Machine Learning based Landslide Prediction System for Hilly Areas

R Archana Reddy¹, R Gobinath¹, Chilupuri Saloni Khanna², G Shyamala³
¹SR Engineering College, Warangal, Telangana, India
²Sumathi Reddy Institute of Technology for Women, Warangal, Telangana, India
³SR Engineering College, Warangal, Telangana, India

Email: archanareddy.srec@gmail.com

Abstract. The recent decade had seen many natural hazards across the globe claiming numerous human lives and caused severe damage to infrastructure creating havoc. Landslide is one such natural disaster which not only creates irreversible damage but also proved to be frequently occurring in hilly areas. Several parts of the world suffer from landslides, and numerous research works were conducted across the globe to predict and manage landslides. In these works, researchers had used a specific Machine Learning-based prediction system to provide early warning before potential landslide occurs. Bountiful research works conducted on landslide generation proved that elevated water content in the soil, which may increase due to continuous and prolonged rainfall occurring in the slopes, leads to most of the landslides. It is evident that measuring the amount of rainfall is the key to predict landslide generation. This is an area where we looked upon and had developed a mechanism to predict landslides using machine learning models. This study uses seven machine learning algorithms that are trained and tested with integrated rainfall and landslide data for 36 meteorological subdivisions of India from the years 2009 to 2019. The results obtained were consistent and reliable, the algorithm that outperformed other algorithms is Logistic Regression with an accuracy of about 94.6%. This predictive model shows better performance than the conventional rainfall threshold method.

1. Introduction
A landslide is a natural calamity that accounts for the downward accelerated moving of any rock, its debris or even the soil/ earth down a sloped section of land [1]. Due to the influence of gravity, the soil or rock moving downwards occurs which is called as Mass wasting.

Landslides can be categorized into five major modes of slope movement: falls, topples, spreads, slides, and flows as per Varnes classification [2,3]. Geological material such as bedrock, debris, earth are further classifications. Majority of landslide types are debris flow or mudflow. (“Landslide”, n.d.) There could be various causes of landslide[4]. When the strength of the composition of materials of the slope is exceeded by the forces acting upon the slope then a visible movement of the slope can be seen[5]. There are plenty of factors that affect the strength and durability of a sloped section of land resulting in escalated down slope forces. Additional to the act of gravity, factors such as rainfall, snowmelt, changes in water level, stream erosion, changes in groundwater, earthquakes, volcanic activity, disturbance by human activities, etc could trigger soil displacement leading to landslides. (“U.S. Geological Survey”, 2019)
Landslides have devastating effect on humans and resources [6,7]. They claim the lives of thousands of people annually and damage property. They have deplorable effect on water supply, fisheries, sewage disposal systems, forest, roadways, and dams. Landslide hazards can be minimized by practicing good engineering, investigating the geological area, effective implementation of land-use management regulations [8,9,10]. It is crucial to avoid such huge losses due to landslide. This study focuses on predicting landslides in the different meteorological subdivisions of India specifically hilly areas [11,12].

According to the available data, nearly 0.42 million km² 12.6% of any Indian land area without adding the snow-covered regions is susceptible for landslide hazard [13,14]. Therefore, it becomes indispensable to devise a system for predicting potential landslides to avert loss of valuable lives and property. This study includes geodatabases namely “Landslide recent incidents” taken from Geological survey of India from the years 2009 to 2019 and sub divisional monthly rainfall dataset from open government data platform India [15,16]. The areas of various hazard zones in different states of India is shown in the landslide hazard map (Fig. 1). The Himalayas of Northwest and Northeast India, and Western Ghats come under highly vulnerable zones to landslides [17].

