Landslide Susceptibility Mapping Using Machine Learning in the Shillong Plateau, Meghalaya

Authors:
Harshita Maurya A011
Priyangi Jain A013
Ananya Sharma A016
Meyhar Sharma A018
Muskan Nagdeo A032

Mentors:
Dr. Rushina Singhi
Dr. Sunayana Sarkar
Prof. Prashant Dhamale

Abstract

Landslides cause immense damage to life and property. In an area like the Shillong Plateau, where landslides are an extremely common occurrence, preparedness, quick response and damage control become very important. Accurate landslide susceptibility mapping is thus crucial. This paper seeks to predict the occurrence or non-occurrence of landslides using a classification algorithm. Three machine learning models (Random Forest, Support Vector Machines, Convolutional Neural Networks) were applied to classify points according to whether a landslide will occur there or not. The end goal was to classify a set of 12 points in the study area (currently being mapped under a project for the Department of Science and Technology). RF and SVM provided predictions for coordinates, while CNN segmented landslide locations based on images. All three models showed a relatively good accuracy, with RF showing the best performance.

Introduction

Landslides are a very common phenomenon in North-East India. They are one of the most devastating natural catastrophes, causing damage to human life as well as infrastructure. Meghalaya, is known for its rugged terrain and high rainfall, which are two of the main factors contributing to the large number of landslides in the region. The state is situated in a geologically complex area, where the underlying rock formations are composed of sedimentary, metamorphic, and igneous rocks. The terrain is characterized by steep slopes, deep valleys, and narrow ridges, which are prone to erosion and weathering. Mawsynram and Cherrapunji, which lie on the Shillong Plateau (and are located very close to the study area) receive the highest average annual rainfall (11,872 mm and 11,430 mm respectively) in the world. This further contributes to the instability of the soil and rock slopes. The rainfall can cause saturation of the soil, leading to soil liquefaction and the movement of the soil and rocks down the slopes. Meghalaya is also located in a seismically active region, with high tectonic activity. All these factors lead to Meghalaya suffering from regular landslides.
For prevention measures to avoid the devastating effect of landslides, landslide susceptibility must first be mapped for the concerned areas to identify areas that are more prone to landslides, based on previous knowledge about the spatial distribution of past occurrences (Ballabio & Sterlacchini, 2012). Also, to ensure development in the North-East, the land and its features need to be studied and mapped. The susceptibility of the area needs to be documented to be able to put prevention measures in place and minimize damage. This process helps in formulating disaster management policies and efficient land use planning.

Keeping this in mind, The Department of Science and Technology is currently conducting ground analysis in the region. The focus of the project is to map different areas throughout the state. One such area is near the Umiam Lake in the Shillong Plateau. The landslide susceptibility of 12 locations here is to be studied.
The aim this research is to provide a statistical model that is able to classify data accurately as 'landslide' and 'non-landslide' and which can thus be used to map landslide susceptibility in the Shillong Plateau. This is done using machine learning in order to add to their process and make mapping easier in the future. Thus, the 12 points that the DST are studying within the area are considered as the prediction set for the study.
Three machine learning models were applied with the objective of classification. Random Forest and Support Vector Machines were used to classify landslides using conditioning factors. Convolutional Neural Networks was used for image segmentation to map locations prone to landslides.

Literature Review

Sr.No.	Title of Research Paper, Year	Name of Authors	Summary
1.	Deep learning-based landslide susceptibility mapping, 2021	Mohammad Azarafza, Mehdi Azarafza, Haluk Akgün, Peter M. Atkinson, Reza Derakhshani	This research is conducted for the Isfahan province in Iran and aims to assess the suitability of a combined CNN-DNN model to find out the main factors that can trigger a landslide. Many algorithms such as random forest, linear regression, SVM etc. were compared to conclude that the combined CNN-DNN model outperforms all and hence is the best for landslide susceptibility mapping.
2.	Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and random forest, 2021	Deliang Sun, Jiahui Xu, Haijia Wen, Danzhou Wang	The research develops and compares two models, logistic regression and random forest for landslide prone areas in China. Bayesian algorithm is used for hyperparameter optimization to conclude that the random forest model outperforms the logistic regression model.
3.	HR-GLDD: A globally distributed dataset using generalized DL for rapid landslide mapping on HR satellite imagery, 2022	Sansar Raj Meena, Lorenzo Nava, Kushanav Bhuyan, Silvia Puliero, Lucas Pedrosa Soares, Helen Cristina Dias, Mario Floris, Filippo Catani	This study takes a dataset for landslide mapping composed of ten different physiographic regions globally on which five deep learning models were tested to know the transferability and robustness of the HR-GLDD.
4.	Integration of convolutional neural network and conventional machine learning classifiers for	Zhice Fang, Yi Wang, Ling Peng, Haoyuan Hong	The aim of this study is to assess landslide susceptibility by integrating a convolutional neural network with three conventional machine learning classifiers of support vector machine, random forest and logistic

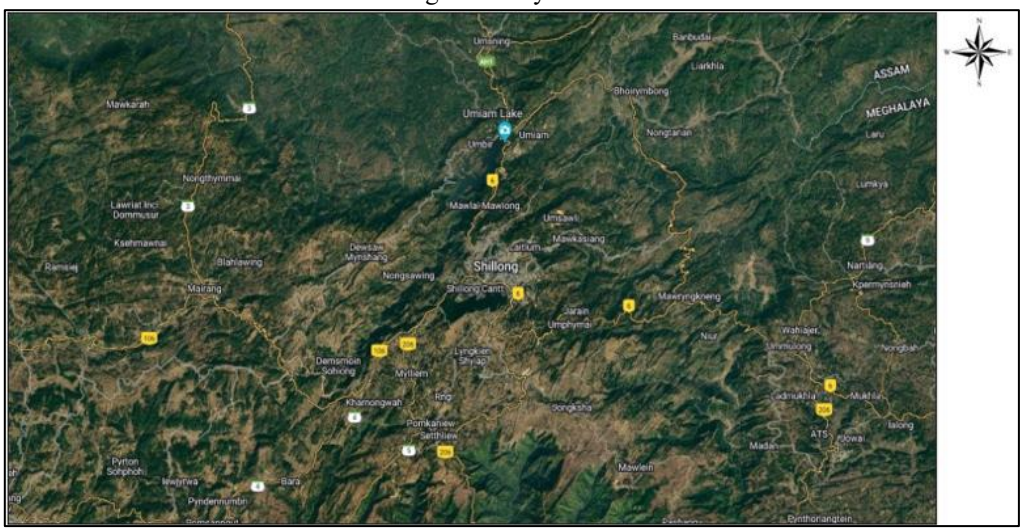
	landslide susceptibility mapping, 2020		regression. The experimental results demonstrated that the performance of these machine learning classifiers effectively improved by integrating CNN and concluded the hybrid models should be recommended for landslide spatial modelling
5.	A Support Vector Machine for Landslide Susceptibility Mapping in Gangwon Province, Korea, 2017	Saro Lee, Soo-Min Hong, Hyung-Sup Jung	This research applied and validated support vector machine by using the geographic information system in order to map landslide susceptibility. Moreover, sensitivity assessment of the factors was performed to conclude SVMs are useful for landslide susceptibility analysis.
6.	Landslide susceptibility mapping using support vector machine and GIS at the Golestan Province, Iran, 2013	Hamid Reza Pourghasemi, Abbas Goli Jirandeh, Biswajeet Pradhan, Chong Xu, Candan Gokceoglu	The main goal of this study was to produce landslide susceptibility maps using GIS-based support vector machine in Kalaleh Township, Iran. A comparison of six different kernel types was done and Based on the results, the differences in the rates (success and prediction) of the six models was not significant hence, the produced susceptibility maps by any of the six kernels will be useful for general land-use planning.
7.	Comparison of four kernel functions used in support vector machines for landslide susceptibility mapping: a case study at Suichuan area (China), 2016	Haoyuan Hong, Biswajeet Pradhan, Dieu Tien Bui, Chong Xu, Ahmed M. Youssef Wei Chen	This research compared four kernel functions used in support vector machines for landslide susceptibility in Suichuan region in China. Based on AUC values of each model, the radial basis kernel function performed the best.
8.	Support Vector Machines for Landslide Susceptibility Mapping: The Staffora River Basin Case Study, Italy, 2012	Cristiano Ballabio, Simone Sterlacchini	This study applied support vector machines to landslide susceptibility mapping in the Staffora river basin (Lombardy, Northern Italy). The model used cross validation of results and then compared performance with logistic regression, linear discriminant analysis, and naive Bayes classifier to conclude SVM outperform other techniques in terms of accuracy and generalization capacity.
9.	Assessment of landslide susceptibility for Meghalaya (India) using bivariate (frequency ratio and Shannon entropy) and multi-criteria decision analysis (AHP and fuzzy-AHP) models, 2022	Navdeep Agrawal, Jagabandhu Dixit	The focus of this paper is on assessing the landslide susceptibility of Meghalaya, India using different models. Models considered were bivariate models based on frequency ratio and Shannon entropy, and multi-criteria decision analysis models based on the analytical hierarchy process and fuzzy-AHP. The AHP showed the highest prediction

			accuracy for Meghalaya based on AUC and F1 Score.
10.	Spatio-statistical comparative approaches for landslide susceptibility modeling: case of Mae Phun, Uttaradit Province, Thailand, 2020	Muhammad Farhan Ul Moazzam, Anujit Vansarochana, Jaruntorn Boonyanuphap, Sittichai Choosumrong, Ghani Rahman, Geraud Poueme Djueyep	The aim of this research was to compare three quantitative techniques - frequency ratio, information value, and weight of evidence - for landslide susceptibility modeling. The results of the FR model indicated that almost 40% of the total study area falls in high to very high landslide susceptibility zones, while in WOE and IFV models, found almost 50% of the total area. The evaluation of landslide density test and seed cell index area indicated that calculated and classified landslide susceptibility maps are in good agreement with the field conditions. From this study, it was found that slope angle, elevation, land use/land cover, and roads play a major influencing role in the occurrence of landslides in the study area.
11.	Mapping landslide susceptibility and types using Random Forest, 2018	Khaled Taalab, Tao Cheng, Yang Zhang	This research discussed the importance of landslide susceptibility maps in managing the environment, urban planning, and minimizing economic losses. This paper presents a data mining approach using a Random Forest algorithm to produce LSMs for a large, diverse region that is susceptible to multiple types of landslides. The method is demonstrated using a case study of Piedmont, Italy, and results in a highly accurate LSM that is easy to interpret without the need for multiple susceptibility assessments.

