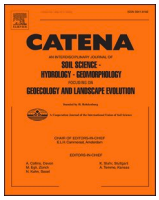
Catena 235 (2024) 107653

Contents lists available at ScienceDirect

Catena

journal homepage: www.elsevier.com/locate/catena



High resolution landslide susceptibility mapping using ensemble machine learning and geospatial big data

Nirdesh Sharma [a], Manabendra Saharia [a,b,*], G.V. Ramana [a]

[a] Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
[b] Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India

ARTICLE INFO

Keywords:
Ensemble learning
Big data
One-sided selection (OSS)
Support Vector Machine Synthetic Minority Oversampling Technique (SVMSMOTE)
High resolution landslide susceptibility

ABSTRACT

Landslide susceptibility represents the potential of slope failure for given geo-environmental conditions. The existing landslide susceptibility maps suffer from several limitations, such as being based on limited data, heuristic methodologies, low spatial resolution, and small areas of interest. In this study, we overcome all these limitations by developing a probabilistic framework that combines imbalance handling and ensemble machine learning for landslide susceptibility mapping. We employ a combination of One -Sided Selection and Support Vector Machine Synthetic Minority Oversampling Technique (SVMSMOTE) to eliminate class imbalance and develop smaller representative data from big data for model training. A blending ensemble approach using hyperparameter tuned Artificial Neural Networks, Random Forests, and Support Vector Machine, is employed to reduce the uncertainty associated with a single model. The methodology provides the landslide susceptibility probability and a landslide susceptibility class. A thorough evaluation of the framework is performed using receiver operating characteristic curves, confusion matrices, and the derivatives of confusion matrices. This framework is used to develop India's first national-scale machine learning based landslide susceptibility map. The landslide database is carefully curated from global and local inventories, and the landslide conditioning factors are selected from a multitude of geophysical and climatological variables. The Indian Landslide Susceptibility Map (ILSM) is developed at a resolution of 0.001° (~100 m) and is classified into five classes: very low, low, medium, high, and very high. We report an accuracy of 95.73 %, sensitivity of 97.08 %, and matthews correlation coefficient (MCC) of 0.915 on test data, demonstrating the accuracy, robustness, and generalizability of the framework for landslide identification. The model classified 4.75 % area in India as very highly susceptible to landslides and detected new landslide susceptible zones in the Eastern Ghats, hitherto unreported in the government landslide records. The ILSM is expected to aid policymaking in disaster risk reduction and developing landslide prediction models.

1. Introduction

Landslides are one of the most devastating geohazards causing acute loss of life and property. According to EM-DAT, landslides are responsible for 17 % of natural disaster-related deaths worldwide and billions of dollars in annual damages (CRED, 2022). Recent studies suggest that landslides' frequency and socio-economic impact are increasing with more communities being exposed to landslides (Alimohammadlou et al., 2013; Highland et al., 1998; Yalcin et al., 2011). Moreover, the impact of landslides is often underestimated due to the absence of systematically collected data (Froude and Petley, 2018; Sim et al., 2022). Identification of landslide-prone areas with a high degree of accuracy is necessary to reduce the risk associated with landslides (Azarafza et al., 2021; Castellanos Abella and Van Westen, 2008). With an improvement in computing technology, there is an opportunity to develop an improved landslide susceptibility map by leveraging remote sensing and ground-based datasets as well as state-of-the-art machine learning techniques.

Landslides occur when gravity forces pushing on hillslope material exceed the frictional forces holding the material in place, causing slope failure. Landslide susceptibility represents this potential of slope failure (Reichenbach et al., 2018) and a landslide susceptibility map divides the terrain into zones based on the likelihood of landslide occurrence. The landslide susceptibility of any area can be determined before the occurrence of a landslide event by assuming that future landslides would occur under identical conditions as previous landslides (Guzzetti et al., 2006). Landslides are caused by complex interactions of geological,

* Corresponding author at: Indian Institute of Technology Delhi, New Delhi 110016, India.

https://doi.org/10.1016/j.catena.2023.107653
Received 15 February 2023; Received in revised form 12 October 2023; Accepted 31 October 2023

geomorphological, hydrological, and meteorological characteristics. Modeling such complex processes requires sophisticated approaches that can map the non-linear relationships between landslide occurrence and landslide governing variables. To this end, various qualitative and quantitative methods have been developed. The qualitative methods include geomorphological analysis and heuristic methods, whereas the quantitative approaches are based on statistics, physics, or numerical equations (Wieczorek, 1996). The qualitative methods are limited since they are simplistic and based on expert judgement. The quantitative methods can further be classified as physics-based methods and statistical methods. The physics-based methods simulate the physical process to capture the processes leading to slope instability (Li et al., 2016). In contrast, the statistical methods employ data driven approaches to model the landslide process.

Machine learning methods are a subset of statistical methods that model the underlying process using data. Machine learning methods are powerful information processing techniques that can augment our understanding of landslide processes. The main advantages of machine learning models include their objectivity, reproducibility, and ability to be continually updated with new data. Once trained, the machine learning models can be used to determine the potential probability of landslides. Earlier studies have compared the utility of various machine learning models in developing landslide susceptibility maps. For instance, Bayesian Network (BN), radial basis function (RBF) classifier, logistic model tree (LMT), and random forest (RF) models (Chen et al., 2018), random forest and XGBoost (Meena et al., 2022), decision tree (DT), support vector machine (SVM), and neuro-fuzzy inference system (ANFIS) (Pradhan, 2013), and Support Vector Machines (SVM), Logistic Regression (LR), Fisher's Linear Discriminant Analysis (FLDA), Bayesian Network (BN), and Naïve Bayes (NB) (Pham et al., 2016a) have compared the predictability of machine learning models for landslides prediction.

Recently ensembles of multiple machine learning models have been used to develop landslide susceptibility maps, which have achieved an increased accuracy compared to base classifiers (Kavzoglu and Teke, 2022; Sahin, 2020). Although there are many studies which use homogenous landslide susceptibility models and simple ensemble strategies, the heterogenous ensemble learning models are not yet extensively explored for landslide susceptibility mapping (Fang et al., 2021).

There have been limited landslide susceptibility studies at high resolution on a national or global scale due to the unavailability of landslide inventories and required computational power (Bălteanu et al., 2010; Okalp and Akgün, 2016). However, with the advent of high-resolution remote sensing and ground mapped geospatial data, it is now possible to develop high resolution landslide susceptibility maps. Currently, most large-scale landslide susceptibility maps are based on heuristic methods and are available at coarse resolution. Some of the large-scale landslide susceptibility maps available in literature are shown in Table 1.

Table 1
Large scale landslide susceptibility maps.

Study Area	Resolution	Methodology	Inventory	Reference
Global	1000 m	Heuristic fuzzy	62,898	(Stanley and Kirschbaum, 2017)
Global	1000 m	Weighted linear combination	3000	(Nadim et al., 2006)
Global	0.25°	Weighted linear combination	555	(Hong et al., 2007)
China	0.01°	Artificial Neural Networks	1200	(Liu et al., 2013)
Georgia	100 m	Weighted linear combination	1350	(Gaprindashvili and Van Westen, 2016)
Europe	1000 m	Analytical Hierarchy processes	102,000	(Günther et al., 2014)
Iran	85 m	Deep Learning (CNN and RNN)	4069	(Thi Ngo et al., 2021)

To overcome all these shortcomings, we present a general framework consisting of 7 steps: data collection, data preprocessing, machine learning based modelling, hyperparameter tuning, ensemble generation, performance evaluation, output generation and visualization. The framework is deployed for the political boundary of India to generate the India Landslide Susceptibility Map (ILSM), which is India's first national-scale landslide susceptibility map at high resolution (0.001°).

2. Study area

The study area covers the political boundary of India. India accounts for nearly 8 % of global landslide fatalities (Ram and Gupta, 2022). From 2001 through 2021, India's average annual landslide death count is 847, and the average annual financial losses amount to $0.3 billion (CRED, 2022). In 2018 heavy rainfall and landslides caused about 500 casualties in Kerala, Karnataka, and Tamil Nadu (Martha et al., 2021). Recently floods and landslides in Assam killed at least 14 people and displaced 1.7 million people across 29 districts (https://reliefweb.int/disaster/fl-2022-000213-ind).

India exhibits some of the highest diversity in geology and topography. Due to steep slopes and heavy rainfall most of the landslides occur in the northwest himalayas followed by the northeast himalayas and the western ghats (Martha et al., 2021).

The Himalayas are composed of sedimentary rocks which are prone to denudation and erosion. Furthermore, the steep slopes and rapid flowing rivers cause a large amount of toe erosion making the slope unstable. Therefore, most of the landslides in himalayas are rockfalls. On the contrary western ghats have basalt rocks, and rivers with gentle slopes thereby resulting in fewer rockfalls. However, weathering due to heavy rainfall has led to a development of thick layer of regolith, thereby leading to mudslides (Martha et al., 2021). Another important geological factor is the slope of the terrain. While most of the landslides in the western ghats are associated with steep slopes, most of the landslides in northeast himalayas are associated with gentler slopes due to a compressed fold and fault sequence which leads to reduced shear strength (Martha et al., 2021).

Apart from geology, environmental and anthropogenic factors play an important role in the spatio-temporal variability of landslides. Rainfall intensity and duration are the most important environmental factor for triggering landslides, consequently most of the landslides occur on the areas towards the windward side of western ghats and Himalayas. However, the western ghats require less rainfall to trigger landslides when compared to Himalayas due to high soil depth which allows more water retention and an increased porewater pressure ultimately leading to landslides, whereas the Himalayas have exposed rocks which require large rainfall to increase porewater pressure (Martha et al., 2021). Among the anthropogenic factors, road development, construction add to slope instability and increase the risk of landslides e.g. In Sikkim most of the landslides occurred in urban areas which have been attributed to urbanization and infrastructure development (Singh et al., 2020). A more comprehensive review of the spatio-temporal variability of geology and landslides can be found in (Martha et al., 2021; Valdiya, 2015).

The landslide maps prepared in India till 2013 are limited to important transportation corridors and discrete locations which witnessed heavy damage due to landslides. In 2013 floods and landslides in Uttarakhand impacted 12 of the 13 districts, caused thousands of deaths, and left nearly 75,000 pilgrims stranded (https://reliefweb.int/report/india/uttarakhand-flash-floods-%E2%80%93-report). In response to the Uttarakhand disaster, the Geological Survey of India (GSI) launched the National Landslide Susceptibility Mapping (NLSM) project, which has since produced a 1:50,000 scale landslide susceptibility map of 85 % of the total target area in landslide-prone areas ("Lok Sabha," 2021). The NLSM project models susceptibility using an analytic hierarchy process

(AHP) where an assigned weight is allocated for various factors based on the Bureau of Indian Standards, (1998) guidelines. The AHP makes the landslide maps subjective and reduces the usability of the model; also since the maps are on a scale of 1:50000, smaller landslides cannot be detected ("NDMA," 2019). Therefore, we need a strategy to generate a finer scale susceptibility map for India using data-based methods.

