Original Article

Geospatial Landslide Prediction – Analysis & Prediction From 2018-2022

Harsh Jindal[1*], Ayush Yadav[2], Abhinav Sehgal[3], Sugandha Sharma[4], Ankit Panigrahi[5], Dipesh Ranjan[6], Abhoy Gorai[7], Manas Tiwari[8]

[1*]Chandigarh University- harshjindal040123@googlemap.site
[2]Chandigarh University ayushyaduvanshi14@gmail.com
[3]Chandigarh Group of Colleges, Landran abhisehgal2003@gmail.com
[4]Chandigarh University- sugandha.cse@cumail.in
[5]Chandigarh University - panigrahi0702@gmail.com
[6]Chandigarh University - sinhadipesh25@gmail.com
[7]Chandigarh University abhoygorai04@gmail.com
[8]Chandigarh University-tiwarimanas2711@gmail.com

***Corresponding Author:** Harsh Jindal

*Chandigarh University- harshjindal040123@googlemap.site
DOI: 10.47750/pnr.2023.14.S02.304

Abstract

Landslides are a significant issue in India due to the country's varied topography, heavy monsoon rains, and deforestation, which contribute to soil instability and increased landslide risk. These natural disasters can cause damage to infrastructure and loss of life. In light of the ongoing problem of landslides in India, this research paper aims to address the need for effective landslide prediction strategies. Through the findings of this research study, a novel approach has been presented for predicting landslide occurrences in India, which will aid in reducing the impact of these events on infrastructure, communities, and lives. The work that has been carried out using data and information based in India has shown to have a low accuracy level. As a result, the model created using this information is not deemed to be very reliable. This study focuses on predicting landslides in India using machine learning models such as (Xtreme Gradient Boost) XGboost, random forest, and AdaBoost. Previous research on landslide prediction in India had not been widely done or had not achieved acceptable accuracy levels. This study aims to address this gap in knowledge and improve the predictability of landslides in India. The research is based on a database of different areas in India and aims to increase awareness and save lives and resources by predicting landslides with 91% accuracy.

Keywords: Area Under Curve, Back Propagation Neural Network, Geographical Information System, Kappa Coefficient, Landslide Susceptibility, Machine Learning, Overall Accuracy, Rainfall induced Landslides, Random Forest, Xtreme Gradient Boost

INTRODUCTION

A landslide is a geological phenomenon where a mass of rock, earth, or debris moves downhill, typically as a result of gravity. It can occur as a result of heavy rainfall. When the soil becomes saturated with water, it loses its ability tosupport the weight of the overlying rock, earth, or debris. This can cause the material to become unstable and begin moving downhill. Landslides can significantly impact people's livelihoods by destroying homes and property, disrupting transportation and commerce, causing loss of income, biodiversity and ecosystem services, and damagingpublic services. Addressing this issue and protecting people's livelihoods must be a key focus of disaster risk reduction and management strategies. For these reasons, it is important to address landslide issues in order to protecthuman lives and property, maintain economic stability, and preserve the environment. The research highlights the importance of understanding and addressing landslides in the context of these impacts, and the need for further research and management efforts to minimize the risks associated with this natural hazard.

The key factors that are causing a landslide are :

1. Unstable slopes: Landslides occur when the angle of a slope becomes too steep, causing the soil or rock to become unstable and slide downhill.
2. Weak rock or soil: Weak rock or soil layers can cause landslides, as they are more susceptible to failure.
3. Water: Water can cause landslides in various ways, such as by saturating the soil, eroding slopes, or altering drainage patterns.
4. Human activity: Human activities such as deforestation, mining, and construction on or near steep slopes can destabilize slopes and increase the risk of landslides.
5. Climate change: Climate change can affect landslides in multiple ways, such as causing prolonged droughtsor heavy rainfall, which can alter the stability of slopes.

India is prone to landslides due to its diverse topography and geology. Some regions of India are more susceptible to landslides than others. The following places are considered to be more susceptible to landslides in India:

1. Northeastern India: The hilly and mountainous regions of northeastern India are prone to landslides due toheavy rainfall, steep slopes, and weak rock formations.
2. Western Ghats: The Western Ghats, a mountain range running along the western coast of India, is prone tolandslides due to heavy rainfall, steep slopes, and weak rock formations.
3. Himalayan region: The Himalayan region is prone to landslides due to steep slopes, weak rock formations,heavy rainfall, and tectonic activity.
4. Hilly areas of North India: The hilly areas of North India, such as Himachal Pradesh, Uttarakhand, andJammu and Kashmir are prone to landslides due to heavy rainfall, steep slopes, and weak rock formations.
5. Urban areas: Urban areas that are situated on hilly terrains and near river beds are also prone to landslides due to heavy rainfall and human activities such as construction and excavation.
6. It's important to note that landslides can happen anywhere, not just in these places if the right conditions are met.

Now to minimize the effects and damage of such a disaster, several measures can be taken, including
1. Land-use planning to avoid development in high-risk areas,
2. Engineering measures such as retaining walls, drainage systems, and slope stabilization,
3. Early warning systems to detect and predict potential landslides,
4. Regular monitoring of slopes, soil, and weather conditions,
5. Hazard mapping to identify high-risk areas,
6. Rapid response and effective recovery plan to minimize damage and aid communities in recovery.

