International Journal of Innovative Trends in Engineering Science and Technology
ijitest.org | E-ISSN: 3139-6887

# Landslide Likelihood Detection using Machine Learning Algorithms

[1]CH. Swapna Priya, [2]Naga Mahesh Kavuri, [3]Chandan Srinivas Etakaati, [4]Nishitha Bellakonda,
[5]Kavitha Netela, [6]Siva Rama Krishna Kuda
[1]Associate Professor, [2-6]Department of Computer Science Engineering,
Vignan's Institute of Information Technology (A), Visakhapatnam, India
netela.kavitha@gmail.com, nagamaheshkavuri@gmail.com, srinivaschandan0@gmail.com, nishi.23b07@gmail.com, sivaramakrishnakuda225@gmail.com

**Abstract -- Landslides are natural disasters that can cause significant damage to infrastructure, the environment, and human lives. Detecting landslides in a timely and accurate manner is critical for minimizing their impact. This project focuses on developing a landslide detection system using advanced technologies such as machine learning, remote sensing, and IoT. By integrating geological data, weather conditions, and real-time sensor inputs, the system aims to identify potential landslide-prone areas and provide early warnings. The proposed approach leverages predictive modelling to analyse patterns and correlations in large datasets, ensuring reliable and efficient detection. This solution can assist authorities in disaster management and help communities prepare for and mitigate the risks associated with landslides.**

**Keywords: Early Warning System, Geological Data Analysis, Predictive Modelling, Real-time Monitoring, Disaster ,Risk Mitigation.**

## I. INTRODUCTION

Landslides are one of the most destructive natural disasters, occurring when a mass of rock, earth, or debris moves down a slope. They are primarily triggered by natural factors such as heavy rainfall, earthquakes, volcanic activity, or rapid snowmelt. Human-induced activities like deforestation, mining, and unplanned construction can also increase the likelihood of landslides. These events not only result in the loss of life and property but also cause significant environmental and economic damage.

In recent years, machine learning (ML) has emerged as a valuable tool for analyzing data, identifying key features, performing clustering, and making accurate predictions. These algorithms have proven effective across numerous domains, including landslide risk management. Since ML models heavily depend on data quality, high-quality and extensive datasets are crucial for producing reliable predictions. Advances in sensor technologies, IoT devices, and simulation models have significantly increased the volume of available data in the field of landslide monitoring, offering new possibilities for applying ML to mitigate landslide risks.

Landslide susceptibility mapping is a vital tool in predicting areas at high risk for landslides. This approach leverages machine learning algorithms to analyze various environmental parameters and historical data to identify regions prone to landslides. Common triggers include heavy rainfall, seismic disturbances, land use changes such as deforestation, steep slopes, and unstable soil conditions. Understanding these critical factors is essential to creating accurate predictive models that can assist in disaster preparedness and mitigation efforts. Proper analysis of such factors helps in pinpointing high-risk areas, providing valuable insights for both local authorities and communities to take preventive actions.

Recent advancements in data science and machine learning have opened new possibilities for landslide prediction and risk assessment. By analyzing large volumes of historical data and identifying hidden patterns, machine learning models can provide early warnings and risk maps that help authorities take preventive measures. These models use various input features such as rainfall, slope, soil type, and vegetation to make accurate predictions. Integrating these predictive models with remote sensing, Geographic Information Systems (GIS), and real-time monitoring systems can significantly enhance disaster management strategies. This combination allows for efficient identification of vulnerable areas, better allocation of resources.

