the weight of each

dependent variable  $X_i$  (Chen et al., 2018; Zhou et al., 2018).

**Support Vector Machine (SVM)**

In a multidimensional space, the SVM algorithm traces a hyperplane that could characterize diverse data points. The

**Table 1** Description of LCFs

LCF	Description	Source of data	Reclassification method	No. of classes
Elevation (altitude)	Snapshot of the terrain along with the features that are present (Saleem et al., 2019)	United States Geological Survey (USGS)	Natural break	5
Slope	First derivative of elevation (Saleem et al., 2019)	DEM derivatives	Natural break	5
Aspect	The direction of the terrain surface's greatest slope (Saleem et al., 2019)		Not reclassified	10
Plan curvature	It is perpendicular to the direction of highest slope (Saleem et al., 2019)		Manual	3
Profile curvature	It is parallel to the direction of highest slope (Saleem et al., 2019)			
Total curvature	Integrates the profile and plan curvatures (Saleem et al., 2019)			
SPI	SPI is calculated with the help of Eq. (1) $SPI = \ln(FlowAcc + 0.001 * \frac{slope}{100} + 0.001) \quad (1)$		Natural break	5
TWI	Explains the spatial variability of particular processes (Mattivi et al., 2019)		Natural break	6
Annual rainfall	Amount of rainfall (Yunus et al., 2021)	Climatic Research Unit (CRU)	Natural break	6
Distance to road	Distance from the highways (Yunus et al., 2021)	Open Street Map (OSM) data	Natural break	5
Depth to bedrock/soil depth	Depth between the ground's surface and the point where coherent (continuous) bedrock is encountered (Guiling et al., 2017)	ISRIC data hub (Hengl et al., 2017)	Manual	3
Soil type	Grouping soils with comparable chemical, physical, and biological characteristics into units (Babitha et al., 2022)	FAO-UNESCO Soil Map	NA	3
Lithology	Type of rock such as clay sand, basalt, argillite, granite, laterite, and limestone	USGS world geological map	NA	31
LULC	Land formations such as water bodies, built-up land, agricultural land, and forest areas	ArcGIS living atlas	NA	7

adoption of a suitable kernel function, such as linear, sigmoid, polynomial, and radial basis functions (RBF), has a substantial impact on the model's ability to forecast outcomes. The gamma parameter measures the influence of a single training example, with low values implying 'far' and large values denoting 'close'. Recognizing training cases is traded off by the C parameter to enhance the tolerance of the decision function (Pourghasemi et al., 2013a, 2013b). Because they yielded better results than those of other choices, the default parameters for the kernel, C, and gamma were retained in our investigation.

### K-Nearest Neighbours (KNN)

The KNN algorithm determines the distance between the target point and the closest points in accordance with the number of points that were picked, taking into account the value specified for K and the highest number of votes cast for these encircling points in the classification setting. The KNN approach skips the density subordinate and goes

directly to a decision rule, assuming that pixels near one another in the trait space should belong to the same class (Avand et al., 2019). In our implementation, the 'Euclidean' distance algorithm with the optimal value of 5 for a number of neighbours has been used.

### Random Forest (RF)

This ensemble learning technique leads to the creation of classification trees that are aggregated to achieve a classification (Youssef et al., 2016). The two parameters that RF must specify in order to create a classification model are the size of trees in the forest and the amount of analysed parameters at each node to promote the growth of the tree (Taalab et al., 2018). In this experiment, utilizing 'entropy' as the function to determine the quality of the division and setting the count of trees in the forest to 30 yielded improved accuracy relative to other settings.

## Assessment of Model Performance

With the aid of a confusion matrix and the area under the receiver operating characteristic curve (AUC-ROC), the effectiveness of the various classifiers outlined above has been examined. The description of these evaluation methods is presented in this section. The number of pixels in the confusion matrix that is successfully identified as landslides is known as true positive (TP), whereas the number of pixels correctly identified as non-landslides is known as true negative (TN). False positive (FP) and false negative (FN) pixels denote the proportion of pixels that are incorrectly classified as landslides and non-landslides, respectively. The different metrics such as accuracy, precision, recall, F1-score, and TSS have been derived from this confusion matrix using Eqs. (3)–(7) (Pourghasemi et al., 2013a, 2013b).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

$$F1 - \text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

$$\text{TSS} = \frac{TP}{(TP + FN)} - \frac{FP}{(FP + TN)} \quad (7)$$

When evaluating or displaying the performance of the classification task, we use the AUC-ROC as the evaluation statistic. Recall and false-positive rate (FP/(FP + TN)) are shown on the *y*-axis and *x*-axis, respectively, to represent the ROC curve (Nhu et al., 2020). AUC falls within a range of 0 and 1. AUC above 0.7 is regarded as fair in studies on landslides. The reliability of LSM is determined by Cohen's kappa statistic, which quantitatively assesses the agreement between predicted and observed values (Kalantar et al., 2020).

