Landslide Likelihood Prediction using Machine Learning Algorithms

1st Vasundhara Acharya
Rensselaer Polytechnic Institute
Troy, USA
acharv2@rpi.edu

2ndAnindita Ghosh
Rensselaer Polytechnic Institute
Troy, USA
ghosha7@rpi.edu

3rd Inwon Kang
Rensselaer Polytechnic Institute
Troy, USA
kangi@rpi.edu

4th Thilanka Munasinghe
Rensselaer Polytechnic Institute
Troy, USA
munast@rpi.edu

5thBinita KC
NASA Goddard Earth Sciences Data and Information
Services Center (GES DISC) /Adnet Systems, Inc
NASA Goddard Space Flight Center
Greenbelt, MD, 20771, USA
binita.kc@nasa.gov

***Abstract*—The supply of electricity via power plants is critical to the operation of many critical infrastructure systems in modern society. Natural hazards can disrupt the power supply, cause power outages that can halt economic growth, and impede emergency response until power is restored. The proposed work aims to predict the landslides likelihood in these critical infrastructure locations in the Northeastern USA using integrated databases of explanatory variables and machine learning algorithms. First, data related to landslides are obtained and merged, including topographic, soil moisture, and precipitation-related data. Five regression algorithms, namely: Random Forest, Extreme Gradient Boosting (XGBoost), K-Nearest Neighbor regression (KNN), Linear Support Vector Regressor (SVR), and Linear regression, are utilized to predict the landslide probability and evaluated on the dataset. The accuracy of the models is assessed by using statistical metrics such as mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). The study results show that Random Forest outperformed other models with the mutual information feature selection method. It achieved an MSE of 0.0011 with mutual information-based feature selection and an MSE of 0.00157 without feature selection. KNN regressor outperformed the other models with an MSE of 0.00139 with correlation-based information selection. The proposed landslide identification model with Random Forest algorithm shows outstanding robustness and great potential in tackling the landslide likelihood prediction by employing ML algorithms.**

***Index Terms*—landslide likelihood, critical infrastructure, machine learning, regression, random forest and k-nearest neighbor**

I. Introduction

The growing number of natural disasters due to climate change is of critical concern. Landslides are one of the predominant geologic hazards that result in massive human and economic losses. A study on “Economic losses, poverty & disasters” conducted by the World Health Organization reveals that between the years 1998 and 2017, landslides affected around 4.8 million people and caused 18000 fatalities [1]. The U.S. Geological Survey estimates the impact of landslides as a fatality between 25 to 50 people each year and an economic loss of approximately 1 billion dollars [2].

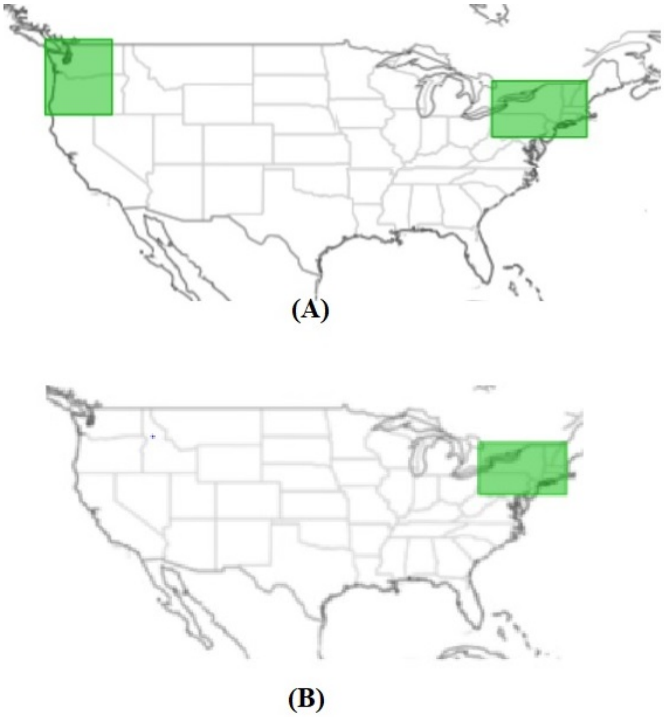
Power plants are one of the critical infrastructures that are vulnerable to landslides. Electricity demand is ubiquitous in our modern world. According to the U.S. Energy Information Administration (EIA), the USA alone consumed 3.8 trillion kWh (kilowatt-hours) of electricity in 2020 [3]. With the rapid advancement of the digital world, the energy demand can only be expected to rise exponentially. To prevent extreme damage to power plants, and improve system resilience against natural hazards, its crucial to identify areas prone to landslide.

Machine learning (ML) algorithms have been recently employed as analysis tools to extract important features that can assist in decision making, perform clustering and prediction tasks. ML has proven to be very helpful in solving challenging tasks in various domains. The domain of landslide prevention has also harnessed the potential of these algorithms to efficiently and precisely solve various problems. Most of the machine learning algorithms are data centric. They require high quality data to generate meaningful predictions. According to a survey [4], [5], data volumes in the landslide prevention domain are increasing exponentially due to the advancement in sensors, the Internet of Things (IoT), and model simulations . The availability of this cumulative data has created various opportunities to apply of ML to solve major problems in this domain.

Possibility of landslide occurrence in a given region is estimated through landslide susceptibility assessment. The proposed work predicted the region prone to landslides based on available data, including conditional factors, by employing ML algorithms. Landslides can occur due to heavy rainfall, earthquake, loss of vegetation and support structures in high elevation, and insufficient soil moisture [6]. Detailed research on factors affecting a landslide is necessary to determine landslide-related predictors.

