



Machine Learning and IoT-Based Approaches to Detect and Predict Rainfall-Triggered Landslides



Abhijit Kumar^{1,3*}, Vinay Kumar Singh², Rajiv Misra¹, Trilok Nath Singh^{1,2}, Tanupriya Choudhury⁴

¹ Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna 801106, India

² Department of Earth Science, Indian Institute of Technology Bombay, Bombay 400076, India

³ School of Computer Science, UPES Dehradun, Dehradun 248007, India

⁴ CSE Department, Symbiosis Institute of Technology, Symbiosis International University, Pune 411042, Maharashtra, India

Corresponding Author Email: abhijitkaran@hotmail.com

<https://doi.org/10.18280/ria.370522>

ABSTRACT

Received: 12 May 2023

Revised: 30 August 2023

Accepted: 4 September 2023

Available online: 31 October 2023

Keywords:

landslide prediction, sensor network, machine learning, rainfall

Landslides, predominantly triggered by rainfall, pose significant global threats, causing extensive loss of life and severe socio-economic disruptions. Among such calamities, the landslide in Uttarakhand stands as a stark exemplar of the severe repercussions these natural disasters can inflict. This study proposes two sophisticated approaches aimed at the detection and prediction of rainfall-induced landslides. Our initial approach presents a comprehensive analysis of the topographic and hydro-meteorological conditions that catalyzed the catastrophic Kedarnath disaster. This method utilizes an innovative algorithm, validated through machine learning models, in conjunction with an IoT-based application designed to collect critical data necessary for model training and validation. Emphasis is placed on rainfall, identified as a pivotal factor influencing debris flow and lake outbursts during the Kedarnath event. Utilizing the standard deviation of landslide data induced by rainfall from 2013-17, a threshold value was calculated to gauge the severity of such scenarios. The second approach employs a range of machine learning and ensemble learning algorithms to enhance the prediction of rainfall-triggered landslides. These proposed methods were investigated using web-scraped datasets acquired from NASA and IMD portals, with under-sampling and oversampling carried out to mitigate any potential dataset bias. Following extensive exploration and exploitation of diverse learning algorithms, it was inferred that oversampling techniques and the random forest model outperformed alternative models consistently across all performance measures, including Accuracy, Precision, Recall, and F1-Score.

1. INTRODUCTION

Landslides are one of the most devastating natural disasters, causing destruction to people's lives, economy, and infrastructures [1]. As a result, forecasting and monitoring are critical in order to avoid massive loss of life and damage to the property. The landslide criticality illustrated in Figure 1. It represents the number of fatalities with re-spect to the total number of landslides across India and neighboring nations till 2013 [2]. The authors [3] proposed extensive instrumental monitoring of landslides is typically difficult due to a lack of scientific expertise or insufficient resources for instruments and subsurface research. Identification and selection of possible hazard sites, as well as the return duration of such events, are critical for landslide monitoring. There are well-established signs that may be responsible for initiating landslides, and they must be investigated before a monitoring location is chosen. These are seismic activity, rate of precipitation, climatic conditions, wind velocity, river incision, and groundwater fluctuation [4]. Landslides have become very frequent in steep slopes, which are anthropogenically modified however, they have also occurred in much gentler slopes making them very difficult to understand. Their ability to cause substantial damage makes them an extremely important subject of research. Himalayan

region is known for having lithologically, topographically and seismically unfavorable conditions, which may cause landslides and floods [5]. The landslides and debris overflow that took place at Kedarnath left it in ruins that created havoc resulting in a lot of people losing their homes, more than 6000 people lost their lives, and a lot of bridges and roads were destroyed where atleast 30 hydropower plants were severely damaged. This unfortunate event left 1,00,000 Pilgrims [6-8] and tourists marooned that needed military support to rescue them. Mass moments (debris flow, rock fall, debris cum rockslide etc.) were caused across Uttarakhand because of the heavy rainfall on June 16th and 17th of 2013. This has been the most devastating disaster that ever occurred over the last century in India [9, 10]. This has particularly grabbed the attention of various researchers and authors as it has been the only incident that has occurred from a glaciated environment [10-12]. The scientific community has focused its attention on heavy rainfall, water discharge of this tragic event, and understanding the other factors that led to this incident [13]. The data regarding this event has been collected from IMD [14]. The scientific community has focused its attention on heavy rainfall, water discharge of this tragic event, and understanding the other factors that led to this incident [15, 16]. In order to find the cumulative rainfall and the standard deviation and their effects have been measured where we

wanted to provide mathematical support to the findings and assess the future happenings that can likely cause a similar effect. The threshold that defines the intensity of the situation is measured based on the values we identified and has helped us create a systematic representation of the condition. It is very important to generate a method that can effectively help us estimate the threshold values. The rainfall or water precipitation most of the time works as the triggering factor towards occurrence of landslide.

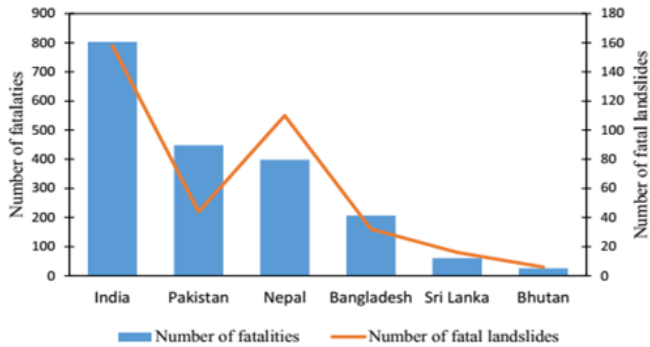


Figure 1. Graph of the number of fatal landslides and resulting fatalities in India and its neighboring countries

Hence, we developed a system that is a highly sophisticated methodology where we accumulated the data through numerous of sensors (Rain Gauge, Temperature, Wind speed, Humidity) and fed into a machine learning environment, which finds the cumulative rainfall and threshold value of rainfall to predict the accuracy and gives out the required output. We mainly deal with shallow landslides and considered a week data that can help us get our desired result. Several technologies have been developed for landslide monitoring in which wireless sensor network has proved to be the most significant. The sensors are also playing an instrumental role in capturing or sensing the real data in a physical environment by deployment it [17, 18]. It is also important because of its application in the field of environmental engineering, space sciences, climate science, electrical engineering, electronics engineering, computer engineering, atmospheric sciences & aeronautical engineering and has been explored by various workers. A wireless sensor network (WSN) in this aspect is one of the emerging technologies in the field of electronics engineering to measure real-world data. Authors obtained different parameter such as rainfall, moisture, pore pressure and movement aiding in better understanding of landslide. The triggering parameters for landslide hazard include rainfall, seismic activity, the rate of precipitation, climatic conditions, wind velocity, river

incision, soil properties, and groundwater fluctuation etc. [19, 20]. Landslides are very frequent in steep slopes, however those also occurred in much gentler slopes and difficult to understand. The study based on Himalayan terrain suggests that the major contributor of occurrence of the landslides are 65% by debris slide, 29% debris cum rockslide, 5% rock fall and remaining 1% by any other triggering factors. Furthermore, the study also suggests that the triggering factor of debris flow, rock slides are the rainfall and the soil properties. The proposed study is considered both these two factors like rainfall and soil properties. Based upon the soil properties, the sensor is deployed on the specific locations and rainfall precipitation has been captured for further processing.