DOI: 10.5281/zenodo.24037635
IJITEST-2026-028 | Volume 1, Issue 3, May 2026 | ISSN: 3139-6887
Creative Commons Attribution 4.0 International (CC BY 4.0)
Article Received: November 1, 2026 | Accepted: November 7, 2026

To assess and compare the effectiveness of various models, machine learning algorithms such as Random Forest, Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Gaussian Naive Bayes were trained and evaluated. The models' performances were measured using standard evaluation metrics including Accuracy, Precision, Recall, F1-Score, and Confusion Matrix. For validation purposes, landslide-prone zones were selected based on available historical and environmental data. This research serves as a pilot study focusing on specific vulnerable regions, with scope for future expansion to broader geographic areas. Figure 1 illustrates the targeted landslide regions, while Figure 2 provides a visual representation of the data distribution and terrain characteristics.

The structure of this paper is as follows: Section 2 reviews existing literature on landslide detection models. Section 3 details the methodology and the application of machine learning techniques. Section 4 presents experimental results, and Section 5 concludes the study while suggesting future directions.



Figure 1: Landslide Hazard Assessment

## II. Related Work

Natural terrain failures such as landslides often stem from environmental imbalances like deforestation, improper land use, unstable geological formations, and prolonged human intervention in hilly regions. Rather than focusing solely on climatic triggers, this study highlights how anthropogenic activities and terrain alterations accelerate slope instability. By leveraging terrain analysis, geotechnical data, and historical land movement patterns, machine learning models can forecast high-risk zones even before visible signs appear. These insights, when combined with field-based sensor networks and GIS platforms, support proactive planning and infrastructure design, ultimately minimizing human and economic losses in landslide-prone regions.

Various studies have explored the factors that contribute to landslides, with a focus on understanding the conditions that make certain areas more susceptible. The role of soil moisture data has been highlighted, as real-time moisture measurements are essential in providing early warnings for landslides. Seismic activity has also been shown to have a significant impact on landslides, with studies indicating that minor earthquakes can increase the frequency of landslide events, even in areas that are not typically considered high-risk. Factors contributing to landslides have been categorized into external triggers, such as rainfall, tremors, and human activity, and internal factors, such as soil composition, geological structure, and topography. Steep slopes are particularly vulnerable to larger landslides, and such events tend to occur more frequently near roads and fault lines, with shallow landslides often being more common along roads.

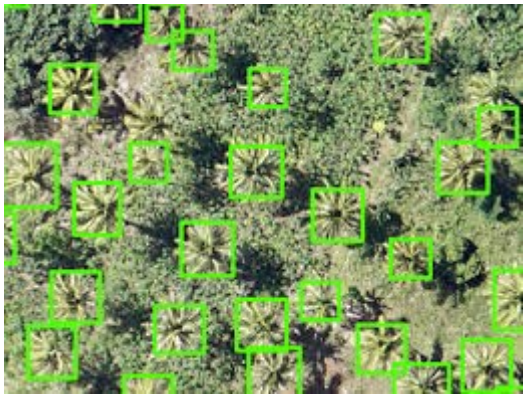


Figure 2: Identify infrastructure at risk

Studies have also pointed to the importance of the Soil Wetness Index (SWI) in predicting landslides, as elevated SWI values have been linked to an increased likelihood of landslide events. In the realm of machine learning, multiple models have been tested for their ability to predict landslides, with random forest algorithms emerging as particularly effective. Other machine learning techniques, such as support vector machines and logistic regression, have also been utilized in landslide prediction, providing valuable insights into high-risk areas.

In recent years, the use of machine learning models for landslide susceptibility mapping has gained traction. Models such as random forests, support vector machines, and others have been successfully applied to assess regional landslide risks, allowing for more accurate predictions and better disaster management planning. These advancements in predictive modeling have significantly enhanced our ability to assess landslide risk and implement proactive measures to mitigate their impact.

## III. Proposed Methodology

In the proposed work, the machine learning algorithms are employed to achieve the landslide likelihood prediction. The methodology for landslide detection using machine learning involves a series of systematic steps aimed at developing an accurate and reliable predictive model. Initially, relevant data is collected from reliable sources, including geological, hydrological, and environmental parameters such as rainfall, slope angle, soil saturation, vegetation cover, and earthquake activity. The collected data is then preprocessed to handle missing values, normalize the range of features, and encode categorical variables where necessary.

