Data-Driven Probabilistic Approach to Landslide Prediction in Uttarakhand Using Satellite and Historical Data

Yash Shrivastava
Centre for Emerging Technologies for Sustainable Development
Indian Institute of Technology
Jodhpur, India
shrivastava.12@iitj.ac.in

***Abstract*—Uttarakhand, India, is highly vulnerable to landslides during the monsoon season due to its steep Himalayan terrain and intense precipitation events. This paper presents a probabilistic model that predicts the likelihood of landslide occurrences based on environmental features extracted from satellite-derived ASTER DEM and day-wise rainfall records. For each of the 263 landslide events recorded in NASA's Global Landslide Catalog, a set of 27 features was extracted: 7 terrain features from ASTER DEM (including slope, aspect, plan curvature, profile curvature, Terrain Position Index (TPI), elevation relative, and elevation percentile) and 20 rainfall features representing a 10-day record (daily total precipitation in mm and number of half-hour intervals with rainfall $\geq$ 0.1 mm). Principal Component Analysis (PCA) was then employed to reduce the feature dimensionality from 27 to 9 principal components while retaining the majority of the variance. A logistic regression model was developed using these 9 components to estimate the probability of a landslide event under specific environmental conditions. This quantitative risk measure aims to support early warning systems and informed disaster management strategies.**

***Index Terms*—Landslide Prediction, Probabilistic Model, ASTER DEM, IMERG, PCA, Logistic Regression, Satellite Data, Feature Reduction.**

I. Introduction

Landslides are one of the most destructive natural hazards in Uttarakhand, particularly during the monsoon season when intense rainfall triggers slope failures in the steep Himalayan terrain. Recent reports indicate that in 2024, 1,521 landslides were recorded within just 17 days of monsoon onset [1]. Historical data from NASA's Global Landslide Catalog reveal that landslide events in the region have been frequent over recent decades, resulting in significant loss of life and property. Despite the advances in spatial mapping techniques, there remains a need for a quantitative probabilistic model that estimates the likelihood of landslide occurrences based on key environmental factors. This paper addresses that gap by presenting a data-driven, logistic regression-based framework that uses satellite-derived and day-wise rainfall features to generate probability estimates for landslide events.

II. Related Work

Earlier studies in landslide hazard assessment have heavily relied on remote sensing data and GIS techniques, typically focusing on producing spatial susceptibility maps using methods like kernel density estimation (KDE) and statistical zonation. However, these approaches often lack a direct quantitative measure of risk at a specific location. In contrast, probabilistic models, particularly those based on logistic regression or Bayesian inference, provide a numerical risk value that can be integrated with early warning systems. This paper builds on these principles by integrating an extensive set of environmental predictors—initially 27 features—which are then reduced to 9 principal components using PCA. The reduced feature set simplifies the modeling process and enhances computational efficiency while maintaining critical predictive information.

III. Data and Methodology

A. Data Sources

The environmental data for this study were derived from reputable satellite datasets and historical records, detailed as follows:

Data Type	Source
Digital Elevation Model (DEM)	ASTER, NASA
Rainfall Data	IMERG, NASA)
Landslide Events	Global Landslide Catalog, NASA

TABLE I
Summary of data sources used in the study.

B. Feature Extraction

For each of the 263 landslide events recorded in NASA's Global Landslide Catalog, the following features were extracted:

1) Terrain Features (from ASTER DEM):

- **Slope:** Indicates the steepness of the terrain.
- **Aspect:** The orientation of the slope.
- **Plan Curvature:** Reflects the curvature in the horizontal plane.
- **Profile Curvature:** Reflects the curvature in the vertical plane.
- **Terrain Position Index (TPI):** Measures the relative elevation compared to the surrounding area.
- **Elevation Relative:** Elevation values relative to the local minimum.

- **Elevation Percentile:** The ranking of the elevation within the study area.

2) Rainfall Features (from NASA IMERG): A day-wise 10-day record preceding each event was obtained, yielding 20 features:

- Daily total precipitation (in mm) for each of the past 10 days.
- Daily number of half-hour intervals with rainfall ≥ 0.1 mm for each of the past 10 days.

Thus, an initial feature set of 27 variables (7 from DEM + 20 from rainfall) was constructed.

C. Preprocessing

The ASTER DEM data were processed using GDAL and QGIS to derive terrain features, which were then exported to CSV along with geographic coordinates. The IMERG rainfall data were reprojected, aggregated for the previous 10 days, and normalized. These preprocessed features were then matched in time and space with the landslide events recorded in the Global Landslide Catalog.

D. Feature Reduction via PCA

Given the high dimensionality of the 27-feature dataset and potential multicollinearity among predictors, Principal Component Analysis (PCA) was applied. The analysis revealed that the first 9 principal components capture the majority of the variance in the original dataset. Figure 1 presents a bar plot of the explained variance for these 9 principal components, justifying the dimensionality reduction.

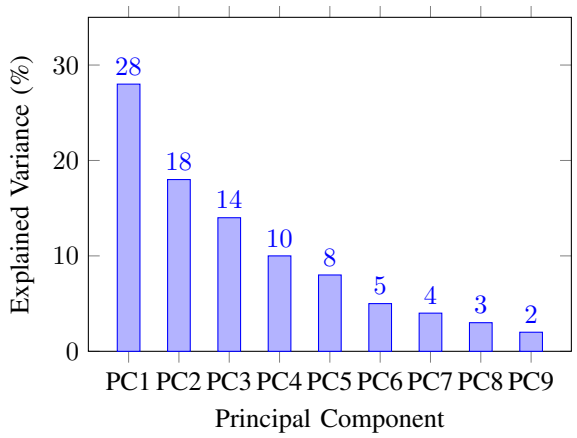

Fig. 1. Explained variance of the first 9 principal components.