Study Area

The study area is a small part of the Shillong Plateau in Meghalaya. It lies within the following range of coordinates: 25o 43' 53"N - 25o 27' 49" N latitude and 91o 33' 59" E - 92o 08' 09"E longitude. Image 1 depicts the study area.

Image 1: Study Area



Source: QGIS

Data Collection – RF and SVM

A total of 219 data points, comprising 99 landslide points and 120 non-landslide points were considered. The landslide points taken within the study area have been obtained from the landslide inventory of Bhukosh Geological Survey of India (Bhukosh GSI). The non-landslide points were mapped using Google Earth. The prediction set of 12 data points has been collected by a group of experts in the field of geology through an on-field investigation in the area near Umiam Lake in Shillong. Data for these points was collected according to the various geological analyses, under an ongoing project by Department of Science and Technology and the aim is to evaluate their susceptibility to landslides.

Various landslide conditioning factors were considered for classifying whether landslides will occur or not. There were 10 variables taken into consideration, of which 3 variables are topographic (elevation, slope, aspect), 3 variables relate to geological conditions (lithology, distance from faults, Watershed), 3 variables relate to environmental conditions (distance from rivers, land cover, VARI), and 1 variable pertains to human activities (distance from Shillong city). These factors exert significant influence on the land and therefore, on its susceptibility to a landslide. One of the most significant factors contributing to a landslide is rainfall. Most landslides in the area are triggered by heavy rainfall. However, since the study area considered is small, there is not much variation in the rainfall distribution among the points considered. Therefore, for this study, average rainfall has not been used as a conditioning factor.

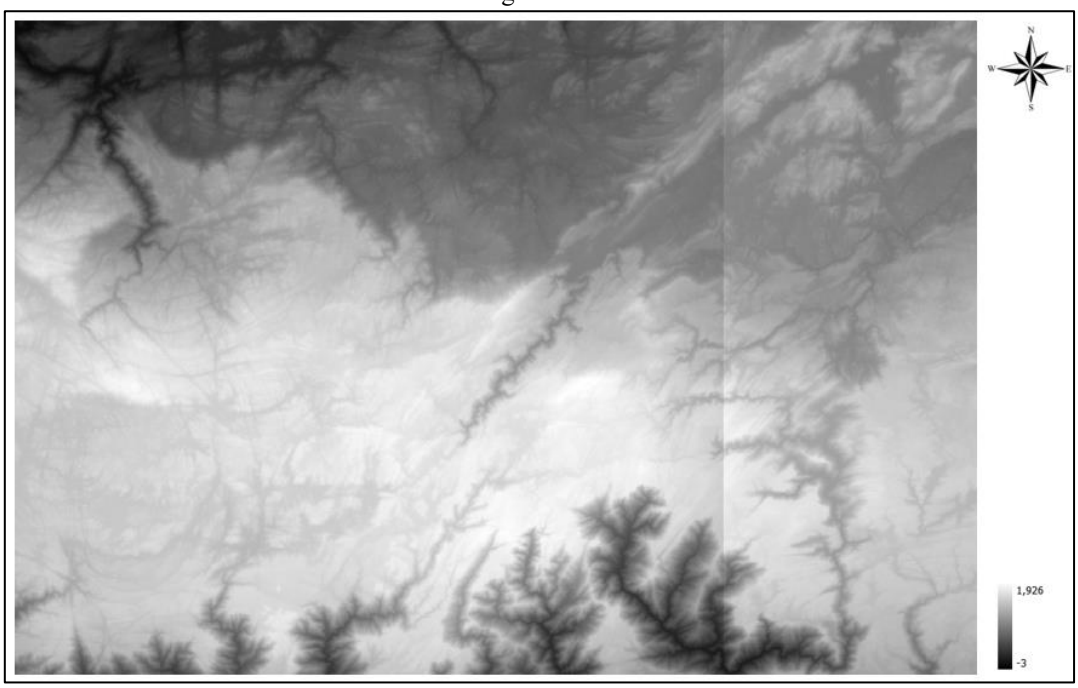
The softwares used to extract values for the conditioning factors were ArcGIS (Arc Geographic Information System), QGIS (Quantum Geographic Information System), Google Earth. The websites that were used to help obtain the raster images for extracting data for the variables were Bhukosh GSI, USGS (United States Geological Survey) and DIVA GIS (DIVA Geographic Information System).

DEM

A Digital Elevation Model (DEM) is a digital cartographic dataset that represents a continuous topographic elevation surface through a series of cells. Each cell represents the elevation (Z) of a feature at its location (X and Y). Digital Elevation Models are a “bare earth” representation because they only contain information about the elevation of geological (ground) features, such as valleys, mountains, and landslides, to name a few. They do not include any elevation data concerning non-ground features, such as vegetation or buildings. Digital Elevation Models can be used to create topographic maps of overland terrain. A digital cartography dataset called a "Digital Elevation Model" (DEM) depicts a continuous topographic elevation surface through a grid of cells. Each cell depicts a feature's height (Z) at its X and Y coordinates. They only provide information on the elevation of geological (ground) features like valleys, mountains, and landslides, to name a few. They don't have any information on the elevation of any above-ground items, including vegetation or structures. Overland topographic maps can be made using digital elevation models. Variables such as slope and aspect of the features can then be derived from these maps.

The DEM for the study area was extracted using the USGS website. Image 2 shows the DEM for the study area. A darker gradient represents lower elevation while a lighter gradient represents higher elevation.

Image 2: DEM



Source: QGIS

Elevation

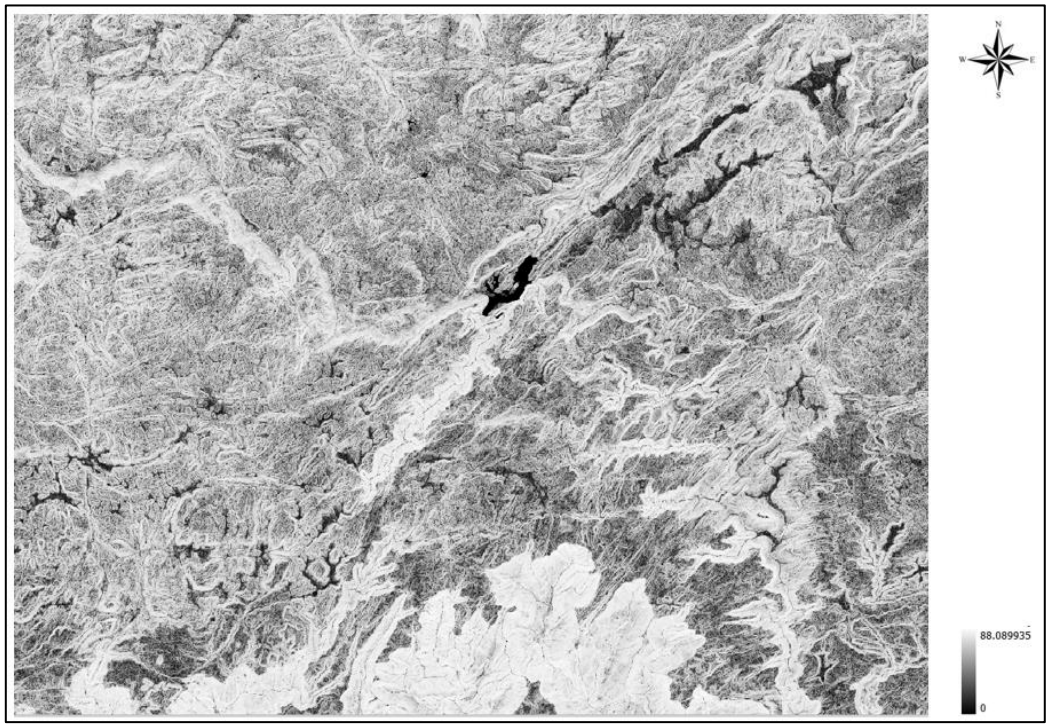
Elevation of a feature is its height above sea level. It is known that with an increase in elevation, the temperature decreases, while rainfall increases with increase in elevation. Thus, there is a higher probability of landslides occurring on high slopes (Moazzam et al., 2020). Elevation values in meters for each data point were extracted from the DEM using QGIS.

Slope

Slope is the percent change in the elevation of a topography over a certain distance.

It is the measurement of the steepness of a surface and is measured in degrees. It has a range of 0-90˚, where 0 represents a flat surface and 90 represents a steep surface. (Yılmaz et al., 2012). Based on gravity, soil water content, soil structure, erosion potential, and hydrological and geomorphological processes, the studies reviewed in the literature demonstrate that slope considerably, either directly or indirectly, affects the velocity of slope surface and subsurface water flow. (Anbalagan, 1992; Wilson and Gallant, 2000). This in turn affects the area's susceptibility to landslide. Slope for the chosen study area's landslide and non-landslide points has been extracted from the DEM of the study area. A slope map was derived from the DEM using QGIS and slope values for each point were extracted from this map. The slope map for the study area is shown in Image 3.