## Result and Analysis

The results obtained for the algorithms described in the previous section are discussed here.

### LCF Selection

According to the results of the Pearson correlation coefficient test, the categorical variables lithology, soil type, and LULC are not correlated with each other as shown in Fig. 5a. As shown by Fig. 5b, among the 11 continuous

variables, total curvature is highly correlated with other variables. Therefore, it has not been considered for further processing. The result obtained after applying the RF feature selection is shown in Fig. 6. The LCF distance to road is the most influencing factor as it is having the highest score of 0.26 as shown in the graph. It is followed by slope angle and annual rainfall with considerable information gain. Eleven LCFs with score of more than 0.03 have been taken into account for preparing LSM. Although it is challenging to establish a connection between many categories and previous landslide locations, category factors such as lithology and LULC do not contribute to predicting future disasters.

### Landslide Susceptibility Maps

Different levels of susceptibility, including very low, low, moderate, high, and very high, exist in the LSM that we have created via our research. These separations were achieved by calculating the prediction probability values for every data point and then interpolating these points over the research region. The reclassified LSMs using LR, SVM, KNN, and RF are shown in Fig. 7a, b, c, and d, respectively. The range of these classes and the distribution of these classes over the study area are shown in Table 2.

Approximately 2.34% of the entire study area is very highly prone to landslides, and almost 11.91% of the study area is incredibly susceptible to landslides, according to the results of LSM derived using the LR algorithm. Approximately 2.34% of the entire study area is very highly susceptible to landslides, and approximately 11.91% of the study area is also vulnerable to landslides, according to the results of LSM created using the SVM algorithm. The KNN classifier produced an LSM that indicates that 2.21% of the land is very sensitive to landslides and 6.90% of Goa is very highly vulnerable to them. As per the results of the RF classifier, 0.25% of Goa is extremely prone to landslides, while 3% of the area is highly vulnerable. On an average, 36.32% of the total land is vulnerable to landslides whereas the remaining portion of the study area does not have the risk of landslides.

From the visual interpretation of Fig. 7b, it is observed that the border areas of Karnataka are very susceptible to landslides, according to an LSM created using the SVM algorithm. In comparison with the other three LSMs, an LSM developed using SVM has more sites that are not at risk of landslides. The LSM generated by the LR classifier has more locations that are at high risk of landslides. The north-western locations of Parnem, Bardez, and Tiswadi are not at risk of landslides, according to the RF algorithm's output, which is displayed in Fig. 7d. There are many locations that are at moderate risk as per the results of RF. The LSMs produced in this study indicate that the