A majority of work has been done to predict the landslide using physical system as well as machine learning models. Although some work has been proposed which combined both systems i.e. physical system as well machine learning model, but most of them are unable to generate the system on real time basis [21-26].

The major contribution of the work is to detect and predict the occurrence of landslides by deploying sensor nodes over the landslide prone region and machine learning models all together on real time basic.

The rest of the paper has been divided into five more sections. The section 2 discusses the related worked done so far to monitor and detect landslides. The study area has been described followed by literature review. The next section proposes the methodology adopted in the study. Afterwards the next section follows the results and discussion. Finally the last section discussed the conclusion of the proposed work carried out in this paper.

2. LITERATURE REVIEW

Monitoring and measurement of superficial displacement have been used by several researchers in the past and is the easiest way to observe and predict a slope failure activity [22-33] Monitoring can also provide useful information to the response of various triggering factors like rainfall, earthquake, freeze & thaw, wind velocity. Although it is a challenging task to fetch physical properties of earth sur-face from real world environment, researchers have proposed various techniques (Table 1) to determine earth surface property (displacement and structural deformation) through continuous monitoring. In recent years machine learning techniques also utilized for forecasting natural hazards [34, 35]. In initial days the broad categories of physical landslide monitoring system have been divided into three basics approached viz. visual, instrumentation and surveying.

Table 1. Literature reviewed on landslide prediction

Method	System	Application	Resolution	Limitation	Stren	Author
Early warning	Instrumentation	Landslide Monitoring	cm to m	Field survey Required	Cheap and longtime monitoring	Mittal et al. (2008) [9]
Weather Radar	Early System	Early warning from rainfall intensity	m to km	Calibration Needed	Helps in Maps spatial distribution	Bhatta and Natarajan (2017) [17]
Photogrammetric	Terrestrial, Airbone	Joint survey, Landslide Detection	cm to m	Image Resolution, Field survey Required	Rapid cheap and long term monitoring	Bhatta and Natarajan (2017) [17]

Machine Learning based EWS	XGBoost based EWS	Applied inclination sensor and capacitive soi moisture sensor	--	Zero False negative indicates overfitting.	High Accuracy	Sreevidya et al. (2021) [24]
LiDAR	Intelligent Wireless Probe (IWP) sensor	Slope instability monitoring	cm to m	Image resolution high	Volume of unstable soil mass estimation	Kumar and Ramesh (2022) [25]
MPU6050 Accelerometer and Gyroscope Sensor	Android based Fuzzy EWS	Fuzzy rule based landslide early warning system	--	Deployment of sensors, Accuracy low	Android based real time monitoring	Fatimah et al. (2020) [26]
Early Warning	Wireless Sensor Network	Landslide Detection	mm to m	Field survey Required	Long time Online Monitor	Hemalatha et al. (2019) [27]
Hybrid Early Warning System	Wireless Sensor Network	Landslide prediction warning	cm to m	Installation of sensors, Field survey required	Automated Velocity and Acceleration computation using Least square method	Bai et al. (2022) [28]
LiDAR	Ground based static, Ground Based Mobile, Airborne	Landslides monitoring, mapping, Temperature and moisture detection	cm to m	Humidity, Temperature, Displacement	High accuracy, high rate of acquisition is possible	Wieczorek and Snyder (2009) [31]
INSAR	Standard	Movement monitoring, detection, warning,	mm to m	Visibility and shadowing require, Refelctance, Limited to low movement	Long time monitoring, large area survey	Lacasse et al. (2008) [32]
Optical Satellite	Landslides inventory	Landslide Time series based analysis	cm to m	It is very expensive poor image quality due to cloud	Landslides survey, large area coverage	Theodoulidi et al. (2006) [33]

3. STUDY AREA

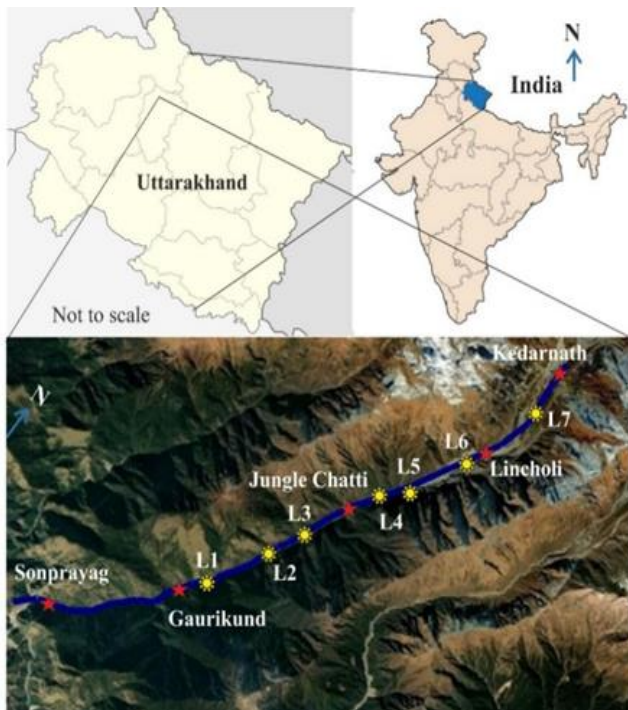


Figure 2. Sensor deployed sites

The Kedarnath has been a great significant area as per Hindu mythological. Its construction details are unknown but, it attracts a great deal of adventurers, pilgrims, and crusaders. It is located in foothill of Himalayas at 3583m above sea-level. It is situated in Uttarakhand region. There are river banks such as Saraswati, Mandakini. It falls under Rudraprayag district. It spread across an area of 1982.09 sq.km approx. It lies between

latitude 30°12'58.132-30°48'27.642N and longitude 79°2'58.649-79°2'0.952E. The study region is riddled with numerous slope stability issues. The slopes may be classified as extremely stable, stable, moderate, low, or very low based on eye assessment and multiple geotechnical investigations. Since, it is a glaciated area, snowmelt has become more noticeable as a result of environmental changes. Over the years from 2003 to 2010, the amount of snow lost as a result of this impact rose with the rise in water levels. The region is composed of a wide variety of soils and rocks, including silicates, granites, and other kinds. The geography and kind of land in the region have been taken into consideration. Mountains, very steep slopes, and two banks of river on top of the temple area that is being marked down by a big moraine are very significant features to be consider for the study area. The study area is depicted in Figure 2.