The main objective of this paper is to propose a landslide likelihood identification method using machine learning algorithms. The training dataset is prepared by stacking various layers composed of landslide explanatory variables. An approach to handle multi-sensor satellite observations having many noise sources, missing data, and outliers is presented. Various machine learning algorithms, namely Random Forest, Extreme Gradient Boosting (XGBoost), K-Nearest Neighbor regression (KNN), Linear Support Vector Regressor (SVR), and Linear regression, are trained and compared to evaluate their performance. The performance of the models was then measured using mean squared error (MSE), Mean absolute error (MAE), and Root Mean squared error (RMSE). The power plants located in regions of the Northeastern USA are used to validate the method. The proposed work is a pilot study, and it currently focuses on the Northeast of the United States of America. It can be further expanded to other parts of the USA or the entire USA. The figure 1 depicts the regions of interest. Figure 2 shows the location of various power plants in the Northeastern region of the USA.

The rest of this paper is organized as follows: Section 2 briefly introduces related works conducted in the domain of landslides identification. Section 3 demonstrates the methodology of the proposed study and the machine learning applications in landslide detection. Section 4 discusses the results of the study. Section 5 concludes the paper and talks about the future work.


(A)

(B)

Fig. 1. Region of Interest.(A). Represents the training set with Northwestern and Northeastern USA as our region of interest. (B). Represents the test set with the Northeastern USA as the region of interest [7]

II. Related Works

Landslides can occur due to heavy rainfall, earthquake, high elevation, and insufficient soil moisture. Various factors that triggered landslides were considered in the model while building the information system. Several works have been done to identify the main predictors that had the potential to induce landslides. Wicki et al. [8] studied the potential of in situ soil moisture data for the regional landslide. The study found that in situ soil moisture data efficiently provides remarkable information that may be used to detect landslides early. Johnston et al. [9] identified the influence of urbanization on precipitation-induced landslide hazards and emphasized the significance of taking urbanization into account when estimating landslide hazards. Abraham et al. [10] established a relationship between landslides, precipitation, and antecedent soil moisture. They also found that less severe rainfalls can trigger landslides when the soil wetness is high. Martino et al. [11] revealed an increase in landslide activity after the low-magnitude earthquake concerning the activities recorded in the same months of the previous years. Nakileza et al. [12] proposed that causal factors such as rainfall, tremors, and land use were external stimuli responsible for the actual initiation of mass movements. At the same time, conditional factors such as geology, weathering, soils, and topography were responsible for inducing slope instability. Hosseini et al. [13] identified that the landslide dimensions increased with increasing slope angle. Most landslides were situated along roads and on faults, and shallow landslides were more frequent along roads than those on faults. Matsuyama et al. [14] identified landslide disasters were likely to occur when SWI in an event exceeded the maximum value observed in the past ten years.

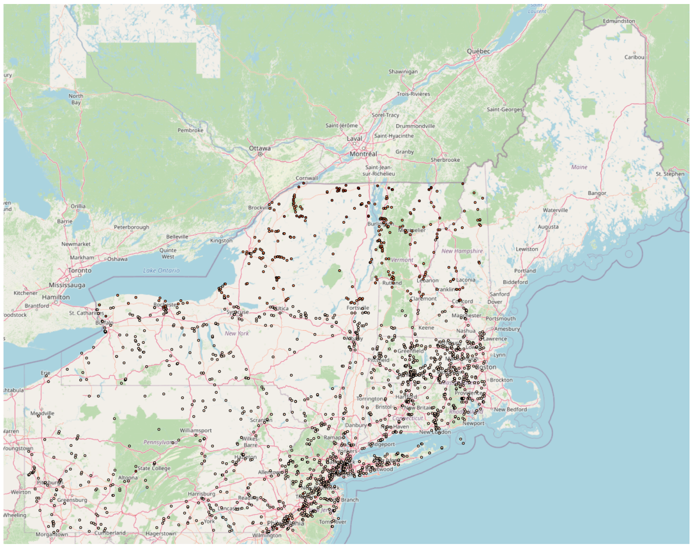


Fig. 2. Locations of power plants in the Northeastern USA

Several works have been done in the past to identify the landslide susceptibility modeling using machine learning techniques. Karianne et al. [15] emphasized the presence of a vast and highly sophisticated geo-dataset that can provide valuable and preventive insights on geo-hazards through machine learning methods. Goetz et al. [16] discussed the performance of several data-driven approaches of which random forest showed the best predictive performance. Chen et al. [17] studied regional landslide susceptibility by implementing various machine learning models like random forests and created landslide susceptibility maps for the perusal of policymakers. Stumpf et al. [18] made use of high-resolution satellite images in combination with object-oriented image analysis to generate features to be used by a random forest ensemble model. However, the accuracy score of around 80% showed that there is still room for improvement. Lei et al. [19] made use of similar imagery along with a newly proposed model, a fully convolutional neural network with pyramid pooling (FCN-PP), to extract features and detected landslide locations in a post-disaster image taken from an airplane. While this approach showed an impressive accuracy score of up to 95%, it cannot be used directly in an early warning system to prevent large-scale disasters as it looks at detecting the landslide regions after the disaster has already taken place. Wei et al. [20] made use of precipitation data in combination with groundwater level and its fluctuation to build a model to predict landslides in a given region. Using an SVM-based model, they achieved an RMSE score of 0.144, suggesting that precipitation data is highly correlated with landslides and can be used as a good predictor.

Despite several efforts to develop landslide detection systems by applying ML techniques, the problem still poses many challenges to the machine learning domain [21]. Class imbalance is one of the main hindrances for the ML models. A class imbalance develops when observation in one class is higher than observation in other classes. Fewer instances in the training set have medium/high landslide probability. Such imbalances lead the model to consider susceptible areas as safe areas. Hence, this problem needs to be mitigated to make the model efficient. The noise in the dataset is another primary concern. Multi-sensor satellite observations have many noise sources, missing data, and outliers. Such complex datasets pose problems to the ML models as they heavily depend on the input dataset.