3. Datasets

A wide variety of factors are significant in influencing landslide susceptibility and have been extensively studied (Guzzetti et al., 2006). In landslide susceptibility studies, these parameters are required to model the shear strength of soil, soil–water interaction, soil vegetation interaction, and the impact of anthropogenic activities. The landslide conditioning factors are available in different formats and spatial resolutions.

Chang et al. (2019) suggests that susceptibility maps with high resolution topographic data may be inaccurate due to noise in the data, meanwhile when mapping landslide susceptibility at coarse resolution, the number of landslide conditioning factors required is more (Gaidzik and Ramírez-Herrera, 2021). Based on the study area, available spatial resolution, and data availability, we selected 16 landslide conditioning factors at a spatial resolution of 0.001°. The datasets developed are discussed below:

3.1. Landslide inventory

Landslide inventory is a systematic record of landside location, extent of landslide and other characteristics. Landslide inventory is important for mapping landslide susceptibility, landslide risk, landslide early warning and understanding the evolution of landscapes especially in hilly regions which are dominated by landslides. Developing a Landslide susceptibility is landslide inventory is the primary step towards developing landslide susceptibility since susceptibility is based on the idea that future landslides will occur in similar conditions as past landslides.

Landslides cause a discernable impact on the terrain which is used to delineate the geographical extent of these landslides. This extent is then mapped using field surveys, aerial photogrammetry, or satellite imagery. The selection of the mapping technique is based on a multitude of factors such as the size of the study area, data availability, and the availability of resources.

In this study, landslide inventory was formed by merging landslide inventories acquired from the Geological Survey of India and the Cooperative Open Online Landslide Repository (COOLR) (Juang et al., 2019). Prior to the widespread availability of high-resolution satellite imagery, the Geological Survey of India (GSI) primarily relied on field surveys to construct landslide inventories. However, this approach posed significant limitations, particularly in regions characterized by mountainous slopes. Following a thorough review of multiple GSI reports, it is evident that contemporary landslide inventories are now predominantly created through multitemporal analysis using Google Earth imagery.

Before the advent of high resolution satellite imagery GSI primarily relied on field surveys to construct landslide inventories. However, this approach was costly and posed severe limitations, particularly in inaccessible mountainous regions. Our analysis of multiple landslide reports from GSI leads us to the conclusion that current landslide inventories are developed using multitemporal analysis on Google Earth imagery. Furthermore, to enhance the accuracy of these inventories, some of the accessible landslides are verified through on-site field visits. (https://www.gsi.gov.in/webcenter/portal/OCBIS/pageQuickLinks/pageLandslideHazardReport).

GSI's landslide inventory contains 49,105 landslides mapped as points and 105,224 landslides mapped as polygons. COOLR repository, on the other hand, is a citizen science-based landslide repository. The COOLR data selection and preparation are carried out by eliminating data entries whose location is not accurate (Stanley and Kirschbaum, 2017). Fig. 1 shows the case study area boundary and the landslide inventory.

3.2. Landslide conditioning factors

A landslide susceptibility model aims to develop a relationship between landslide inventory and landslide conditioning factors. Based on extensive literature review our initial experiments consisted of 19 landslide conditioning factors. Three variables, namely LULC data, distance from faults and drainage density of rivers had zero random forest feature importance. The feature importance was also validated using the global feature importance method Partial Dependence Plots for SVM and ANN. Therefore, we removed these features from landslide conditioning factors database. Finally we develop 16 landslide conditioning which are shown in Table 2.

3.2.1. DEM (Digital Elevation Model)

Digital Elevation Models represent the terrain of an area, in the form of a raster grid. DEMs help to identify intricate terrain features for landslide hazards. The input DEM is extracted from MERIT DEM, a high accuracy global DEM at three arc-second resolution developed using multisensor data fusion (Yamazaki et al., 2017). MERIT is designed to improve upon the original SRTM DEM by using advanced techniques to reduce errors and fill gaps in the data. MERIT DEM is significantly improved over flat regions, along with a better representation of rivers and valleys, which is essential for landslide mapping (Yamazaki et al., 2017). DEMs are also used to derive secondary inputs like slope, aspect, curvature, and topographic wetness index. The slope dictates the shear stress and hydrological process, whereas the curvature dictates the weathering and erosion processes. The aspect shows the slope direction and is an essential and complex variable for landslide susceptibility. Aspect is also closely related to climatic conditions especially the variations in solar radiation, soil moisture, and temperature distribution which impact the occurrence of landslides (Cellek, 2021). Since 0° aspect and 360° aspect are identical, the numerical values of the aspect do not accurately depict the aspect value. The aspect was divided into nine directional groups to account for the impact of the aspect (Youssef and Pourghasemi, 2021).

The topographical wetness index (TWI) estimates locations where water will accumulate. Consequently, an area with higher TWI will be associated with more landslides. We also include upslope and downslope curvature of pixels to model landslide susceptibility. The average of all hydrologically upslope pixels is represented by upslope curvature, whereas the average of all downslope pixels is represented by downslope curvature. The datasets for slope, aspect, curvature, and Topographic Wetness Index are derived from MERIT DEM in Google Earth Engine (Gorelick et al., 2017).

3.2.2. NDVI

The Normalized Difference Vegetation Index or NDVI is developed using near-infrared and infrared imagery reflectance since green vegetation strongly absorbs red light and reflects infrared rays. NDVI is an indicator of healthy vegetation and is used to incorporate the vegetation-soil interaction, which is a crucial factor influencing landslides. To develop NDVI map we first calculate the median NDVI for monsoon months from sentinel 2 data after cloud masking. Afterwards we use month wise median of images to get monthly NDVI. This process is repeated for years 2015–2020. The final NDVI map is an average of all the NDVIs.

3.2.3. Soil composition

The soil composition, which determines the soil's shear strength and drainage capacity, is an essential factor influencing landslide susceptibility. Fine-grained soils have smaller particle sizes and a higher surface

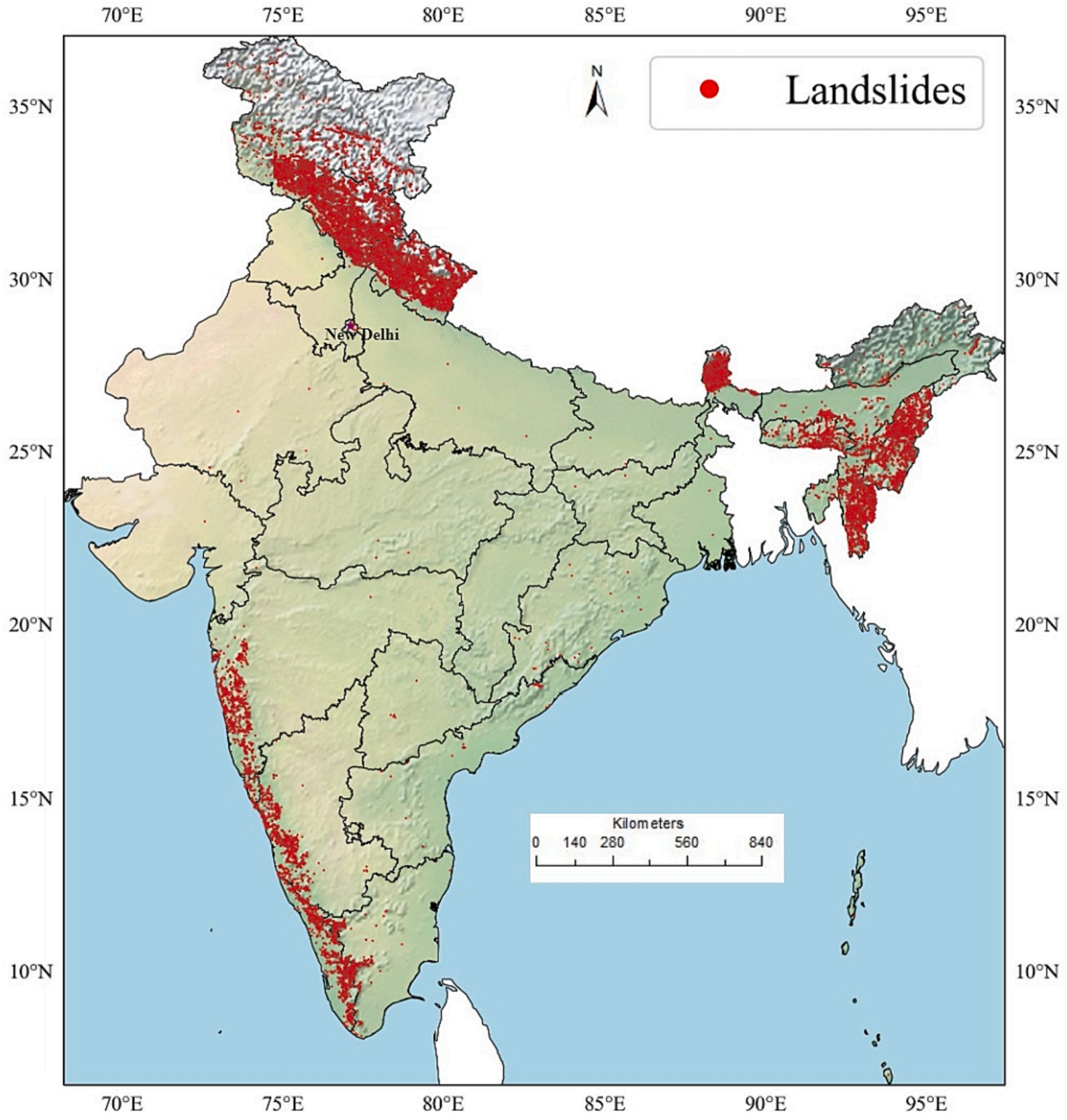


Fig. 1. Landslide inventory for India.

area, making them more prone to erosion and slope failure. They also tend to have lower permeability, meaning they can absorb and drain less water. As a result, they can become saturated with water and lose their stability, leading to landslides.

On the other hand, coarse-grained soils have larger particle sizes and a lower surface area, making them more resistant to erosion and slope failure. They also tend to have higher permeability, implying they can absorb and drain water better, which helps stabilize the soil and reduce the risk of landslides.

Soil properties, namely sand fraction, silt fraction, and clay fraction, are incorporated from the International Soil Reference and Information Centre (ISRIC) soil grids dataset. ISRIC develops soil properties and classes using global covariates and globally fitted models (Laura and de Sousa, 2020; Poggio and de Sousa, 2020b, 2020a).