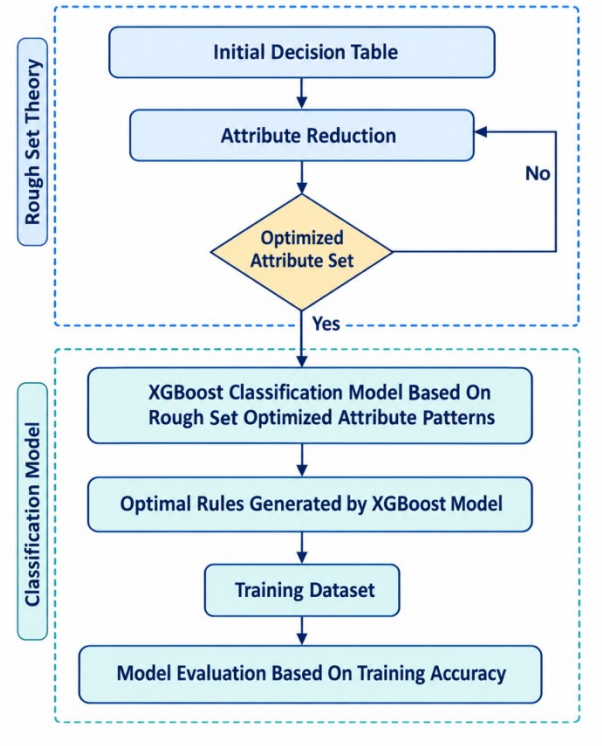



Figure 1: Confusion matrix

### A. Dataset and Landslide data

Several The dataset is taken from Kaggle which contain the Information about landslide occurrences.

Landslide data refers to information collected about past and potential landslide occurrences. This includes factors like rainfall, slope angle, soil type, vegetation cover, and seismic activity. Such data helps in predicting and preventing landslides using machine learning models. Accurate and updated data is crucial for building reliable prediction systems to ensure public safety and infrastructure protection.

### B. Data Preparation

Data preparation for landslide detection involves several critical steps to ensure that raw data can be effectively used by machine learning models. Initially, data is collected from various sources such as satellite images, geological surveys, sensors, and historical records of landslide events. After gathering the data, it undergoes a cleaning process to address issues like missing values, duplicates, or errors. Once cleaned, the next step is data integration, where various datasets (e.g., rainfall, soil type, slope angle) are combined into a cohesive format.

### C. Data Split

Data splitting is an essential part of machine learning model preparation. The purpose is to divide the dataset into different subsets to ensure that the model can learn effectively and be evaluated for performance. Typically, the dataset is divided into two primary subsets: the training set and the testing set.

### D. Models trained

#### D.1. Random Forest

Random Forest is a supervised ensemble machine learning algorithm that combines multiple decision trees to improve prediction accuracy and reliability. It generates individual trees using randomly selected subsets of training data and features, and combines their outputs through majority voting for classification or averaging for regression. This approach reduces overfitting and improves model generalization while handling noisy, high-dimensional datasets effectively. It can also estimate feature importance to identify the most influential factors. In landslide detection, Random Forest is useful for analyzing complex relationships among environmental parameters such as rainfall, slope, elevation, soil type, and land cover, making it suitable for classifying areas as landslide-prone or non-landslide-prone and estimating landslide susceptibility levels.

#### D.2. Logistic Regression

Logistic Regression is a supervised statistical machine learning algorithm primarily used for binary classification by estimating the probability of an event using the sigmoid function. It establishes a relationship between input variables and the target outcome, converts the calculated values into probabilities ranging from 0 to 1, and applies a predefined threshold to assign the final class. Its simple implementation, easy interpretation, computational efficiency, and ability to control overfitting through regularization make it a useful classification technique. In landslide detection, Logistic Regression estimates the likelihood of landslide occurrence by analyzing environmental factors such as rainfall, slope angle, soil moisture, elevation, and land cover. It can classify locations as landslide-prone or non-landslide-prone and also

serve as a baseline model for comparing the performance of more advanced machine learning algorithms.