III. Methodology

In the proposed work, the machine learning algorithms are employed to achieve the landslide likelihood prediction. The flow is depicted in the figure 3.

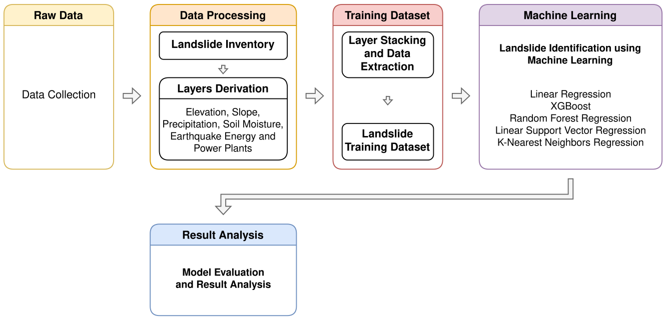

Raw Data
Data Collection
Data Processing
Landslide Inventory
Layers Derivation
Elevation, Slope, Precipitation, Soil Moisture, Earthquake Energy and Power Plants
Training Dataset
Layer Stacking and Data Extraction
Landslide Training Dataset
Machine Learning
Landslide Identification using Machine Learning
Linear Regression
XGBoost
Random Forest Regression
Linear Support Vector Regression
K-Nearest Neighbors Regression
Result Analysis
Model Evaluation and Result Analysis

Fig. 3. Proposed model to predict landslide probability

A. Explanatory variables derived from the data sources

Several explanatory variables were derived from globally available datasets. These variables are soil moisture, earthquake energy, precipitation, elevation, and slope. The sources of these variables are tabulated in Table I. Slope is the measurement of surface steepness. The formation, development, and vulnerability of landslides are all significantly influenced by slope. This was obtained from the Google Earth Pro. The primary rationale behind utilizing earthquake energy is that earthquakes trigger landslides. The amount of energy in the seismic waves generated during the earthquakes determines its potential to create landslides. The variable was derived from the earthquake hazards program dataset. Soil moisture conditions play a vital role in the initiation of landslides. Also, multiple studies have obtained a strong correlation between satellite soil moisture and landslide events. The soil moisture data is extracted from the NASA-USDA soil moisture dataset in this study. Precipitation is the fundamental triggering variable in this work. It affects slope stability as it percolates through soil and rocks and weakens the slope. This variable is obtained from Goddard Earth Sciences Data and Information Services Center.

The data layers obtained from each of the explanatory variables are stacked over one another to form the final dataset. The stacking of data layers is shown in the figure 5.

TABLE I
Explanatory variables that were chosen based on their relevance and contribution to the predictive strength of the model.

Derived variable	Dataset	Reference
Elevation (in feet) and Slope Data (in degrees)	Google Earth Pro, GPS Visualizer	[22]
Precipitation (in mm/hr)	Goddard Earth Sciences Data and Information Services Center (GES DISC)	[7]
Soil Moisture (in mm/hr)	NASA-USDA Global Soil Moisture Data	[23]
EarthQuake Energy (in joules)	The USGS Earthquake Hazards Program	[24]
Landslide Probability	Global_Landslide_Nowcast	[25]

B. Landslide data

NASA's earth observatory provided the landslide probability value, which is the target variable. They were extracted from the global landslide susceptibility map [25]. The values of these probabilities ranged from 0 to 1.

An example of sample data obtained from the training set is shown in the figure 4. It represents a location in the Washington state. The elevation is in feet, the soil moisture and precipitation are measured in mm/hr, earthquake energy is in joules, and the run and slope are measured in degrees. The landslide probability computed in the region is 0.033.

C. Data Preparation

The next stage is to form the training set. Data extraction, layer stacking, and training set preparation are the different sub stages involved. Only the necessary variables are obtained from each dataset during the data extraction. For instance, the latitude, longitude, date, and precipitation (in mm/hr) are

Fig. 4. Sample taken from the training set. The latitude of the location is 48.783393 and the longitude is -120.983333.

retained from the precipitation dataset. Each of the data layers obtained is later stacked as shown in Figure 5. Every layer in the figure denotes a predictor, which is combined to form an n-dimensional training dataset. The layers are stacked upon one another based on latitude and longitude values. For training, the data from Northwestern USA and Northeastern USA are utilized. The landslide probabilities (target variable) extracted from the dataset are later merged with this dataset.

The feature vector computed for the figure 4 is listed in the table II.

1) Handling Class Imbalance and Data Pre-processing: Unbalanced datasets harm the performance of the regression algorithms. Imbalance issues are challenging to handle in the case of regression problems as the target values are continuous and can have an infinite number of values. In the proposed study, the target value is landslide probability, a continuous variable.To better understand the under-sampled instances, the class category was generated by distributing the probability values into three different bins (after train-test split): Low, Medium, and High. These bins were generated using threshold values obtained from the domain expert from NASA for our use case. This column was only utilized for knowing about instance distribution and was dropped during training and testing.

We have utilized the SMOGN algorithm to solve this problem in this study. SMOGN [26] combines the power of random undersampling and two oversampling techniques, namely SMOTER and the introduction of Gaussian Noise. It ensures that the samples generated are highly diverse. The authors identified that the minority classes were the probabilities that generated high and medium landslide risks during the study. Before, the training set had 3401550 instances. After applying the SMOGN algorithm, each class (generated using pandas cut) had 3388518 instances (total of 10,165,554). The outliers are the unusual values that can distort the statistical results. The authors conducted a small experiment to determine if the outliers elimination for the proposed study is a legitimate step. The correlation matrix of the dataset was plotted to see the correlation between the predictors and the target variable. Unfortunately, due to the presence of the outliers, the precipitation and soil moisture was negatively correlated to landslide probability which is not the case in reality. Hence, the outliers in the dataset were eliminated using two different techniques, namely: Winsorization [27], and the boxplot approach [28]. The rows with null values (missing values) were also eliminated.