Image 3: Slope Map

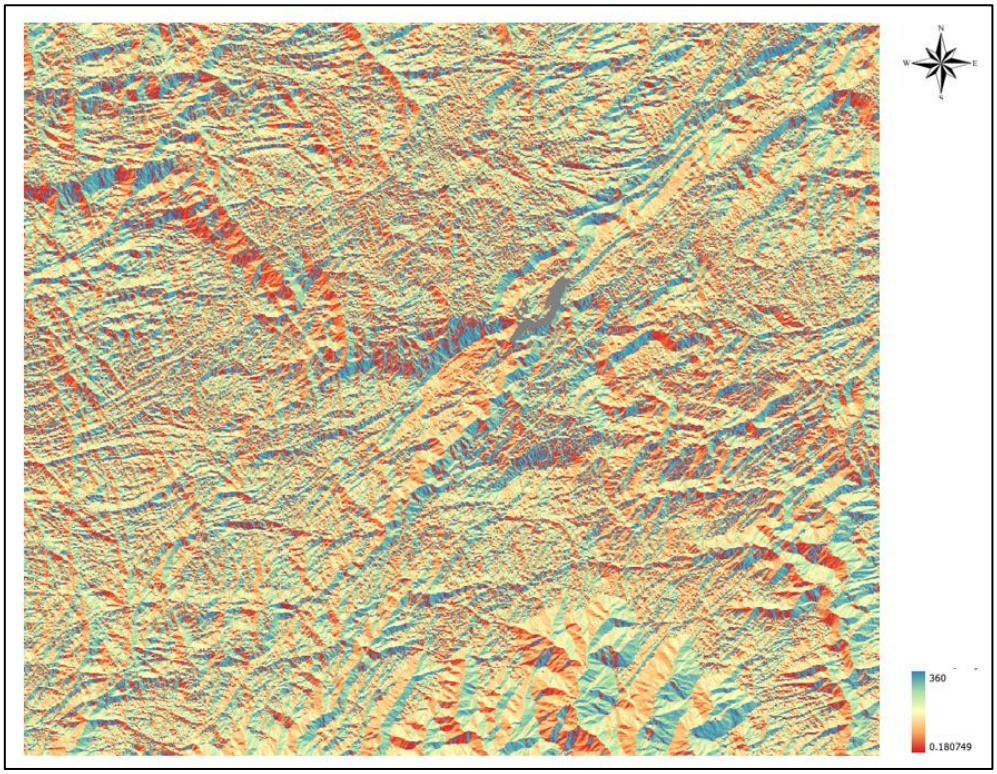


Source: USGS, ArcGIS

Aspect

The maximum direction of the slope of a land surface is represented by aspect. It ranges from 0 to 360, and is measured in degrees from the north direction. Aspect map shows both the direction and grade of a terrain at the same time. (Cellek, 2021). It alters how much sunlight, wind, and precipitation is exposed to, which indirectly alters other elements that cause landslides such soil moisture, vegetation cover, and soil thickness. (Clerici et al. 2006). Image 4 shows the aspect map for the study area. It must be noted that to analyse directional data, they must first be transformed into rectangular polar coordinates. First, a 'unit circle' is specified, i.e., a circle with a radius of 1. The polar location is then defined as the angular measurement and its intersection with the unit circle. The cosine and sine functions are then used to place this location into a standardized Cartesian space. The transformed values are used in the models as input.

Image 4: Aspect Map



Source: QGIS

Lithology

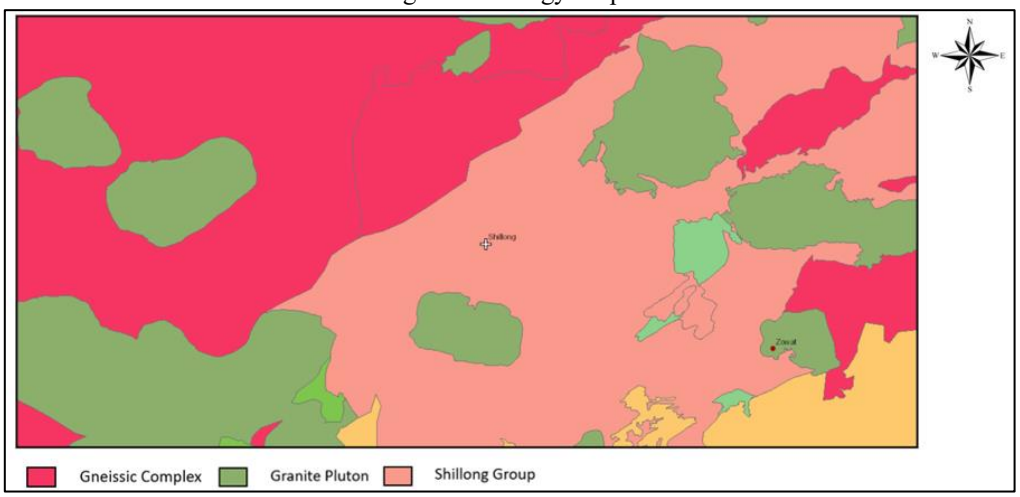
The study of the intrinsic physical properties of rocks is known as lithology. The lithology of the study area is broadly classified into three groups mentioned in table 1. The lithology map of the study area is shown in Image 5.

Table 1

Gneissic Complex	Precambrian rocks of Gneissic composition are the oldest rocks found here and considered to be basement complex. This basement complex consists of biotite-gneiss, biotite-hornblende gneiss, granitic gneiss, mica-schist, biotite granulite-amphibolite, pyroxene granulite, gabbro and diorite. The rocks in this class belong to the age of Archean to Proterozoic era.
Granite Plutons	The formation of this class consists of Kyrdem granite, Nongpoh granite, Mylliem granite, South Khasi granite. These rocks consist of Porphyritic coarse grain granite, pegmatite, apatite/quartz vein traversed by epidiorite, dolerite and basalt dykes. The rocks in this class belong to the Neo-Proterozoic to Early Paleozoic age group.
Shillong Group	This is the class of rocks that are majorly found in the study area. This group is divided into two broad classes, the Upper Shillong Group (also known as the Upper Quartzite Formation) and the Lower Shillong Formation (also known as the Tyrsad/Barapani Formation). The rocks in this area date back to Mid-Proterozoic age, which are mainly composed of Quartzites intercalated with Phyllite and conglomerate, and Early Proterozoic age, which are mainly composed of schists with Calc- Silicate rocks, carbonaceous phyllite and thin quartzite layers, schist, slate, conglomerate.

Source: Singh & Singh, 2018

Image 5: Lithology Map



Source: bhukosh.gsi

Land Cover

Land cover maps show the different classes of physical coverage of an area's surface like tree cover, shrub cover, grasslands, water bodies and so on. Certain changes in land cover due to time or human activity, such as deforestation or construction of roads and buildings have a significant impact on that area's propensity to a landslide occurring. (Reichenbach et al., 2014).

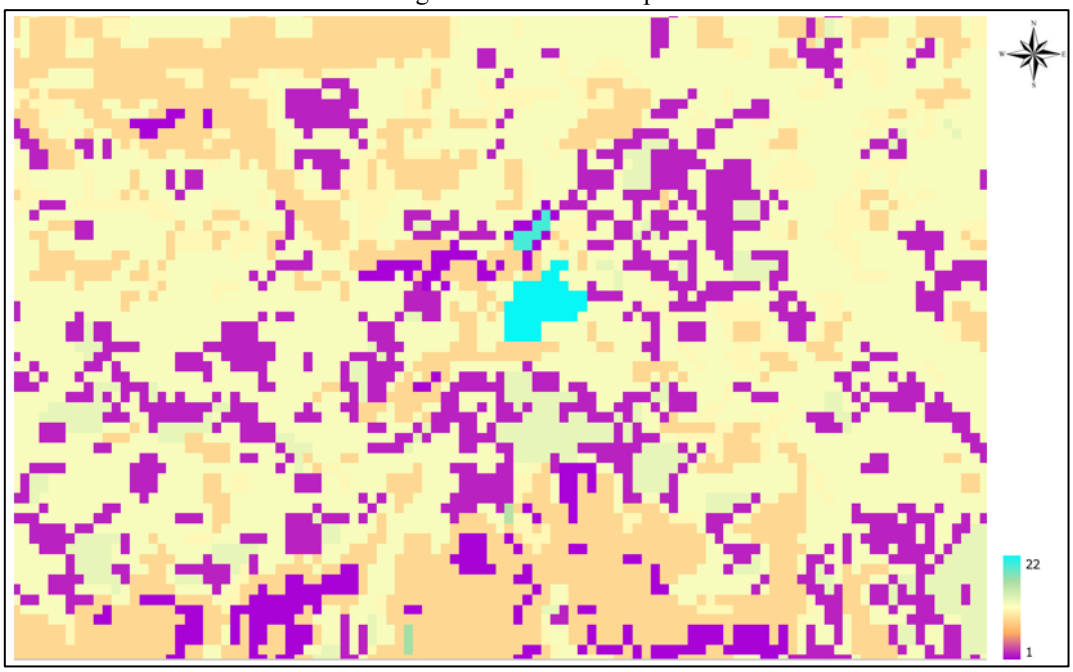
The land cover for our study area was extracted from the DIVA GIS website, which was a raster layer input in QGIS. The specific land cover class values were extracted for each of the points selected in the study area. According to the global land cover classification system (LCCS), the land cover classes range from 1-22, but only 1,2,9,11,12,13,20 and 22 are present in the study area. Table 2 shows the interpretation of the classes present in the study area. Image 6 shows the land cover map of the study area.