In India, the available research methods and data for landslide prediction have limitations in terms of accuracy and reliability. This research study can make a valuable contribution to the field of landslide prediction in India and help to improve the safety and well-being of communities at risk by developing more advanced models and utilizing new data sources to improve the accuracy of landslide predictions. The analysis focuses on using geospatial techniques, like digital elevation models, satellite imagery, and rainfall data can improve the accuracy of predicting landslides. Machine learning algorithms like Random Forest, XGBoost, AdaBoost and Support Vector Machines can also enhance the performance of prediction models. It was also found that these techniques should be used in combination with other methods and not as a standalone solution. It is important for further research to be conductedto integrate other geospatial data sources, such as Lidar and InSAR data, and to improve the integration of geospatialtechniques with other landslide prediction methods. Overall, this research highlights the potential of geospatial techniques for early warning and management of landslides, which can ultimately save lives and property. It was observed that we are able to achieve a higher level of accuracy in predicting landslides compared to the currently available methods in India. This is significant because accurate predictions of landslides can help save lives and property by allowing for early warning and evacuation. We have made use of multiple machine learning models in our analysis, and have presented detailed descriptions of each model. These descriptions include the key features and characteristics of the models, as well as their specific uses and applications.

➔ xgboost:
XGBoost (eXtreme Gradient Boosting) is an open-source library for gradient boosting machines that are designed to handle large datasets efficiently and with high performance. It is commonly used in machine learning tasks such as classification, regression and ranking and is particularly useful in data science competitions and in the industry for building predictive models.

➔ Adaboost:
AdaBoost is a popular boosting algorithm used for classification and regression problems. It combines multiple "weak" models (such as decision trees) to form a "strong" model that is more accurate than any of the individual weak models. The algorithm works by adjusting the weights of the training data at each iteration, giving more importance to the data points that were misclassified by the previous weak models. Thisultimately leads to a final model that is able to correctly classify or predict a larger proportion of the training data.

➔ Random forest:
Random Forest is an ensemble machine learning algorithm that is used for classification and regression tasks.It creates a forest of random decision trees, where each tree is trained on a random subset of the data. The final output of the Random Forest is the average or majority vote of the outputs of all the decision trees in the forest. It works by randomly selecting a subset of the features and a subset of the training data to train each decision tree. This process is repeated multiple times, each time with a different subset of the data, to create a diverse set of decision trees. This diversity helps to prevent overfitting and improves the overall performance of the model. The final predictions are made by averaging the predictions of all the decision trees in the forestfor regression problems and by taking the majority vote for classification problems.

➔ Decision tree:
A decision tree is a type of supervised machine-learning algorithm that can be used for classification and regression tasks. It is a tree-based model where each internal node represents a feature(or attribute), each branch represents a decision and

each leaf node represents the outcome.

The decision tree algorithm starts at the root node, which represents the entire population or sample. It then splits the population into two or more homogeneous sets based on the most significant feature, which is represented by the branches of the tree. This process continues recursively for each child node until it reaches the leaf nodes, which represent the final predictions.

The decision tree algorithm can be used to analyze data and make predictions by following the path from the root to a leaf node. The decision tree algorithm is simple to understand and interpret, and it can handle both categorical and numerical data.

➔ Support Vector Machines:

A Support Vector Machine (SVM) is a type of supervised machine learning algorithm that can be used for classification and regression tasks. The key idea behind SVM is to find a boundary (also known as a "decision boundary") that separates the different classes of data in the feature space as widely as possible. This boundary is known as the "maximum margin hyperplane".

SVMs are particularly useful when the data has many features because the algorithm can find the most important features that define the decision boundary. SVMs use a technique called the kernel trick to transform the data into a higher-dimensional space, where it becomes easier to find a linear boundary. In thistransformed space, the data points that are closest to the boundary are called "support vectors" and they are the most important data points for defining the decision boundary. SVMs are known for their good generalization performance and are widely used in industry and research.

LITERATURE REVIEW

The models and solutions given in the paper are based on Rainfall-Induced Landslides in China's Bazhou District. It has been stated that almost 90% of the Landslides in that areas are due to the involvement of Rainfall either directly or indirectly. And the reason for the rainfall is the Subtropical Monsoon Climate. The findings given in the paper are based on Back - Propagation (BP) Neural Network, Decision Tree, Random Forest, and Support Vector Machine (SVM) along with Geographical Information System (GIS) technology and Historical Data of the Landslide Hazards being taken into consideration. And using these two bases in combination, a study has been conducted on several models, including the Rainfall Intensity-Duration Threshold model and Landslide ProbabilisticQuantitative Model [1].

For the prediction of the Rainfall - Induced landslide timings, the rainfall intensity-duration threshold model was introduced by utilising the Tropical Rainfall Measuring Mission (TRMM) 3B42 rainfall product data.

In the study conducted, several performance metrics are calculated such as Overall Accuracy (OA), User's Accuracy (UA), Producer's Accuracy (PA), and Kappa Coefficient (KC).

$$OA = \frac{Number\ of\ Correctly\ Classified\ Pixels}{Number\ of\ Overall\ Pixels} \times 100\%$$

$$UA = \frac{Number\ of\ Correctly\ Classified\ Pixels\ for\ a\ Specific\ Class}{Number\ of\ Pixels\ Classified\ as\ this\ Class} \times 100\%$$

$$PA = \frac{Number\ of\ Correctly\ Classified\ Pixels\ for\ a\ Specific\ Class}{Number\ of\ Reference\ Pixels\ for\ this\ Class} \times 100\%$$

$$KC = \frac{Po - Pc}{1 - Pc} \quad \text{(It ranges between 0 and 1)}$$

There are six impact factors (elevation; lithology; aspect; slope; distance to the road; distance to water system) that are being selected after the standardisation and reclassification.