D. Data Split

The dataset was split into train and test sets in a 70-30 ratio. The feature values were scaled using the standard scaling method [29]. The scaling was performed after the dataset was split to prevent the data snooping [30]. The training set had 10,165,554 instances, and the test set had 1,457,808 instances which focused mainly in the Northeastern region of the USA. The power plant details (latitude, longitude and name) were integrated with the test set to facilitate the prediction of landslides in respective power plant locations.

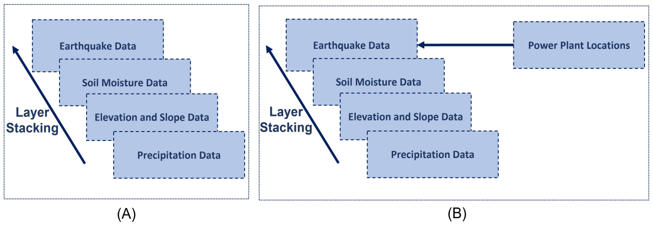


Fig. 5. Data layers stacked. (A). Layers in Training Set. (B). Layers in Test Set

E. Machine Learning Models

Having established the training dataset, the next step is to train the ML models and conduct landslide prediction. In this study, landslide identification is a regression problem; Five ML algorithms are chosen to evaluate the feasibility of machine learning in landslide risk prediction.

1) Linear Regression: The linear regression is utilized to understand the linear relationship between the dependant variable (landslide probability) and the independent variable (various predictors). In the study, the multiple linear regression is employed as a single predictor is not enough to explain the landslide probability. The linear function is shown in equation 1.

$$\mathrm{Y} = \mathrm{a} + \mathrm{b_1} \times \mathrm{X_1} + \mathrm{b_2} \times \mathrm{X_2} + \ldots + \mathrm{b_n} \times \mathrm{X_n} \tag{1}$$

where, Y is the dependent variable (landslide probability in our case) , X_i represents independent variables (all predictors), a is the constant and b_i is the regression coefficient of the variable X_i.

2) Random Forest Regression: Random forest regression (RFR) is a machine learning that ensembles a bag of trees to achieve the overall prediction. The trees are trained by a number of bootstrap samples created from the main training set. Bagging is responsible for the reduction in variance in the ensemble and prevents overfitting. The regression trees are characterized by low bias and high variance. In regression tasks, it outputs the mean prediction of K regression trees. The mean of the predictions is computed using the equation 2.

TABLE II
EXPLANATORY VARIABLES FOR THE SAMPLE INSTANCE SHOWN IN THE FIGURE 4

Precipitation	Energy	Soil_moisture	Latitude	Longitude	Elevation	Run	Slope	Landslide probability
0.396931	16.9862	25.39641	48.783393	-120.983333	5997.200195	4.870978	46.412987	0.033333

$$\text{RFR prediction} = \frac{1}{K}\sum_{k=1}^{K} h_k(x) \tag{2}$$

3) XGBoost Regression: Extreme Gradient Boosting, also known as XGBoost, is another learning algorithm that uses an ensemble of decision trees. Each tree is trained to learn the decision function by aiming to minimize the loss function using gradient descent. Each tree aims to correct the prior one in the learning step. The training of the ensemble of K trees can be characterized as in equation 3, where the first term denotes the training loss that is being minimized, and the second term refers to the regularization parameters, which acts as a limit to the trees complexity:

$$Obj = \sum_{i=1}^{n} loss(y_i, \hat{y}) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

The final prediction of the ensemble is calculated using the equation 4,

$$Y_{\text{xgb}}(x) = \sum_{k=1}^{K} tree_k(x),\ tree_k \in T \tag{4}$$

4) Linear SVR: Support vector regression (SVR) is a machine learning algorithm employed to predict a quantity. It operates on either discrete-valued or real-valued inputs. In the proposed work, multivariate regression is utilized due to multiple input variables. The output of the SVR can be computed using the equation 5.

$$Y_{\text{svr}}(x) = \sum_{i=1}^{n} \beta_i K\left(x; x_i\right) + b \tag{5}$$

where β_i and x_i are respectively the weight and the position of each SVs. In addition, n is the number of SVs, b is the bias, and $K\left(x; x_i\right)$ is the kernel function corresponding to x_i. Due to the large training dataset, a linear kernel implemented in liblinear, which means that K is a linear function, defined as $(x \cdot x_i)$.

5) K-Nearest Neighbors Regression: K-Nearest Neighbors regression (KNN) is a subclass of clustering algorithms that aims to group the samples of similar 'neighborhoods' based on their feature values to find a correlation between the features and the label value. The distance between each sample is decided using the Euclidean distance of the features.

$$Y_{\text{KNN}} = \frac{\sum_{i=1}^{K} N_i}{K}\ ;\ N = \text{X sorted by Euclidean distance} \tag{6}$$

To achieve prediction, the KNN algorithm will find K points closest to the input value, and outputs the average of their labels as can be seen in equation 6.