3.2.4. *Anthropogenic factors*

Anthropogenic interventions such as construction and toe cutting destabilize the slopes leading to landslides. In hilly regions, most landslides occur along transportation corridors since transportation corridors are developed by cutting hills, destabilizing the slope, and thereby causing an increase in the number of landslides. In this study, we use road maps from the GeoSadak platform (https://geosadak-pmgsy.nic.in/), which have been meticulously ground mapped under the directive of the Government of India and contain 1,027,269 major and minor roads mapped as polylines. This data explicitly records rural habitations in distant areas not adequately covered by previous datasets. The road polyline is used to prepare a raster containing the euclidean distance from the nearest road, and this is the first time this detailed road network data has been used for landslide susceptibility studies.

3.2.5. *Meteorological factors*

Most slope failures in India are triggered by meteorological events, specifically intense and prolonged rainfall. Heavy precipitation increases soil weight and reduces shear resistance, making the slope vulnerable to failure. To incorporate the impact of precipitation on slope failure, the precipitation of the wettest month from worldclim data is

Table 2
Landslide Conditioning Factors.

Attribute Name	Description	Data Type	Data Source
Elevation	Height in meters	Raster	MERIT DEM (Digital Elevation Model) (Yamazaki et al., 2017)
Slope	Rate of change of elevation	Raster	
Aspect	Direction of slope in degrees	Raster	
TWI	Location of water accumulation	Raster	
Curvature	Shape of slope	Raster	
Upslope Curvature	Average curvature of upslope pixels	Raster	
Downslope Curvature	Average curvature of downslope pixels	Raster	
Sand	Fraction of Sand in the soil	Raster	ISRIC (Laura and de Sousa, 2020; Poggio and de Sousa, 2020a, 2020b)
Silt	Fraction of Silt in the soil	Raster	
Clay	Fraction of clay in the soil	Raster	
Roads	Distance from urban and rural roads	Vector (line)	PMGSY
Rivers	Minor and Major rivers of India	Vector (line)	CWC (Central Water Commission)
Precipitation	Precipitation of the wettest month	Raster	World Clim (Fick and Hijmans, 2017)
NDVI	Area of green vegetation	Raster	MODIS
Landslide Lineament	Zone of faults and Fractures	Vector (line)	GSI
Erosivity factor	India Rainfall Erosivity Dataset	Raster	IIT Delhi (Raj et al., 2022)
Landslide inventory	Database of historical landslides	Vector (Points, Polygons)	GSI

used (Fick and Hijmans, 2017; Hijmans et al., 2005). The precipitation of the wettest month from worldclim has been previously utilized in multiple landslide susceptibility studies to account for the spatial distribution of rainfall (Dinanta et al., 2020; Ramachandra et al., 2013).

3.2.6. Distance from rivers

Moving water continuously causes soil erosion, removal of toe support, and water seepage from rivers thereby reducing the soil strength and increasing the likelihood of landslides along riverbanks. The euclidean distance to river map is developed using a line map for India's major rivers, minor rivers, and rivulets acquired from the Central Water Commission. The line map is processed to create a raster map of 0.001° (100 m) resolution where every pixel denotes the euclidean distance from the nearest river.

3.2.7. Rainfall erosivity

Rainfall erosivity represents the kinetic energy of rainfall intensity. Falling raindrops exert pressure on the surface and cause instability of the soil surface, leading to soil erosion and landslides. Higher rainfall erosivity increases the chance of landslides (Ahmad et al., 2019). Despite its importance, erosivity is a highly underutilized variable in landslide research. In this study, we utilize the Indian Rainfall Erosivity Dataset (IRED), the first national-scale mapping of rainfall erosivity factor (R factor) over India, to incorporate the impact of erosivity on landslides (Raj et al., 2022).

3.2.8. Landslide lineament

The structural geology of an area considerably impacts the occurrence of landslides (Anbalagan and Singh, 1996; Ramli et al., 2010). To incorporate the impact of structural geology, we use landslide lineaments which are features caused by joints and faults. The landslide lineament data has been developed by GSI and is available at bhukosh (https://bhukosh.gsi.gov.in/), India's national geological data portal.

4. Methodology

The landslide inventory consists of 105,224 landslides mapped as polygons and 49,105 landslides mapped as points from GSI. The COOLR database contains 1820 landslide points, but only 489 landslides are selected for model development since other data points are not mapped accurately enough for high-resolution landslide mapping. The landslide conditioning factors and landslide inventory is divided into pixels of 0.001° *0.001°. If any pixel has a point landslide, landslide conditioning factors of the pixel are assumed to be landslide causing. In case of polygon data, all the pixels covered by landslide polygon were assumed to be landslide pixels. The overall methodology followed to develop a high-resolution landslide susceptibility map is shown in Fig. 2.

4.1. Data pre-processing

All the landslide conditioning datasets are resampled to a spatial resolution of 0.001°. There are no missing data points for 10 variables, and the missing data points for the rest of the 6 variables are less than 0.1 % of the total data. The closest pixels in four directions are weighed using inverse distance weighing to interpolate for the missing data points.

4.1.1. Data encoding

All the variables except the aspect are in numerical format. We need to convert the categorical aspect values into numerical features to be used as input for machine learning models. We benchmark multiple encoding methods like one-hot encoding, leave-one-out encoding, James Stein encoding, mean encoding, and catboost encoding using a subset of landslide data. Based on the results, James Stein encoding (Stein, 1956) is selected for encoding aspect.

4.1.2. Data split and normalization

If any landslide point lies inside the pixel, the pixel is assumed to be a landslide pixel. In the case of landslide polygons, the polygons are first separated into training, testing, and validation polygons and then transformed into pixels; this is important because neighborhood pixels of a single polygon might be separated into the training and validation set causing the model estimates to be over-optimistic (Emberson et al., 2021; Peña and Brenning, 2015). The testing data and validation data non-landslide pixels are randomly sampled from the entire database such that they are representative of the total data. After splitting, the datasets are scaled between 0 and 1 using a min–max scaler. Data normalization is done after splitting the data into training, validation, and testing data to prevent data leakage.

4.2. Imbalance handling of training data

Machine learning algorithms for classification are designed around the assumption of equal data for all classes. However, since landslides are localized and rare event, the landslide database is highly imbalanced. The number of data points representing the occurrence of landslides is significantly lower than those representing non-occurrence of landslides.

Even the most common performance metrics, such as accuracy, assume balanced class distribution. In case of imbalanced data, these algorithms lead to poor predictive performance, especially for minority classes which are more important since multiple highly susceptible landslide points identified as low susceptibility would degrade the usability of the model. A machine learning model trained on unbalanced data will exhibit poor predictive performance, especially in the case of landslide pixels, since it tends to be overfitted and biased towards non-landslide points. Therefore, we remove the imbalance only from the

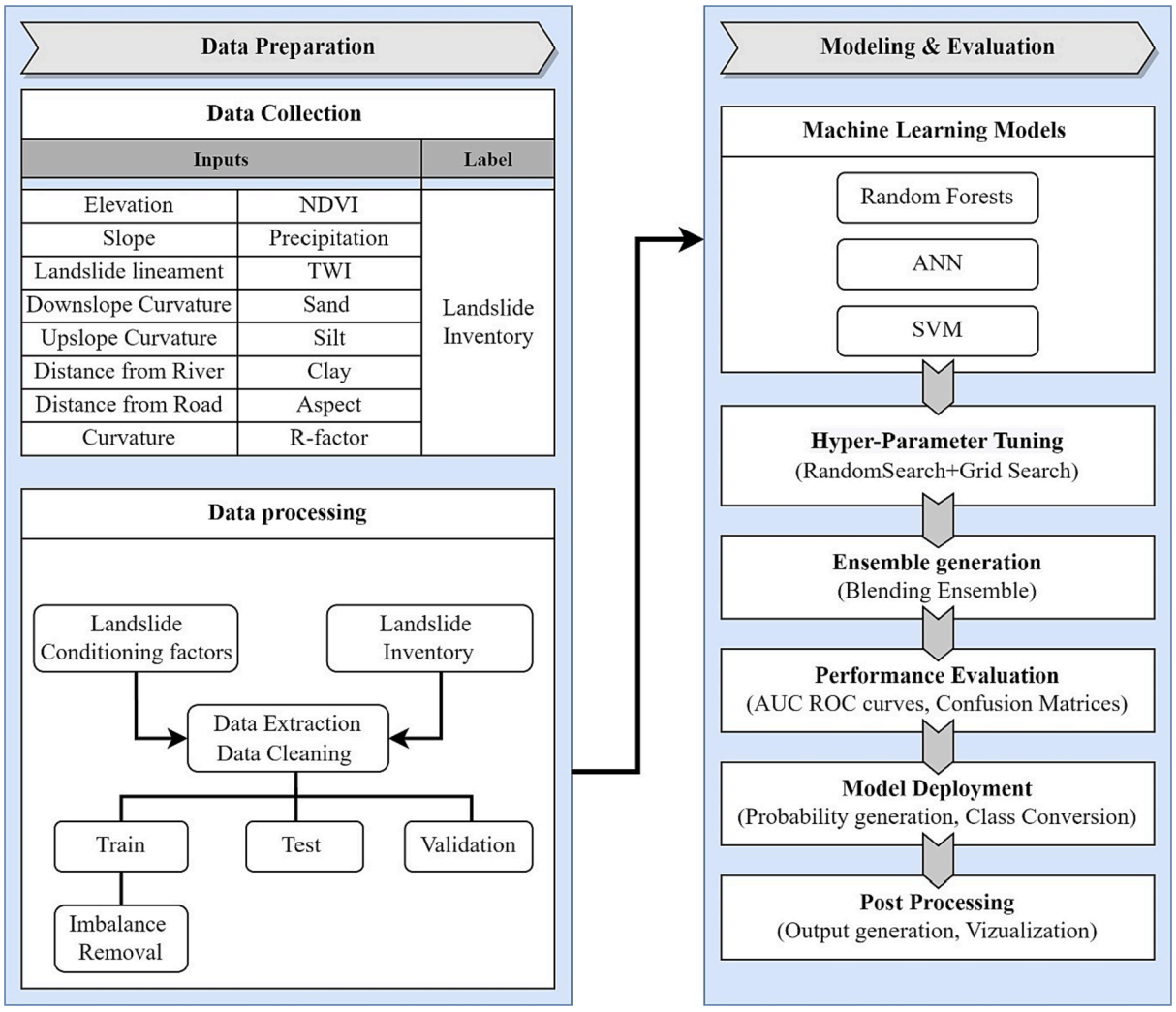


Fig. 2. Methodology for developing a high-resolution landslide susceptibility map.

training data. As explained in section 4.1.2, the validation and testing data are supposed to be a subset of the original data without including synthetic data; hence no imbalance removal techniques are applied to the testing and validation data.