Table 2

GLC Global Class (according to LCCS terminology)	
1	Tree Cover, broadleaved, evergreen LCCS >15% tree cover, tree height >3m (Examples of sub-classes at regional level* : closed > 40% tree cove; open 15-40% tree cover)
2	Tree Cover, broadleaved, deciduous, closed
9	Mosaic: Tree cover / Other natural vegetation
11	Shrub Cover, closed-open, evergreen (Examples of sub-classes at reg. level *: (i) sparse tree layer)
12	Shrub Cover, closed-open, deciduous (Examples of sub-classes at reg. level *: (i) sparse tree layer)
13	Herbaceous Cover, closed-open (Examples of sub-classes at regional level: (i) natural, (ii) pasture, (iii) sparse trees or shrubs)
20	Water Bodies (natural & artificial)
22	Artificial surfaces and associated areas

Source: LCCS by FAO, USA

Image 6: Land Cover Map

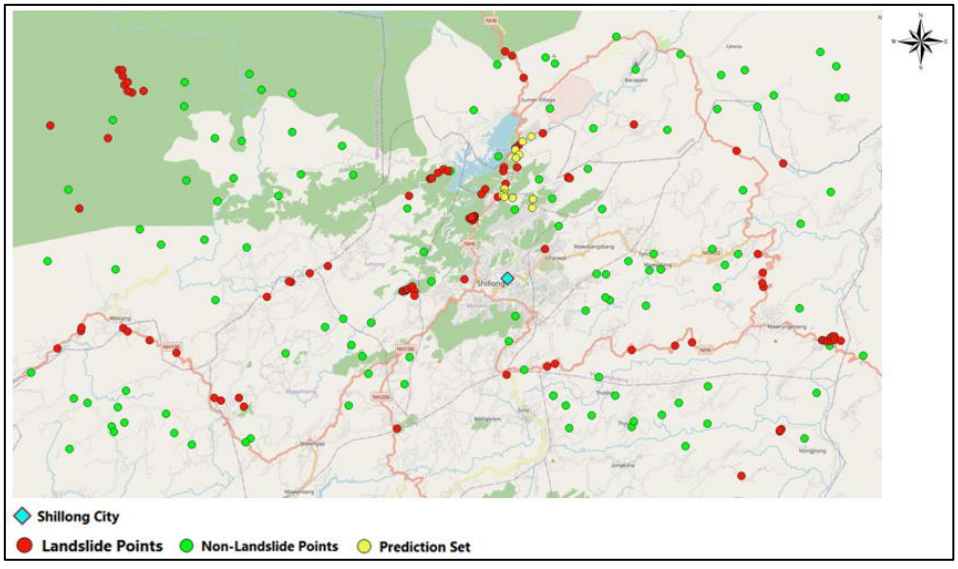


Source: DIVAGIS, QGIS

Distance from Shillong City

One of the factors affecting slope stability is urbanization, which consists of construction of man-made structures like roads and road cuts, buildings, and cities. It significantly changes surface water run-off and slope stability. According to a study, urban areas are at a greater risk of rainfall induced landslides than rural areas (Johnston et al., 2021). The only major city in the study area is Shillong. If the point is closer to the city (or lying within it), it may be at a higher risk of landslide occurrence because of steep road cuts and other infrastructure. Thus, the shortest distance in kilometers from all points to the midpoint of Shillong city is calculated using QGIS. Image 7 shows Shillong City and all the data points in the study area.

Image 7: Shillong City and Observations



Source: QGIS

Distance from Nearest River

The slope of an area becomes more unstable, the closer it is to any river. This is because proximity to river streams increases the amount of moisture in soil, thus making it more susceptible to erosion, and eventually, slope failure. (Pourghasemi et al., 2012a)

Three rivers have been considered in the study area, namely Kulsi river, Umtyngar river and Umiam river. For each data point, the shortest distance was measured in kilometers between the point and the nearest river. This was done using QGIS.

Distance from Nearest Fault

Fault represents structural discontinuities. They indicate tectonic breaks where the rock strength has been decreased. Generally, an area that is closer to a Faultline is more prone to landslide occurrence (Chen & Li, 2020).

Three major faults can be identified within the study area. They are the Barapani-Tyrsad Shear Zone, Kulsi Fault and Um Ngot Lineament.

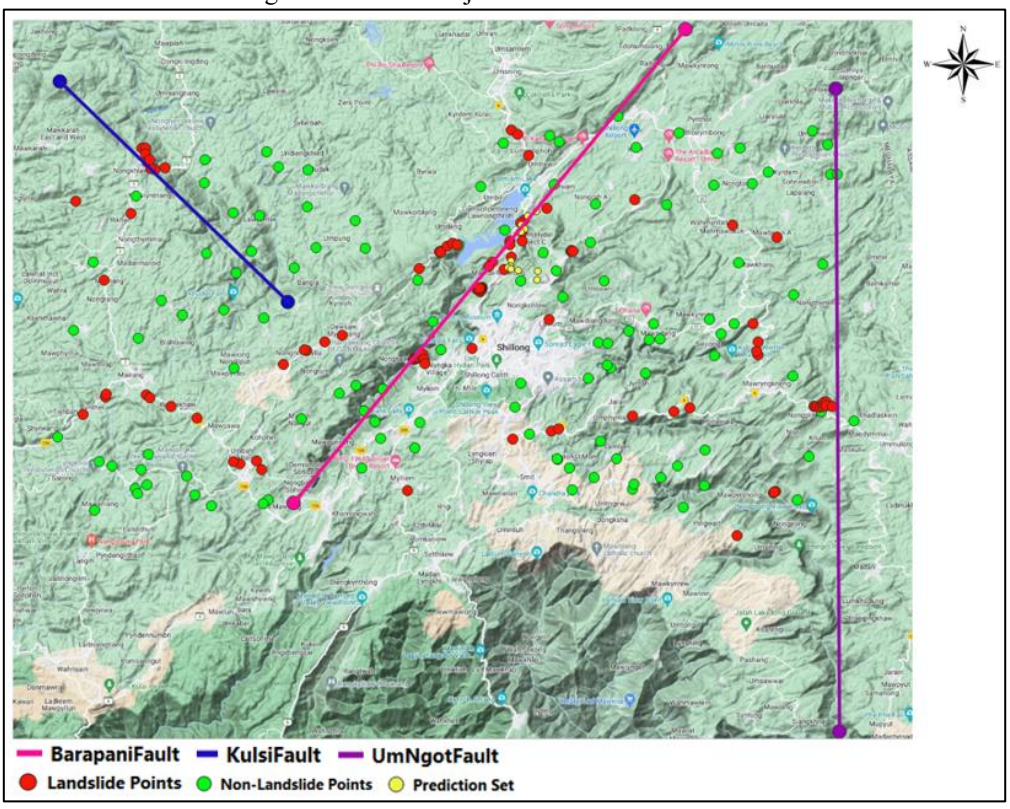
Barapani-Tyrsad Shear Zone is a lateral slip fault that has a North-East trend. It passes through central and eastern part of Shillong Plateau.

Kulsi fault is an active fault with a NNE-SSW trend. It passes through northern margin of Shillong Plateau.

Um Ngot lineament cuts across the Shillong Plataeu with a NE–SW trend. This lineament developed during the late Jurassic– Early Cretaceous period and contains several alkaline intrusive bodies, including the Sung Valley complex. (Singh & Singh, 2018)

For each data point, the shortest distance was measured in kilometres between the point and the nearest fault. This was done using QGIS. Image 8 shows parts of the three fault lines and the data points within the study area.

Image 8: Parts of 3 major fault lines and Observations



Source: QGIS

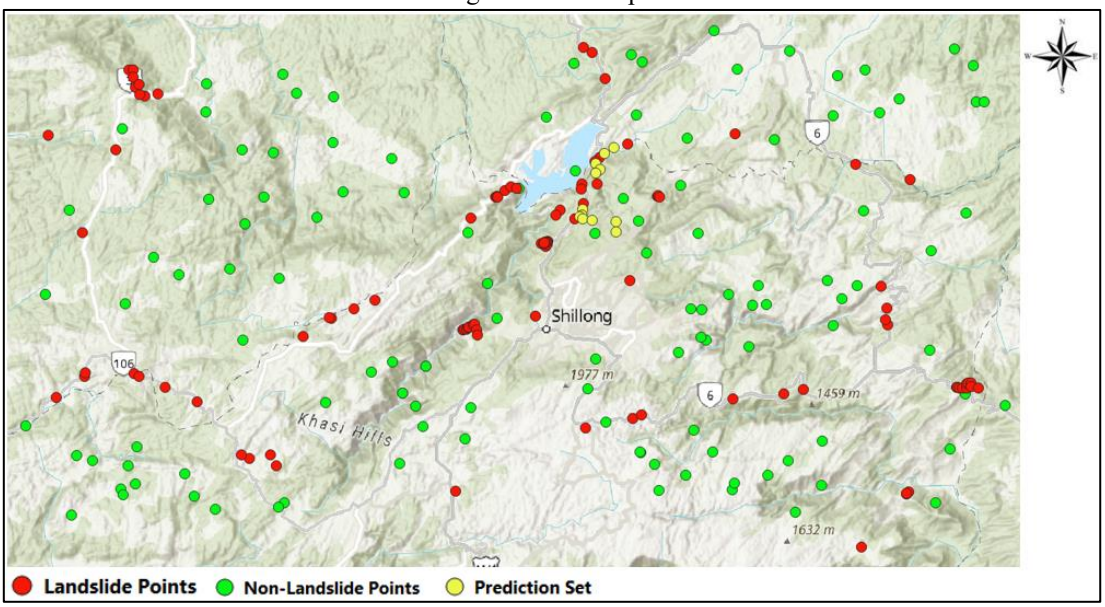
Visual Atmospheric Resistance Index (VARI)

The Visual Atmospheric Resistance Index (VARI) is an index for statistically assessing the fraction of vegetation by utilizing just the visible spectrum. It is one of the most often applied metrics for gathering information from vegetation using simply RGB cameras. It was created as a crop index that is based on RGB. It can be calculated by applying the following equation:

$$VARI = \frac{R_{Green} - R_{Red}}{R_{Green} + R_{Red} - R_{Blue}}$$

The RGB values were extracted from a vegetation map created using ArcGIS. This map is in Image 9.