F. Feature Selection

Feature selection methods are utilized to eliminate unimportant features. The focus is on the features that contribute the most to the target variable. This step helps in reducing the cost involved in modeling and improves the model performance. In this study, SelectKBest with the Correlation feature selection [31] and Mutual Information [32] were used to extract the best features from the dataset. The SelectKBest function uses these methods as a score function to determine a score and the correlation between each feature and the target feature. The score between each feature and target variable is determined using these two score functions. A lower score means that the feature is independent of the target variables. If the resulting value is lower, the feature is independent of the target feature, while the higher resulting value indicates that the feature is related to the target feature.

G. Performance Evaluation Criteria

In the proposed study, standard statistical measures namely: mean squared error (MSE), root mean squared error (RMSE) and mean absolute error (MAE) are used to evaluate the performance of regression model. Root mean square error (RMSE) denotes the square root of the mean of the square of all errors . It is computed using the equation 7. Mean squared error represents the mean of square of the errors. It is calculated using the equation 8. Mean absolute error denotes the absolute value of differences between true and predicted values. Equation 9 is used to calculate this entity.

$$RMSE = \sqrt{\frac{\sum_{i-1}^{n}(Y_i - Yi)^2}{n}} \tag{7}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{8}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=0}^{n-1}\left(|y_i - \hat{y}_i|\right) \tag{9}$$

Here, y_i denotes the expected value and $\hat{y}_i$ is the predicted value.

IV. RESULTS AND ANALYSIS

This section presents the landslide probability prediction results using various ML models. In the modeling, latitude, longitude, soil moisture ratio, precipitation value, earthquake magnitude, elevation, run, and slope were input variables, whereas landslide probability was the output variable. The hyperparameters of the models were obtained via hyperparameter optimization.

The GridSearchCV [33] with three folds (chosen after experimental analysis) was employed for this task. The optimized hyperparameters are tabulated in Table III. The hyperparameters of linear regression are not present in the table as there are no such hyperparameters that can be tuned. As a first step, the influence of feature selection on the performance of the ML models is presented, followed by comparing the performance of models with and without the feature selection method.

A. *Outlier Elimination*

Figure 6 shows the box plot of explanatory variables before eliminating outliers. Figure 7 shows the box plot after eliminating outliers using the Winsorization and the boxplot approach. The correlation matrix was constructed to understand the dependence between different variables and their relation to landslide probability. Figure 8 shows the correlation matrix. We can see that attributes such as precipitation, energy, elevation, and soil moisture are all positively correlated to landslide probability which supports our initial assertion. Also, the slope is negatively correlated to landslide likelihood. Strong rocks make up land regions with high slope values, and these kinds of rocks are stable. Hence, the probability of a landslide is less. These results were also supported in another study [34]

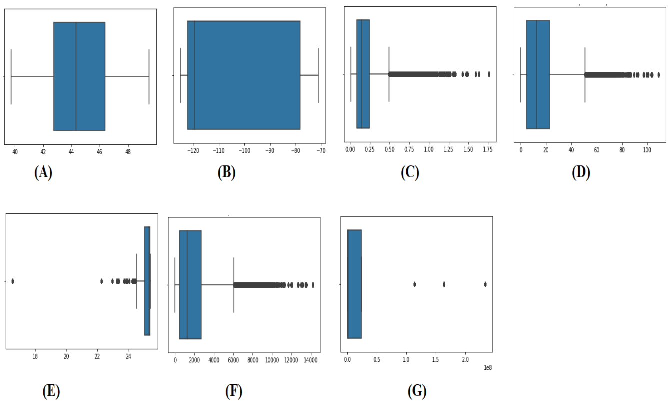
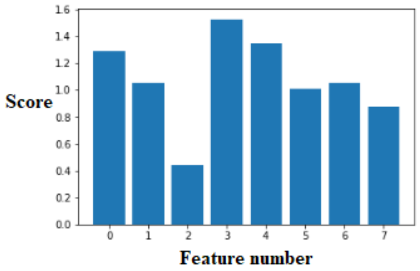


Fig. 6. With Outliers: A. Box plot of latitude. B. Box plot of longitude. C.Box plot of Precipitation. D. Box plot of Slope. E. Box plot of soil moisture. F. Box plot of Elevation. G. Box plot of Energy

B. *Feature Scores obtained using Correlation Feature selection and Mutual Information Feature Selection method*

Figure 4 describes how the features with score values returned by the SelectKBest function with Correlation feature selection. The highest score was assigned to precipitation, and the lowest score was assigned to the slope. Figure 10 describes the score values returned by the SelectKBest function with Mutual Information feature selection. The highest score was assigned to latitude, and the lowest score was assigned to soil moisture.



Feature_Names	Number	Score
Precipitation	Feature 0	1.29437
Energy	Feature 1	1.053706
Soil Moisture	Feature 2	0.438734
Latitude	Feature 3	1.529635
Longitude	Feature 4	1.350683
Elevation	Feature 5	1.010534
Run	Feature 6	1.053442
Slope	Feature 7	0.880499

Fig. 10. Score values of the features obtained using Mutual Information based feature selection method

C. *Performance of the models*

All the models were run by employing Correlation and Mutual information based feature selection methods. In the proposed study, due to a small feature vector, only the feature that received the lowest score was eliminated. The models were also run with all the features (without selecting features) to evaluate its impact in predicting the landslide probability. As a result of the correlation-based feature selection method, slope received the lowest score and was eliminated. After the Mutual Information feature selection, soil moisture received the lowest score, and hence it was eliminated while running the models.

Algorithm 1 represents Linear Regression, 2 denotes Random Forest, 3 indicates XGBoost, 4 stands for KNN Regressor, and 5 denotes Linear SVR. The performance of the models on the test set can be seen in figure 11-13. After analyzing the metrics, it can be stated that Random Forest outperforms the other algorithms with Mutual Information based feature selection and when all the features are present. KNN regression also performs equally well, and it outperforms the other algorithms with correlation-based feature selection method.