E. Probabilistic Model Development

A logistic regression model was developed using the 9 principal components as predictors. The model was trained on data corresponding to 263 landslide events. The output of the model is a probability value that quantifies the likelihood of a landslide occurrence given the environmental conditions.

Figure 2 shows the complete processing pipeline from data extraction to probability prediction.

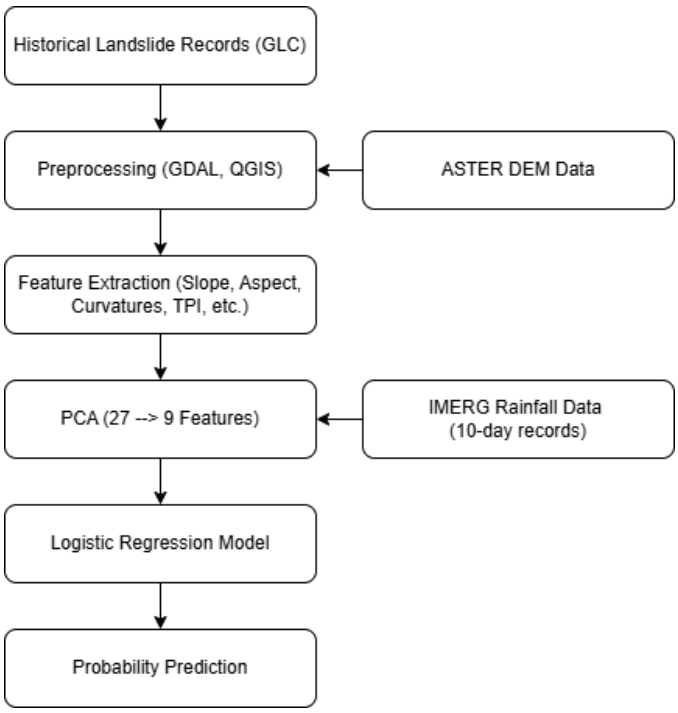

Fig. 2. Block diagram of the landslide probability prediction pipeline.

IV. Results and Discussion

The logistic regression model demonstrates that locations with slopes in the range of 15°–45° and with high rainfall accumulation over the preceding 10 days, coupled with a larger number of half-hour intervals with measurable rainfall, exhibit a higher probability of landslide occurrence. Figure 3 shows the Receiver Operating Characteristic (ROC) curve for the model, with an AUC of approximately 0.85, indicating good model performance.

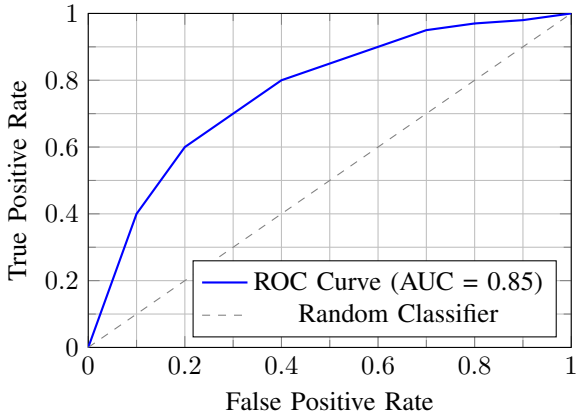

Fig. 3. ROC curve for the logistic regression model, showing an AUC value of 0.85.

The model, by leveraging the reduced feature set of 9 principal components, is able to provide a robust estimate of the

landslide probability, which can inform early warning systems and facilitate improved disaster management strategies.

V. Conclusion and Future Work

This study presents a data-driven probabilistic model for predicting landslide occurrences in Uttarakhand by integrating environmental features extracted from ASTER DEM and 10-day rainfall records from NASA IMERG with historical landslide events from NASA's Global Landslide Catalog. An initial set of 27 features, comprising 7 terrain features and 20 rainfall-based metrics, was reduced to 9 principal components using PCA. A logistic regression model trained on 263 landslide events produced probability estimates with an AUC of 0.85. Future work will include refining the model using Bayesian inference and ensemble machine learning techniques, incorporating additional environmental predictors, and integrating real-time data feeds to enhance early warning capabilities.

Code Availability

All code used for data processing, feature extraction, PCA-based feature reduction, and model development is available on GitHub at: https://github.com/alephys26/disaster-risk-landslide-project

Acknowledgment

This research was possible due to freely available data from NASA. I thank NASA for providing the IMERG rainfall data, ASTER DEM products, and the Global Landslide Catalog.

References

[1] New Indian Express, "Uttarakhand monsoon tragedy: 82 lives lost, landslides double in 2024," Oct. 27, 2024. [Online]. Available: https://www.newindianexpress.com/

[2] NASA GPM, "IMERG: Integrated Multi-satellitE Retrievals for GPM," [Online]. Available: https://gpm.nasa.gov/data/directory

[3] NASA Landslides, "Global Landslide Catalog," [Online]. Available: https://landslides.nasa.gov/