Image 9: VARI Map

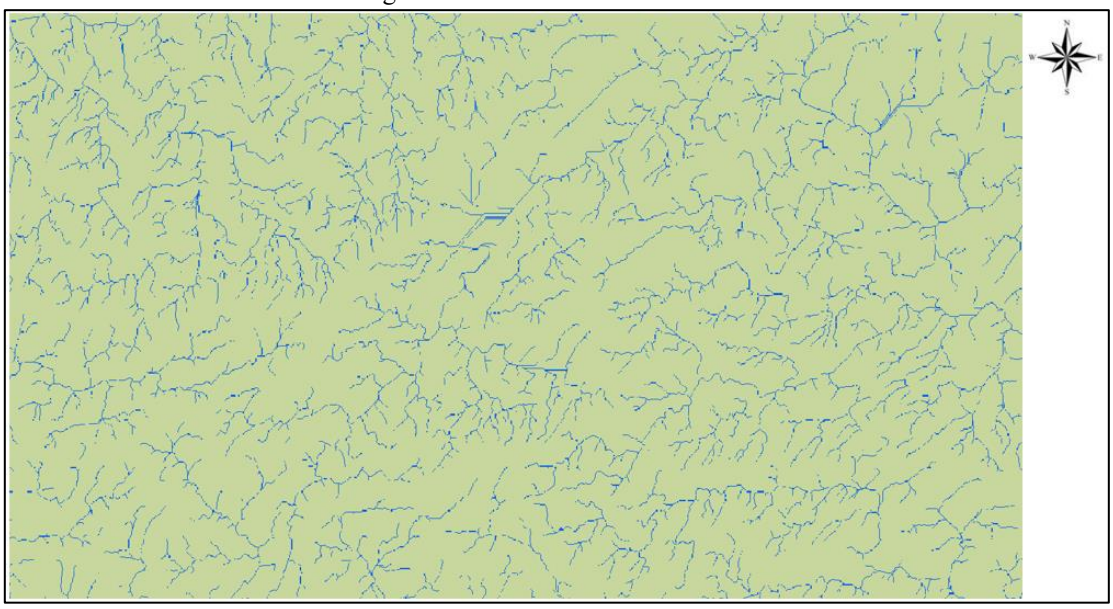


Source: ArcGIS, QGIS

Watershed (Flow Direction)

A watershed is a type of natural hydrologic entity that includes a 'specific area' of land surface where runoff or rainfall originates and travels to a particular drain. Rainfall and snowmelt are channelled into streams and rivers by watersheds. Lakes, bays, and oceans are among the bigger bodies of water that these smaller ones flow into. Watershed offers a framework for sensible land use planning and enables tackling upstream-downstream connection concerns like landslides. The watershed in the study area and the directional flow of these entities was mapped using ArcGIS and directional classes were extracted using QGIS. Image 10 shows the watershed distribution in the study area.

Image 10: Watershed Distribution



Source: ArcGIS

Data Input

The conditioning factors were tabulated for each point. Landslide points and non-landslide points were tabulated in two separate sets. The landslide and non-landslide sets were randomly split in the 80:20 ratio. The training set was formed by merging 80% of the landslide set and 80% of the non-landslide set. The test set was formed by merging the remaining 20% points from each set.

Methodology – RF and SVM

The objective of the study was to determine whether a landslide will occur at a given point or not. Two machine learning classification algorithms: Random Forest (RF) and Support Vector Machines (SVM) were employed. Python was used to build the models. Compared with other tree-ensemble methods, RF is computationally light. Therefore, RF is commonly used for large-scale mapping and classification applications in ecology, soil science and flood mapping (Taalab et al., 2018). Support Vector Machine is one of the popular models for classification and requires little tuning. It employs a linear hyperplane that separates data patterns in an optimum manner. It can be used for non-linear data by transforming it into a linearly separable form in a higher dimension. SVM can be used efficiently for landslide susceptibility analysis and may be used widely for the prediction of various spatial events (Lee et al., 2017).

Random Forest

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). It is an ensemble machine learning algorithm that constructs multiple decision trees in training, and gives its output as the class which is selected by a majority of the decision trees.

Random Forests employs CART algorithm (Classification and Regression Trees) to build multiple classification trees. The algorithm helps reduce overfitting of datasets and increase precision, overcoming the limitations of a decision tree algorithm. CART algorithm uses Gini's impurity index, which is a measure of the purity of a specific class, to build a decision tree. The best split increases the purity of the sets resulting from the split (Tangirala, 2020).

$$Gini(t) = 1 - \sum_{i=1}^{j} P(i|t)^2$$

Where j represents the number of classes in the label, and P represents the ratio of class at the ith node.

The model runs new input down every decision tree in the forest. It then makes predictions by assigning a class to that input based on majority vote. The proportion of votes that a class receives is used to attribute a probability of class membership (Bostrom, 2007).

Random Forest has no formal assumptions about distributions. The algorithm does not follow a parametric approach and can thus handle skewed and multi-modal data. It can also handle ordinal as well as non-ordinal categorical data.

To predict landslide susceptibility, a binary Random Forest Classifier needs to be trained in order to determine the class of the predictor variable, which is the presence or absence of landslides. The base model was created using Random Forest Classifier algorithm by taking the earlier mentioned variables as inputs on the training dataset. The next step is to improve the performance of the model by conducting hyperparameter tuning. The values of the hyperparameters greatly influence the random forest model's performance. Therefore, the optimum value for these parameters must be found (Sun et al., 2020). A range of values for the hyperparameters was obtained after using the Randomized Search algorithm. This range was further used as the input for the Grid Search algorithm. This algorithm produced the optimal hyperparameters. These hyperparameters were then used to fit a best model and run a 10-fold cross validation on the training dataset. Finally, this model was used to predict landslide susceptibility on the test dataset. Table 3 explains the hyperparameters that RF uses.

Table 3

Hyperparameter	**Explanation**
n_estimators	No. of decision trees
min_sample_split	Minimum no. of samples to be split
min_samples_leaf	Minimum no. of samples required at a leaf node
max_features	Maximum no. of features to consider when splitting a node
max_depth	Maximum no. of splits the tree makes until all the samples in leaves are pure samples
bootstrap	Random sampling with replacement

Source: Authors

Support Vector Machines

A Support Vector Machine (SVM) is a supervised learning algorithm based on the principle of structural risk minimization (Lee et al., 2017). Basically, the model capacity is tuned to match data complexity. Thus, SVM models are data-dependent (Ballabio & Sterlacchini, 2012). The training data is taken as input and the goal is to separate it into two classes. SVM uses a decision surface to separate them with the goal of maximizing the margin between the classes. In other words, a decision boundary is created in such a way that the separation between the classes is as wide as it can be. This decision boundary is called the optimal

hyperplane and the training points that are closest to the hyperplane are called support vectors (Pourghasemi et al., 2013).

SVM can handle a wide range of classification problems. In the case of landslide susceptibility mapping, SVM aims to separate the classes 'landslide' and 'non-landslide'. Some of the most commonly used classification techniques for susceptibility mapping are based on linear functions (such as Logistic Regression and Discriminant Analysis). However, the linear model is just an approximation for real-world data, which is often linearly inseparable. The SVM approach can solve this problem. If the data is non-linear, they can be projected to a higher dimension where they become linearly separable. This projection may become a computationally costly task in the real world since there might be many features in the data which require complex polynomial transformations and combinations. To simplify this task, the kernel trick is employed. The "trick" represents the data through pairwise similarity comparisons, instead of explicitly transforming the data and representing these transformed coordinates. The appropriate kernel can be chosen for a given task from among various kernel functions like linear, polynomial, sigmoid, Gaussian radial basis function, etc. The objective of this paper is to classify 'landslide' versus 'non-landslide'. Before building an SVM model, the data needs to be scaled. SVMs are sensitive to the scale of the input features, and if the features are on different scales, it can lead to poor performance of the model. Standardization was chosen as the preferred feature scaling technique since it yielded a better performing model. Standardization involves subtraction of the mean from each value and its division by the standard deviation. The scaled data consists of each feature, now with a mean of 0 and standard deviation of 1.

SVM models were then built using the linear, polynomial degree 2, 3, 4 and Gaussian radial basis function kernels. The parameters of the SVM greatly influence the performance of the model. Thus, they need to be fine-tuned. These include the cost parameter C and the gamma parameter γ. C helps adjust the training error and the margins. A small cost allows more misclassifications and therefore, a large (soft) margin is created. On the other hand, a large cost allows for fewer misclassifications and thus, a narrow (hard) margin is created. γ is simply the radius of the kernel function. If gamma is too small, the model cannot truly capture the shape (complexity) of the data. It can be said that the model is too constrained. If gamma is too large, the model will overfit the training data and reliable classifications cannot be obtained using new data. Thus, the correct definition of these user-defined parameters significantly increases the accuracy of the SVM solution (Pourghasemi et al., 2013). In this study, the optimum parameters were obtained using the Grid Search Algorithm. The best performance was shown by the model using the Gaussian radial basis function kernel.