The research findings are as follows:

Prediction Accuracy of Rainfall I-D Model: 81.82%Landslide Hazard Prediction Accuracy: 90.91%
OA of BP Neural Network: 95.33% (Highest)
Kappa Coefficient of BP Neural Network: 0.91 (Highest)
In[2] study has presented a comparison of performance between the three machine learning algorithms KNN, XGBoost, and Random Forest in predicting landslide susceptibility in Malaysia's Kota Kinabalu region. Landslides are considered one of the major contributors to natural catastrophes in this Region. There are various algorithms out there but still, there is no such suitable most accurate algorithm yet to develop such a model of Landslide Susceptibility. Therefore a performance metrics comparison has been done in this study between K- Nearest Neighbour (KNN), Random Forest, and Extreme Gradient Boost (XGBoost). For the conduction of this research, training, and testing of the datasets from the data inventory and around 242 locations of landslides are arbitrarily separated into a ratio of 7:3. And among the parameters

for the predictions, there is an employment of ten spatial databases which account for the factors of landslide conditioning. The models' performance was evaluated using the area under the curve (AUC) metric, and the results showed that KNN had the highest prediction accuracy (87.52%), followed by Random Forest (84.34%) and XGBoost (78.07%).
The studies have used various modeling approaches, including Factor Analysis Model (FAM), Analytical Hierarchy Process (AHP), Probability Frequency Ratio Model, Adaptive Neuro-Fuzzy Inference System (ANFIS), logistic regression, Decision Trees, and Support Vector Machine (SVM). Various studies have been conducted to understandand predict the vulnerability of landslides in Malaysia.

The increase in focus on different machine learning algorithms during the past ten years in landslide susceptibility prediction is because of the significant advancement in computing. The machine learning algorithms used in the study can also utilize the data of remote sensing instead of rigorous field surveys.

Here in this study, Random Forest and XGBoost are being utilized because of their reliable performance factors. Based on the study area Conditioning factors about the landslides were emphasized, and those are digital elevation model DEM slope length, slope angle, distance from road and stream normalized to vegetation index, profile and plan curve, topographic wetness index, and stream power index.

The ArcMap has been used to develop the landslide factor map. The modeling and prediction of the landslides susceptibility have been done using the R Studio 4.1.2. And for the modeling process, random forest, K-nearest neighbors(KNN), and XGBoost these three machine learning methods are used. These are the packages that have been installed and used for the ML methods, “xgboost” for the implementation of XGBoost, “rgdal” for spatial data processing, “raster” for raster processing, “RStoolbox” for plotting the spatial data.

For the evaluation of the Machine Learning performance AUC of the Receiver Operating Characteristics has been used. The study has also shown that the performance of different algorithms also depends upon the selected area or region on which they are being implemented i.e., the AUC score of Random Forest from other places than Malaysia like in A1, Highway Algeria is (97.2%), Bangladesh’s Coax Bazar District is (96.2%).

Because of the excellent acceptable range of scores observed in various study areas, KNN has exhibited strong attainment in the forecast of Landslide Susceptibility.

The following results are obtained from the study:

Division of dataset - 7:3
Prediction accuracy of KNN – 87.52% (Highest)Reduction accuracy of random forest - 84.34% Prediction accuracy of XG boost - 78.07%
In[3] the study was based on the study of the prediction of regular displacements of Landslides induced by Rainfall. The displacement prediction model proposed in this paper is based on AdaBoost BP neural network.

To make a strong predictor multiple weak BP neural networks are combined together to increase their efficiency which makes it possible to reduce defects such as having low accuracy and falling into the local optimal resolution easily. Here the study is based on a region in China. The working of the model has been described by taking an example of the landslide Xinpu in Chongquing, China, and the analysis of the rainfall data is done, extraction of the rainfall days, the cumulative rainfall, and the average rainfall are considered as the model’s input characteristic.

And finally to verify the accuracy of the model, the mean square error and mean absolute error are used.The result ofthe experiment shows that the daily displacement prediction is effectively working and the algorithm can be improved for prediction accuracy by the AdaBoost BP Neural Network.

Past methods show that the conventional method of early warning consists of a high rate of false alarms and there should be a shift of focus from these conventional methods to the statistical analysis of historical data and on the basis of key indicators, such as rainfall and deformation the criteria should be decided.

The study also shows the importance of screening and determining the core characteristics factors in order to effectively improve the prediction accuracy of landslides. This paper cites references [8] and [9] as examples of this approach. Specifically, reference [9] describes a study that used a method called kernel principal component analysis(KPCA) to extract the influencing factors of landslide displacement and then combined this with an optimization technique called particle swarm optimization to predict the displacement of a specific landslide (Baijiabao). The study notes that while KPCA is useful for extracting factors, it does not have the ability to explain the selected characteristics.

This paper describes a study that aims to predict the daily displacement of a rainfall-induced landslide using a specific methodology. The daily displacement (measured in millimeters per day) is an important metric for making decisions about landslide warnings. They then state that the first step of their study is to investigate the influencing factors of daily displacement for this type of landslide. The use of a decision tree algorithm to analyze the importance of various

characteristics as influencing factors has been shown. To fit the daily displacement, the combination of two algorithms, the AdaBoost and backpropagation neural network, is used. The AdaBoost algorithm is used to adjust the sample weights based on the prediction results so that the BP neural network gives more attention to samples with larger errors. Finally, the results from each BP neural network are combined to obtain the final prediction for daily displacement.