A simple way to deal with class imbalance is to resample the original data into balanced data representative of the overall problem. Resampling includes oversampling the minority class or undersampling the majority class. Undersampling means reducing the number of data points in the majority class. Undersampling can be done by randomly removing the data points from the majority class, also known as random undersampling. However, random undersampling has the disadvantage of discarding potentially useful information essential for the model. To be more discerning regarding the deletion of the majority class sample, a learning model should be trained to identify redundant examples for deletion. In this study, we use One-Sided Selection (OSS) technique to reduce most data points (Kubat and Matwin, 1997). OSS first uses Tomek Links to reduce the number of ambiguous data points in the class boundary, and then it uses Condensed Nearest Neighbours (CNN) to remove redundant data points far from the class boundary. This method reduces the redundant data points from the majority class significantly.

Oversampling data means increasing the number of data points in the minority class. Random oversampling can be easily done by replicating the data points multiple times. Since similar data points are repeated in random oversampling, this technique leads to the overfitting of the model. A model fitted with random oversampling has high training accuracy but low testing accuracy.

There are several approaches to informed oversampling, out of which Synthetic Minority Oversampling Technique (SMOTE) is the most widely used (Chawla et al., 2011). SMOTE randomly selects an example from the minority class and identifies k nearest neighbors of that sample. A random nearest neighbor is then selected, and a line segment is drawn between the example and the selected nearest neighbor; a synthetic example is then created at a random point between the line. This method can be used to generate any number of synthetic samples.

SMOTE creates synthetic instances without considering the majority class, which might result in incorrect synthetic data if the majority and minority classes overlap. Hence, we also compare multiple extensions of SMOTE that consider the decision boundary to generate synthetic samples. The Borderline-SMOTE1 (BLSMOTE) (Han et al., 2005) can generate samples near the decision boundary. The SVMSMOTE (Nguyen et al., 2011) uses an SVM classifier on the original dataset to create a decision boundary and then creates synthetic samples along lines joining the minority class support vector and the data points. The Adaptive Synthetic Sampling Approach (ADASYN) (He et al., 2008) creates synthetic samples that are inversely proportional to the density of the minority class examples. All these approaches are compared using fivefold cross-validation on raw landslide data using Random Forests Classifier and accuracy as a metric. SVMSMOTE has the highest oversampling accuracy, whereas random oversampling has the lowest accuracy. The combination of undersampling and oversampling methods improves the

overall performance of machine-learning models (Chawla et al., 2011). Since the data imbalance in the case of landslides is extremely high, we firstly use OSS to undersample the data and SVMSMOTE to oversample the data to create an accurate representation of the entire dataset and enhance the performance of machine learning models. The combination of undersampling and oversampling makes the model more robust. In the case of large datasets, it reduces the majority class significantly, thereby reducing the total data required for model training.

4.3. Machine learning based modelling

Machine learning based classification methods map the interactions between the various input datasets to the label data to model the underlying fundamental processes. For modelling a spatially heterogenous area such as India with multiple landslide types, it is critical to use approaches that do not assign apriori weights to input parameters (Emberson et al., 2021). There is a growing preference for machine learning methods to avoid arbitrary parameterization (Reichenbach et al., 2018; Ahmed et al., 2020; Youssef and Pourghasemi, 2021). Machine Learning models such as Support Vector Machines (Pham et al., 2016b; Pradhan, 2013), Random Forests (Chen et al., 2018; Emberson et al., 2021), XGBoost (Kavzoglu and Teke, 2022), ANN (Hong et al., 2020) have been extensively used in landslide studies. Ensemble models that combine information from individual models to generate a more robust and accurate final model have recently been used for landslide studies (Pham et al., 2017; Felsberg et al., 2021).

The machine learning classifiers used in this study are discussed below.

4.3.1. Random Forest classifier

The idea behind bagging is that by training multiple models on different subsets of the data, each model will be able to learn from different variations in the data. Bagging classifiers use a random subset of the dataset with replacement to train various weak models. The majority voting aggregates these final predictions of all weak models, thereby reducing the model's variance, which can improve the model's ability to generalize to new data making more accurate predictions. We use Random Forest model (Breiman, 2001), where bagging is used to train various decision tree models. The trees are independent of each other and have equal weightage in the final output. Since each dataset is randomly sampled, Random Forest has a higher variance and low bias than Decision Trees.

4.3.2. Artificial neural networks

Artificial Neural Network (ANN) is a machine learning model inspired by the brain's neural pathways. ANNs are interconnected nodes that receive input from other neurons and then process the input using an activation function. The output is then passed on to other neurons in the network. The strength of the connections between the neurons, called weights, which determines the importance of the input received by each neuron. The weights are adjusted during the learning process by the training data. We use feedforward neural networks with a back-propagation algorithm. The neural network has rectified linear units as activation functions and Adam optimizer.

4.3.3. SVM classifier

The SVM classifier assumes that data not linearly separable will become linearly separable if transformed into a higher dimension. The transformation of data into higher dimension is done using a kernel function. SVM uses the transformed data to define the decision boundary known as a hyperplane (Cortes and Vapnik, 1995). SVMs have several advantages, including high accuracy, the ability to handle high-dimensional data, and the ability to perform well on small datasets. However, they are computationally intensive because their time complexity increases more than quadratically. Hence with large samples, it becomes difficult to scale. This study uses an SVM classifier with a radial basis function to define a hyperplane.

4.4. Hyperparameter tuning

Machine learning models learn the model parameters when provided with training data. Hyperparameters, on the other hand, are a set of external parameters of a model that decide the model's architecture and are not derived from data. Hyperparameters are set before training the model and vary from model to model. To identify the hyperparameters, we use coarse to fine-tuning, which uses random search in collaboration with grid search to find the optimal set of hyperparameters. This methodology aids in the discovery of hyperparameters while incurring the least amount of computational expense. Firstly, a random search narrows the range for each hyperparameter, followed by a grid search to precisely specify and assess the parameter combination. In the case of random forests, we use maximum tree depth, number of estimators, and maximum features as hyperparameters. In the case of the SVM kernel, the function is kept as a radial basis function, and C and gamma hyperparameters are used for tuning. For ANN, we vary the number of neurons and hidden layers to find the optimum architecture.

4.5. Ensemble Machine learning

Ensemble machine learning fuses the results of individual models to enhance the overall predictability and robustness. Ensemble machine learning works best when individual models are not only accurate but also diverse i.e vulnerable to different kinds of noise. In this study after developing multiple models, we select SVM, ANN, and Random forest for ensemble generation since they were not only highly accurate are based on different underlying principles. Traditional methodologies, such as the average and weighted average rules, can be used to build ensembles, but these tactics are overly simplistic and may not be as accurate, therefore, newer techniques, such as stacking are becoming more prominent. Stacking optimally combines the results of multiple classifiers, also known as base classifiers, using a meta classifier. Stacking uses k-fold cross-validation of the training data, the base models are trained on k-1 folds, whereas the *meta*-model is trained on out-of-fold predictions. We use a case of stacking ensemble known as blending ensemble where initial training data is split into base model training and *meta*-model validation data; this leads to less information leakage than a stacking ensemble. The blending ensemble has been found to outperformed stacking, averaging and weighted averaging of machine learning and deep learning models for landslide susceptibility mapping (Fang et al., 2021).

In this study, we create a blending ensemble using SVM, ANN, and Random Forest models as base level models and Logistic Regression as the *meta*-model. The base level models have different underlying principles; therefore, the ensemble model is free from the biases of a single model, making the ensemble more robust. Since the aim of meta model in a blending ensemble is to optimally combine the outputs of the base model, hence we use a simple meta model without hyperparameter tuning of meta model parameters. Using a complex meta model and hyperparameter tuning would require additional computation power as well as additional data split without adding much information to model.

4.6. Model validation

The testing data must be randomly sampled from the overall dataset. Since the overall dataset is highly imbalanced, the testing data will also be imbalanced. The techniques like undersampling and oversampling change the overall data structure, therefore, they cannot be applied for testing. The final dataset had total of 1,282,908 data points with 641,454 of landslide and non-landslide points. The testing data contains randomly sampled 15,000 landslide points and 300,000 non-landslide points.

The metrics used to test the accuracy of our model are explained

below:

a) The Area Under the Receiver Operating Characteristic curve (AUCROC) indicates how well the model predicts 0 values as 0 and 1 value as 1. When presented with imbalanced data sets, AUCROC cannot give an accurate picture of the skill of the classifier; therefore, it is used in conjunction with other methods to evaluate model efficacy.
b) Accuracy is the ratio of correct predictions to total predictions and is an overall metric to check how well the model can differentiate between classes. Accuracy works well only in cases where positive samples are equal to negative ones.
c) Sensitivity is the rate of positivity and is the ratio of True Positives to the total number of positive samples. Sensitivity is the most important metric in the case of landslides classification as it shows the model's capability to identify the existence of landslides (True Positives) since a landslide prone area classified as a low probability is much worse than an area without landslides being classified as landslide susceptible.
d) Precision is the ratio of True Positives and total data points classified as positives. Precision shows the ability of the model to identify relevant data points from total data points.
e) Matthews correlation or MCC is a reliable measure based on the mean square contingency coefficient that produces a high score only if the prediction performs well in all four confusion matrix categories and is used to check the overall model performance, which is crucial for imbalanced datasets.

4.7. Output generation

After completing the training, validation, and testing of individual machine learning models and the ensemble model, the ensemble model to estimate the probability of landslide for each pixel. These probabilities lie between 0 and 1 and are treated as a quantitative estimate of susceptibility. These values are transformed into a single metric of susceptibility classes using Jenks natural breaks classification method. Jenks method has been widely used for landslide susceptibility classification from landslide probability (Piacentini et al., 2012; Sterlacchini et al., 2011). The Jenks method finds optimal class breaks by minimizing the sum of the squared deviations within each class, thereby minimizing inter-class variance and maximizing intra-class variance.

5. Results

5.1. Accuracy metrics

The ensemble model outperforms all individual machine learning models. Fig. 3 (a) shows the AUC ROC curve of individual models and the ensemble. The normalized confusion matrix for the ensemble model is shown in Fig. 3 (b). Due to a high imbalance in the testing data, the metrics will be highly skewed towards non-landslide points if a normal confusion matrix is used. On the other hand, a normalized confusion matrix transforms the values such that the sum of each row is 1, thus giving a more accurate representation of the model performance for imbalanced data.