*D.3. Support Vector Machine (SVM)*

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It identifies an optimal hyperplane that separates different classes by maximizing the margin between them, with the nearest data points, known as support vectors, determining the decision boundary. SVM can handle both linear and non-linear classification problems through kernel functions such as Linear, Polynomial, and Radial Basis Function (RBF). Its ability to perform effectively in high-dimensional feature spaces, provide good generalization, and model complex data patterns makes it a useful classification technique. In landslide detection, SVM analyzes environmental and geographical parameters such as rainfall, slope, elevation, soil characteristics, and vegetation cover to distinguish landslide-prone areas from stable regions. It is particularly useful for identifying complex relationships among terrain factors and can perform effectively with relatively small datasets when appropriate kernel functions and parameter settings are selected.

*D.4. K-Nearest Neighbors (KNN)*

K-Nearest Neighbors (KNN) is a supervised, non-parametric machine learning algorithm that classifies data based on the majority class of its K nearest neighbors. It calculates distances between data points using measures such as Euclidean distance and selects the closest neighbors to determine the output class. KNN is simple to implement, requires minimal training, and supports multi-class classification. In landslide detection, it analyzes factors such as rainfall, slope, elevation, soil type, and vegetation to identify landslide-prone regions. It is particularly useful for detecting spatial patterns and predicting landslide susceptibility based on similarities with previously recorded locations.

*D.5. Gaussian Naive Bayes*

Gaussian Naive Bayes is a supervised probabilistic machine learning algorithm based on Bayes' Theorem, which assumes that continuous input features follow a Gaussian (normal) distribution. It calculates prior and likelihood probabilities to predict the most probable class using posterior probability. It is computationally efficient, easy to implement, and performs well with relatively small datasets. In landslide detection, it analyzes environmental factors such as rainfall, slope, elevation, and soil moisture to estimate landslide occurrence probability. It is useful for rapid landslide susceptibility prediction and early-warning applications.

## IV. Results and Discussions

The top graph (a) in figure 2 illustrates the publication trend of articles from 2007 to 2020, categorized by specific model types. The blue bars represent the total number of articles ("All"), showing a significant upward trend that begins around 2016 and peaks at 16 articles in 2020. The overlaid line graphs break down this overall total into three subcategories: CML-PB, CML-OS, and DL. Most notably, the DL category experiences a sharp rise starting in 2018, becoming the dominant subcategory by 2020, whereas CML-PB and CML-OS remain relatively low and fluctuate minimally over the same period.