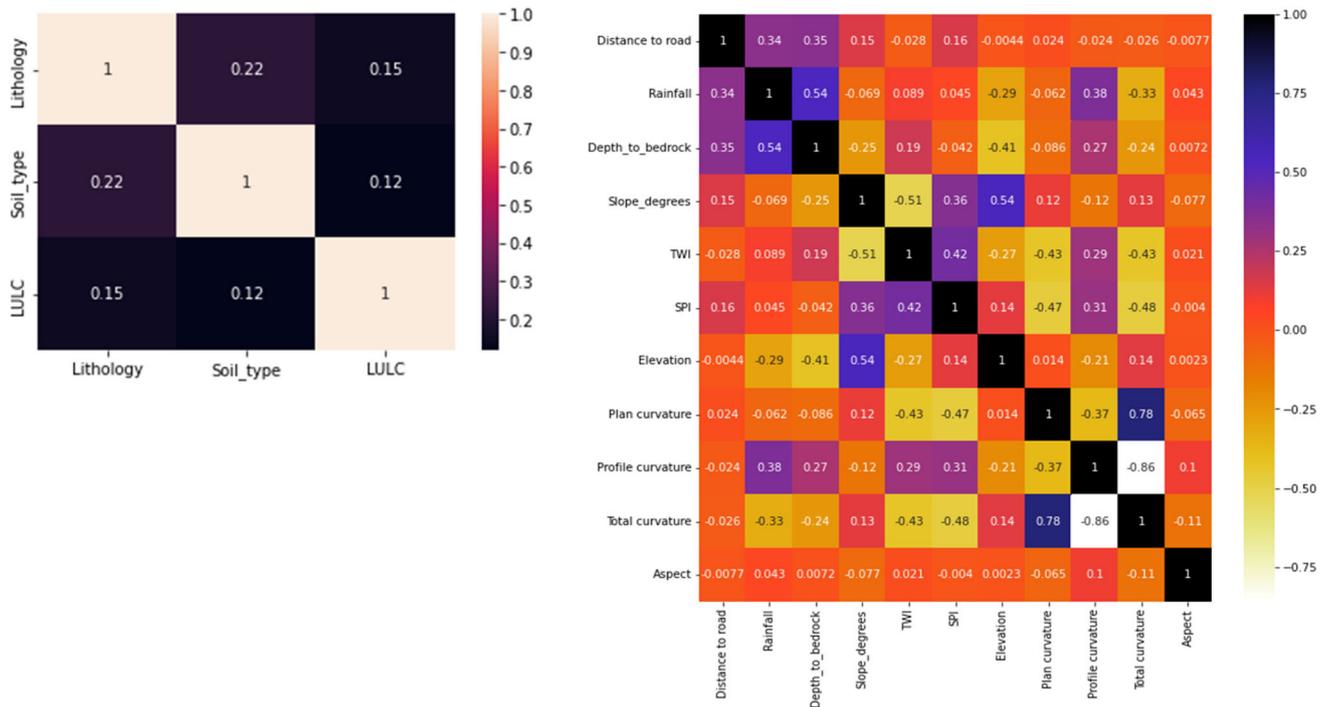


Fig. 5 Results of multi-collinearity test a for categorical variable and b for continuous variables