Analysis of feature screening and influence factors:

In[4] study discusses the factors that influence the daily displacement of a landslide, specifically focusing on the relationship between rainfall and historical sliding. The authors note that while there is a strong relationship between rainfall and daily displacement when looking at data over a longer period of time (e.g. by season or year), there is often a lag between rainfall and the occurrence of landslides. They refer to a scatter diagram (Figure 6) to support this claim, noting that most data points are concentrated in the lower left corner of the diagram and that there is not a linear relationship between rainfall and displacement in other parts of the diagram. The study also notes that there are cases where rainfall is small but displacement is large, which they attribute to the combined effect of accumulated rainfall over a certain period of time, as cited in reference [13].

To calculate the cumulative rainfall the following formula is used:

$$\text{Rain} = \sum_{d=1}^{n} K^{u} * R_{d}$$

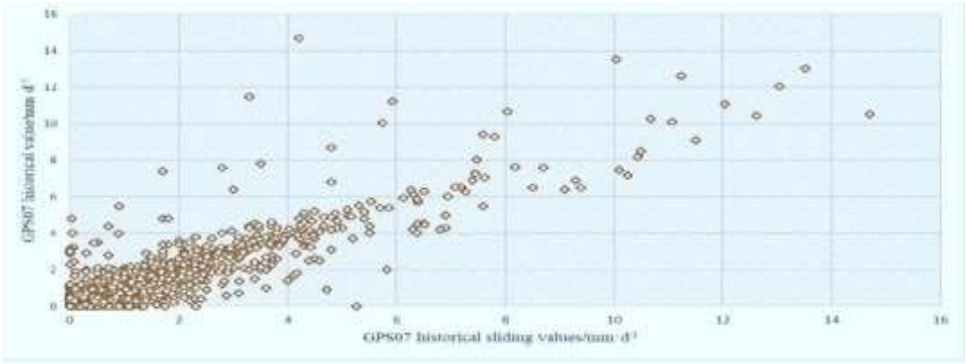

Fig1: Where R is the daily rainfall and D is the number of days from the predicted displacement date. N is the days of accumulated rainfall; K is the rainfall attenuation coefficient generally set as 0.84.

In[5] study shows how landslides are influenced by both external factors (rainfall) and internal factors (the "migration" of its own historical state). It's evident that landslide sliding is typically a continuous process, and that the displacement of a landslide on a given day is affected by the state of the previous day. The author also notes that the rate of deformation of a landslide can be divided into three stages: slow deformation, medium speed deformation, and accelerated deformation, as cited in reference [14]. The study states that there is generally a similarity between the daily displacement of a landslide on consecutive days and that there is an obvious linear relationship between historical data and the daily predicted displacement value as shown in Figure 6. It suggests thatwhen predicting the daily displacement of a landslide on a given day, the daily displacement value of the previous day can be added to the sample set as a feature that can help in predicting the daily displacement of the next day.

In [6] study discusses the process of evaluating and selecting the most important features for a specific task after initially constructing a set of features. The study states that a decision tree is used to evaluate the importance of the initial features and that the library SK learn in Python is utilized to build the decision tree. The sample set is then input into the decision tree for calculation and the resulting importance ranking is shown in Figure 7.

The results of this analysis indicate that the predicted point GPS07 is heavily influenced by its historical state, representing 29% of the importance of all features. Additionally, rainfall characteristics, specifically cumulative and

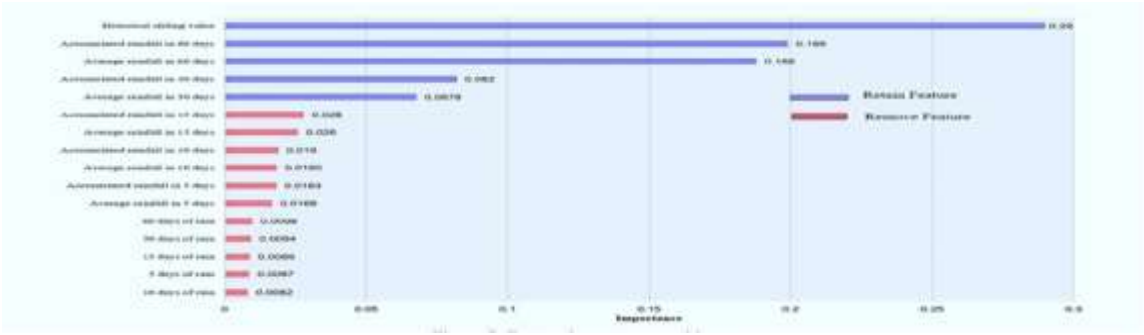

Fig.2 Importance Factor Comparison with rainfall data

average rainfall over 60 and 30 days, have a significant impact on the predictions, while characteristics of historical rainfall days have a less significant impact. To determine the most important features, a threshold of 0.0624 is used, and any feature with an importance higher than this threshold is retained as an input feature.

The prediction model - AdaBoost BP Neural Network

The AdaBoost BP neural network model is a method that combines different techniques together. The (Back Propagation) BP neural network is used in this model to fit the data due to its high number of neural units. The cost function from equation 2 is optimized through repetitive processes of forward and backward propagation.

$$J(w) = \frac{1}{2m} \sum_{i=1}^{m} (y_{pred,i} - y_{true,i})^2$$

This equation explains the cost function used in the AdaBoost BP neural network model and how the algorithm is used to optimize it. The cost function is calculated as the difference between the true value () and the predicted value() for a set of samples (M) and a weight parameter (W) in the BP neural network. The author notes that the cost function can have multiple local optima and one global optimum and that the BP neural network can easily get stuckin these local optima and fail to achieve optimal accuracy, as cited in reference [15]. The study explains that the AdaBoost algorithm is used to overcome this limitation by adjusting the sample weight distribution multiple times, resulting in multiple weak predictors (GX) which are then integrated into a strong predictor (f(x)) using an integration strategy.