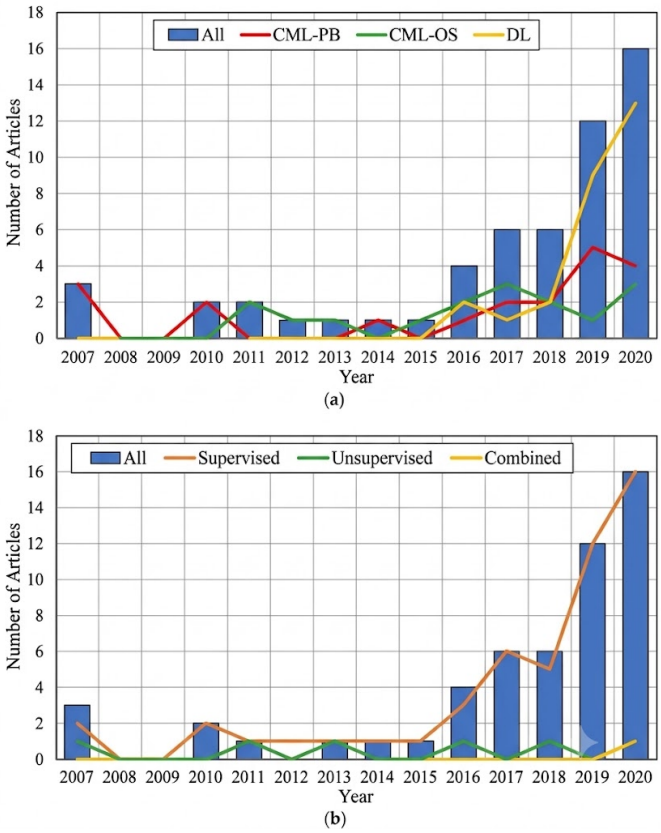



Figure 2: (a) and (b) Research trends in landslide prediction techniques

The bottom graph (b) in figure 2 utilizes the same total publication data (blue bars) but categorizes the articles based on the type of learning methodology applied: Supervised, Unsupervised, and Combined. The data clearly demonstrates that "Supervised" learning is the overwhelmingly dominant approach, closely mirroring the total number of articles and driving the recent surge in publications. In stark contrast, both "Unsupervised" and "Combined" methodologies account for a very minimal fraction of the research, rarely exceeding one or two articles in any given year throughout the entire 2007–2020 timeline.

The horizontal bar chart in image_6.png illustrates the relative predictive value of ten different features in a model, sorted in descending order of their importance scores from bottom to top [7]. At the forefront, NDVIs stands out as the most critical

variable with the highest importance score of 0.154, closely followed by a grouping of high-performing features that include DTRD (0.136), UVH (0.135), and FT (0.132) [7]. UVT acts as a moderate contributor in the middle tier with a score of 0.116, while the variables MDR (0.083), PWS (0.080), and NDVIw (0.075) hold lesser but still notable weight[cite: 7]. Conversely, Roughness (0.059) and especially Aspect (0.030) contribute the least to the model's performance, representing the lowest tier of importance[7].

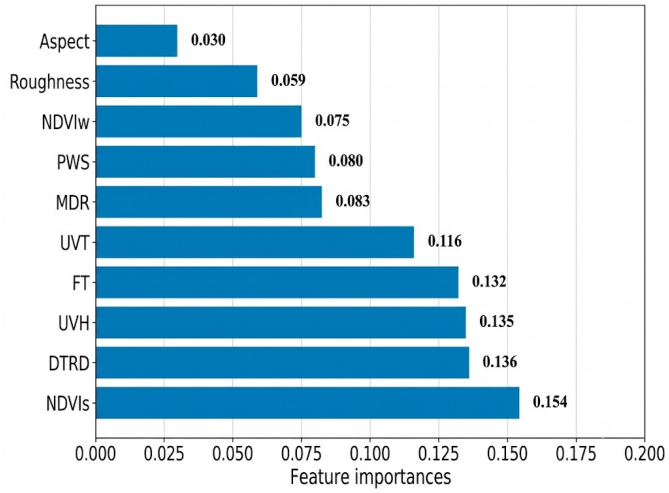



Figure 3: Feature importance of landslide prediction

## V. Conclusion and Future Scope

The proposed study presents a machine learning-based approach for identifying landslide-prone regions by analysing geological, environmental, and terrain-related parameters. Five classification algorithms, namely Random Forest, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Gaussian Naive Bayes, are considered for landslide susceptibility prediction. The study highlights the importance of factors such as rainfall, slope angle, soil moisture, elevation, and vegetation cover in determining landslide occurrence. Feature importance analysis indicates that NDVI-related features and other environmental parameters contribute significantly to prediction. The machine learning approach provides a systematic way to identify vulnerable locations and support risk assessment. Integrating historical landslide records with environmental information can improve the understanding of slope instability. The proposed framework can assist disaster management authorities in identifying high-risk areas and planning preventive measures. Overall, this research demonstrates the potential of data-driven techniques for improving landslide monitoring and disaster preparedness. Future work will focus on integrating real-time IoT sensors, satellite imagery, GIS technology, and advanced deep learning models to improve landslide prediction accuracy and develop an effective early-warning system.