![Figure 1. Zone map of Landslide prone areas in various states of India.](image1)

Landslide susceptibility maps can be strategically used to identify areas subjected to landslide hazards[18]. The map depicts where landslides take place and the potential causes of them that include, but not limited to, the composition of soil, slope, and the amount of precipitation. (Fig.2)

![Figure 2. Landslide susceptibility map of India. The landslide vulnerable areas are marked orange.](image2)
2. Methodology

2.1 Rainfall Frequency Analysis

The primary focus of this study is to predict the landslides triggered by prolonged rainfall. Rainfall is a crucial factor to be considered while studying landslides as they directly influence the strength of the soil that eventually makes the soil erode down a slope. Several landslides occurred across the globe, one of those being the landslide that struck the Philippines village of Guinsaugon on February 17, 2006. This led to the destruction of over 350 houses and an elementary school, burying more than 1,100 people. Unfortunately, the locals had no knowledge that such a catastrophe was going to take place. There was no early warning and no sufficient time to evacuate. The experts could not identify a direct trigger for the landslide, but the investigations suggested that the unusually heavy rainfall could be a driving factor of the landslide as the heavy rainfall had saturated the mountainside prior to the slide. (Laura Naranjo, 2020)

A study conducted by NASA scientists Robert Alder and Yang Hong suggested that the slope and soil type are paramount while determining the cause of landslides. The more susceptible types of slopes are steep in nature whereas soil of coarser type and bear soil have major contribution to landslides [19]. According to their research and its outcome mostly rainfall induced landslides are occurring in the hilly regions such as Philippines, Central and South America, and South-eastern Asia which received the wrath of landslides frequently. Heavy rains were predominantly brought by monsoon, hurricanes, and typhoons. The combination of steep terrain and heavy rains resulted in tremendous calamity. Their study was based on a reliable satellite-based system that provided data from an array of satellites. This data was merged to estimate whether remote sensing instruments can provide the location of potential rainfall triggered landslides [20]. Their research confirmed and concluded that extremely intense rainfall, especially when there is continuous rainfall which is moderate to heavy within a short period of time (less than 12 hours, overwhelmed the thresholds for each zone and induced slides. The results obtained from the satellite corresponded with the previously deduced measurements made by the rainfall-gauge-based threshold [21].

Therefore, it is evident by their study that landslides are a consequence of heavy and prolonged rainfall. Another study suggested that the hazard of rainfall triggered landslides could vary according to the catchment because of variations in rainfall behaviour. The rainfall-frequency analysis was performed and an atmospheric general circulation model (AGCM) was used to determine the temporal and spatial rainfall behaviour in an adopted catchment in Central Taiwan. The results of their study indicated the inconsistency in the trend of predicted rainfall and landslide susceptibility that differ according to the catchment. Additionally, their study hinted at the presence of uncertainties in rainfall predictions. [22]. This study uses the rainfall data accumulated over the years 2009 to 2019. The variation of rainfall during southwest monsoon is demonstrated in (Fig.3).

Figure 3. Variation in the amount of rainfall during southwest monsoon in India.
2.2 Acquisition of data
After the acquisition of rainfall dataset, the landslide data is obtained from the years 2009 to 2019, from the Geological survey of India. The landslide data involves the dates with any reported mode of landslide including mudflow, rockfall, and road slip with each landslide incident indicated by 1 and each non-landslide incident indicated by 0. Later a machine learning approach is used to make predictions on this binary classification of rainfall induced landslides that might occur in hilly areas of India[23].

2.3 Predicting Landslides using Machine Learning
Machine learning enables computer systems to learn automatically through experience and enhance their performance eventually rather than being explicitly programmed. Machine learning (ML) is a subset of Artificial Intelligence (AI) [24]. Advancement of computer programs that can access and learn from data is at the core of machine learning. There are several ML models and algorithms that can be used for predictive modelling. This study uses algorithms such as Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Boosting method such as Adaboost (AdaB), Decision Tree (DT), K-Nearest Neighbour (KNN), and Gaussian Naïve Bayes (GNB). These algorithms have been used for model training and testing.