To improve the classification performance of the SVM (especially since the sample size is small), the SVM ensembles with bagging were used. An ensemble of classifiers combines the decisions of several individual classifiers. An ensemble often shows much better performance than the individual classifiers that make it up (Dietterich, 1997). One such ensemble method is bagging, which stands for bootstrap aggregation. It can help reduce variance is a dataset that has too much nosie. The process involves selecting several random samples from a training set. It is essential to note that replacement is allowed, i.e., an individual point can be chosen more than one time. Models are then trained independently on these samples and the final result is the combination or majority of the various models. Bagging can improve performance and help avoid overfitting. A bagging classifier was used with the RBF kernel model that showed best performance.

Results – RF and SVM

The metrics used to evaluate model performance are accuracy, precision, recall and f1 score. Higher the accuracy of the model, the lower is its misclassification rate. Misclassifications should be reduced as much as possible. Usually, there is a tradeoff between precision and recall of the model. In the case of landslide susceptibility mapping, a higher recall would be preferable and a lower precision might be acceptable. The f1 score combines precision and recall. It is a reliable metric to judge model accuracy and performance when the dataset is relatively class balanced. The training dataset consists of 45% landslide points and 55% non-landslide points, which satisfies this criterion. The performance of the model is considered to be better the closer the f1 score is to 1.

Random Forest

The best model was made with the optimized hyperparameters that was obtained through the Grid Search algorithm. The optimum parameters obtained were number of trees = 600, minimum sample split = 4, minimum samples leaf = 2, maximum features = 6, and maximum depth = 10. Cross validation results yielded a mean training accuracy of 81.4%. The test accuracy for this model was 89%. The misclassification rate is 11%. Table 4 shows the confusion matrix.

Table 4

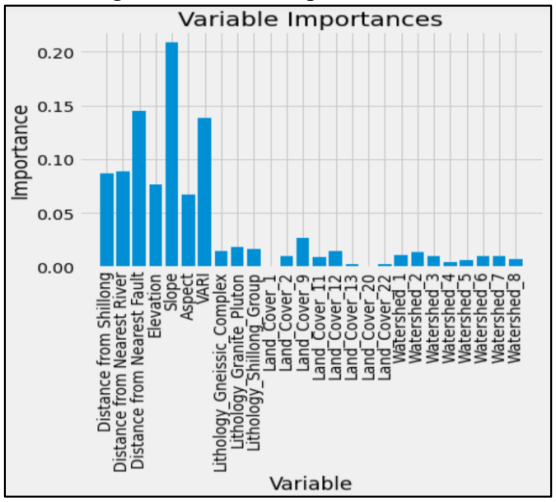
Actual	Predicted	
	Non-Landslide	Landslide
Non-Landslide	22	2
Landslide	3	17

Source: Authors

The precision is 0.89, i.e., 89% of the predicted landslides were actual landslides. The recall is 0.85, i.e., 85% of the actual landslides were identified correctly. The f1 score is 0.87. The model can be considered a good fit.

Variable importance relates to how much the model uses each variable to give accurate predictions. The chart in Image 11 shows that slope, distance from the nearest fault and VARI were the three most important variables.

Image 11: Variable Importance Chart



Source: Authors

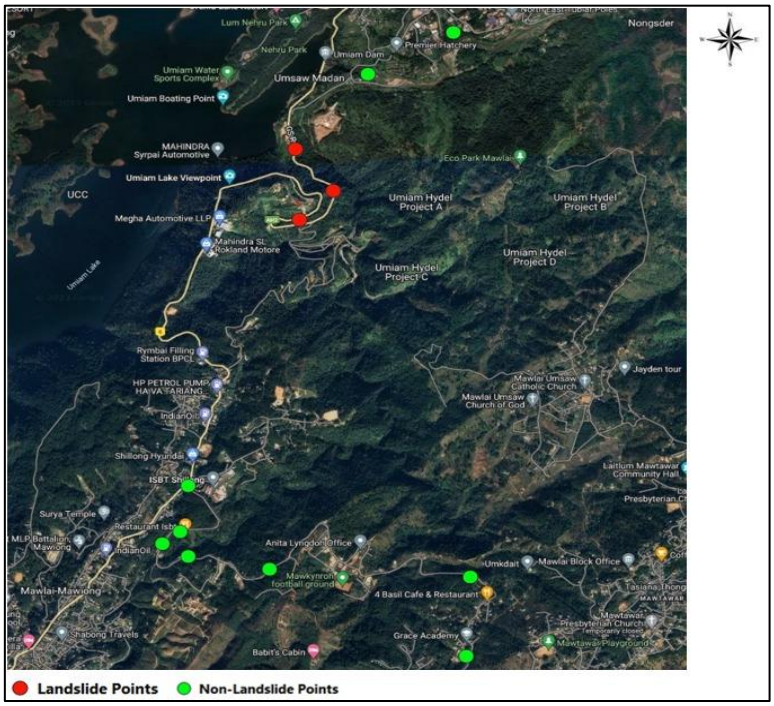
This model was then used to classify the 12 data points with unknown classes (T1-T12). The output is given in Table 5. The predictions are mapped in Image 12.

Table 5

Sr. No.	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
Landslide	No	No	Yes	Yes	Yes	No	No	No	No	No	No	No

Source: Authors

Image 12: Predictions – RF



Source: Authors

Support Vector Machines

The model with the best accuracy was the SVM using the Gaussian radial basis function kernel and bagging classifier. The optimum parameters were obtained as C = 5, γ = 0.1. The bagging classifier was used with this model. A training accuracy of 75.1% and a test accuracy of 86.4% were obtained. The misclassification rate is 13.6%. Table 6 indicates the confusion matrix.

Table 6

Actual	Predicted	
	Non-Landslide	Landslide
Non-Landslide	22	2
Landslide	4	16

Source: Authors

The precision is 0.89, i.e., 89% of the predicted landslides were actual landslides. The recall is 0.8, i.e., 80% of the actual landslides were identified correctly. The f1 score is 0.84. Thus, the model can be considered a good fit.

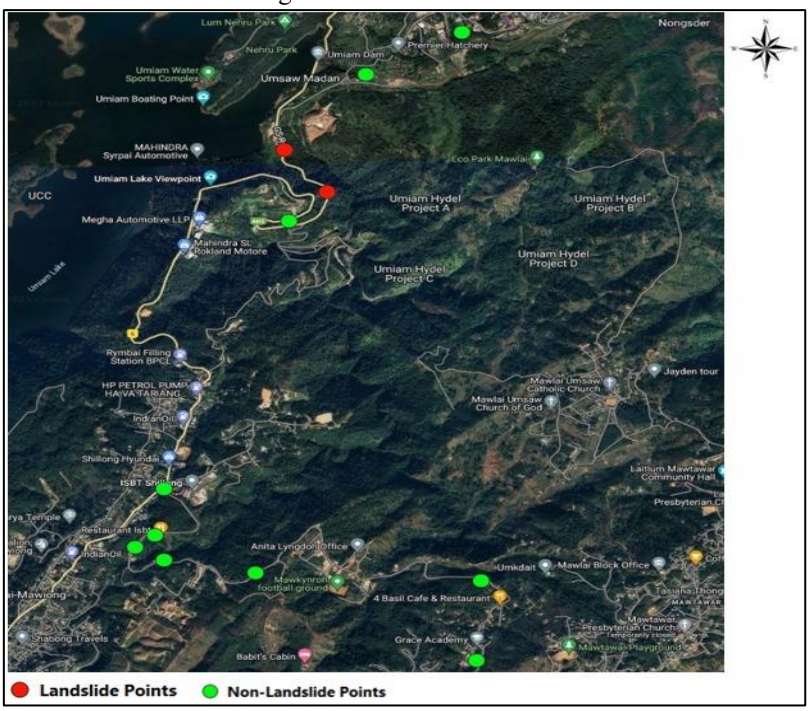
This model was then used to classify the 12 data points with classes unknown (T1-T12). The output is given in Table 7. The predictions are mapped in Image 13.

Table 7

Sr. No.	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
Landslide	No	No	Yes	Yes	No	No	No	No	No	No	No	No

Source: Authors

Image 13: Predictions – SVM



Source: Authors

Data Collection – CNN

The objective is to predict whether a landslide will occur and where. To answer the question of where the landslide will occur, image segmentation is done using CNN. To train a CNN model for understanding where a landslide can occur, Sentinel 2 satellite data is collected from their site (LandSlide4Sense, n.d.), this dataset has 3500 images. The dataset is divided into two subparts – training set and mask set.

The training set contains 3500 satellite images in a matrix form with 14 variables (bands), hence the matrix shape for one image is 128 x 128 x 14. These bands have the RGB values along with 11 more variables captured by the satellite.