3.1 Sensor node deployment

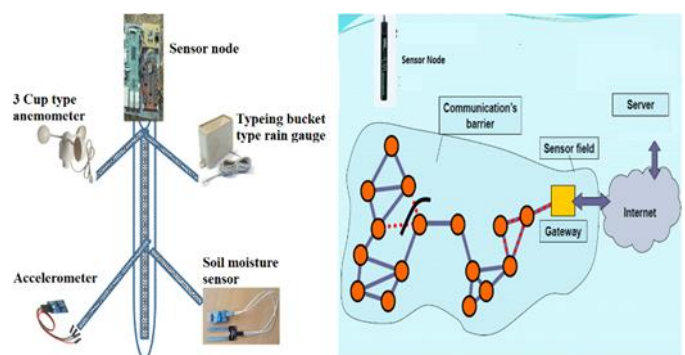


Figure 3. Architecture of the autonomous landslide monitoring system

The schematic depiction of the sensor node architecture is

presented in Figure 3. In this study, a comprehensive sensor system was employed, encompassing a variety of sensors for detecting parameters like temperature, moisture, rainfall, and acceleration. These parameters are instrumental in establishing individual threshold values, which, in turn, aid in the prediction of slope movements. Central to the landslide monitoring system, the sensor node's core component is the microcontroller. The pivotal role in data acquisition and transmission is assumed by the ARM7 microcontroller. The amalgamation of sensors, GPS, and Wi-Fi is facilitated through this central unit. By gathering data from the sensors, the ARM7 microcontroller effectively communicates this information to a remote server. The communication is established via a ZigBee wireless module. The choice of the ARM7 microcontroller stems from its favorable attributes, namely its simplicity, cost-effectiveness, low power consumption, and versatility.

The ARM7 microcontroller boasts a 32-bit architecture, packaged in a configuration with 64 pins. It incorporates two 32-bit timers, complemented by a chip oscillator that operates within the frequency range of 1MHz to 20MHz. A significant feature of this microcontroller is its 500Kb on-chip flash program memory, further augmented by in-system and application programming capabilities.

3.2 Rainfall threshold prediction using cumulative rainfall and standard deviation

Initially, the study area has been divided into 7 small regions. The division performed according to several physiographic and environmental factors, including lithological, topographical, and meteorological data).

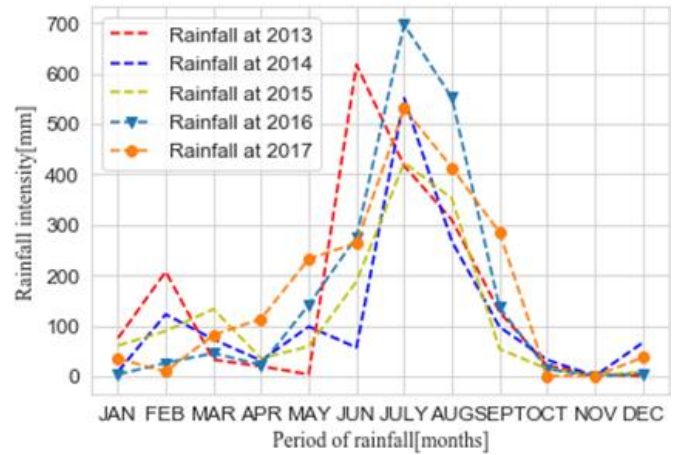


Figure 4. Month wise rainfall intensity in mm for landslides modeling

The proposed work incorporates 4-year historical rainfall triggered landslide data from IMD (Table 2) has been employed to calibrate the threshold rainfall during the year 2013-2017 for study. According to Table 2, the months of May through August and the years 2013 through 2017 exhibit the most significant rainfall value from the point of view of landslides. The heaviest rainfall was recorded in the months of June and July.

The graph plot of cumulative rainfall intensity vs period of rainfall (in terms of month and year wise) is illustrated in Figure 4. The cumulative rainfall was recorded as highest in the years 2013, 2016, and 2017. The years 2013, 2016, and 2017 each display the maximum value of the standard deviation, and the median.

Table 2. Rainfall data downloaded from IMD

YEAR	JAN	FEB	MAR	APR	MAY	JUN	JULY	AUG	SEPT	OCT	NOV	DEC	mean	Standard Deviations
2013	75.2	207	32.4	19.4	3.4	617.4	416.7	309.6	126.8	20.7	3.1	0	1831.7	190.9489742
2014	8.8	122.4	72.8	30.7	98.2	56.5	550.9	266.9	97.2	32.1	0	66	1402.5	147.1338683
2015	60.5	89.3	132.8	34.7	58.9	187.5	423.3	352.1	53.5	14.4	0.8	8.7	1416.5	131.7185602
2016	3.3	26	45.4	22.2	141.4	275.8	696.7	551.9	136.1	13.7	0	4.2	1916.7	223.9547243
2017	34.5	8.9	81.3	113.6	232.6	263.7	530.4	413.6	284.4	0.3	0	37	2000.3	169.7136778

4. METHODOLOGY

The proposed study discusses two frontiers of work carried out. The first work carried is based on the data captured by deploying of sensor to detect rainfall and cumulative rainfall. A detailed discussion about the study area and the sensor deployment has been elaborated in earlier section of the paper. While, the second approach employed over the rainfall triggered landslide dataset across the globe. This study employed various machine learning and ensemble learning algorithms. The datasets have been scrapped from NASA and IMD portal. The detailed discussion about the datasets and performance analysis of the algorithm has been elaborated in further section of the paper.

4.1 Proposed approach I

The first approach employed supervised learning based logistic regression algorithm to predict the rainfall-triggered landslide. We proposed an algorithm to figure out and evaluate

the threshold value, which is a key part of figuring out how much damage has been done and what steps can be taken to fix it.

$$CumRain_{1-7} = [\sum_{n=1}^k P(t-1+i)]_{k=1,2,3...} \geq [\delta_k(\sigma)]_{k=1,2,3...} \quad (1)$$

where, $CumRain_{1-7}$ be the vector of cumulative frequency of rainfall, while P be the precipitation deposited at the point time of analysis t.

The thresholds with respect to the standard deviation σ be $\delta_k(\sigma)$ and n be the cumulative number of days of observation. Here, $[\delta_k(\sigma)]_{k=1,2,3...}$ is a vector.