V. Conclusion & Future Work

Landslide identification is essential for risk assessment. Prediction of landslide probability in critical infrastructure locations using various ML models might help in hazard monitoring and mitigation. In this study, five popular ML regression models, Random Forest, XGBoost, KNN regressor, Linear SVR, and Linear regression, were applied and compared to predict the landslide probability using explanatory variables. Based on the statistical analysis, a ratio of 70/30 for training and testing datasets was considered as the best ratio for training and testing of models. In addition, the performance of these models was also investigated under the influence

TABLE III
OPTIMIZED HYPERPARAMETERS OF THE MODEL

Algo	Random Forest	XGBoost	KNN Regression	SVR
Parameters	bootstrap: False max_depth: sqrt max_features: sqrt min_samples_leaf: 1 min_samples_split: 2 n_estimators: 10	objective: reg:linear colsample_bytree : 0.3 learning_rate: 0.1 max_depth: 5 alpha: 10 n_estimators:100	Leaf size:1 p:1 N_neigbors:3	C:20 tol:1

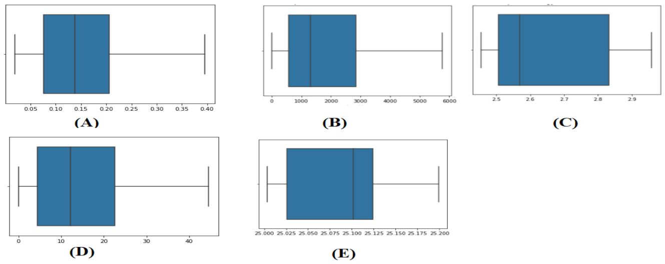
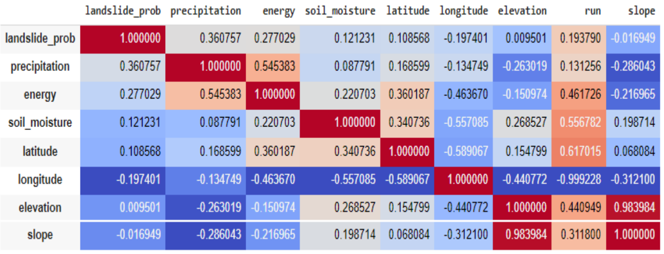
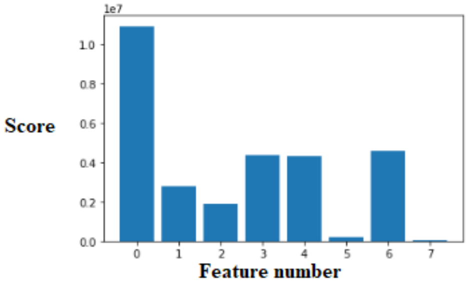


Fig. 7. Without Outliers: A. Box plot of Precipitation. B. Box plot of Elevation. C. Box plot of Energy. D. Box plot of slope. E. Box plot of soil moisture



	landslide_prob	precipitation	energy	soil_moisture	latitude	longitude	elevation	run	slope
landslide_prob	1.000000	0.360757	0.277029	0.121231	0.108568	-0.197401	0.009501	0.193790	-0.016949
precipitation	0.360757	1.000000	0.545383	0.087791	0.168599	-0.134749	-0.263019	0.131256	-0.286043
energy	0.277029	0.545383	1.000000	0.220703	0.360187	-0.463670	-0.150974	0.461726	-0.216965
soil_moisture	0.121231	0.087791	0.220703	1.000000	0.340736	-0.557085	0.268527	0.556782	0.198714
latitude	0.108568	0.168599	0.360187	0.340736	1.000000	-0.589067	0.154799	0.617015	0.068084
longitude	-0.197401	-0.134749	-0.463670	-0.557085	-0.589067	1.000000	-0.440772	-0.999228	-0.312100
elevation	0.009501	-0.263019	-0.150974	0.268527	0.154799	-0.440772	1.000000	0.440949	0.983984
slope	-0.016949	-0.286043	-0.216965	0.198714	0.068084	-0.312100	0.983984	0.311800	1.000000

Fig. 8. Correlation Matrix



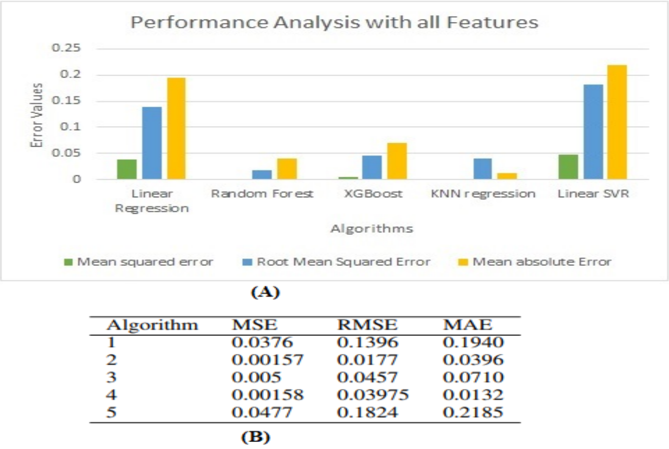
Feature_Names	Number	Score
Precipitation	Feature 0	10934728.78
Energy	Feature 1	2763313.393
Soil Moisture	Feature 2	1874710.191
Latitude	Feature 3	4370009.482
Longitude	Feature 4	4297308.305
Elevation	Feature 5	212471.9967
Run	Feature 6	4570840.708
Slope	Feature 7	65511.44098

Fig. 9. Score values of the features obtained using Correlation based feature selection method

of the SMOGN algorithm with two feature selection methods, including Correlation-based feature selection and Mutual information-based selection method. Results showed that the performance of all models was acceptable as they achieved reasonable MSE scores, and the Random Forest regressor outperformed other models with the mutual information-based feature selection method. KNN regressor was found to be good under the influence of the Correlation-based feature selection method. This study shows the potential of using ML in predicting landslide likelihood. Future work includes-expanding to other parts of the USA, and many other attributes, such as the impact of climate change and vegetation cover, can be added to improve the model's prediction capability. An information dashboard can be developed to display the predictions generated for each critical infrastructure.