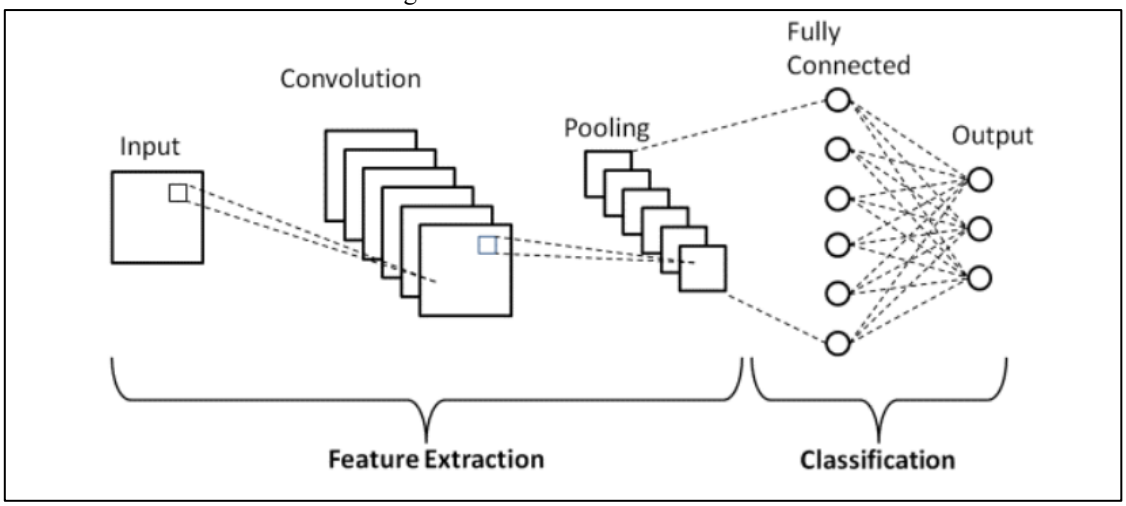
The mask dataset contains 3500 images showing the actual prediction of where the landslide occurs by mapping it in a different colour for a respective train dataset image. It acts as the actual y variable. Each image in the mask set is given in the shape 128 x 128 x 1 indicating each image is of 128 pixels and has one channel, which denotes whether a landslide has occurred or not in a specific pixel.

The final predictions were made for 32 drone images of the 12 points from the DST project.

Methodology – CNN

CNNs are a type of artificial neural network. They have been applied in the project to predict whether and where a landslide will occur.
CNNs are designed to process data that can be represented in a matrix structure, such as images represented as a matrix of pixel values. There are broadly two parts to a CNN architecture, the first part being the layers that identify, separate, and learn to detect the various features in the image data which comes to be known as feature extraction and the second part a fully connected/dense layer which uses the results of feature extraction to predict the label or class of the image. Image 14 shows a basic CNN architecture.

Image 14: Basic CNN Architecture



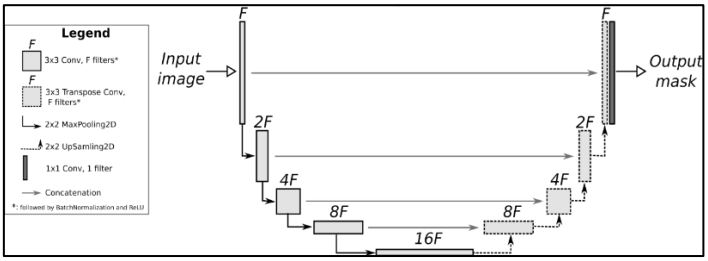
Source: https://www.analyticsvidhya.com/blog/2022/03/basic-introduction-to-convolutional-neural-network-in-deep-learning/

CNNs are called "Convolutional" because of the application of the convolution mathematical function in their layers, which is a special linear function that expresses the shape of one function when it is modified by another. Mathematically, it is an integral that expresses the amount of overlap of a function g as it is shifted over another function f, written as (Weisstein, n.d.):

$$c = (f * g)(t) = \int_0^t f(\tau)g(t-\tau)d\tau$$

Where,
c: convolution output function
f: input function
g: function that shifts over input
t: variable representing range of shift
τ: shifting against t
Other than the convolution layers, there is pooling done under feature extraction to reduce the size of the pixel matrices, a combination of convolution and pooling layers can be applied depending on the use case. For this research, a U-Net architecture is followed by CNN. U-Net gets its name from the "U" shape of the layers as seen in Image 15.

Image 15: U-Net Architecture



Source: https://nchlis.github.io/2019_10_30/page.html

This architecture is particularly useful in image segmentation, which not only classifies the image, but also gives information on where the classification is occurring. In the case of landslide mapping, U-Net architecture will help us understand whether a landslide will occur in a particular image or not, as well as where the landslide will occur. The model will give an image as the output which will highlight the area where the landslide will occur.

The first layer in the U-Net model is the input layer, where the training images in the form of pixel matrices are entered into the model. The next levels together are called the "contraction" path. Each level has a convolution layer followed by a dropout layer (a dropout layer is added to prevent the model from over-fitting) then a convolution layer again and finally a pooling layer.

After the contraction path, comes the "expansion" path. Here, the levels start with a transpose convolution layer (a transpose convolution layer is used to bring back the pooled data to its original size) followed by a convolution layer, subsequently a dropout layer and finally another convolution layer. The number of filters set for the convolution layer and the dropout rates increases in each level in the contraction path and both decrease with each level in the expansion path.

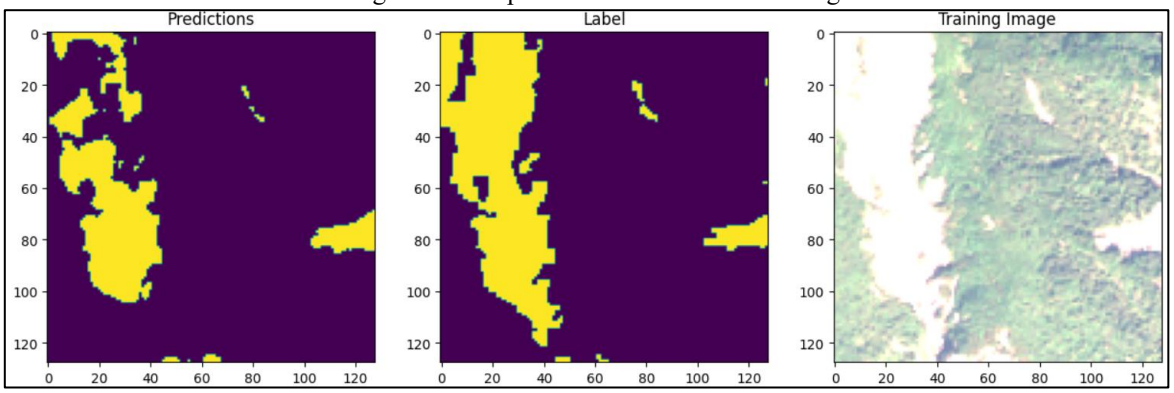
For this research the model is defined only on the RGB bands of the satellite data. The U-Net model is trained using the Relu activation function and sigmoid transformations based on literature review. In addition, padding has been included to reduce loss of information at the borders of the image. For the objective of landslide susceptibility, the best model is chosen based on the highest recall value among 100 epochs run.

Results – CNN

The best U-Net model was found to be the one with seven levels (each level comprising the layers mentioned in the methodology), four in contraction path and three in expansion path based on trial and error. The sentinel satellite data of 3500 images was divided into 80% training data, i.e., 2800 images and 20% validation data, i.e., 700 images. Using this division and the RGB values the model is trained.

An example of the output for the test images can be seen in Image 16. The yellow areas indicate occurrence of a landslide, while the purple indicates areas where landslides are not likely to occur.

Image 16: Example of Prediction on a test image

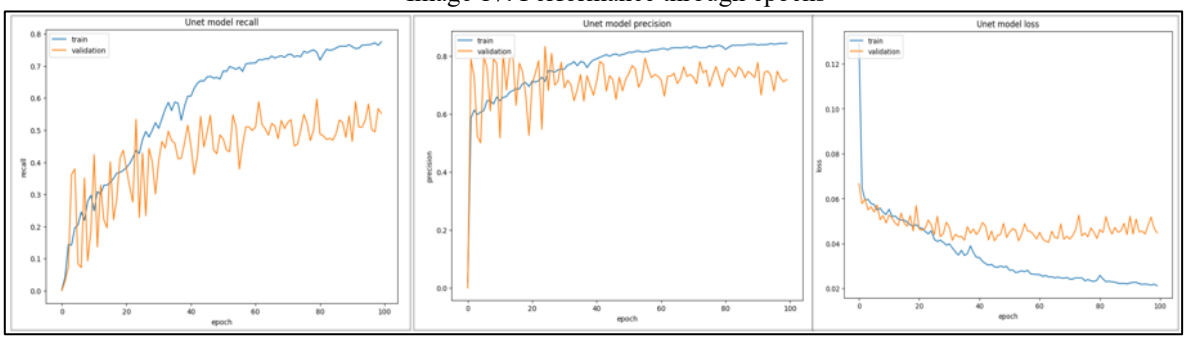


Source: Authors

Similarly, the model predicts the output for all 700 images in the test set.

The model performance throughout the training epochs can be observed from the graphs in Image 17.

Image 17: Performance through epochs



Source: Authors

The performance of the best model is summarized in Table 8.