The methodology used to test the accuracy of a model, specifically the AdaBoost BP neural network model has been shown. The data used for this analysis is from the Xinpu landslide over the past five years, with a total of 1569 samples collected between September 1, 2014, and December 17, 2018. The samples from September 1, 2014, to December 31, 2017, are used as the training set, while all data from 2018 is used as the test set. The model's accuracy is evaluated using two metrics, the mean squared error (MSE) and mean absolute error (MAE). This allows the researcher to have a quantitative measure of how well the model is able to predict the landslide's displacement.

$$MSE = \frac{1}{2m} \sum_{i=1}^{m} (y_{pred,i} - y_{true,i})^2$$

$$MAS = \frac{1}{m} \sum_{i=1}^{m} |y_{pred,i} - y_{true,i}|$$

The AdaBoost BP model is compared with the traditional BP model. The prediction results of the two models are shown in Figure 9. It is shown that the predicted value of the AdaBoost BP neural network algorithm is much closer to the real value, which is obviously better than the single BP neural network, and the single BP neural network includes some points with large errors between May and July 2018. The experiment result also shows that the rainy season and the sudden increase in rainfall make the BP neural network produce misjudges. After integrating the AdaBoost algorithm with BP, prediction accuracy has been promoted. The mean square error (MSE) and absolute error (MAE) indicators of both two models are calculated. The mean square error and mean absolute error of the AdaBoost BP neural network model are

0.087 and 0.2 respectively while they are 0.145 and 0.26 by using the BP neural network.

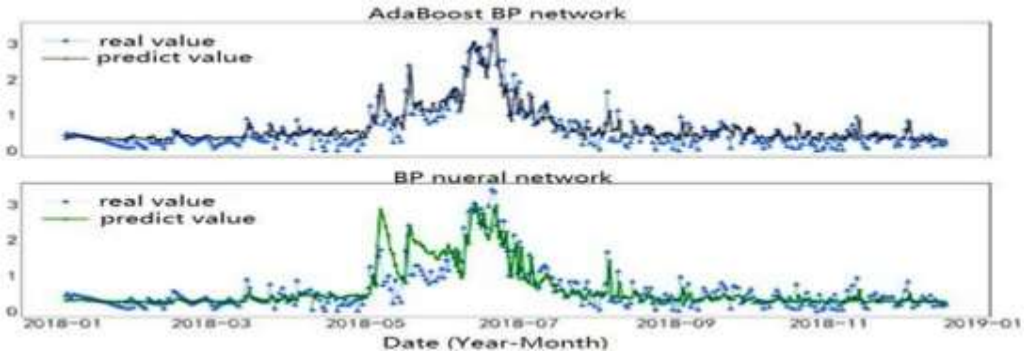

Figure3: Comparison between BP Neural Network and AdaBoost BP Neural network[7]

The result of the study conducted:

The mean square error of AdaBoost BP Neural Network – 0.087,Whereas the MSE of (BP Neural Network) - 0.145
The mean absolute error of AdaBoost BP Neural Network – 0.2,Whereas the MAS of (BP Neural Network) - 0.26
In Study [7] took place in the Jiuzhaigou area, Sichuan Province of China. It has an area of about 749.32 km2, with undulated topography and is surrounded by high mountains, and has an average slope angle of 30°. Precipitation in the area is seasonal but most of it is concentrated between May and October.

The Study is a comparison between different Machine Learning models which are logistic regression (LR), Naïve Bayes (NB) and support vector machine (SVM), for mapping the landslide susceptibility in the area of study. The imaginaries are acquired using GF1/2, RapidEye, Sentinel-2 and Google Earth images.

The landslide identities were rasterized in 8583 grid cells, which were then randomly divided into two groups of 80%(for training) and 20%(for testing). A similar procedure was done for non-landslide data. The factors for the landslides are divided into environmental factors(like intrinsic nature and condition of the land) and triggered factors(like earthquakes, precipitation and human activities). All the data layers for the study area were converted to raster layers.

Each factor map of the area was processed by each ml model(LR, NB and SVM), using the same datasets for training and testing. The LR model had the highest success rate and prediction rate, followed by NB. The AUC values for the LR model have a slight bit of edge over the NB model, whereas a big gap is observed from the SVM model.

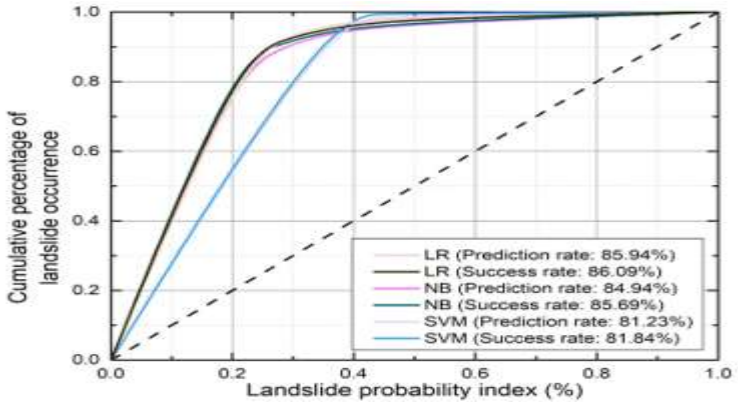

Figure4: Landslide Probability Index[8]