VI. ACKNOWLEDGEMENTS

We want to thank Dr. Jennifer Wei, the lead scientist at the NASA Goddard Earth Science (GES) Data and Information Services Center, for her high-level guidance during this project. We also want to thank Thomas Stanley, Aerospace Technologist at NASA Stennis Space Center, for his guidance on probability threshold values related to this project and for pointing us to relevant reference material and research papers.

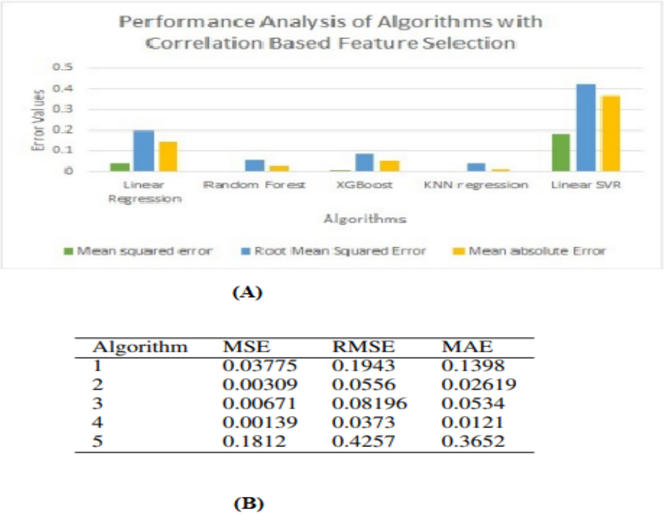


(A)

Algorithm	MSE	RMSE	MAE
1	0.0376	0.1396	0.1940
2	0.00157	0.0177	0.0396
3	0.005	0.0457	0.0710
4	0.00158	0.03975	0.0132
5	0.0477	0.1824	0.2185

(B)

Fig. 11. (A). Graph depicting error values achieved by algorithms using all the features. MSE, RMSE and MAE values

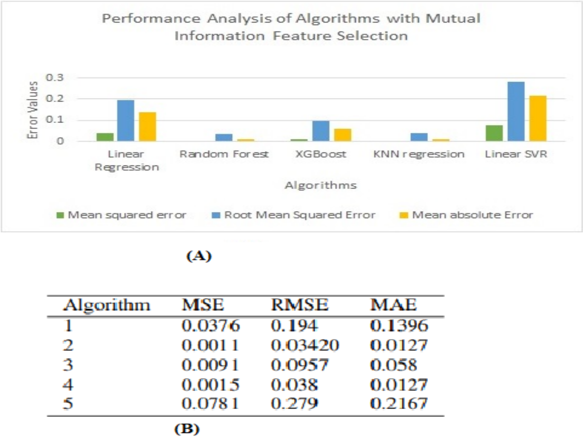


(A)

Algorithm	MSE	RMSE	MAE
1	0.03775	0.1943	0.1398
2	0.00309	0.0556	0.02619
3	0.00671	0.08196	0.0534
4	0.00139	0.0373	0.0121
5	0.1812	0.4257	0.3652

(B)

Fig. 12. (A). Graph depicting error values achieved by algorithms after employing Correlation Based feature selection. MSE, RMSE and MAE values.



(A)

Algorithm	MSE	RMSE	MAE
1	0.0376	0.194	0.1396
2	0.0011	0.03420	0.0127
3	0.0091	0.0957	0.058
4	0.0015	0.038	0.0127
5	0.0781	0.279	0.2167

(B)

Fig. 13. (A). Graph depicting error values achieved by algorithms after employing Mutual Information feature selection. MSE, RMSE and MAE values.

We sincerely thank the School of Science at Rensselaer Polytechnic Institute (RPI) for their support during the project.

We thank Institute for Data, Applications, and Exploration (IDEA) and Tetherless World Constellation (TWC) at RPI for the resources provided during the project.

References

[1] "Landslides." [Online]. Available: https://www.who.int/health-topics/landslides

[2] U. G. Survey, "Landslides 101," 2021. [Online]. Available: https://www.usgs.gov/programs/landslide-hazards/landslides-101

[3] U. E. I. A. (EIA), "Use of electricity," 2022. [Online]. Available: https://www.eia.gov/energyexplained/electricity/use-of-electricity.php

[4] Z. Ma, G. Mei, and F. Piccialli, "Machine learning for landslides prevention: a survey," *Neural Computing and Applications*, vol. 33, no. 17, pp. 10 881–10 907, 2021.

[5] F. S. Tehrani, M. Calvello, Z. Liu, L. Zhang, and S. Lacasse, "Machine learning and landslide studies: recent advances and applications," *Natural Hazards*, pp. 1–49, 2022.

[6] T. A. Stanley, D. B. Kirschbaum, G. Benz, R. A. Emberson, P. M. Amatya, W. Medwedeff, and M. K. Clark, "Data-driven landslide nowcasting at the global scale," *Frontiers in Earth Science*, p. 378, 2021.

[7] NASA, "Global precipitation measurement," 2021. [Online]. Available: https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGM_06/summary

[8] A. Wicki, P. Lehmann, C. Hauck, S. I. Seneviratne, P. Waldner, and M. Stähli, "Assessing the potential of soil moisture measurements for regional landslide early warning," *Landslides*, vol. 17, no. 8, pp. 1881–1896, 2020.