Sensitivity, precision, and Matthew's correlation coefficient are derived from the confusion matrices to ascertain the skill of the models and are shown in Table 3. The ensemble model performs well in all spheres and has the highest accuracy, sensitivity, precision and MCC than the individual models showcasing the superiority of ensemble based model.

5.2. Relative importance of landslide conditioning factors

Fig. 4 shows the feature importance of the Random Forests model, which indicates the predictive potential of individual landslide conditioning factors. All the input variables contribute to the model development since the redundant variables are removed during model development using feature selection. The slope, TWI, and elevation have the highest feature importance value for the Random Forest model. India's newly developed road database also contributes highly to the model since many landslides in hilly areas are detected along the roadside. The erosivity factor, although a less used landslide conditioning factor, has high feature importance, indicating that it should be incorporated in regional landslide prediction models and developed at high spatial resolution.

5.3. Developing the India landslide susceptibility map (ILSM)

The India Landslide Susceptibility Map is developed by dividing the ensemble landslide probability into 5 susceptibility classes according to the procedure described in section 4.7. The landslide susceptibility map is transformed into five classes, namely Very Low (probability <= 0.1), Low (0.10 < probability<=0.30), Medium (0.30 < probability<=0.54), High (0.54 < probability<=0.78), Very High (probability > 0.78). Fig. 5 shows the India Landslide Susceptibility Map (ILSM).

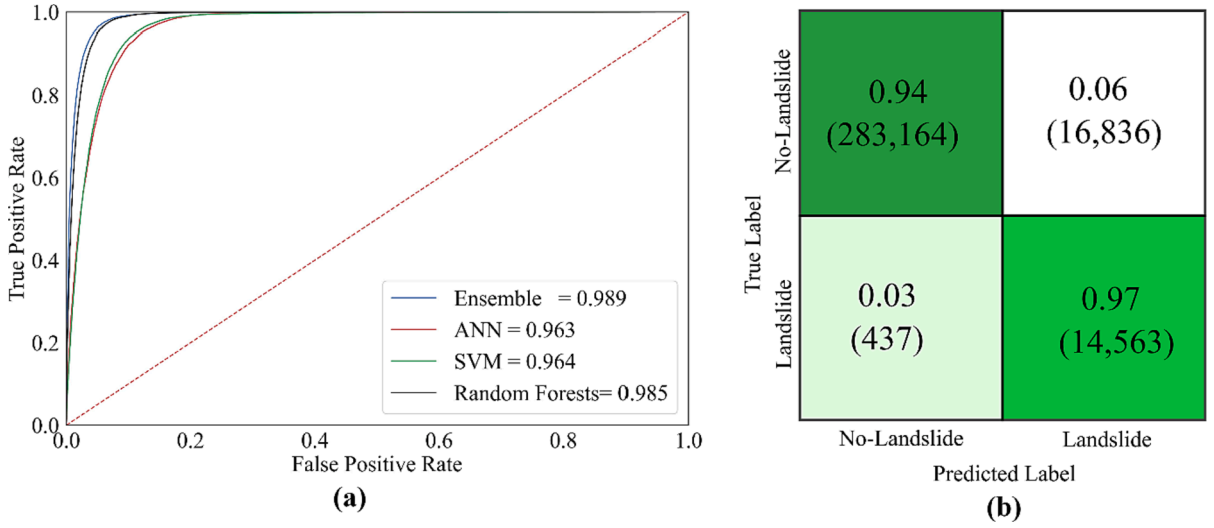


Fig. 3. (a) AUCROC curves for individual machine learning models and blended ensemble model (b) Normalized Confusion Matrix for the ensemble model.

Table 3
Accuracy metrics for various models.

Metric	Expression	Random Forests	SVM	ANN	Ensemble
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	95.24	91.58	90.90	95.73
Sensitivity	$\frac{TP}{TP+FN}$	96.21	92.46	91.62	97.08
Precision	$\frac{TP}{TP+FP}$	94.39	90.86	90.31	94.53
Matthews Correlation	$\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	0.905	0.831	0.818	0.915

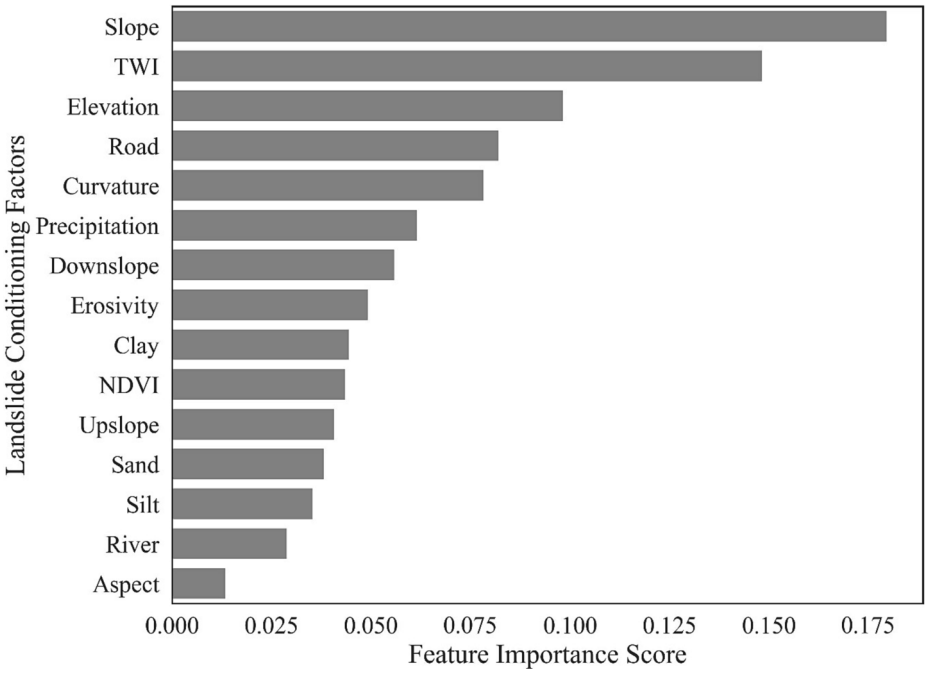


Fig. 4. Feature Importance using Random Forest model.

Table 4 shows the distribution of landslide susceptibility by class. The India Landslide susceptibility map shows that 4.745 % of India is very highly susceptible to landslides. The ILSM shows that the total area susceptible to landslides is 13.17 %, whereas the Geological Survey of India estimates that around 12.6 % area of India is susceptible to landslides. The top 10 landslide susceptible states in India, according to ILSM, are shown in Fig. 6. The Himalayas are the most affected due to landslides often associated with heavy rainfall. The Western Ghats, characterized by steep slopes and thick soil cover is India's second most landslide-affected region. Fig. 6 shows state wise distribution of landslide susceptibility. Sikkim has the highest percentage land area (57.6 %) susceptible to landslides, whereas Arunachal Pradesh, 31845 km2, has the highest area susceptible to landslides. Among the non-Himalayan regions, Kerala has the highest area susceptible to landslides, with 14.32 % in the very high susceptibility zone and 15.73 % in the high susceptibility zone.

5.4. Newly identified landslide zones

The current ILSM and NLSM identify similar areas as highly susceptible to landslides, but ILSM identifies a larger landslide susceptible area, especially in the eastern ghats, as shown in Fig. 7.

The presence of landslides in the eastern ghats is further validated using other global landslide inventories (Froude and Petley, 2018). The identification of new landslide regions by ILSM highlights the superiority of the methodology presented in this study and the inherent advantage of machine learning based frameworks to learn meaningful information about landslide conditioning factors from other areas with similar conditions. Additionally, this map points towards the need for a thorough study with an updated landslide inventory to understand the landslides' behavior in the eastern ghats.

5.5. Comparison with an independent landslide inventory

To assess the quality of our framework and demonstrate the reliability of the ILSM, we compare the final landslide susceptibility map results with an independent landslide inventory, the global fatal landslide database (Froude and Petley, 2018) which contains fatal landslides recorded between 2001–2017. Fig. 8 shows the locations of landslides and their corresponding landslide susceptibility class according to ILSM. Most of the landslides fall in the very high landslide category. Global fatal landslide catalogue also shows some landslide in the eastern ghats region in line with ILSM. Although there are some landslides in very low category they are rare and localized to a few events.

6. Discussion

In this study, we have developed a comprehensive framework for high resolution landslide susceptibility mapping using a combination of multiple machine learning models. We use the framework to develop a pan-India landslide susceptibility map.

This study improves upon the earlier reported studies by:

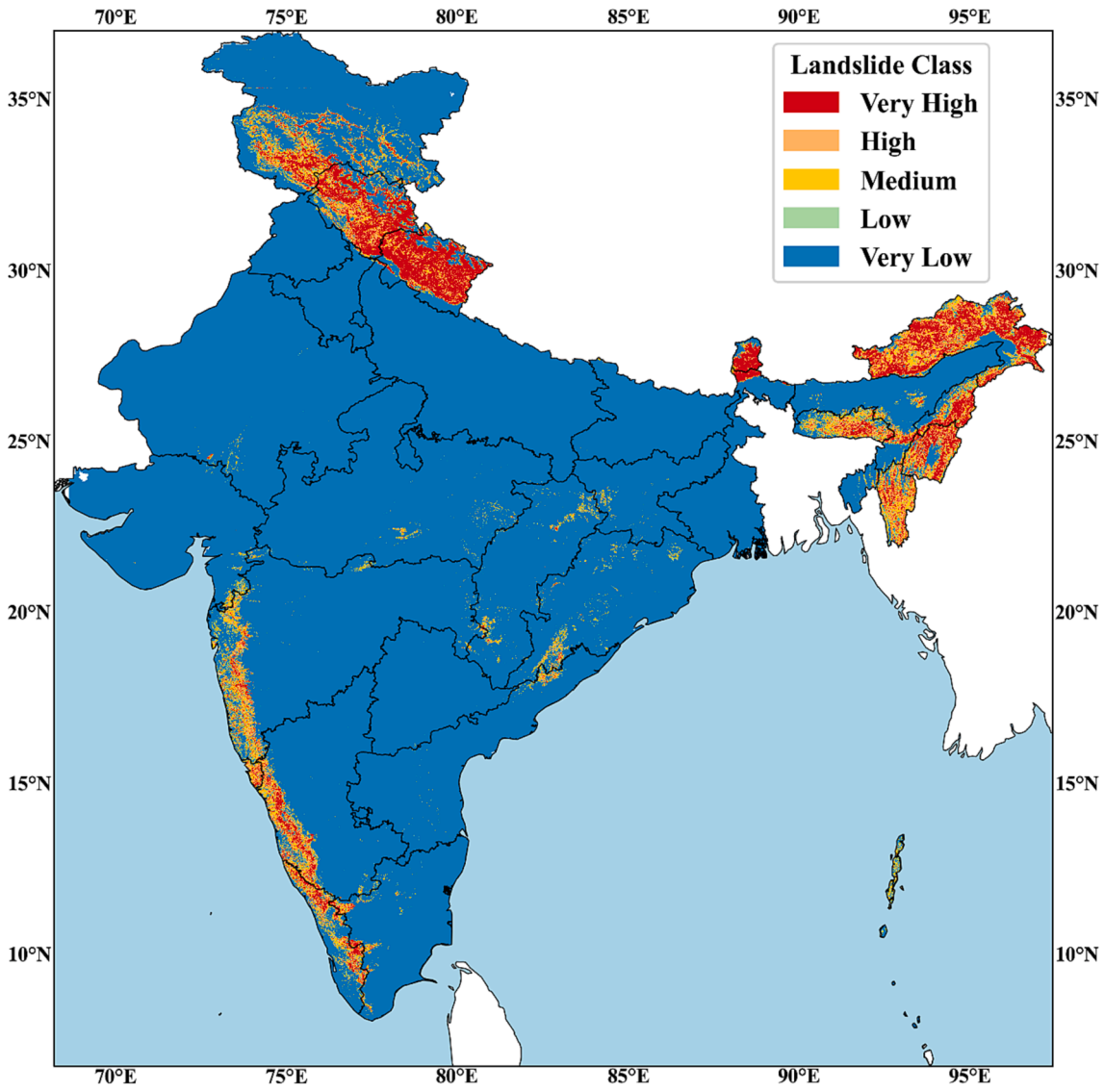


Fig. 5. India Landslide Susceptibility Map (ILSM).

Table 4
Distributions of landslide percentage by class.

Landslide Susceptibility Class	Area in percentage
Very High	4.745
High	3.529
Medium	3.196
Low	1.694
Very Low	86.835

a. Embracing big data: Rather than using coarse resolution data, we use big data to consider the spatiotemporal heterogeneity which enables our machine learning model to generalize and produce accurate landslide susceptibility map. Most of the landslide research is based on limited landslide inventories which makes it challenging to develop data driven models. This study employs 154,329 landslide points in a national landslide inventory meticulously mapped by GSI and 489 landslide points from the global landslide repository. Additionally, we also move away from random sampling techniques towards sophisticated resampling techniques to create a representative and balanced training dataset for the machine learning models.

b. Ensemble methodology: A single machine learning model is based on a single principle and can produce inaccurate results for specific cases. In this study, we use an ensemble of multiple machine learning models with different underlying principles, making our model more robust and freer from individual model bias.

c. Improved input data: Machine learning techniques heavily depend on input data quality (Geiger et al., 2020). In this study, we used a combination of Earth observation data with ground data to generate India Landslide Susceptibility Map (ILSM). We utilize the improved MERIT DEM in conjunction with national datasets for roads and rivers, which have been ground mapped extensively under the directive of the Government of India and significantly improve upon the previous datasets. We also use rainfall kinetic energy (rainfall erosivity) and TWI data, which are often overlooked in landslide susceptibility studies; these datasets are found to significantly contribute to the model.

d. Identifies new landslide zones: This study identifies a much larger area is susceptible to landslides especially in the eastern ghats, which is not considered for developing landslide susceptibility in India's official landslide susceptibility map NLSM.

e. Usability: The explicit landslide probability and class provided in this study can be used for prioritizing landslide research, studying

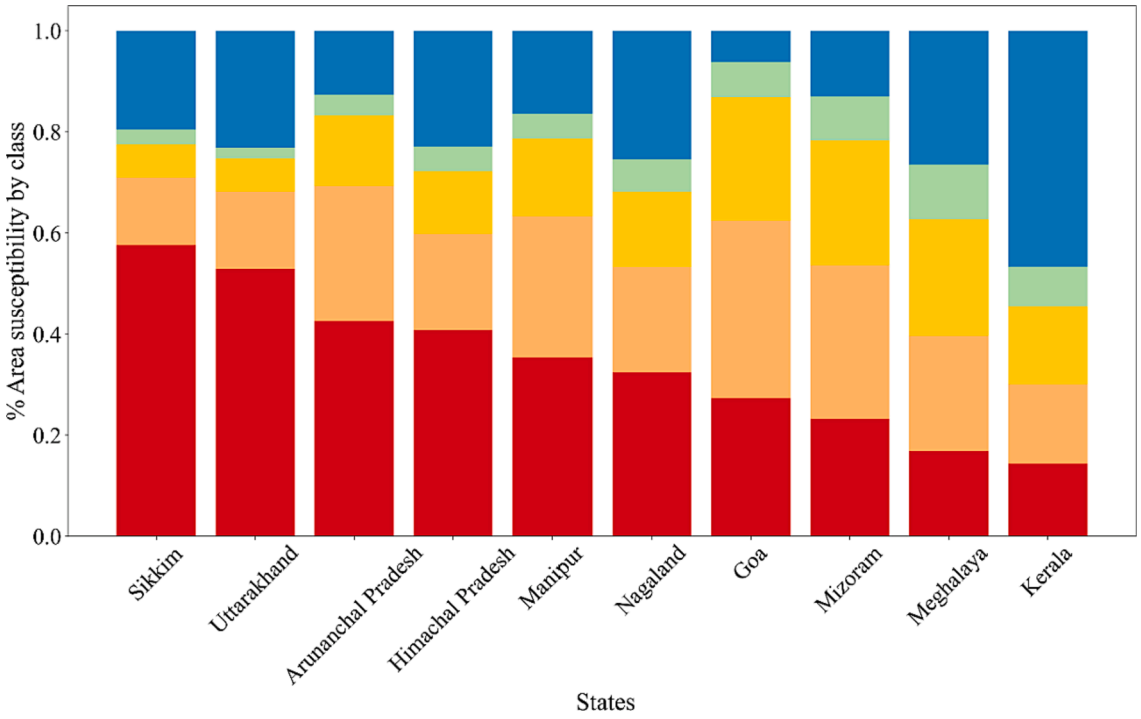


Fig. 6. Top 10 landslide susceptible states in India (by the percentage of area).

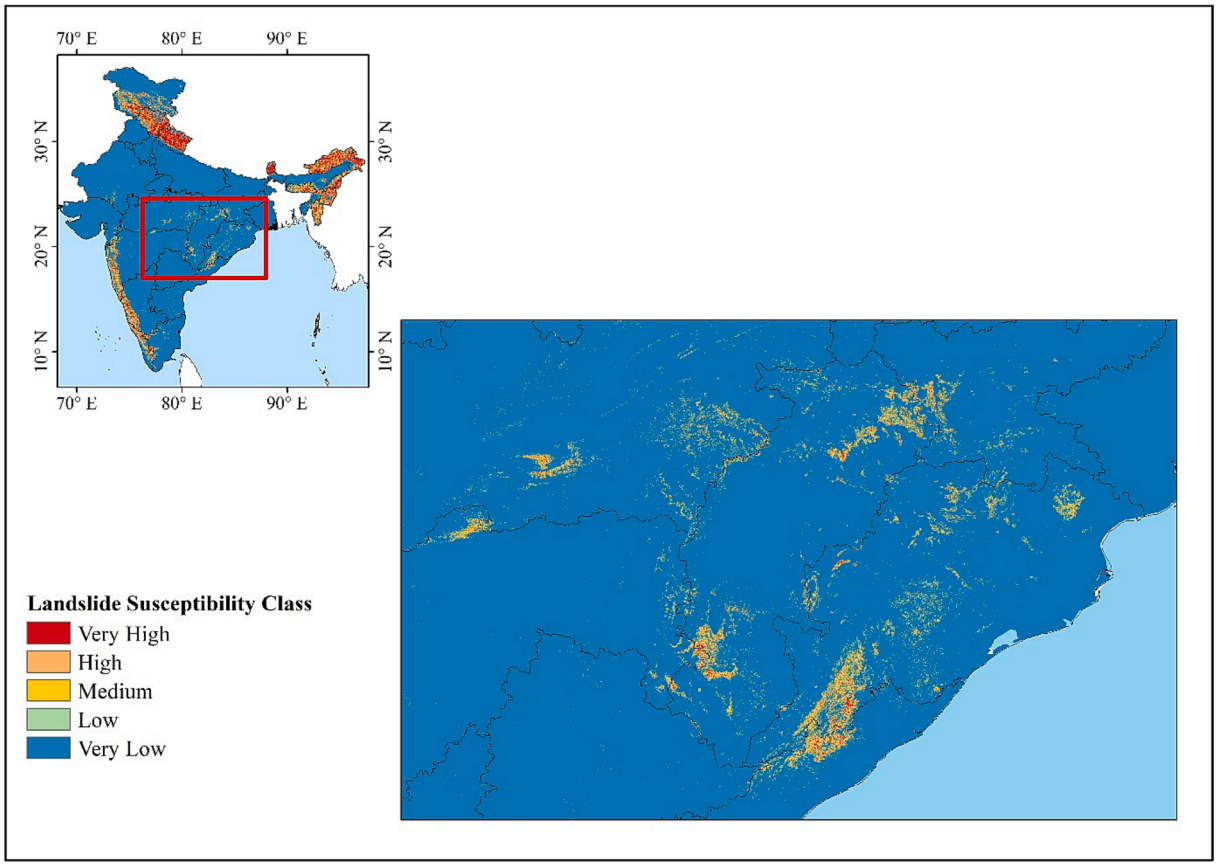


Fig. 7. Newly identified medium landslide susceptibility zones in eastern ghats.