In other words, the algorithm takes into account that the cumulative rainfall from first day to up to sixth day (day before analysis day i.e. the seventh day will be the analysis day) for the above equation measured the cumulative frequency of rainfall ($CumRain_{1-7}$) for the entire week's data within the interval of time-period under consideration. The standard deviation is being computed in order to analyze the value at

which the rainfall intensity changes.

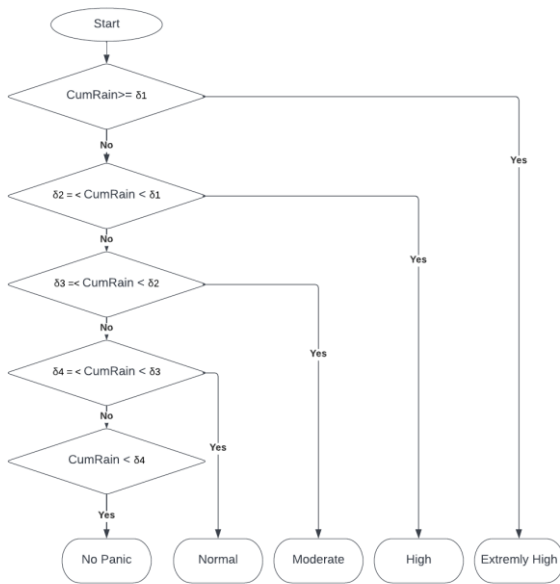


Figure 5. Flowchart of threshold rainfall prediction algorithm

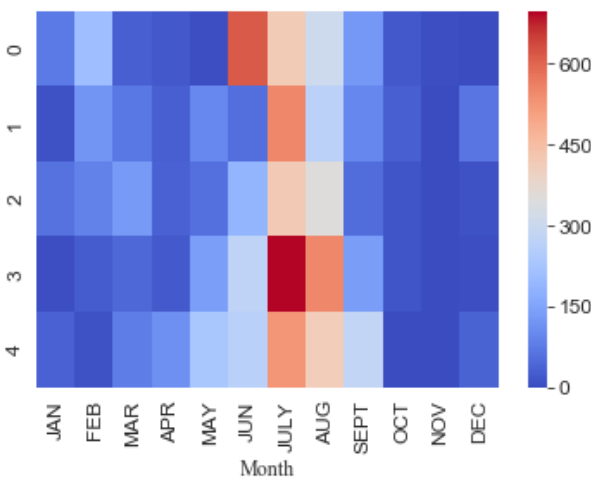


Figure 6. Month wise rainfall intensity 2013

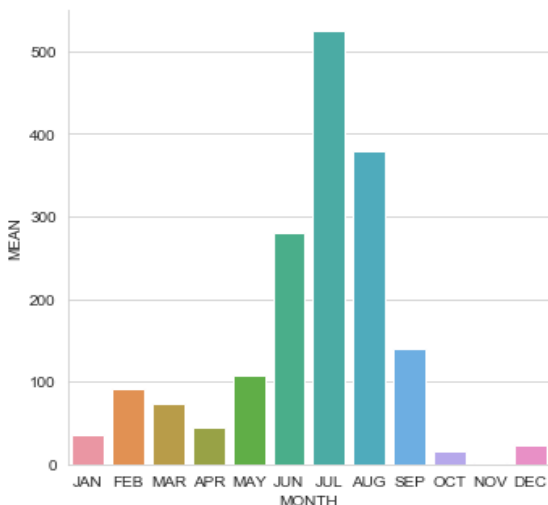


Figure 7. Mean of rainfall data year 2013

The flowchart in Figure 5 is used the measure of cumulative rainfall frequency (CumRain) with a computed value of standard deviation (σ) week's data within the interval of the time period. A relation between cumulative rainfall with the standard deviation has been established in order to predict the threshold rainfall intensity (δ_k).

Logistics Regression: Logistic Regression have employed to predict landside occurrence based on standard deviation coefficient. Principally, logistic regression is employed to describe the relationship between a binary dependent variable and one or more nominal, ordinal, interval, or ratio-level independent variables.

$$p = 1/[1 + \exp(-a - \beta X)] \quad (2)$$

Here, p is the likelihood that a landslide will occur.

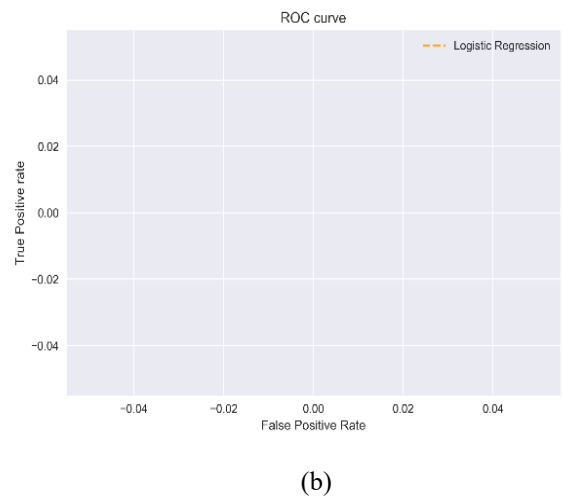
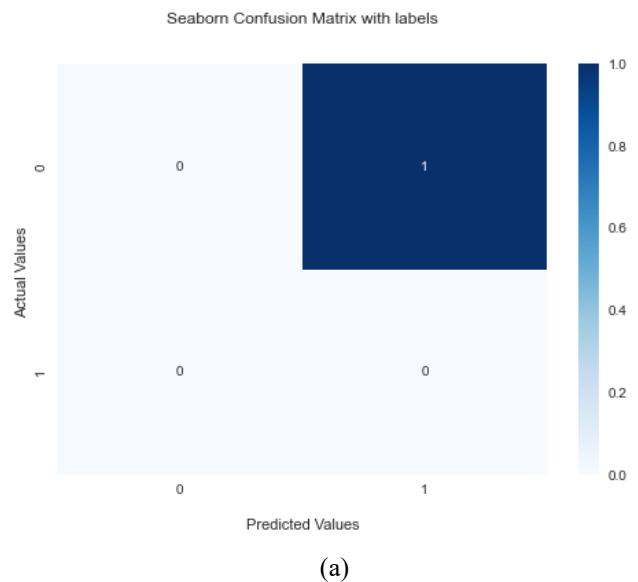


Figure 8. a) Confusion matrix; b) ROC curve of the applied logistic regression model

We combined the IoT technology and Machine learning to get the results accurate and to help us understand the susceptibility of the area for the landslides. We collected the data of cumulative rainfall as depicted in Figure 6 and Figure 7, which is the main contributor and tried to study and find relations between the standard deviations and their changes over the year. We developed a logistic regression based

machine-learning model that has been trained with the data captured by deployed sensor nodes. Further, we have drawn in Figure 8 the confusion matrix and ROC curves, which boost the results, achieved from the trained model. A flowchart has been generated that provides us with a simpler understanding of the situation that can help us contain and manage it in a more effective and smart manner. The obtained data has been captured through sensor has fed into the model and whenever at any potential threat of landslide will be predicted by the model as illustrated in the Figure 5. There are five category of warning signals will be generated: No panic, normal, moderate, high and extremely high. As per the warning, the action may be taken through the use of backend software functionality will be incorporated into the system. The integration of backend software functionality has been achieved using deploying the sensor node across the study area and the data generated by nodes will be captured and being sent to the developed model using cloud based services on real time basis.