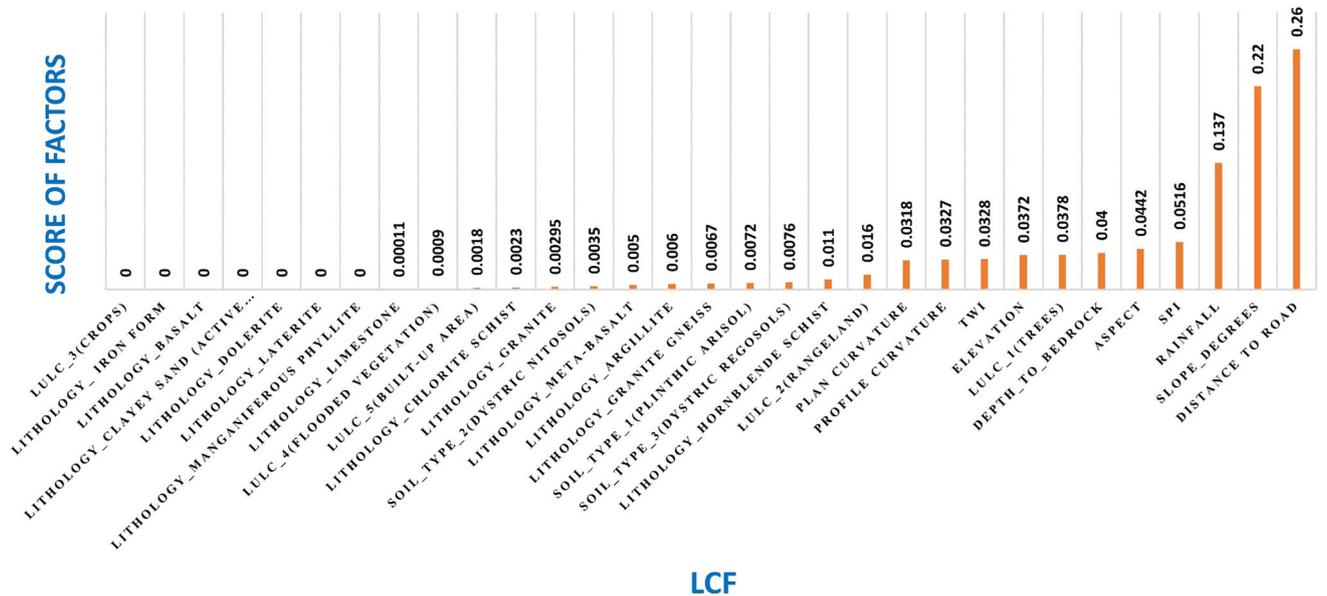


Fig. 6 Landslide influencing factor selection

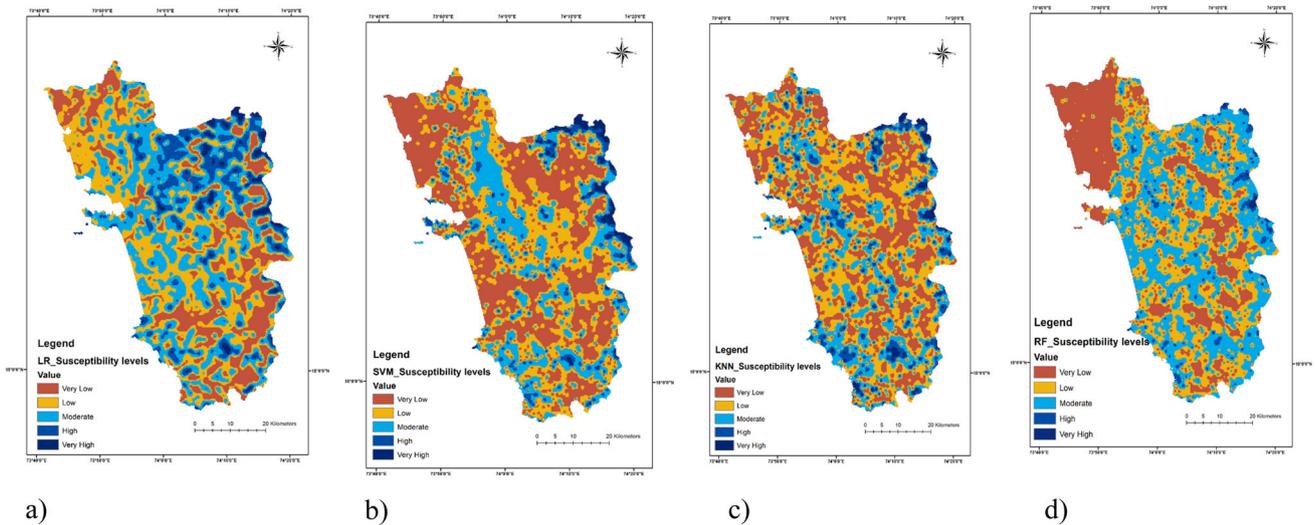
Sattari, Dharbandora, Sanguem, and Canacona regions have more locations that are dangerous compared to other regions.

**Validation of LSMs**

Different metrics such as accuracy, precision, recall, and F1-score were derived from the confusion matrix of the test

dataset. The values of these metrics shown in Table 3 used to compare the performance of classifiers LR, SVM, KNN, and RF are presented in this section. This section also shows a comparison of LSMs produced by these classifiers using the AUC-ROC parameter.

The findings of the experiments show that all of these models delivered good-quality outcomes. The highest accuracy of 86.11% reveals that the RF classifier can



**Fig.7** LSM using different ML algorithms **a** LSM using LR, **b** LSM using SVM, **c** LSM using KNN, and **d** LSM using RF

**Table 2** Distribution of percentage of area over different classes of LSMs

Susceptibility levels	ML classifiers Range of values	ML classifiers			
		LR Area (%)	SVM Area (%)	KNN Area (%)	RF Area (%)
Very low	0–0.2	20.12	38.78	32.37	27.48
Low	0.2–0.4	33.70	32.42	37.03	32.75
Moderate	0.4–0.6	31.91	22.01	21.49	36.52
High	0.6–0.8	11.91	4.62	6.90	3.00
Very high	0.8–1	2.34	2.15	2.21	0.25