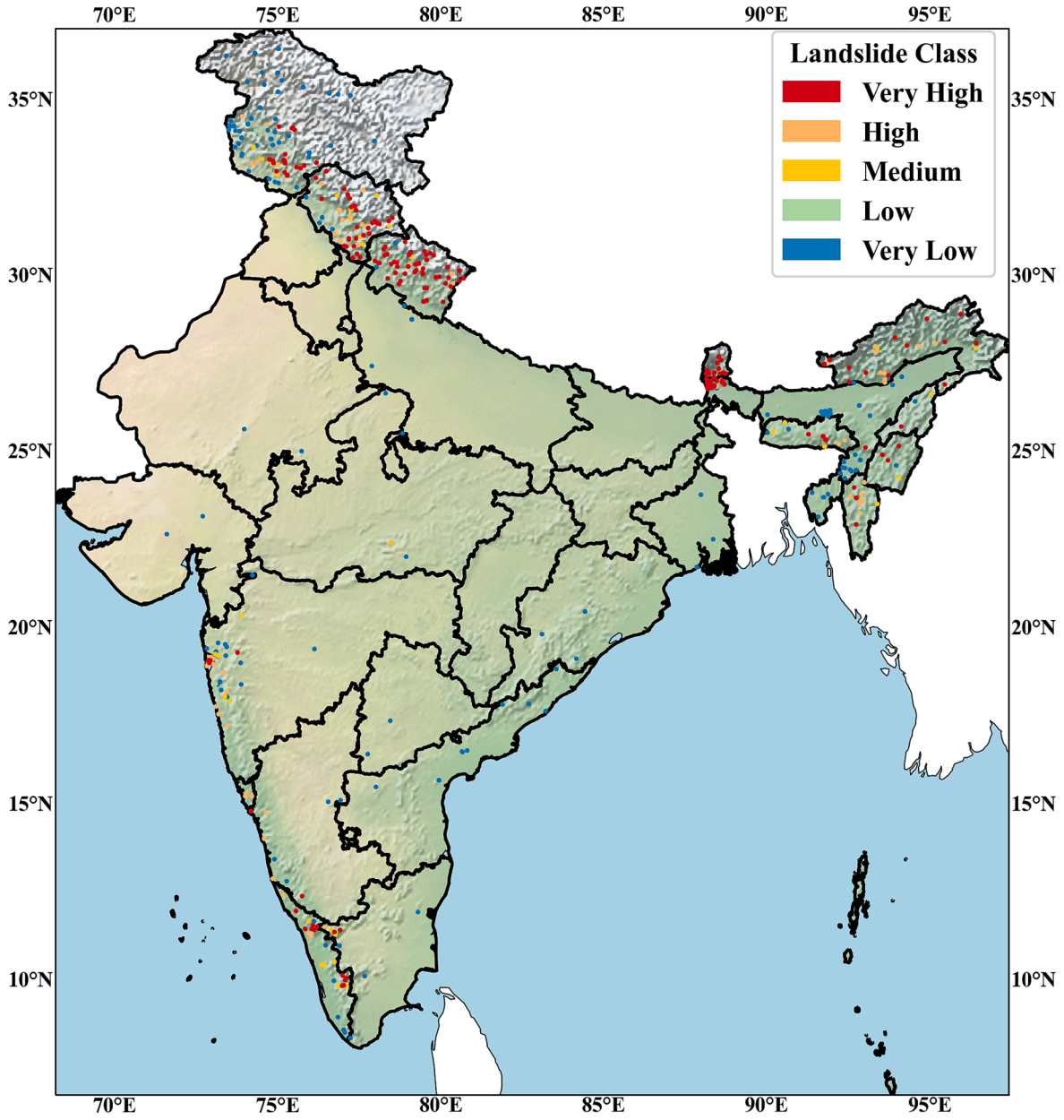


Fig. 8. Location of global fatal landslides database by ILSM class.

compound hazards, and understanding the impact of landslides on the environment. Comprehensive landslide susceptibility maps can help policy makers to design effective mitigation measures against landslides.

7. Conclusion

Every year landslides claim many lives and cause losses amounting to billions of dollars. Understanding the areas prone to landslides and the factors driving them is critical to reducing the impact of landslides. This study aims to develop a framework that leverages big data and uses state of the art data curation and machine learning methods to map landslides at a high spatial resolution. Machine learning models are flexible on datasets and resolution, therefore the framework suggested in this study can be implemented on a variety of datasets, given the datasets are consistent. This methodology is especially useful in developing and underdeveloped countries where ground datasets are absent and landslide models are based on earth observation data. The framework also provides ensemble probabilistic estimates for landslide susceptibility which can help develop landslide models. Using the framework, we develop the India Landslide Susceptibility Map which is not only able to replicate the landslide zones identified in the existing national and global maps but is also able to identify new landslide zones. The ILSM can therefore be used for awareness of present and future landslide hotspots, prioritization of future landslide research, and designing development strategies in the landslide prone regions. However, in this methodology all types of rainfall induced landslides are considered equivalent, but in the real world, the factors driving landslides are more complex. To understand how a machine learning frameworks model different kinds of landslides spatially as well as by type requires an extensive study using local model interpretation methods. Another important factor in machine learning based studies is the quality of input datasets. Most the datasets used in this study are developed using remote sensing which is susceptible to sensor and calibration noise. Therefore, an improvement in input datasets can help improve machine learning based landslide susceptibility maps further. Also, an area where

landslides are more frequent is more susceptible to landslides than an area with a single landslide. These shortcomings will be a subject for future research. The products of this study are freely available in google earth engine.

Data availability

The Indian Landslide Susceptibility Map (ILSM) probability and class data is freely available from

1) Zenodo: https://zenodo.org/doi/10.5281/zenodo.10085271
2) Google Earth engine:
- var ILSM_class = ee.Image("projects/ee-nirdeshsharmanith1/assets/ILSM")
- var ILSM_probability = ee.Image("projects/ee-nirdeshsharmanith1/assets/ILSM_probability")
3) Code: https://github.com/hydrosenselab/ILSM

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data is available publically

Acknowledgments

This research was conducted in the HydroSense Lab (https://hydrosense.iitd.ac.in/) at the Indian Institute of Technology Delhi. Manabendra Saharia acknowledges the financial support for this research through ISRO Space Applications Center (STC0374/RP04139) and DST IC-IMPACTS (RP04558). This work was also supported by the IIT Delhi IoE funded project (PLN12/02CE). The authors gratefully acknowledge the Geological Survey of India (GSI), National Remote Sensing Center (NRSC), and Pradhan Mantri Gram Sadak Yojana (PMGSY) for the datasets used in this study. The authors acknowledge the IIT Delhi High Performance Computing facility for providing computational and storage resources.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.catena.2023.107653.

References

Ahmad, A., Lopulisa, C., Imran, A., Baja, S., 2019. Rainfall erosivity in climate changes and the connection to landslide events. In: IOP Conference Series: Earth and Environmental Science, p. 012007. https://doi.org/10.1088/1755-1315/280/1/012007.

Ahmed, N., Firoze, A., Rahman, R.M., 2020. Machine learning for predicting landslide risk of Rohingya refugee camp infrastructure. J. Inf. Telecommun. 4, 175–198. https://doi.org/10.1080/24751839.2019.1704114.

Alimohammadlou, Y., Najafi, A., Yalcin, A., 2013. Landslide process and impacts: A proposed classification method. CATENA 104, 219–232. https://doi.org/10.1016/j.catena.2012.11.013.

Anbalagan, R., Singh, B., 1996. Landslide hazard and risk assessment mapping of mountainous terrains — a case study from Kumaun Himalaya, India. Eng. Geol. 43, 237–246. https://doi.org/10.1016/S0013-7952(96)00033-6.

Azarafza, M., Azarafza, M., Akgün, H., Atkinson, P.M., Derakhshani, R., 2021. Deep learning-based landslide susceptibility mapping. Sci. Rep. 11, 24112. https://doi.org/10.1038/s41598-021-03585-1.

Bălteanu, D., Chendeş, V., Sima, M., Enciu, P., 2010. A country-wide spatial assessment of landslide susceptibility in Romania. Geomorphol., Recent Adv. Landslide Invest. 124, 102–112. https://doi.org/10.1016/j.geomorph.2010.03.005.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Bureau of Indian Standards, 1998. IS 14496-2: Guidelines for preparation of landslide - Hazard zonation maps in mountainous terrains, Part 2: Macro-zonation.

Castellanos Abella, E.A., Van Westen, C.J., 2008. Qualitative landslide susceptibility assessment by multicriteria analysis: A case study from San Antonio del Sur, Guantánamo, Cuba. Geomorphol., GIS Technol. Models Assessing Landslide Hazard Risk 94, 453–466. https://doi.org/10.1016/j.geomorph.2006.10.038.

Cellek, S., 2021. The Effect of Aspect on Landslide and Its Relationship with Other Parameters, in: Landslides. IntechOpen. https://doi.org/10.5772/intechopen.99389.

Chang, K.-T., Merghadi, A., Yunus, A.P., Pham, B.T., Dou, J., 2019. Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-based machine learning techniques. Sci. Rep. 9, 12296. https://doi.org/10.1038/s41598-019-48773-2.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2011. SMOTE: Synthetic Minority Over-sampling Technique. https://doi.org/10.48550/ARXIV.1106.1813.

Chen, W., Peng, J., Hong, H., Shahabi, H., Pradhan, B., Liu, J., Zhu, A.-X., Pei, X., Duan, Z., 2018. Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. Sci. Total Environ. 626, 1121–1135. https://doi.org/10.1016/j.scitotenv.2018.01.124.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

CRED, 2022. Centre for Research on the Epidemiology of Disasters (CRED).

Dinanta, G.P., Cassidy, D.P., Octariady, J., Fernando, D., Yusuf, M.D., 2020. Assessing landslide susceptibility using ANN and ANFIS to forecast landslides in Sumatera Indonesia, in: 2020 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS). In: Presented at the 2020 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS), pp. 1–11. https://doi.org/10.1109/AGERS51788.2020.9452781.

Emberson, R.A., Kirschbaum, D.B., Stanley, T., 2021. Landslide Hazard and Exposure Modelling in Data-Poor Regions: The Example of the Rohingya Refugee Camps in Bangladesh. Earths Future 9, e2020EF001666. https://doi.org/10.1029/2020EF001666.

Fang, Z., Wang, Y., Peng, L., Hong, H., 2021. A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping. Int. J. Geogr. Inf. Sci. 35, 321–347. https://doi.org/10.1080/13658816.2020.1808897.

Felsberg, A., Poesen, J., Bechtold, M., Vanmaercke, M., De Lannoy, G.J.M., 2021. Estimating global landslide susceptibility and its uncertainty through ensemble modelling. Nat. Hazards Earth Syst. Sci. Discuss. 1–30 https://doi.org/10.5194/nhess-2021-360.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. https://doi.org/10.1002/joc.5086.

Froude, M.J., Petley, D.N., 2018. Global fatal landslide occurrence from 2004 to 2016. Nat. Hazards Earth Syst. Sci. 18, 2161–2181. https://doi.org/10.5194/nhess-18-2161-2018.

Gaidzik, K., Ramírez-Herrera, M.T., 2021. The importance of input data on landslide susceptibility mapping. Sci. Rep. 11, 19334. https://doi.org/10.1038/s41598-021-98830-y.

Gaprindashvili, G., Van Westen, C.J., 2016. Generation of a national landslide hazard and risk map for the country of Georgia. Nat. Hazards 80, 69–101. https://doi.org/10.1007/s11069-015-1958-5.

Geiger, R.S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., Huang, J., 2020. Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? Proc. 2020 Conf. Fairness Account. Transpar. 325–336. https://doi.org/10.1145/3351095.3372862.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2017.06.031.