Logistic regression is used to predict the probability of a categorical dependent variable. Random forests build an aggregation of decision trees at the training time and gives the output of the mean prediction of the individual trees. SVM assigns and represents the examples as points in space. The examples of different category are separated by a wide gap. Adaboost when used with other algorithms enhances the performance. A decision tree uses tree-like model of decisions and their possible consequences to act like a decision support tool. KNN maps all available cases and new cases are segregated based on the measure of affinity. GNB follows a normal distribution of continuous values associated with each feature. In this study, Scikit-learn is used to build predictive model. Scikit-learn is a python-based module for machine learning. It consists of numerous classifications, regression, and clustering algorithms which is used widely by several researchers in their work.

2.4 Predictive Modelling using Scikit-learn
The predictive modelling is done by first collecting the main dataset i.e., landslide and rainfall dataset from the year 2009 to 2019. Then the data is split into training set and testing set. Training set is used to train the ML algorithm or model and testing set is used to test for the prediction performance of the model. Finally, the overall performance is evaluated by comparing the predicted response values with the true response values. After the evaluation of the model, predictions are made accordingly. Predictive modelling is performed using python by first importing the required libraries and estimator model. An instance of the estimator is created and tested by training dataset. The summary of the dataset is studied and the identifying, target, categorical, numerical, and other variables are identified. The variables with missing values are flagged. The missing values are then imputed. Then the data set is split to train and test. The flagged and imputed missing values are passed into the modelling process. Finally, the predictions are made after evaluating the model.

3. Results and Discussion
The classification algorithms such as Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Boosting methods like Adaboost (ADA), Decision tree (DT), K-Nearest Neighbour (KNN), and Gaussian Naïve Bayes (GNB) are used to build predictive model of the landslides for different subdivisions of India. The dataset had been split into training and testing set, by importing the train_test_split method, in the ratio 70:30, respectively. The classifiers had been trained, tested, and split several times and 10-Fold Cross Validation was applied. The following evaluation metrics were used to measure the performance of each individual algorithm. Accuracy is one of the classification metrics. A Confusion Matrix was used to characterize the predictions from the testing set as true positive, tp, (actual landslide event correctly predicted), true negative, tn, (actual non-landslide event correctly predicted), false positive, fp, (false alarm) and false negative, fn, (missed prediction).
The Precision and Recall are calculated from the ratios \( \frac{tp}{tp+fp} \) and \( \frac{tp}{tp+fn} \) respectively. The F1 score can be interpreted as a weighted average of the precision and recall given by the harmonic mean \( F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \). The means of the accuracy, precision, recall, F1 score, and area under the curve for various classification algorithms is listed in (Tab. 1). The stacked histograms of True Positive Rates (Recall or Sensitivity) and False Positive Rates (Specificity) are demonstrated in (Figs. 4 and 5). For accurate prediction, the TPR must be high and FPR must be low. The Receiver Operating Characteristic (ROC) curve typically features true positive rate on the Y axis, and false positive rate on the X axis. It is used to estimate the Area Under the Curve (AUC). Larger area under the curve is preferable for better performance of the algorithm.\(^4\)

**Figure 4.** Stacked histogram of True Positive Rates (TPR) for the various machine learning algorithms evaluated.

**Figure 5.** Stacked histogram of False Positive Rates (FPR) for the various machine learning algorithms evaluated.
Table 1. Predictive performance of various Machine Learning Models

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1 Score (%) | AUC (%) |
|-------|--------------|---------------|------------|--------------|---------|
| SVM   | 93           | 90.9          | 96.1       | 93.4         | 93      |
| DT    | 91.5         | 92.2          | 91.3       | 91.7         | 92      |
| GNB   | 93           | 91.9          | 94.8       | 93.3         | 93      |
| LR    | 94.6         | 93.2          | 96.7       | 94.9         | 95      |
| RF    | 9.6          | 93.3          | 94.5       | 93.9         | 94      |
| KNN   | 87.3         | 86.4          | 89.7       | 88           | 87      |
| ADA   | 92.8         | 93.5          | 92.6       | 93           | 93      |

The ROC curve of different classifiers and their respective areas under the curve are shown in (Fig.6). A Precision-Recall curve is a useful measure of success of prediction when the classes are very imbalanced. A high area under the curve represents both high precision and high recall.\cite{4} This curve for Logistic Regression is plotted in (Fig.7).