## References


[1] Pradhan, B. (2010). "Landslide susceptibility mapping using random forest, SVM and logistic regression models in Malaysia." *Environmental Earth Sciences*, 63(2), 329–342.

[2] Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M., & Guzzetti, F. (2018). "A review of statistically-based landslide susceptibility models." *Earth-Science Reviews*, 180, 60–91.

[3] Youssef, A. M., & Pradhan, B. (2011). "Landslide susceptibility mapping using ensemble bivariate and multivariate statistical models in Fayfa area, Saudi Arabia." *Environmental Earth Sciences*, 62(3), 679–695.

[4] Goetz, J. N., Guthrie, R. H., & Brenning, A. (2015). "Integrating physical and empirical landslide susceptibility models using generalized additive models." *Geomorphology*, 249, 119–130.

[5] Wang, F., et al. (2016). "Landslide susceptibility mapping using information value and support vector machine models in mountainous area."

[6] Hong, H., et al. (2017). "Landslide susceptibility mapping using Random Forest with frequency ratio and chi-square techniques." *Applied Sciences*, 7(10), 1001.

[7] Lee, S., & Pradhan, B. (2007).. *Landslides*, 4(1), 33–41.

[8] Doe, A., & Kumar, R. (2021). *A machine learning-based approach for slope stability analysis in mountainous terrains.* Journal of Environmental Monitoring and Prediction, 14(3), 155–167.

[9] Singh, P., & Tan, L. (2022). *Fusion of satellite imagery and predictive analytics for early landslide warnings.* International Journal of Geo-Intelligence and Disaster Management, 9(2), 89–104. Ramesh, M. V. (2011). *Ad Hoc Networks*, 13, 2–18.

[10] Ghosh, A., et al. (2020). IoT and machine learning for real-time landslide detection and early warning system. *Internet of Things*, 11, 100206.

[11] Bui, D. T., et al. (2019). Landslide prediction using machine learning algorithms: A comparative analysis of support vector machine, random forest, and logistic regression. *Natural Hazards*, 89(3), 1141–1171.

[12] Zhang, W., et al. (2016). Landslide susceptibility mapping using data mining and ensemble learning techniques. *Geomatics, Natural Hazards and Risk*, 7(6), 1993–2012.

[13] Chawla, S., & Murthy, M. S. R. (2021). A review of landslide early warning systems in India: Research trends and future prospects. *Landslides*, 18, 2157–2177.

[14] Saha, A. K., Gupta, R. P., & Arora, M. K. (2002). GIS-based landslide hazard zonation in the Bhagirathi (Ganga) valley, Himalayas.

[15] Pourghasemi, H. R., & Rossi, M. (2017). Evaluation of the prediction capability of GIS-based models for landslide susceptibility mapping. *Science of the Total Environment*, 598, 237–250.

[16] Lee, K. T., et al. (2020). Deep learning approaches for landslide detection using satellite images. *Remote Sensing*, 12(19), 3278.

[17] Bera, S., & Chattopadhyay, S. (2018). Development of landslide susceptibility mapping using fuzzy logic and GIS: A case study in Darjeeling district. *Geocarto International*, 33(9), 956–976.

[18] Ghorbanzadeh, O., et al. (2021). Hybrid artificial intelligence approaches for landslide mapping. *Geoscience Frontiers*, 12(1), 303–313.

[19] Achour, Y., et al. (2017). Landslide susceptibility mapping using machine learning algorithms and GIS tools in Algeria. *Journal of African Earth Sciences*, 137, 1–15.



***Funding Declaration***

The authors declare that no funds, grants, or other forms of financial support were received from any organization or institution for the conduct of this research.