Table 8

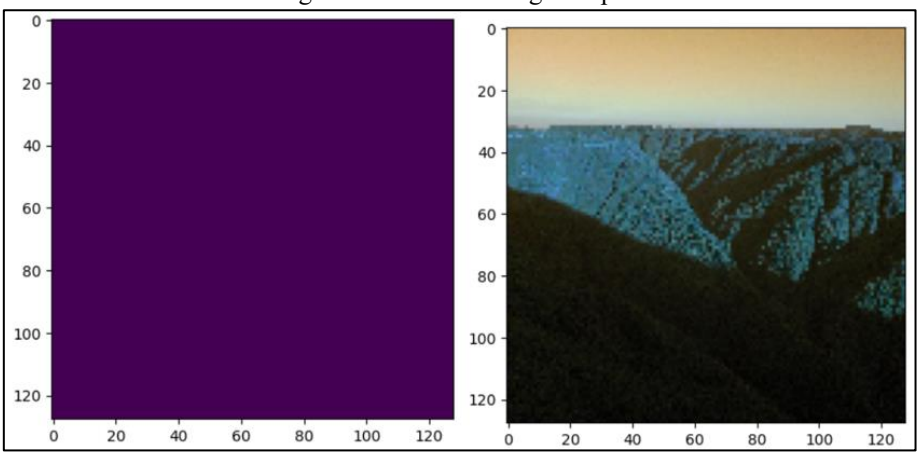
	Training	Testing
Accuracy	99.04%	98.54%
Recall	74.28%	71.93%
Precision	83.29%	55.29%
F1 Score	78.53%	62.53%
Loss	2.37%	4.47%

Source: Authors

A high accuracy of 98.54% is observed. This is usually interpreted as the model predicts 98.54% of the pixels in each image correctly. However, when it comes to landslide susceptibility, accuracy may not be the most reliable metric to judge the model on; because the accuracy does not measure if the location of the landslide is mapped correctly. In other words, the model may be predicting the proportion of landslide to non-landslide pixels with high accuracy however, it still might not be performing very well in terms of which area should have a higher probability of landslide. To overcome this, we look at the recall of the model. Another reason for choosing recall would be considering landslide susceptibility

mapping false positives would be preferable over false negatives. Hence, the recall is observed.
A recall of 71.93% indicates that the model identifies 71.93% of the landslides correctly. This metric can still be improved. However, since only RGB values were used to build a model, the metrics can be considered to indicate a good fit.
This best model is then applied to images from the 12 points considered for prediction in the study area. An example of a prediction on one of these images is given in Image 18.

Image 18: Prediction Image Output



Source: Authors

For this image, the model predicts there will be no landslide.

Conclusion

Landslide susceptibility mapping is a valuable tool for land use planning and management. Even when a planning policy is to be implemented, such a map proves to be useful for hazard studies.
The Shillong Plateau is highly prone to landslides and suffers a lot of damage. Mapping susceptibility in this area is the first step to being able to implement prevention and control measures.
The DST has identified this as an important area for documentation and analysis. This study aims to contribute to the mapping process.
The machine learning models in the study run classification algorithms to separate the data into the classes 'landslide' versus 'non-landslide'. All the chosen models show relatively good accuracy. They can be used for further research, mapping and planning.
The best model in this study was SVM and can be used going forward to create accurate susceptibility maps. This can prove beneficial in efforts towards detection and prevention of landslides.
The misclassification rate of the best RF model was 3% lower than that of the best SVM model. RF gave an output with 2 false negatives and 3 false positives, while SVM gave an output with 2 false negatives and 4 false positives. Thus, the models are not too far apart in their performance. It can be concluded that in the case of this study, the two models are comparable. However, judging purely by the test accuracy and recall rates of the models, RF is the better fit. It can be used to accurately map landslide susceptibility in the area.

The CNN model's performance metrics should be examined. The current model cannot be considered as capable of mapping landslide susceptibility. This is because it only uses RGB bands for segmentation. For landslide susceptibility mapping, other bands must be included to give accurate results.
The area of predictions for the three models can be compared.
CNN is the only one whose predictions are not restricted to the Shillong area. This is because CNN was trained on global satellite data. The RF and SVM models in this study have been trained using factor ranges specific to the Shillong Plateau and their predictions will remain accurate only for this area.

Limitations

- The terrain can be highly complex, and landslides can occur due to a wide range of factors. It can be difficult to account for all these factors and develop a model that accurately reflects the real-world conditions.
- Models trained on data from past landslides may not be able to account for changes in land and structure for the locations where landslides have already occurred.
- Using a larger dataset for RF and SVM models can generally lead to better training and potentially result in a more accurate model. The models will have more diverse and representative samples to learn from if the dataset size is larger. This can help capture a wider range of patterns and relationships, leading to improved generalization performance. However, the data set considered for RF and SVM models is comparatively small, and the models have a smaller sample to train with.
- For landslide susceptibility mapping to be precise and informative using a CNN model, the model needs to be trained and tested with variables influencing occurrence of landslides like slope, curvature, land cover, etc. The CNN model in this study simply uses RGB. This limits its scope.

Future Scope

- To improve the accuracy of landslide prediction, it may be beneficial to use more extensive models that take into account a larger number of variables that exert influence on the occurrence of landslides.
- The models in this study only predict the occurrence or non-occurrence of a landslide, but the objective of the study can further be expanded to include identification of the type of landslide and the severity of the potential landslides. This can provide valuable insights to help with better preparedness and response planning.
- It is imperative to consider the influence of time on the geology of an area affected by, or prone to landslides. Each time a landslide occurs, certain features of the land are affected. A model incorporating the dimension of time for each factor contributing towards the occurrence of landslide should thus be considered. A time series model can be used to forecast the propensity of a landslide occurrence over a period of time, as older landslides can be given a lower weightage than newer landslide while predicting.

Acknowledgements

This study was possible thanks to the Department of Science and Technology and to Dr. Sunayana Sarkar, a geologist part of the ongoing DST project.

References

Azarafza, M., Azarafza, M., Akgün, H., Atkinson, P. M., & Derakhshani, R. (2021). Deep learning-based landslide susceptibility mapping. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-03585-1

Costa, L., Nunes, L., & Ampatzidis, Y. (2020). A new visible band index (vNDVI) for estimating NDVI values on RGB images utilizing genetic algorithms. Computers and Electronics in Agriculture, 172, 105334. https://doi.org/10.1016/j.compag.2020.105334

Dolidon, N., Hofer, T., Jansky, L., & Sidle, R. (n.d.). Watershed and forest management for landslide risk reduction. In Landslides – Disaster Risk Reduction (pp. 633–649). Springer Berlin Heidelberg. Retrieved May 5, 2023, from http://dx.doi.org/10.1007/978-3-540-69970-5_33

Sun, D., Xu, J., Wen, H., & Wang, D. (2021). Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and random forest. Engineering Geology, 281, 105972. https://doi.org/10.1016/j.enggeo.2020.105972

Meena, S. R., Nava, L., Bhuyan, K., Puliero, S., Soares, L. P., Dias, H. C., Floris, M., & Catani, F. (2022). HR-GLDD: A globally distributed dataset using generalized DL for rapid landslide mapping on HR satellite imagery. https://doi.org/10.5194/essd-2022-350

Fang, Z., Wang, Y., Peng, L., & Hong, H. (2020). Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping. Computers & Geosciences, 139, 104470. https://doi.org/10.1016/j.cageo.2020.104470

Lee, S., Hong, S.-M., & Jung, H.-S. (2017). A Support Vector Machine for Landslide Susceptibility Mapping in Gangwon Province, Korea. Sustainability, 9(1), 48. https://doi.org/10.3390/su9010048

POURGHASEMI, H. R., JIRANDEH, A. G., PRADHAN, B., XU, C., & GOKCEOGLU, C. (2013).

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

Hong, H., Pradhan, B., Pradhan, B., Xu, C., Youssef, A. M., & Chen, W. (2017). Comparison of four kernel functions used in support vector machines for landslide susceptibility mapping: a case study at Suichuan area (China). Geomatics, Natural Hazards and Risk, 8(2), 544–569. https://doi.org/10.1080/19475705.2016.1250112

Ballabio, C., & Sterlacchini, S. (2012). Support Vector Machines for Landslide Susceptibility Mapping: The Staffora River Basin Case Study, Italy. Mathematical Geosciences, 44(1), 47–70. https://doi.org/10.1007/s11004-011-9379-9

Guyon, I., Vladimir Vapnik, Boser, B. E., Léon Bottou, & Solla, S. A. (1991). Structural Risk Minimization for Character Recognition. Neural Information Processing Systems, 4, 471–479.

Kim, H.-C., Pang, S., Je Hongmo, Kim, D., & Bang, S.-Y. (2002). Support Vector Machine Ensemble with Bagging. Lecture Notes in Computer Science, 397–408. https://doi.org/10.1007/3-540-45665-1_31

Henrik Boström. (2008). Calibrating Random Forests. International Conference on Machine Learning and Applications. https://doi.org/10.1109/icmla.2008.107
Breiman, L. (n.d.). Random Forests [Review of Random Forests]. Machine Learning, 45, 5–32.

Agrawal, N., & Dixit, J. (2022). Assessment of landslide susceptibility for Meghalaya (India) using bivariate (frequency ratio and Shannon entropy) and multi-criteria decision analysis (AHP and fuzzy-AHP) models. All Earth, 34(1), 179–201. https://doi.org/10.1080/27669645.2022.2101256

Muhammad Moazzam, Anujit Vansarochana, Jaruntorn Boonyanuphap, Sittichai Choosumrong, Rahman, G., & Geraud Poueme Djueyep. (2020). Spatio-statistical comparative approaches for landslide susceptibility modeling: case of Mae Phun, Uttaradit Province, Thailand. SN Applied Sciences, 2(3). https://doi.org/10.1007/s42452-020-2106-8

Tangirala, S. (n.d.). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm [Review of Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm]. International Journal of Advanced Computer Science and Applications, 11(2), 612–619.

Singh, L. S., & Singh, K. M. (n.d.). GEOLOGY, STRUCTURE AND TECTONICS OF SHILLONG PLATEAU [Review of GEOLOGY, STRUCTURE AND TECTONICS OF SHILLONG PLATEAU].

Journal of Emerging Technologies and Innovative Research (JETIR), 5(8), 1043–1051.
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