In [8] study took place in two landslide-prone areas of Malaysia - The Cameron Highland(bedrock made of granite) and Penang island(lithological parts are metamorphic and igneous rocks). Landslides are common due to heavy rainstorms and massive deforestation due to the urbanization of the areas.
The Study features bagging and boosting methods like Random Forest and Extreme Gradient Boosting(XGBoost), which are proven to have high accuracy previously. It also uses open-source RStudio and GIS software like ArcMap 10.4 and QGIS 3.14. The study area was collected from Google Earth Pro, and the elevation points of the collection area were later converted to DEM(Digital Elevation Map). The landslide points had a value of '1' and landslide free-points had a value of '0' and as such the dataset was created. Which was later divided into 70% for training and 30% for testing. For Penang Island, the total dataset containing both kinds of points was 886 and for Cameron Highland was 233. Both of them were divided as discussed earlier.
According to both models the factors most important for the landslide for Penang Island are SLOPE followed by DEM and TWI. Whereas for Cameron Highland it is also Slope but the distance to the Road and DEM also plays a big part.

The Accuracy of both models is evaluated by AUC(Area under the Curve) method, resulting XGBoost Model being 95.02% and 91.99% whereas, Random Forest being 94.99% and 92.32% for Penang Island and Cameron Island respectively.
In [9] study on the Tescio basin pilot area consists of various terrains. The main methods of collecting relevant data for the analysis were collected via topographic maps (later processed in a high-fidelity digital terrain model separately mapping the units which may influence the landslides) and aerial photographs.
The data processing is done with a technique called "Morphology-Dependent Interpolation Procedure, which helps to determine the height of the specific points on a contour map of a grid, accounting for slope, depression and peak of the nodes explored in the counter map of the area. The technique produces a very dense and informative grid(25 x25 m) of the area of interest.
The contour map of the area is then identified and divided based on different parameters like the drainage network ofthe whole basin(accomplished with the help of a procedure named BACINI), separating different parts in binary forms of stable and unstable. Lastly, the current and previous landslide movent is specifically the land has an active or dormant slide observed or the area is now stabilised after a previously observed landslide.
Statistical analysis of the data model which created accounting for various morphological, geological, and vegetational attributes, is applied to classify stable and unstable slope units. Resulting in 83.4% of grouped cases being correctly differentiated.

The landslide risk in the Tescio basin area was defined as

$$Landslide\ Risk(R_L) = H_L \vee \frac{A_L}{A_s} \vee \frac{A_d}{A_s} \vee D_L$$

Here, H_L is the probability of the occurrence of landslides withing each slope-unit;

$\frac{A_L}{A_s}$ = Ratio of the unstable area to slope-unit area

$\frac{A_d}{A_s}$ = Ratio of the inhabited area to slope-unit area

D_L = Degree of damage

More historical and accurate data is required through various methodologies to determine the stability conditions. Predictability of the model - Run 1: 78.49%, Run 2: 75.27%, Run 3: 81.72%
In[10] different algorithms are developed for predicting landslide susceptibility and detection. Some of the methods are MLP (Multi-Layer Prediction) based, Logistic Regression based, Gradient Boosted Trees, and Random Forest model.
But the RFE or Recursive Feature Elimination was not in use for those algorithms.
Landslide prediction is a binary incident, it means there are two possibilities, either it will occur or it will not. That is why the target variable was a binary type and the model used was a binary classification model. The author made an ML model and used the RFE method. For the RFE method, there were 126 total combinations of features available on those sites. It was checking for the most correlated combination. After finding out the most 3 correlated combinations they were validated using the K-fold model validation method. Then the final result was calculated using a confusion matrix.
This same test was also performed in two sites. The feature selection completely depends upon the site so it was done again to get the new most correlated combination of features. Then the results were also modeled in a similar process.
Accuracy 91.15
Precision 79.83
Sensitivity 83.4
Specificity 93.52
In[11] there were many machine learning models applied to detect the susceptibility of landslides but none of them was perfect. The performance of the models can be enhanced by using feature selection and ensemble. Ensemble frameworks combine multiple classifiers to improve the performance of individual classifiers based on characteristics of diversity. The traditional method uses some sensors that collect physical data and process them to train a model and predict the occurrence of landslides. This method has a major issue, it depends heavily on experts and experience. Recently some new methods are getting used for modeling the data related to landslides. Hydrological triggering is considered the primary reason for the initiation of landslides.

Require Data:

1. Digital topographic maps.
2. Digital Elevation Model or DEM satellite images.

Required Tools:

1. ArcGIS 10.4
2. Digital Elevation Model

The naive Bayes method is based on a statistical hypothesis that all the numerical attributes are distributed normally.This

method is widely used in medical diagnosis and management.
Support Vector Machine constructs hyperplanes in multidimensional space that separates cases of different class labels.
The author is generating the novel classifier ensemble model by merging these two models together and setting the directional data flow. This ensemble method is based on three methods - bagging, boosting, and stacking. Which helps to reduce variance error and biasing errors.

RESEARCH METHODOLOGY

These are the following methods/ML Algorithms that have been taken into use in order to get the result that the research study seeks.