**Table 3** Performance metrics values for different classifiers

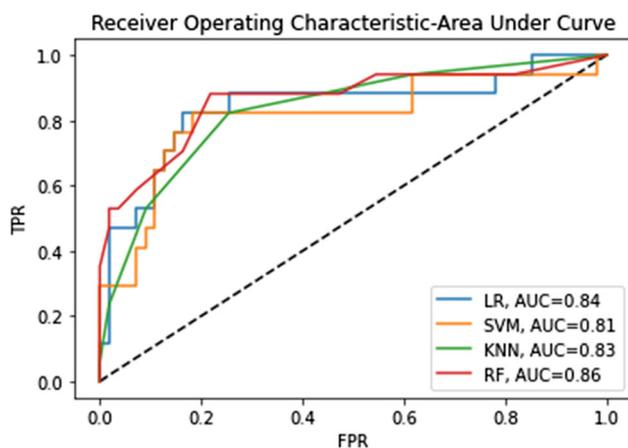
Classifier	Accuracy	Precision	Recall	F1-score	Kappa Index	TSS
LR	81.90	0.611	0.647	0.629	0.50	0.52
SVM	83.33	0.92	0.765	0.684	0.57	0.62
KNN	81.94	0.643	0.529	0.581	0.46	0.43
RF	86.11	0.733	0.647	0.74	0.59	0.58

predict the landslide and non-landslide points accurately compared to the other three models. SVM classifier can predict the positive (landslide sites) more precisely than other classifiers, as shown by the highest precision values of 0.92. Recall values of 0.64, 0.76, 0.52, and 0.64, respectively, for the various models LR, SVM, KNN, and RF show the proportion of accurate landslide predictions to all landslide spots. The SVM classifier has a maximum recall value of 0.76. The F1-score considers both landslide and non-landslide points while assessing the performance. F1-score can be utilized as a helpful indicator when comparing classifiers because of the imbalance between landslides (78 points) and non-landslides (200 points) in the given dataset. The greatest F1-score of 0.74 shows that RF outperforms the other classifiers considered in this study. When comparing the LSMs produced by the four distinct

classifiers taken into consideration in our study, the ensemble RF classifier performs slightly well when compared to the other three classifiers, according to the highest AUC-ROC value of 0.86 as shown in Fig. 8. The AUC-ROC values of 0.84, 0.81, and 0.83 for LR, SVM, and KNN, respectively, indicate that these algorithms behave almost similarly in terms of their performance.

### Discussion

Regional landslide mapping has become a popular topic due to the regular occurrence of landslides in various places of the world. Mapping the susceptibility to future landslides is a crucial task in the field of environmental protection. Due to the nonlinear nature of the LSM



**Fig. 8** Comparison of different classifiers in terms of AUC-ROC

datasets, it is difficult to execute in the field and requires significant resources (Jennifer, 2022). Parts of the Western Ghats of India have been witnessing several landslides in recent years. Goa is one of the states that lie within the Western Ghats which has been experiencing landslides during the rainy season. In recent years, there have been multiple landslides in the Indian Western Ghats. Landslides have been happening in Goa, one of the states in the Western Ghats, during the rainy season. It is evident from the literature that Goa has been a region that has not been thoroughly researched when it comes to creating LSM employing machine learning techniques. One of the causes of this can be a lack of data availability. In this work, we used the freely released landslide inventory to assess the elements that contribute to landslide fatalities in the state of Goa and then produced a landslide susceptibility map that illustrates the likelihood of landslides in a specific area. This is a result of the interaction between existing landslides and the triggering factors associated with the environment. We investigated fourteen landslide conditioning factors that affect landslide disasters in the state of Goa and created LSM utilizing the most important parameters that have the greatest impact on the forecasting of upcoming landslides. Slope, aspect, distance from road, lithology, curvature, LULC, rainfall, curvature, Topographic Wetness Index (TWI), and Stream Power Index (SPI) are examples of factors that are widely employed (Achu et al., 2022a, 2022b; Jennifer, 2022). By noticing that the majority of the landslides in the research area are rain-induced mud flows, we have added soil depth and depth to bedrock as additional factors to prepare LSM.

From the distribution of landslide points, it is clear that 55% of the total landslides lie in locations that are having a distance of 0–195.5 m from the road. Therefore, areas close to roadways have a significant probability of slope failure. Another factor that affects the occurrence of landslides in the research area is the slope. The majority of