[9] E. C. Johnston, F. V. Davenport, L. Wang, J. K. Caers, S. Muthukrishnan, M. Burke, and N. S. Diffenbaugh, "Quantifying the effect of precipitation on landslide hazard in urbanized and non-urbanized areas," *Geophysical Research Letters*, vol. 48, no. 16, p. e2021GL094038, 2021.

[10] M. T. Abraham, N. Satyam, B. Pradhan, and A. M. Alamri, "Forecasting of landslides using rainfall severity and soil wetness: a probabilistic approach for darjeeling himalayas," *Water*, vol. 12, no. 3, p. 804, 2020.

[11] S. Martino, M. Fiorucci, G. Marmoni, L. Casaburi, B. Antonielli, and P. Mazzanti, "Increase in landslide activity after a low-magnitude earthquake as inferred from dinsar interferometry," *Scientific reports*, vol. 12, no. 1, pp. 1–19, 2022.

[12] B. R. Nakileza and S. Nedala, "Topographic influence on landslides characteristics and implication for risk management in upper manafwa catchment, mt elgon uganda," *Geoenvironmental Disasters*, vol. 7, no. 1, pp. 1–13, 2020.

[13] S. A. Hosseini, R. Lotfi, M. Lotfalian, A. Kavian, and A. Parsakhoo, "The effect of terrain factors on landslide features along forest road," *African journal of Biotechnology*, vol. 10, no. 64, pp. 14 108–14 115, 2011.

[14] H. Matsuyama, H. Saito, and V. Zemtsov, "Application of soil water index to landslide prediction in snowy regions: sensitivity analysis in japan and preliminary results from tomsk, russia," *Progress in Earth and Planetary Science*, vol. 8, no. 1, pp. 1–13, 2021.

[15] M. V. D. H. KARIANNE J. BERGEN, PAUL A. JOHNSON and G. C. BEROZA, "Machine learning for data-driven discovery in solid earth geoscience," *Science*, vol. 363, 2019.

[16] H. P. J.N. Goetz, A. Brenning and P. Leopold, "Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling," *Computers Geosciences*, vol. 81, pp. 1–11, 2015.

[17] R.-q. N. C. J. L. P. Tao Chen, Li Zhu and T. Lei, "Mapping landslide susceptibility at the three gorges reservoir, china, using gradient boosting decision tree, random forest and information value models," *Journal of Mountain Science*, vol. 17, pp. 670–685, 2020.

[18] A. Stumpf and N. Kerle, "Combining random forests and object-oriented analysis for landslide mapping from very high resolution imagery," *Procedia Environmental Sciences*, vol. 3, pp. 123–129, 2011.

[19] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, Nandi, and A. K, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 982–986, 2019.

[20] Z.-l. Wei, Q. Lü, H.-y. Sun, and Y.-q. Shang, "Estimating the rainfall threshold of a deep-seated landslide by integrating models for predicting the groundwater level and stability analysis of the slope," *Engineering Geology*, vol. 253, pp. 14–26, 2019.

[21] Z. Ma, G. Mei, and F. Piccialli, "Machine learning for landslides prevention: a survey," *Neural Computing and Applications*, vol. 33, no. 17, pp. 10 881–10 907, 2021.

[22] G. Visualizer, "Gps visualizer elevation data," 2021. [Online]. Available: https://www.gpsvisualizer.com/elevation

[23] NASA, "Nasa-usda global soil moisture data," 2021. [Online]. Available: https://earth.gsfc.nasa.gov/hydro/data/nasa-usda-global-soil-moisture-data

[24] USGS, "Earthquake hazards program," 2021. [Online]. Available: https://earthquake.usgs.gov/earthquakes/search/

[25] T. A. S. Dalia B. Kirschbaum, "Global landslide nowcasts from lhasa l4 1 day 1 km x 1 km version 1.1," 2020. [Online]. Available: https://disc.gsfc.nasa.gov/datasets/Global_Landslide_Nowcast_1.1/summary

[26] P. Branco, L. Torgo, and R. P. Ribeiro, "Smogn: a pre-processing approach for imbalanced regression," in *First international workshop on learning with imbalanced domains: Theory and applications*. PMLR, 2017, pp. 36–50.

[27] R. Chambers, P. Kokic, P. Smith, and M. Cruddas, "Winsorization for identifying and treating outliers in business surveys," in *Proceedings of the Second International Conference on Establishment Surveys*. American Statistical Association Alexandria, Virginia, 2000, pp. 717–726.

[28] A. Kolbaşi and A. Ünsal, "A comparison of the outlier detecting methods: an application on turkish foreign trade data," *J. Math. Stat. Sci*, vol. 5, pp. 213–234, 2015.

[29] M. M. Ahsan, M. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, vol. 9, no. 3, p. 52, 2021.

[30] H. White, "A reality check for data snooping," *Econometrica*, vol. 68, no. 5, pp. 1097–1126, 2000.

[31] J. Mielniczuk and P. Teisseyre, "Model selection in logistic regression using p-values and greedy search," in *International Joint Conferences on Security and Intelligent Information Systems*. Springer, 2011, pp. 128–141.

[32] M. A. Sulaiman, Labadin, and Jane, "Feature selection with mutual information for regression problems," in *2015 9th International Conference on IT in Asia (CITA)*, 2015, pp. 1–6.

[33] P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm: a big comparison for nas," *arXiv preprint arXiv:1912.06059*, 2019.

[34] S. Çellek, "Effect of the slope angle and its classification on landslide," *Natural Hazards and Earth System Sciences Discussions*, pp. 1–23, 2020.