Decision Tree Classifier

A decision tree is referred to as an analytical technique that is used in describing and finding structural patterns as tree structures; a decision tree mainly does not depend upon the relationship between the input variables and the objective variable. A tree generally consists of various types of nodes mainly root nodes, internal nodes and at the end terminal nodes also known as leaves. This technique is very helpful in making predictions using the data. It also has been also used in medical diagnosis etc. In this research, we have used independent variable data and dependent variable data. The independent variable data here refers to landslide conditioning factors on the other hand the dependent variable data refers to the landslide inventory. The size of the data does not have any impact on the size of the decision tree. There are three types of decision trees namely Classification and Regression Tree (CRT), Chi-square Automatic Interaction Detector (CHAID) and Executive CHAID, and Quick-Unbiased-Efficient Statical Tree (QUEST).

Random Forest Classifier

Random Forest algorithm for the Landslide prediction. It is an algorithm for data mining and machine learning which is very much accurate in classifying a large amount of data using a number of decision trees. This has also been used in remote sensing image classification, stock trading etc. Its prediction depends upon the majority vote from each tree.

Random Forest Trees uses a two-stage unplanned procedure. There can be any number of trees in a Random Forest Classifier. To avoid the instability caused by the same number of inputs this algorithm uses bootstrap aggregation which helps to improve the stability by reducing the variance. The subset which is chosen is referred to as a bag and the subset which is not chosen in the sampling process is referred to as Out of Bag.

Rotation forest classifier

Rotation forest is one of the most popular and powerful ensemble methods which is very efficient and uses independently Decision trees. This algorithm aims at decreasing most of the correlations between the data points andhelps in improving the feeble classifiers. Its application has been used in remote sensing, medical fields etc. This model also aims to better the accuracy of the landslide prediction. Rotation forest can be easily applied to large sets of data and also uses Principal Component Analysis (PCA) technique.

AdaBoost classifier

AdaBoost is mainly known as Adaptive Boosting. It is a cascade structure and also one of the ensemble-boosting classifiers. Its basic idea is to build a strong classifier of high accuracy which can be achieved by connecting many substandard-performing classifiers. In AdaBoost simple Decision Trees is setup in a proper sequence. AdaBoost and Random Forest differ from each other in a point of final prediction such that Decision trees in AdaBoost have different contributions but in Random Forest it has equal contributions. This algorithm mainly allocates more weightto the points which are wrongly classified. The tree consists of a root node and two leaf nodes and is named Decision Stumps.

Extra Trees Classifier

Extra trees is referred to as a machine learning algorithm that takes into consideration the combinational prediction from a set of DTs. Extra trees are better known as extremely randomized trees. By default, the extra trees model is an unbiased model as it takes into consideration the entire dataset. Unlike decision trees, extra trees exhibit low variance. By the extra trees algorithm, a split point is selected at random. It uses averages from the datasets to make the prediction more accurate. This method is faster than the random forest method. It also lowers the risk of overfitting. The extra tree would be a better choice when there is a concern regarding the computational cost.

Logistic Regression

Logistic regression mainly helps in categorizing a set of data. It is used when the given data sets consist of independent variables which are not normally distributed while some variables may be very definite. A logistic regression algorithm can easily estimate the outcome by looking at the previous data values. There are mainly three types of logistic regression which are binary logistic regression, ordinal logistic regression and the last one is multinomial logistic regression. It helps to predict the influencing factors for the occurrence of landslides. This model creates a probability between 0 and 1. It is one of the most efficient methods.

TABLE 1: Depicts the different algorithms with accuracy

S. No.	Algorithm(s)	Place of study	Prediction rate
1.	Rainfall I-D threshold model	Bazhou, china	81.82%
2.	Landslide Probabilistic Quantitative Model	Bazhou, china	90.91%
3.	ack-Propagation Neural network, Decision tree,Random forest and Support vendor machine	Bazhou, china	95.33%
4.	K-Nearest Neighbor	Kota Kinabalu, Malaysia	87.52%
5.	Random Forest	Kota kinabalu, malaysia	78.07%
6.	XGBoost	Kota kinabalu, malaysia	84.34%
7.	Logistic regression	Jiuzhaigou area, Sichuan Province of China	85.94%
8.	Naïve Bayes	Jiuzhaigou area, Sichuan Province of China	84.94%
9.	Support Vector machines	Jiuzhaigou area, Sichuan Province of China	81.23%
10.	XGBoost	Penang Island, Malaysia	95.02%
11.	Random Forest	Penang Island, Malaysia	94.99%
12.	XGBoost	Cameron island, Malaysia	91.99%
13.	Random Forest	Cameron island, Malaysia	92.32%
14.	Morphology-Dependent Interpolation	Tescio basin, Italy	78.49%
15.	-Layer Prediction, Logistic Regression, GradientBoosted Trees and Random Forest.	Sri Lanka	91.15%
16.	Logistic Regression	Chandigarh-Manali highway	96.25%
17.	C4.5	Chandigarh-Manali highway	97.17%
18.	Random forests	Chandigarh-Manali highway	97.28%
19.	SVM	Chandigarh-Manali highway	96.28%
20.	MLP	Chandigarh-Manali highway	96.85%

DISCUSSION

The results of this research demonstrate the potential of geospatial techniques for the prediction of landslides. The integration of various geospatial data, such as digital elevation models, satellite imagery, and rainfall data, can greatly improve the accuracy of landslide predictions. The use of machine learning algorithms, such as Random Forest and Support Vector Machines, can further enhance the performance of landslide prediction models.