the study area has a slope ranging from 0 to 17 degrees which can be referred to as a low-to-moderate slope. However, the eastern and southern portions of the study area that belong to WG have the steepest slopes up to 70 degrees. In areas where the slope angle ranges from 10.20 to 17.38 degrees, there is a significant density of landslides. The average annual rainfall of the study area in millimetres has been classified into 6 classes. The locations that receive rainfall of 3044–3213 mm have more landslide points compared to all other classes. SPI is also one of the factors that play a role in predicting future landslides. The values of SPI range from – 13.81 to 12.83. Elevation is another element that is thought to be a trigger. The majority of the landslide points lie in those study areas with SPI ranging from – 13.81 to – 6.08. From the inventory of landslides, it is revealed that the majority of the landslides are rainfall induced mass movements. So, depth of bedrock or soil depth has been considered in our study. Most of the landslide locations lie in a moderate class of soil depth in which depth varies from 20 to 100 m. The locations of 47% of the landslide points are between 64 and 167 m above sea level, and 29% are between 332 and 541 m. The landslide density of the TWI factor is high in the range of 2.5–5.9 and low at 15.47–24.59. LCFs such as total curvature, LULC, and lithology have not contributed much in predicting landslides of the future. It is challenging to establish a relationship between landslide points and LCFs with numerous categories because there are not enough landslide points in the provided study area.

The ML algorithms LR, SVM, KNN, and RF use a portion of the points from landslide inventories and the accompanying LCF values as training data. Based on the results of these classifiers, the LSM is prepared. The landslide susceptibility maps of the study area were classified into a fivefold classification scheme of very low, low, moderate, high, and very high susceptibility classes. The test dataset which was compiled randomly is used to determine the AUC. Other criteria were utilized to compare the suggested models, including accuracy, precision, recall, F1-score, Kappa Index, and TSS.

RF had the greatest prediction rate when the four models were validated making it the most trustworthy approach. The outcomes could also derive from the existence of linear and nonlinear correlations between the numerous geographical data that RF could trustworthily take into account. The advantages of the RF model include the fact that it has robust and accurate ML algorithms and is considered a highly accurate classifier for many datasets. It can run efficiently on large databases, manage enormous input variables without elimination, estimate effective factors in the classification, and generate an internal, unbiased estimate of the generalization error as the forest building progresses. RF algorithm is suitable for developing LSMs

of a larger scale (Taalab et al., 2018). The result of our study justifies the nature of ensemble algorithms over single classifiers. The land use and disaster management policies in the research region can be decided upon using the LSM produced by the RF algorithm.

## Conclusion

LSM is required to reduce the disaster, property losses, and human casualties brought on by landslides. In parts of the Western Ghats, landslides have recently grown to be a particularly sensitive subject. This study sheds light on ML-based LSM modelling, which is based on ten conditioning elements of the study area. The combination of GIS and ML techniques helps to identify the associated relationship between the historical landslide points and landslide conditioning elements. Further, these associations can be used to forecast the landslides that may occur in the future. According to the study's findings, the ensemble RF method is more effective than the other three strategies and is a resilient algorithm to prepare the LSM of the study area. The study shows that some of the regions such as Sattari, Dharbandora, Sanguem, and Canacona require some actions to be taken to mitigate the risk of landslide disasters as it includes more number of landslide-prone locations. Also, the study reveals that North Goa has less risk associated with landslides. Planning agencies are very interested in creating LSM for early hazard analyses, particularly when a regulation planning policy is going to be enacted. Landslide prediction was made simpler by the use of ArcGIS software and the availability of Python modules.

In subsequent research, we intend to use cutting-edge ML and DL approaches to increase the classifiers' accuracy and generate more reliable findings. The performance would also be improved by conducting additional research on the study region to generate more landslide conditioning components and by using a landslide inventory with more landslide points.

**Funding** Open access funding provided by Manipal Academy of Higher Education, Manipal.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate

if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Achu, A. L., Aju, C. D., Pham, Q. B., Reghunath, R., & Anh, D. T. (2022a). Landslide susceptibility modelling using hybrid bivariate statistical-based machine-learning method in a highland segment of Southern Western Ghats, India. *Environmental Earth Sciences*, 81(13), 1–18. <https://doi.org/10.1007/s12665-022-10464-z>
- Achu, A. L., Aju, C. D., & Reghunath, R. (2020). Spatial modelling of shallow landslide susceptibility: a study from the Southern Western Ghats Region of Kerala, India. *Annals of GIS*, 26(2), 113–131. <https://doi.org/10.1080/19475683.2020.1758207>
- Achu, A. L., Thomas, J., Aju, C. D., Remani, P. K., & Gopinath, G. (2022b). Performance evaluation of machine learning and statistical techniques for modelling landslide susceptibility with limited field data. *Earth Science Informatics*, 16, 1025–1039. <https://doi.org/10.1007/s12145-022-00910-8>
- Avand, M., Janizadeh, S., Naghibi, S. A., Pourghasemi, H. R., Bozchaloei, S. K., & Blaschke, T. (2019). A comparative assessment of random forest and K-Nearest Neighbor Classifiers for Gully Erosion susceptibility mapping. *Water (switzerland)*, 11(10), 2076. <https://doi.org/10.3390/w11102076>
- Babitha, B. G., Danumah, J. H., Pradeep, G. S., Costache, R., Patel, N., Prasad, M. K., Rajaneesh, A., Mammen, P. C., Ajin, R. S., & Kuriakose, S. L. (2022). A framework employing the AHP and FR methods to assess the landslide susceptibility of the Western Ghats Region in Kollam District. *Safety in Extreme Environments*, 4(2), 171–191. <https://doi.org/10.1007/s42797-022-00061-5>
- Chen, W., Peng, J., Hong, H., Shahabi, H., Pradhan, B., Junzhi Liu, A., Zhu, X., Pei, X., & Duan, Z. (2018). Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. *Science of the Total Environment*, 626, 1121–1135. <https://doi.org/10.1016/j.scitotenv.2018.01.124>
- Chen, Y., Ming, D., Ling, X., Lv, X., & Zhou, C. (2021). Landslide susceptibility mapping using feature fusion-based CPCNN-ML in Lantau Island, Hong Kong. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 3625–3639. <https://doi.org/10.1109/JSTARS.2021.3066378>
- Dias, H. C., Hölbling, D., & Grohmann, C. H. (2021). Landslide susceptibility mapping in Brazil: A review. *Geosciences (switzerland)*, 11(10), 1–15. <https://doi.org/10.3390/geosciences11100425>
- Dou, J., Yunus, A. P., Bui, D. T., Merghadi, A., Sahana, M., Zhu, Z., Chen, C. W., Han, Z., & Pham, B. T. (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>
- Feby, B., Achu, A. L., Jimnisha, K., Ayisha, V. A., & Reghunath, R. (2020). Landslide susceptibility modelling using integrated evidential belief function based logistic regression method: A study from Southern Western Ghats, India. *Remote Sensing*