**Figure 6.** True and False positive rates plotted in a Receiver Operating Characteristic (ROC) plot for computing the Area Under the Curve (AUC).

**Figure 7.** Precision and Recall curve plotted for Logistic Regression with Average Precision (AP).
By observing the above-mentioned metrics and their scores, the best predictive classifier is chosen. Results show that the Logistic Regression (LR) model has the highest AUC, accuracy, recall, and F1 score among other models. (Tab.1). Therefore, LR can be considered as the best model among those evaluated for predicting landslides in the study area.

4. Conclusion
This study uses a Machine Learning (ML) approach to predict landslides for different subdivisions of India from the years 2009 to 2019. The Logistic Regression (LR) algorithm has the best predictive performance among the evaluated models, with mean Area Under the Curve being 95 %, mean Accuracy being 94.6 %, and mean Precision, Recall and F1 score being 93.2 %, 96.7 %, 94.9 % respectively. This model also outperforms the traditionally used 1-day cumulative rainfall thresholds, for any value of the cumulative rainfall. The predictive performance of the machine learning algorithms can be further enhanced by considering the optimal threshold value and making further optimizations of the data features. This study shows the potential of machine learning for characterizing temporal patterns in rainfall-triggered landslides using minimal data input. This approach can be extended for predicting other natural hazards by obtaining observational dataset and to reduce the damage caused by such catastrophes.

5. References
[1] Keefer DK Landslides caused by earthquakes. Geological Society of America Bulletin. 1984 Apr 1 95(4)406-21
[2] Malamud BD, Turcotte DL, Guzzetti F and Reichenbach P 2004 Landslides, earthquakes, and erosion. Earth and Planetary Science Letters Dec 30 229(1-2)45-59
[3] Collins BD and Znidaric D2004Stability analyses of rainfall induced landslides Journal of geotechnical and geoenvironmental engineering Apr 130(4)362-72
[4] Tsaparas I, Rahardjo H, Toll DG and Leong EC 2002 controlling parameters for rainfall-induced landslides Computers and geotechnics Jan 1 29(1)1-27
[5] Wang G and Sassa K 2003 Pore-pressure generation and movement of rainfall-induced landslides: effects of grain size and fine-particle content Engineering geology Apr 1 69(1-2)109-25
[6] Chandrasekaran SS, Owais RS, Ashwin S, Jain RM, Prasanth S and Venugopalan RB 2013 Investigation on infrastructural damages by rainfall-induced landslides during November 2009 in Nilgiris India Natural hazards Feb 1 65(3)1535-57
[7] Senthilkumar V, Chandrasekaran SS and Maji VB 2018 Rainfall-induced landslides: case study of the Marappalam landslide Nilgiris District Tamil Nadu India International Journal of Geomechanics 18(9)05018006
[8] Dikshit A, Sarkar R, Pradhan B, Segoni S and Alamri AM 2020 Rainfall induced landslide studies in Indian Himalayan region: A critical review Applied Sciences Jan 10(7)2466
[9] Sengupta A, Gupta S and Anbarasu K 2010 Rainfall thresholds for the initiation of landslide at Lanta Khola in north Sikkim India Natural hazards Jan 152(1)31-42
[10] Chen W, Peng J, Hong H, Shahabi H, Pradhan B, Liu J, Zhu AX, Pei X and Duan Z 2018 Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County Jiangxi Province China Science of the total environment Jun 1 626121-35
[11] Zhu X, Xu Q, Tang M, Nie W, Ma S and Xu Z 2017 Comparison of two optimized machine learning models for predicting displacement of rainfall-induced landslide: A case study in Sichuan Province China Engineering geology 23(218) 213-22
[12] Pham BT, Pradhan B, Bui DT, Prakash I and Dholakia MB 2016 A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India) Environmental Modelling & Software 1 84 240-50
[13] Thai Pham B, Shirzadi A, Shahabi H, Omidvar E, Singh SK, Sahana M, Talebpour Asl D, Bin Ahmad B, Kim Quoc N and Lee S 2019 Landslide susceptibility assessment by novel hybrid machine learning algorithms Sustainability 11(16) 4386