Our approach is mainly based on shallow landslides since the region we work on is of that kind.

4.2 Proposed approach II

The second approach employed the various machine learning and ensemble learning algorithms trained and tested over the global rainfall triggered landslide data. The proposed model has been trained and tested on the rainfall triggered landslide data across the globe during 2015-2021. A detailed discussion on Dataset, Preprocessing of the datasets, Feature selection and the deployed model has been discussed as follows:

4.2.1 Dataset

The dataset for deep Learning Based approach has been web scrapped. The data consists of landslide prone areas rainfall triggered landslide information across the globe during 2015-2021. The global data is web scrapped from NASA USA, which includes landslide prone areas rainfall information while the India's data has been scrapped from IMD India. There are 23 features in the dataset and having 145461 records. Out of these 23 features some among these obsolete features which has been discarded by using Univariate Feature Selection algorithm. Some of the important features are cumulative rainfall over 10 days, types of soil, distance from river, distance from road and human interaction etc. The dataset contains 31877 Yes values and 110316 NO values. The Yes value signifies the occurrences of landslide triggered by rainfall while NO value indicates non-occurrences of landslide. The dataset has been precisely analyzed and appropriate preprocessing techniques has applied on it, which has been discussed below. Furthermore, the dataset has been splitted into 80:20 ratio of training and test data.

4.2.2 Preprocessing

Before inputting the data to the model preprocessing is utmost required [29]. In this study, initially categorical data is converted into numeric data. Afterwards, label encoder approach has applied to encode some of the features, while in place of the missing values, some dummy values are inserted using python function `get_dummies` in pandas. Moreover, the Null/NA values are replaced with imputed using mean imputation. Afterwards, as the dataset is heavily biased towards non landslide categories. Thus, it lead towards the necessity of data balancing. Hence, it requires to prepare the

dataset to be balanced prior to fed the dataset to the model.

i) Data Balancing: There are two types of Data Balancing Techniques:

- Undersampling
- Oversampling

Undersampling: Undersampling is a strategy employed to reduce the volume of instances within the majority class in the training dataset. Consequently, the overall size of the training set experiences a significant reduction. This reduction directly translates into accelerated training times during classification processes. Among the varieties of undersampling techniques available, the present study adopts the Random undersampling approach. This technique is characterized by its straightforward nature, where entries from the majority class within the training dataset are randomly eliminated. This process continues iteratively until an appropriate balance between the minority and majority class ratios is achieved. While seemingly simplistic, the efficacy of random undersampling is well-founded. Numerous scientific studies substantiate its high effectiveness as a resampling strategy. By strategically eliminating surplus instances from the majority class, this technique successfully addresses class imbalance, leading to improved model performance despite its inherent simplicity.

Oversampling: The primary objective of oversampling revolves around augmenting the representation of minority samples within the training dataset. This is achieved by increasing the proportion of minority instances in the dataset. One notable advantage of oversampling lies in its ability to retain entire of both majority of classes and minority of classes. This ensures that no original data points are sacrificed in the process. However, a notable drawback emerges as the training dataset expands significantly due to this technique. Within the realm of oversampling, a multitude of techniques is at our disposal. In the context of this study, the chosen method is SMOTE, which stands for Synthetic Minority Oversampling Technique. SMOTE is characterized by its simplicity while maintaining effectiveness, particularly in scenarios with low-dimensional data. An interesting facet of SMOTE is its consistent ability to outperform random oversampling and random undersampling across various performance metrics. In the SMOTE process, each minority class item within the training set is engaged with the algorithm. This algorithm identifies the five nearest neighbors from other minority class instances present in the training set. Subsequently, one neighbor is randomly selected out of the five, serving as the basis for generating a new minority class instance. The newly created data point resides on the hyperplane connecting the two randomly chosen minority class instances. This methodology ensures a balanced and enriched training dataset.

4.2.3 Feature selection

Now, our dataset has 23 features, out of these features some of them are either redundant or least significant features that we need to remove from the data before we train our model. Henceforth, We incorporated feature extraction mechanism to identify the non redundant and significant features So, we need to apply feature extraction method to get important features. The nvariate feature selection approach has been explored and exploited in the proposed work and we obtained 16 best features of 23 features from the dataset.

Univariate Feature Selection: The process of univariate feature selection involves the identification of the most

impactful attributes. This is achieved through the application of univariate statistical tests, wherein each attribute is individually evaluated in relation to the target variable. The purpose is to ascertain if a statistically significant correlation exists between the attribute and the target variable. This methodology is also known as analysis of variance. During this evaluation, the focus is exclusively on the connection between a single feature and the target variable. All other features are temporarily disregarded, thus justifying the term “univariate.” Subsequently, a test score is assigned to each individual feature, reflecting its relationship with the target variable. In a final step, all test outcomes are thoroughly examined, leading to the selection of features exhibiting the highest scores. To achieve this objective, we have adopted the SelectKBest feature selection approach, leveraging the chi-square test to maximize the potential test scores. This method ensures the extraction of attributes that wield the most influence, facilitating an enhanced understanding of their impact on the target variable.

The SelectKBest feature selection mechanism is based upon the values of k-score. The greater value of k-score resembles the greater significance of the feature. The chi-square approach has been explored as a scoring function for classification tasks. Here, the parameter k signifies the desired number of characteristics. Afterwards, fit and transform method has been incorporated to perform training of x & y data. In this approach finally, 16 top features has been identified as most significant. Henceforth, the final dataset we obtained to fed into the model.

4.2.4 Machine learning models

The proposed work involves several state-of-the-art model viz Random Forest, K-Nearest Neighbour (KNN), Decision Tree, SVM (Support Vector Machine), XGBoost, Logistic Regression, Naïve Bayes. The datasets has been trained and test on some supervised learning models and their performance has been evaluated. The description of the models employed are as follows:

Random Forest: Random Forest makes many different decision trees, which are then combined to make a much efficient prediction model. The fundamental idea of Random Forest model is that many different models that don't depend on each other (the individual decision trees) work better together than they do by their own. It tries to find the best way to split the node at hand instead of taking into account how that split affects the whole tree. It uses bagging and feature randomness to build each individual tree. This is done by making a forest of trees that are not related to each other and whose predictions are more accurate as a whole than those of any single tree.