- Applications: Society and Environment*, 20, 100411. <https://doi.org/10.1016/j.rsase.2020.100411>
- Ganesh, B., Vincent, S., Pathan, S., & Benitez, S. R. G. (2023). Machine learning based landslide susceptibility mapping models and GB-SAR based landslide deformation monitoring systems: growth and evolution. *Remote Sensing Applications: Society and Environment*, 29, 100905. <https://doi.org/10.1016/j.rsase.2022.100905>
- Guiling, W., Kazi, F. A., Liangzhi, Y., Miao, Y., Jeremy, P., & Zhenming, J. (2017). Mapping the global depth to bedrock for land surface modeling. *Journal of Advances in Modeling Earth Systems*, 8, 1180–1209. <https://doi.org/10.1002/2016MS000686>. Received
- Hengl, T., Jesus, J. M. D., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., et al. (2017). SoilGrids250m: global gridded soil information based on machine learning. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0169748>
- Jennifer, J. J. (2022). Feature elimination and comparison of machine learning algorithms in landslide susceptibility mapping. *Environmental Earth Sciences*, 81(20), 1–23. <https://doi.org/10.1007/s12665-022-10620-5>
- Kalantar, B., Ueda, N., Saeidi, V., Ahmadi, K., Halin, A. A., & Shabani, F. (2020). Landslide susceptibility mapping: Machine and ensemble learning based on remote sensing big data. *Remote Sensing*, 12(11), 1–23. <https://doi.org/10.3390/rs12111737>
- Kessarkar, P. M., Srinivas, K., Suprit, K. & Chaubey, A. K. (2011). Proposed landslide mapping method for canacona region. Indian Environmental Portal Report.
- Lee, S. (2019). Current and future status of gis-based landslide susceptibility mapping: A literature review. *Korean Journal of Remote Sensing*, 35(1), 179–193.
- Mascarenhas, A., & Kalavampara, G. (2009) *Natural Resources of Goa: A Geological Perspective*. <https://drs.nio.org/drs/handle/2264/4493>.
- Mattivi, P., Franci, F., Lambertini, A., & Bitelli, G. (2019). TWI computation: A comparison of different open source GISs. *Open Geospatial Data, Software and Standards*. <https://doi.org/10.1186/s40965-019-0066-y>
- Mhaddolkar, P. A. (2019). Case study of forensic investigation of landslide in Mardol Goa. Proceedings of 5th International Conference on Forensic Geotechnical Engineering, Dec 8 to 10, 2016, IISc Bangalore.
- Nagarajan, R., Roy, A., Vinod Kumar, R., Mukherjee, A., & Khire, M. V. (2000). Landslide hazard susceptibility mapping based on terrain and climatic factors for tropical monsoon regions. *Bulletin of Engineering Geology and the Environment*, 58, 275–287.
- 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), 1–23. <https://doi.org/10.3390/ijerph17144933>
- Pourghasemi, H. R., Jirandeh, A. G., Pradhan, B., Chong, Xu., & Gokceoglu, C. (2013a). Landslide susceptibility mapping using support vector machine and GIS at the Golestan Province, Iran. *Journal of Earth System Science*, 122(2), 349–369. <https://doi.org/10.1007/s12040-013-0282-2>
- Pourghasemi, H. R., Pradhan, B., Gokceoglu, C., & Moezzi, K. D. (2013b). 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. <https://doi.org/10.1080/19475705.2012.662915>
- Ramani, S. E., Pitchaimani, K., & Gnanamanickam, V. R. (2011). GIS based landslide susceptibility mapping of Tevankarai Ar Sub-Watershed, Kodaikkal, India using binary logistic regression analysis. *Journal of Mountain Science*, 8(4), 505–517. <https://doi.org/10.1007/s11629-011-2157-9>
- Rane, P. R., & Vincent, S. (2022). Landslide susceptibility mapping using machine learning algorithms for Nainital, India. *Engineered Science*, 17, 142–155. <https://doi.org/10.30919/es8d600>
- Saleem, N., Huq, M. E., Twumasi, N. Y. D., Javed, A., & Sajjad, A. (2019). Parameters derived from and/or used with digital elevation models (DEMs) for landslide susceptibility mapping and landslide risk assessment: A review. *ISPRS International Journal of Geo-Information*, 8(12), 545. <https://doi.org/10.3390/ijgi8120545>
- Taalab, K., Cheng, T., & Zhang, Y. (2018). Mapping landslide susceptibility and types using random forest. *Big Earth Data*, 2(2), 159–178. <https://doi.org/10.1080/20964471.2018.1472392>
- Tsangaratos, P., Iliu, I., Hong, H., Chen, W., & Chong, Xu. (2017). Applying information theory and GIS-based quantitative methods to produce landslide susceptibility maps in Nancheng County, China. *Landslides*, 14(3), 1091–1111. <https://doi.org/10.1007/s10346-016-0769-4>
- Vijith, H., Krishnakumar, K. N., Pradeep, G. S., Ninu Krishnan, M. V., & Madhu, G. (2014). Shallow landslide initiation susceptibility Mapping by GIS-based weights-of-evidence analysis of multi-class spatial data-sets: A case study from the natural sloping terrain of Western Ghats, India. *Georisk*, 8(1), 48–62. <https://doi.org/10.1080/17499518.2013.843437>
- Westen, Van C. J. (1993). Application of geographic information systems to landslide Hazard Zonation-Wetern.Pdf
- Wubalem, A. (2021). Landslide susceptibility mapping using statistical methods in Uatzau Catchment Area, Northwestern Ethiopia. *Geoenviron Disasters*, 8(1), 1–21. <https://doi.org/10.1186/s40677-020-00170-y>
- Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., & Al-Katheeri, M. M. (2016). Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides*, 13(5), 839–856. <https://doi.org/10.1007/s10346-015-0614-1>
- Yunus, A. P., Fan, X., Subramanian, S. S., Jie, D., & Qiang, X. (2021). Unraveling the drivers of intensified landslide regimes in Western Ghats, India. *Science of the Total Environment*, 770, 145357. <https://doi.org/10.1016/j.scitotenv.2021.145357>
- Zhou, C., Yin, K., Cao, Y., Ahmed, B., Li, Y., Catani, F., & Pourghasemi, H. R. (2018). Landslide susceptibility modeling applying machine learning methods: a case study from longju in the three gorges reservoir Area, China. *Computers and Geosciences*, 112(April 2017), 23–37. <https://doi.org/10.1016/j.cageo.2017.11.019>