[14] Chen W, Pourghasemi HR, Kornejady A and Zhang N 2017 Landslide spatial modeling: Introducing new ensembles of ANN MaxEnt and SVM machine learning techniques Geoderma Nov 1(305)314-27

[15] Gobinath R, Ganapathy GP and Akinwumi II 2020 Stabilisation of natural slopes using natural plant root as reinforcing agent Materials Today Proceedings In press https://doi.org/10.1016/j.matpr.2020.08.227

[16] Reddy RA, Sivakrishna A, Gobinath R and Babu DR 2020 A novel method to predict pozzolanic nature of concrete with sintered clay using soft computing techniques In IOP Conference Series Materials Science and Engineering Jun 1 872(1)012159 IOP Publishing

[17] Gobinath R, Raja G, Prasath E, Shyamala G, Viloria A and Varela N 2020 Studies on strength characteristics of black cotton soil by using novel SiO2 combination as a stabilizing agent Materials Today: Proceedings 27(1) 657-663 https://doi.org/10.1016/j.matpr.2020.01.597

[18] Prabhanjan N, Gobinath R, Raja G, Prasath E and Sahithi G 2020 Seismic Response of root reinforced soil slope at Nilgris district Soil Tamilnadu Int Jr Sci Tech and res 9(2) 3666-3674

[19] Awoyera PO, Akinmusuru JO, Shiva Krishna A, Gobinath R, Arunkumar B and Sangeetha G 2020 Model Development for Strength Properties of Laterized Concrete Using Artificial Neural Network Principles In: Das K, Bansal J, Deep K, Nagar A, Pathipoornam P and Naidu R (eds) Soft Computing for Problem Solving Advances in Intelligent Systems and Computing 1048 Springer, Singapore https://doi.org/10.1007/978-981-15-0035-0_15

[20] Veeramsetty V, Singal G and Badal, T 2020 Coinnet: platform independent application to recognize Indian currency notes using deep learning techniques Multimed Tools Appl 7922569–22594 https://doi.org/10.1007/s11042-020-09031-0

[21] Veeramsetty V and Deshmukh R 2020 Electric power load forecasting on a 33/11 kV substation using artificial neural networks SN Applied Sciences May 2(5)1-0

[22] Kumar BA, Sangeetha G, Srinivas A, Awoyera P, Gobinath R and Ramana VV 2020 Models for predictions of mechanical properties of low-density self-compacting concrete prepared from mineral admixtures and pumice stone In: Das KN, Bansal JC, Deep K, Nagar AK, Pathipoornam P and Naidu RC (eds) Soft computing for problem solving Springer Berlin 677–690

[23] Poongodi K, Murthi P, Shivaraj M, Arun Kumar Beerala, Sangeetha Gaikadi, Srinivas A and Gobinath R 2018 ANN based prediction of Bond and Impact Strength of Light Weight Self Consolidating Concrete with coconut shell International Conference on Intelligent Computing and Communication for Smart World (I2C2SW) Erode India 364-370 doi: 101109/I2C2SW4581620188997421

[24] Kumar KR, Gobinath R, Shyamala G, Viloria E and Varela N 2020 Free thaw resistance of stabilized and fiber-reinforced soil vulnerable to landslides Materials Today: Proceedings 2020 Mar 4 27(1)664-670 https://doi.org/10.1016/j.matpr.2020.02.041