Günther, A., Van Den Eeckhaut, M., Malet, J.-P., Reichenbach, P., Hervás, J., 2014. Climate-physiographically differentiated Pan-European landslide susceptibility assessment using spatial multi-criteria evaluation and transnational landslide information. Geomorphology 224, 69–85. https://doi.org/10.1016/j.geomorph.2014.07.011.

Guzzetti, F., Paola, R., Ardizzone, F., Cardinali, M., Galli, M., 2006. Estimating the quality of landslide susceptibility models. Geomorphology 81, 166–184. https://doi.org/10.1016/j.geomorph.2006.04.007.

Han, H., Wang, W.-Y., Mao, B.-H., 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, in: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (Eds.), Advances in Intelligent Computing, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 878–887. https://doi.org/10.1007/11538059_91.

He, H., Bai, Y., Garcia, E.A., Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). Presented at the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969.

Highland, L.M., Godt, J.W., Howell, D.G., Savage, W.Z., 1998. El Nino 1997–98; damaging landslides in the San Francisco Bay area. US Dept. of the Interior, US Geological Survey, National Landslide.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978. https://doi.org/10.1002/joc.1276.

Hong, Y., Adler, R., Huffman, G., 2007. Use of satellite remote sensing data in the mapping of global landslide susceptibility. Nat. Hazards 43, 245–256. https://doi.org/10.1007/s11069-006-9104-z.

Hong, H., Tsangaratos, P., Ilia, I., Loupasakis, C., Wang, Y., 2020. Introducing a novel multi-layer perceptron network based on stochastic gradient descent optimized by a

meta-heuristic algorithm for landslide susceptibility mapping. Sci. Total Environ. 742, 140549 https://doi.org/10.1016/j.scitotenv.2020.140549.
Juang, C.S., Stanley, T.A., Kirschbaum, D.B., 2019. Using citizen science to expand the global map of landslides: Introducing the Cooperative Open Online Landslide Repository (COOLR). PLOS ONE 14, e0218657. https://doi.org/10.1371/journal.pone.0218657.
T. Kavzoglu A. Teke Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest 2022 Arab. J. Sci. Eng Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost) https://doi.org/10.1007/s13369-022-06560-8.
Kubat, M., Matwin, S., 1997. In: Addressing the Curse of Imbalanced Training Sets: One-Sided Selection, in. Morgan Kaufmann, pp. 179–186.
Laura, P., de Sousa, L., 2020. SoilGrids250m 2.0 - Clay content. https://doi.org/10.17027/ISRIC-SOILGRIDS.713396F7-1687-11EA-A7C0-A0481CA9E724.
Li, W., Liu, C., Hong, Y., Saharia, M., Sun, W., Yao, D., Chen, W., 2016. Rainstorm-induced shallow landslides process and evaluation – a case study from three hot spots, China. Geomat. Nat. Hazards Risk 7, 1908–1918. https://doi.org/10.1080/19475705.2016.1179685.
Liu, C., Li, W., Wu, H., Lu, P., Sang, K., Sun, W., Chen, W., Hong, Y., Li, R., 2013. Susceptibility evaluation and mapping of China's landslides based on multi-source data. Nat. Hazards 69, 1477–1495. https://doi.org/10.1007/s11069-013-0759-y.
Lok Sabha (http://loksabhaph.nic.in/Questions/QResult15.aspx?qref=22874&lsno=17) [WWW Document], 2021. URL http://loksabhaph.nic.in/Questions/QResult15.aspx?qref=22874&lsno=17 (accessed 5.12.22).
Martha, T.R., Roy, P., Jain, N., Khanna, K., Mrinalni, K., Kumar, K.V., Rao, P.V.N., 2021. Geospatial landslide inventory of India—an insight into occurrence and exposure on a national scale. Landslides 18, 2125–2141. https://doi.org/10.1007/s10346-021-01645-1.
Meena, S.R., Puliero, S., Bhuyan, K., Floris, M., Catani, F., 2022. Assessing the importance of conditioning factor selection in landslide susceptibility for the province of Belluno (region of Veneto, northeastern Italy). Nat. Hazards Earth Syst. Sci. 22, 1395–1417. https://doi.org/10.5194/nhess-22-1395-2022.
Nadim, F., Kjekstad, O., Peduzzi, P., Herold, C., Jaedicke, C., 2006. Global landslide and avalanche hotspots. Landslides 3, 159–173. https://doi.org/10.1007/s10346-006-0036-1.
Nguyen, H.M., Cooper, E.W., Kamei, K., 2011. Borderline over-sampling for imbalanced data classification. Int. J. Knowl. Eng. Soft Data Paradig. 3, 4. https://doi.org/10.1504/IJKESDP.2011.039875.
Okalp, K., Akgün, H., 2016. National level landslide susceptibility assessment of Turkey utilizing public domain dataset. Environ. Earth Sci. 75, 847. https://doi.org/10.1007/s12665-016-5640-3.
Peña, M.A., Brenning, A., 2015. Assessing fruit-tree crop classification from Landsat-8 time series for the Maipo Valley, Chile. Remote Sens. Environ. 171, 234–244. https://doi.org/10.1016/j.rse.2015.10.029.
Pham, B.T., Pradhan, B., Tien Bui, D., Prakash, I., Dholakia, M.B., 2016a. A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). Environ. Model. Softw. 84, 240–250. https://doi.org/10.1016/j.envsoft.2016.07.005.
Pham, B.T., Tien Bui, D., Prakash, I., Dholakia, M.B., 2017. Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. CATENA 149, 52–63. https://doi.org/10.1016/j.catena.2016.09.007.
Piacentini, D., Troiani, F., Soldati, M., Notarnicola, C., Savelli, D., Schneiderbauer, S., Strada, C., 2012. Statistical analysis for assessing shallow-landslide susceptibility in South Tyrol (south-eastern Alps, Italy). Geomorphology 151–152, 196–206. https://doi.org/10.1016/j.geomorph.2012.02.003.
Poggio, L., de Sousa, L., 2020a. SoilGrids250m 2.0 - Sand content. https://doi.org/10.17027/ISRIC-SOILGRIDS.713396FA-1687-11EA-A7C0-A0481CA9E724.
Poggio, L., de Sousa, L., 2020b. SoilGrids250m 2.0 - Silt content. https://doi.org/10.17027/ISRIC-SOILGRIDS.713396FB-1687-11EA-A7C0-A0481CA9E724.
Pradhan, B., 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput. Geosci. 51, 350–365. https://doi.org/10.1016/j.cageo.2012.08.023.
Raj, R., Saharia, M., Chakma, S., Rafieinasab, A., 2022. Mapping rainfall erosivity over India using multiple precipitation datasets. CATENA 214, 106256. https://doi.org/10.1016/j.catena.2022.106256.
Ram, P., Gupta, V., 2022. Landslide hazard, vulnerability, and risk assessment (HVRA), Mussoorie township, lesser himalaya, India. Environ. Dev. Sustain. 24, 473–501. https://doi.org/10.1007/s10668-021-01449-2.
Ramachandra, T.V., Aithal, B.H., Kumar, U., Joshi, N.V., 2013. Prediction of shallow landslide prone regions in undulating terrains. Disaster Adv 6 (1), 54–64.
Ramli, M.F., Yusof, N., Yusoff, M.K., Juahir, H., Shafri, H.Z.M., 2010. Lineament mapping and its application in landslide hazard assessment: a review. Bull. Eng. Geol. Environ. 69, 215–233. https://doi.org/10.1007/s10064-009-0255-5.
Reichenbach, P., Rossi, M., Malamud, B.D., Mihir, M., Guzzetti, F., 2018. A review of statistically-based landslide susceptibility models. Earth-Sci. Rev. 180, 60–91. https://doi.org/10.1016/j.earscirev.2018.03.001.
Sahin, E.K., 2020. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Appl. Sci. 2, 1308. https://doi.org/10.1007/s42452-020-3060-1.
Sim, K.B., Lee, M.L., Wong, S.Y., 2022. A review of landslide acceptable risk and tolerable risk. Geoenvironmental Disasters 9, 3. https://doi.org/10.1186/s40677-022-00205-6.
Singh, A., Ranjan, R.K., Tewari, V.C., 2020. Spatio-temporal Variability of Landslides in Sikkim Himalaya, India, in: Pal, I., von Meding, J., Shrestha, S., Ahmed, I., Gajendran, T. (Eds.), An Interdisciplinary Approach for Disaster Resilience and Sustainability, Disaster Risk Reduction. Springer, Singapore, pp. 219–234. https://doi.org/10.1007/978-981-32-9527-8_13.
Stanley, T., Kirschbaum, D.B., 2017. A heuristic approach to global landslide susceptibility mapping. Nat. Hazards 87, 145–164. https://doi.org/10.1007/s11069-017-2757-y.
Stein, C., 1956. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. Proc. Third Berkeley Symp. Math. Stat. Probab. Vol. 1 Contrib. Theory Stat. 3.1, 197–207.
Sterlacchini, S., Ballabio, C., Blahut, J., Masetti, M., Sorichetta, A., 2011. Spatial agreement of predicted patterns in landslide susceptibility maps. Geomorphology 125, 51–61. https://doi.org/10.1016/j.geomorph.2010.09.004.
Thi Ngo, P.T., Panahi, M., Khosravi, K., Ghorbanzadeh, O., Kariminejad, N., Cerda, A., Lee, S., 2021. Evaluation of deep learning algorithms for national scale landslide susceptibility mapping of Iran. Geosci. Front. 12, 505–519. https://doi.org/10.1016/j.gsf.2020.06.013.
Valdiya, K.S., 2015. The Making of India: Geodynamic Evolution. Springer.
Wieczorek, G.F., 1996. LANDSLIDES: INVESTIGATION AND MITIGATION. CHAPTER 4 - LANDSLIDE TRIGGERING MECHANISMS. Transp. Res. Board Spec. Rep.
Yalcin, A., Reis, S., Aydinoglu, A.C., Yomralioglu, T., 2011. A GIS-based comparative study of frequency ratio, analytical hierarchy process, bivariate statistics and logistics regression methods for landslide susceptibility mapping in Trabzon, NE Turkey. CATENA 85, 274–287. https://doi.org/10.1016/j.catena.2011.01.014.
Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J.C., Sampson, C.C., Kanae, S., Bates, P.D., 2017. A high-accuracy map of global terrain elevations. Geophys. Res. Lett. 44, 5844–5853. https://doi.org/10.1002/2017GL072874.
Youssef, A.M., Pourghasemi, H.R., 2021. Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. Geosci. Front. 12, 639–655. https://doi.org/10.1016/j.gsf.2020.05.010.