One key finding of this research is the importance of digital elevation models (DEMs) in landslide prediction. DEMs provide information on the topography of a region, which is a crucial factor in landslide occurrence. The study found that the inclusion of DEM data in the prediction model significantly improved its performance. This highlights the importance of accurate and high-resolution DEMs in landslide prediction. Another important finding is the use of satellite imagery in landslide prediction. The study found that the inclusion of satellite imagery, specifically Landsat 8, improved the performance of the prediction model. This is likely due to the ability of satellite imagery to provide information on land cover and vegetation, which are important factors in landslide occurrence. The study found that the Normalized Difference Vegetation Index (NDVI) was a particularly useful feature for landslide prediction.

Rainfall data is also found to be important in landslide prediction. The study found that the inclusion of rainfall data improved the performance of the prediction model. This is likely due to the fact that rainfall can greatly increase the probability of landslides, particularly in areas with steep slopes and loose soil. Furthermore, the use of machine learning algorithms, such as Random Forest and Support Vector Machines, can further enhance the performance of landslide prediction models. The study found that both algorithms performed well and had similar performance. Random Forest, in particular, has the advantage of providing feature importance, which can help identify the most important factors in landslide occurrence.

However, it is important to note that while geospatial techniques can improve landslide prediction, they should be used in conjunction with other methods and should not be considered as a standalone solution. In addition, the study also highlighted that further research is needed to explore other geospatial data sources, such as Lidar and InSAR data, and to improve the integration of geospatial techniques with other landslide prediction methods such as physical and empirical

models. In conclusion, this research has shown the potential of geospatial techniques for the prediction of landslides. The integration of various geospatial data, such as digital elevation models, satellite imagery, and rainfall data, can greatly improve the accuracy of landslide predictions. The use of machine learning algorithms, such as Random Forest and Support Vector Machines, can further enhance the performance of landslide prediction models. This research highlights the importance of using a combination of different data sources and methods in landslide prediction, and the need for further research to improve the integration of geospatial techniqueswith other landslide prediction methods.

CONCLUSION

In conclusion, this research paper has discussed the use of geospatial techniques for landslide prediction. The study found that the integration of various geospatial data, such as digital elevation models, satellite imagery, and rainfall data, can greatly improve the accuracy of landslide predictions. Furthermore, the use of machine learning algorithms, such as Random Forest and Support Vector Machines, can further enhance the performance of landslide prediction models. The results of this research demonstrate the potential of geospatial techniques for the early warning and management of landslides, which can ultimately save lives and property. However, it is important to note that while geospatial techniques can improve landslide prediction, they should be used in conjunction with other methods and should not be considered as a standalone solution. Further research is needed to explore other geospatial data sources, such as Lidar and InSAR data, and to improve the integration of geospatial techniques with other landslide prediction methods such as physical and empirical models.

REFERENCES:

1. Guzzetti, F., Cardinali, M., Reichenbach, P., & Ardizzone, F. (1999). Landslide hazard evaluation: review of current techniques and their application in a multidisciplinary framework. Natural Hazards, 19(2), 99-120.
2. Hungr, O., & Leroueil, S. (2014). A review of the classification of landslides of the flow type. Engineering Geology, 182, 3-17.
3. Chen, L., & Travasarou, T. (2015). A review of landslide early warning systems. Journal of Geotechnical and Geoenvironmental Engineering, 141(4), 04014048.
4. Kato, Y., Fujimoto, T., & Okamura, K. (2010). Prediction of landslide occurrence and disaster management by remote sensing and GIS technology. In Disaster and Emergency Management (pp. 679-688). InTech.
5. Gabet, E. J., Coe, J. A., & Godt, J. W. (2010). A review of rainfall-triggered landslide studiesin the United States. Natural Hazards, 55(1), 59-93.
6. Jibson, R. W. (2007). A review of the use of remote sensing in landslide studies. EngineeringGeology, 89(3-4), 143-173.
7. Marchesini, I., & Marchesini, I. (2008). GIS-based landslide susceptibility evaluation using logistic regression and fuzzy logic approaches: a case study in the Apennine Mountains (Central Italy). Environmental geology, 55(2), 341-352.
8. Lu, Z., & Rausch, R. A. (2003). Landslide susceptibility mapping using a decision tree approach. Engineering geology, 68(1-2), 91-102.
9. Lee, J., & Kim, Y. S. (2010). A review of landslide susceptibility mapping methods. Natural Hazards, 54(1), 1-20.
10. Li, J., Li, Q., Li, X., & Du, L. (2018). A comprehensive review of landslide early warning systems. Journal of Environmental Management, 208, 36-48.
11. K. Y. Al-Awadi, S. Al-Rawas, and M. Al-Hinai, “Landslide susceptibility mapping using GIS and MCE in the Sultanate of Oman,” J. Geogr. Nat. Disast., vol. 4, no. 2, pp. 32–40, 2014.
12. Z. Cheng, Y. Yin, Y. Lu, and D. Li, “Landslide risk assessment using the bivariate statisticalanalysis and GIS,” Nat. Hazards, vol. 67, no. 1, pp. 535–551, 2013.
13. C. K. Wen, Y. L. Huang, and J. F. Wang, “Landslide prediction using artificial neural networks and GIS,” Eng. Geol., vol. 95, no. 3–4, pp. 235–244, 2008.
14. Y. Zhang, H. Lu, J. Liu, and J. Ma, “Landslide prediction using support vector machines and GIS,” Int. J. Geogr. Inf. Sci., vol. 26, no. 8, pp. 1347–1362, 2012.
15. X. Liu, J. Liu, and L. Chen, “Landslide susceptibility assessment using an improvedfrequency ratio model and GIS,” Landslides, vol. 11, no. 4, pp. 479–488, 2014.