KNN: A non-parametric supervised learning classifier, the k-nearest neighbour algorithm (also written as KNN or K-NN) makes use of geographical proximity to determine whether or not a given data point belongs to a given group. Although it can be used to solve regression problems, it is more commonly put to use in the latter, classification problems. It's based on the principle that things of a kind tend to cluster together. The algorithm's strength lies in the fact that it makes no assumptions about the data, making it especially applicable to nonlinear data.

Decision Tree: The Decision Tree is a supervised learning method that may be used to handle both classification and regression issues; however, it is most commonly utilized for the classification. It is organized in the form of a tree, with the

core nodes reflecting the features of a dataset, the branches representing the rules for making decisions, and the leaf nodes indicating the outcome of those decisions.

SVM: The Support Vector Machine (SVM), a popular technique in Supervised Learning, is commonly used for classification and regression problems. Its main focus lies in classification tasks. The SVM algorithm aims to create an optimal line or decision boundary that separates classes within n-dimensional space, making it easier to classify subsequent data points. Mathematically speaking, this ideal decision boundary is referred to as a hyperplane. The SVM selects crucial extreme points and vectors necessary for constructing the hyperplane, which are known as support vectors. Hence, the term Support Vector Machine corresponds to this technique.

XGBoost: The researchers at the University of Washington came up with the concept of Extreme Gradient Boosting, or XgBoost. It enhances the learning process for gradient-boosting algorithms. In this particular algorithm, the decision tree is being constructed using a sequential and methodical way. The consideration of weights plays significant performance in the XGBoost algorithm. The procedure is terminated when the weights have been assigned to each of the independent variables and the information has been fed into the decision tree which predicts the results. Following these two steps, the process is complete. The variables whose outcomes the tree had mistakenly expected had their weights increased, and the tree's predictions are then sent along to the second decision tree along with these updated outcomes. After that, all of these distinct classifiers and predictors are integrated to create a model that is more resilient and accurate. It can work on regression, classification, ranking, and user-defined prediction issues.

Naïve Bayes: The Naive Bayes algorithm is a supervised learning technique that is utilized for the purpose of resolving classification issues. This algorithm is based on Bayes' theorem. When it comes to developing fast machine learning models that can reliably predict outcomes, Naive Bayes Classifier is among the simplest and most effective Classification algorithms. It is a probabilistic classifier, which means that it makes its predictions based on the likelihood of the occurrence of an object.

5. RESULTS & DISCUSSIONS

5.1 Result & discussion of approach I

Uttarakhand had significant rainfall in June 2013, which led to numerous landslides in the Kedarnath area. Day wise rainfall data of Kedarnath region has been captured over the period 7 days by deploying sensors. Research takes place in the region of Uttarakhand, India, between the cities of Sonprayag and Kedarnath. Located at the confluence of the Mandakini and Vasukiganga rivers is the holy city of Sonprayag. Sonprayag, along with Gaurikund and Sitapur, are significant locations because they serve as rest stops for pilgrims and tourists enroute to the renowned Kedarnath Temple, where tens of thousands of tourists converge annually. Different types of slope stability problems can be found throughout the area under examination. From a visual inspection and other geotechnical procedures, these slopes can be ranked as extremely stable, stable, moderate, low, or shallow.

The proposed model was validated using the prior number of landslides that occurred and historical rainfall in recordings for the time period (2013). The validation of the proposed model has been simulated with the recorded historical rainfall data with reference to installed rain gauges for the period 2013 as illustrated in Figure 6 and Figure 7. In the sensor deployed region, the daily basis alarm level was delivered by the developed decision algorithm and the same was compared with stored data and Geo-registered landslides for tested period, which were continuously organized and updated in the developed Geo database. The results obtained from the model validation has shown acceptable relation between number of days with landslides considered for the validation and different alarm levels used by the proposed landslides prediction algorithm. The “short period” (1 to 6 days) allow to correctly predict the shallow landslides movements.

In summary, an optimized methodology for shallow landslide prediction has been developed by considering data from Kedarnath region in Uttarakhand, India. An IoT-based technology was developed with integrated software for slope monitoring and alarm in case of failure. The IoT-based sensor has been developed with the integration of top-notch relevant sensors to acquire the required data needed at any deployed region. The developed algorithm helps us process and interpret the parameters that can lead to landslide occurrence and is helpful to predict the level of criticality through the use of a threshold. This system helps us to ensure civil safety in indicating the alert level.

5.2 Result & discussion of approach II

The performances of the models have been evaluated in terms of Precision, Recall, Accuracy and F1-score in Table 3 and Table 4 and graphically illustrated using Figure 9 and Figure 10.

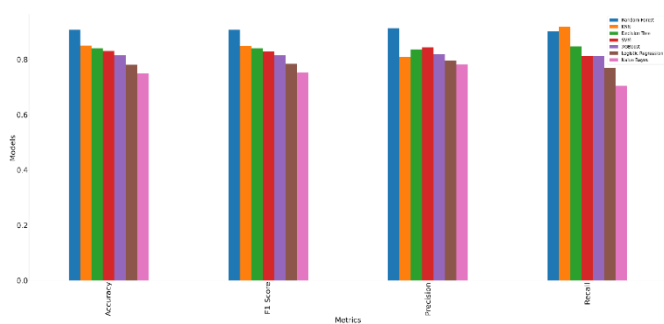


Figure 9. Performance evaluation of different models after applying oversampling

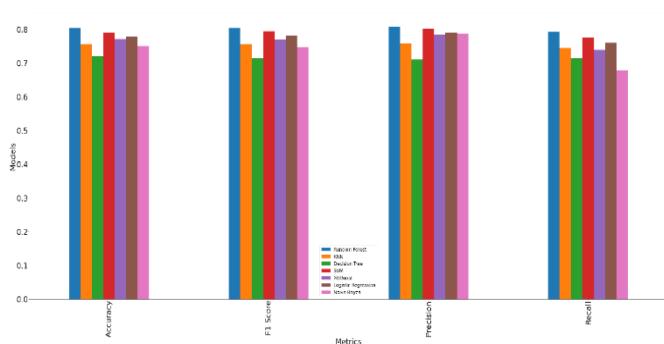


Figure 10. Performance evaluation of different models after applying undersampling

Table 3. Performance evaluation of different models after applying oversampling

Model	Precision	Accuracy	Recall	F1-Score
Random Forest	0.913	0.907	0.901	0.907
KNN	0.808	0.849	0.918	0.848
Decision Tree	0.836	0.840	0.847	0.840
SVM	0.843	0.830	0.812	0.829
XGBoost	0.818	0.815	0.813	0.815
Logistic Regression	0.795	0.780	0.769	0.784
Naive Bayes	0.782	0.750	0.704	0.752

Table 4. Performance evaluation of different models after applying undersampling

Model	Precision	Accuracy	Recall	F1-Score
Random Forest	0.808	0.804	0.792	0.804
SVM	0.802	0.790	0.776	0.794
Logistic Regression	0.791	0.780	0.760	0.781
XGBoost	0.785	0.771	0.740	0.770
KNN	0.758	0.756	0.745	0.756
Naive Bayes	0.787	0.750	0.679	0.748
Decision Tree	0.711	0.720	0.714	0.715

6. CONCLUSIONS

The landslide is one of the most devastating Disaster among all natural hazards. Thus to save human life, money, nature is utmost priority. The study proposes a prototype of landslide prediction triggered by rainfall. The paper explores and exploits machine learning dimensions to predict landslide. The paper illustrates two approaches, in first approach a real time data has been captured by deploying the sensor nodes at around the landslide prone area of Uttarakhand region. Due to limited data size, only logistics regression machine learning model has been explored and exploited. An optimized methodology developed for shallow landslide prediction. It has been developed by considering data from various regions like Uttarakhand, Himachal Pradesh, Rudraprayag, and Kedarnath. With the help of IoT based technologies, we have developed a prototype that can help us to prevent major loss of lives and property across the study region. In order to collect the information we require from any location, the IoT application comes equipped with high-quality sensors. We have suggested the software integration with the proposed algorithm to analyze the parameters that can impact the meteorological conditions that cause landslip occurrence and to estimate the level of critical range via threshold and notify the outcome. The relation between the threshold and alert system can only be considered by referring to the data spatially obtained on rainfall and landslides that actually happened. This system helps us to ensure civil safety in indicating the alert level. In the present work, only rainfall was considered a key triggering factor for landslides. There is another factor also, which are contributors in landslide triggering viz. (Soil moisture, temperature, and ground vibration). Sensors data other than rainfall sensors can be used further for better landslides monitoring and prediction as a future work. In second approach, rainfall triggered landslide datasets have been collected from the NASA and IMD websites. After preprocessing and feature selection techniques, the dataset has been feed to several machine learning algorithms like Random Forest, SVM, KNN, Naive Bayes etc. After exploration and exploitation of diverse domain it may infer that Oversampling

performs better than Undersampling for the considered dataset, and provided the best outcome among for all the models. But still forecasting of weather is very much a tough task. We require more detailed data, to predict future conditions. But as for now, the Random Forest algorithm outperforms all the models in terms of every evaluation Parameters like, accuracy, precision, recall, and F1-score. In future, we can apply different models on the data “Location” wise, because rainfall is highly depending upon the atmosphere of particular locations.

REFERENCES

- [1] Baum, R.L., Godt, J.W. (2010). Early warning of rainfall-induced shallow landslides and debris flows in the USA. *Landslides*, 7: 259-272. <https://doi.org/10.1007/s10346-009-0177-0>
- [2] Bayo, A., Antolín, D., Medrano, N., Calvo, B., Celma, S. (2010). Early detection and monitoring of forest fire with a wireless sensor network system. *Procedia Engineering*, 5: 248-251. <https://doi.org/10.1016/j.proeng.2010.09.094>
- [3] Bhardwaj, G.S., Metha, M., Ahmed, Y., Chowdhury, M.A.I. (2014). Landslide monitoring by using sensor and wireless technique: A review. *International Journal of Geomatics and Geosciences*, 5(1): 1-8.
- [4] Georgieva, K., Smarsly, K., König, M., Law, K.H. (2012). An autonomous landslide monitoring system based on wireless sensor networks. In *Computing in Civil Engineering*, 2012: 145-152. <https://doi.org/10.1061/9780784412343.0019>
- [5] Hloupis, G., Stavrakas, I., Triantis, D. (2010). Landslide and flood warning system prototypes based on wireless sensor networks. In *EGU General Assembly Conference Abstracts*, 14166.
- [6] Hou, S.S., Li, A., Han, B., Zhou, P.G. (2013). An early warning system for regional rain-induced landslide hazard. *International Journal of Geosciences*, 4(3): 584-587. <https://doi.org/10.4236/ijg.2013.43053>
- [7] Huang, A.B., Lee, J.T., Ho, Y.T., Chiu, Y.F., Cheng, S.Y. (2012). Stability monitoring of rainfall-induced deep landslides through pore pressure profile measurements. *Soils and Foundations*, 52(4): 737-747. <https://doi.org/10.1016/j.sandf.2012.07.013>
- [8] Mittal, S., Singh, B. (2014). An instrumental study for the development of time based predictive algorithm for Jhakri (Bari village) landslide site. In *Indorock-2014: 5th Indian Rock Conference*, 588-594.
- [9] Mittal, S.K., Singh, M., Kapur, P., Sharma, B.K., Shamshi, M.A. (2008). Design and development of instrumentation for landslide monitoring and issue. *Journal of Scientific and Industrial Research (JSIR)*, 67(5): 361-365.
- [10] Montrasio, L., Schiliro, L., Terrone, A. (2016). Physical and numerical modelling of shallow landslides. *Landslides*, 13: 873-883. <https://doi.org/10.1007/s10346-015-0642-x>
- [11] Nguyen, C.D., Tran, T.D., Tran, N.D., Huynh, T.H., Nguyen, D.T. (2015). Flexible and efficient wireless sensor networks for detecting rainfall-induced landslides. *International Journal of Distributed Sensor Networks*, 11(11): 235954. <https://doi.org/10.1155/2015/235954>
- [12] Shukla, S.K., Chaulya, S.K., Mandal, R., Kumar, B., Ranjan, P., Mishra, P.K., Prasad, G.M., Dutta, S., Priya, V., Rath, S., Buragohain, K., Sarmah, P.C. (2014). Real-time monitoring system for landslide prediction using wireless sensor networks. *International Journal of Modern Communication Technologies and Research (IJMCTR)*, 2(12): 14-19.
- [13] Qiao, G., Lu, P., Scaioni, M., Xu, S.Y., Tong, X.H., Feng, T.T., Wu, H.B., Chen, W., Tian, Y.X., Li, R.X. (2013). Landslide investigation with remote sensing and sensor network: From susceptibility mapping and scaled-down simulation towards in situ sensor network design. *Remote Sensing*, 5(9): 4319-4346. <https://doi.org/10.3390/rs5094319>
- [14] Prakasham, C., Arvanith, R., Kanwar, V.S., Nagarajan, B. (2021). Design and Development of Real-time landslide early warning system through low cost soil and rainfall sensors. *Second International Conference on Aspects of Materials Science and Engineering (ICAMSE 2021)*, 45: 5649-5659. <https://doi.org/10.1016/j.matpr.2021.02.456>
- [15] Rosi, A., Berti, M., Bicocchi, N., Castelli, G., Corsini, A., Mamei, M., Zambonelli, F. (2011). Landslide monitoring with sensor networks: Experiences and lessons learnt from a real-world deployment. *International Journal of Sensor Networks*, 10(3): 111-122. <https://doi.org/10.1504/IJSNET.2011.042195>
- [16] Lee, C.F., Huang, C.M., Tsao, T.C., Wei, L.W., Huang, W.K., Cheng, C.T., Chi, C.C. (2016). Combining rainfall parameter and landslide susceptibility to forecast shallow landslide in Taiwan. *Geotechnical Engineering Journal of the SEAGS & AGSSEA*, 47(2): 72-82.
- [17] Bhatta, N.P., Natarajan, T. (2017). A review on environmental sensors used for landslide prediction and detection. *Journal of Environmental Engineering and Studies*, 2(2): 1-8.
- [18] Tsangaratos, P., Ilia, I. (2014). A supervised machine learning spatial tool for detecting terrain deformation induced by landslide phenomena. In *Proceedings of the 10th International Congress of the Hellenic Geographical Society*, 22-24.
- [19] Rachel, N., Lakshmi, M. (2016). Landslide prediction with rainfall analysis using support vector machine. *Indian Journal of Science and Technology*, 9(21): 1-6. <https://doi.org/10.17485/ijst/2016/v9i21/95275>
- [20] Aggarwal, S., Mishra, P.K., Sumakar, K.V.S., Chaturvedi, P. (2018). Landslide monitoring system implementing IOT using video camera. In *2018 3rd International Conference for Convergence in Technology (I2CT)*, IEEE, pp. 1-4. <https://doi.org/10.1109/I2CT.2018.8529424>
- [21] Othman, A.A., Gloaguen, R., Andreani, L., Rahnama, M. (2018). Improving landslide susceptibility mapping using morphometric features in the Mawat area, Kurdistan Region, NE Iraq: Comparison of different statistical models. *Geomorphology*, 319: 147-160. <https://doi.org/10.1016/j.geomorph.2018.07.018>
- [22] Singh, V.K., Vishal, V., Angara, K.P., Singh, T.N. (2019). Shallow landslides monitoring using the internet of things and machine learning technique. In *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, pp. 609-613. <https://doi.org/10.1109/ICSSIT46314.2019.8987809>
- [23] Tien Bui, D., Ho, T.C., Pradhan, B., Pham, B.T., Nhu, V.H., Revhaug, I. (2016). GIS-based modeling of rainfall-induced landslides using data mining-based

- functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. *Environmental Earth Sciences*, 75: 1-22. <https://doi.org/10.1007/s12665-016-5919-4>
- [24] Sreevidya, P., Abhilash, C.S., Paul, J., Rejithkumar, G. (2021). A machine learning-based early landslide warning system using IoT. In 2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE), IEEE, pp. 1-6. <https://doi.org/10.1109/ICNTE51185.2021.9487669>
- [25] Kumar, N., Ramesh, M.V. (2022). Accurate IoT based slope instability sensing system for landslide detection. *IEEE Sensors Journal*, 22(17): 17151-17161. <https://doi.org/10.1109/JSEN.2022.3189903>
- [26] Fatimah, P., Irawan, B., Setianingsih, C. (2020). Design of landslide early warning system using fuzzy method based on android. In 2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE), IEEE, pp. 350-355. <https://doi.org/10.1109/ICITEE49829.2020.9271676>
- [27] Hemalatha, T., Ramesh, M.V., Rangan, V.P. (2019). Effective and accelerated forewarning of landslides using wireless sensor networks and machine learning. *IEEE Sensors Journal*, 19(21): 9964-9975. <https://doi.org/10.1109/JSEN.2019.2928358>
- [28] Bai, D.X., Lu, G.Y., Zhu, Z.Q., Zhu, X.D., Tao, C.Y., Fang, J. (2022). A hybrid early warning method for the landslide acceleration process based on automated monitoring data. *Applied Sciences*, 12(13): 6478. <https://doi.org/10.3390/app12136478>
- [29] Labade, A., Gupta, B., Gupta, R.K., Kumar, A. (2023). Machine Learning-Based Prototype Design for Rainfall Forecasting. In: Ramdane-Cherif, A., Singh, T.P., Tomar, R., Choudhury, T., Um, JS. (eds) *Machine Intelligence and Data Science Applications. MIDAS 2022. Algorithms for Intelligent Systems*. Springer, Singapore. https://doi.org/10.1007/978-981-99-1620-7_13
- [30] Kumar, A., Misra, R., Singh, T.N., Singh, V. (2023). Landslide Detection with Ensemble-of-Deep Learning Classifiers Trained with Optimal Features. In: *Advances in Data Science and Artificial Intelligence ICDSAI 2022. Springer Proceedings in Mathematics & Statistics*, vol 403. Springer, Cham. https://doi.org/10.1007/978-3-031-16178-0_21.
- [31] Wieczorek, G., Snyder, J.B. (2009) Monitoring Slope Movements. In Young, R., and Norby, L., *Geological Monitoring: Boulder, Colorado*, Geological Society of America, pp. 245-271.
- [32] Lacasse, M., Cornick, S.M., Lacasse, M.A. (2007). Simulation of wind-driven rain effects on the performance of a stucco-clad wall. In *Proceedings of Thermal Performance of the Exterior Envelopes of Whole Buildings X International Conference*, pp. 2-7.
- [33] Theodoulidis, N., Roumelioti, Z., Panou, A.A., Savvaidis, A., Kiratzi, A.A., Grigoriadis, V.N., Dimitriu, P., Chatzigogos, T. (2006). Retrospective prediction of macroseismic intensities using strong ground motion simulation: The case of the 1978 Thessaloniki (Greece) earthquake (M6. 5). *Bulletin of Earthquake Engineering*, 4(2): 101-130. <http://doi.org/10.1007/s10518-006-9001-6>
- [34] Senthivel, S., Chidambaranathan, M. (2022). Machine learning approaches used for air quality forecast: A review. *Revue d'Intelligence Artificielle*, 36(1): 73-78. <https://doi.org/10.18280/ria.360108>
- [35] Parthiban, S.N., Amudha, P., Sivakumari, S.P. (2022). Exploitation of advanced deep learning methods and feature modeling for air quality prediction. *Revue d'Intelligence Artificielle*, 36(6): 959-967. <https://doi.org/10.18280/ria.360618>