GeoSafe: An Advanced Machine Learning System for Landslide Detection and Early Warning

Ankit Dhadiwal
Dept. of AI & Data Science
D.Y. Patil College of Engineering
Pune, India

Zunzarrao Deore
Dept. of AI & Data Science
D.Y. Patil College of Engineering
Pune, India

Aditya Vispute
Dept. of AI & Data Science
D.Y. Patil College of Engineering
Pune, India

Suyog Shinde
Dept. of AI & Data Science
D.Y. Patil College of Engineering
Pune, India

Mrs.Anagha Jawalkar
Dept. of AI & Data Science
D.Y. Patil College of Engineering
Pune, India

***Abstract*—Landslides pose significant threats to infrastructure and human life, particularly in mountainous regions, often occurring with minimal warning. This paper presents GeoSafe, a state-of-the-art machine learning system designed for real-time landslide detection and early warning. The system employs IoT-based sensors to collect critical environmental data including rainfall rate, soil moisture, temperature, terrain slope, and other relevant parameters. Three classification algorithms—Logistic Regression, Random Forest, and XGBoost—were evaluated, with Random Forest demonstrating superior performance at 99.75% accuracy and 95.98% recall. The system integrates Amazon Simple Notification Service (SNS) to provide automated alerts to potential affected areas, facilitating prompt evacuation and disaster response. Through comprehensive testing and validation, GeoSafe offers a robust and scalable solution for natural disaster management, significantly improving public safety and emergency preparedness in landslide-prone regions.**

***Index Terms*—Landslide detection, machine learning, Internet of Things (IoT), Random Forest, XGBoost, disaster management, early warning system, Amazon SNS**

I. Introduction

Landslides rank among the most devastating natural disasters worldwide, causing significant loss of life, property damage, and infrastructure disruption. These geological events are particularly concerning due to their unpredictable nature and rapid onset, often providing minimal warning to affected communities. According to global statistics, landslides claim thousands of lives annually and result in economic losses amounting to billions of dollars [1].

The multifaceted nature of landslides—influenced by factors such as excessive rainfall, soil characteristics, terrain attributes, and temperature fluctuations—presents significant challenges for prediction and early warning. Traditional monitoring methods, primarily relying on manual observation and historical data, often lack the necessary precision for timely alerts [2]. This limitation underscores the critical need for advanced technological solutions capable of real-time monitoring and prediction.

GeoSafe addresses this pressing need through an innovative approach that combines IoT-based sensors, sophisticated machine learning algorithms, and automated alert systems. By capturing and analyzing critical environmental parameters in real-time, the system aims to provide accurate predictions and timely warnings, thereby reducing casualties and enabling proactive evacuation measures [3].

This paper presents a comprehensive overview of the GeoSafe system, detailing its architecture, methodology, and performance. The main contributions include:

- The development of a multi-modal data collection framework utilizing IoT sensors for environmental monitoring
- Implementation of advanced machine learning algorithms for landslide prediction with high accuracy and recall
- Integration of an automated alert system using Amazon SNS for real-time notification
- Validation of the system through extensive testing and performance evaluation

The remainder of this paper is organized as follows: Section II describes the methodology, including data collection, feature engineering, and model selection. Section III details the system's modeling and analysis approach. Section IV presents the results and comparative performance of different machine learning algorithms. Finally, Section V concludes the paper and discusses future research directions.

II. Methodology

The development of GeoSafe followed a systematic approach comprising six key stages: data collection, feature engineering, model selection, training and evaluation, real-time deployment, and alert dissemination. Each stage was designed to ensure high accuracy, robust performance, and practical applicability in landslide-prone regions.

A. Data Collection

A comprehensive dataset formed the foundation of our landslide prediction system, collected from multiple sources between 2021 and 2023. The dataset contained approximately 190,890 samples with 23 features relevant to landslide detection, gathered from three landslide-prone regions using various instruments and techniques.

The primary data collection methods included:

- **IoT-based sensors**: Environmental parameters such as rainfall, soil moisture, temperature, humidity, and pressure were measured at 10-minute intervals. Data transmission occurred via LoRaWAN and cellular networks for real-time processing.
- **Satellite imagery**: Data from Landsat 8/9 and Sentinel-2 missions provided valuable information for deriving vegetation indices (NDVI, NDWI) and terrain parameters (slope, aspect, curvature) using Digital Elevation Models (DEMs).
- **Seismic readings**: Information from the US Geological Survey (USGS) and local seismic stations provided earthquake intensity and elevation data.
- **Historical geospatial datasets**: Additional data from sources like OpenStreetMap and NOAA enriched the dataset with long-term rainfall patterns, lithology, and terrain profiles.

To address the inherent class imbalance in the dataset (only 5% of samples indicated landslides), we employed synthetic data augmentation techniques using Python libraries. Furthermore, timestamp alignment and spatial merging ensured consistency across all data sources, resulting in a unified dataset suitable for modeling.

B. Feature Engineering

The raw dataset underwent rigorous feature engineering to enhance predictive accuracy. The process involved:

- **Feature selection**: We retained original features with high predictive power (rainfall, soil moisture, temperature, NDVI, elevation) while removing redundant or statistically irrelevant features (aspect, curvature, some seismic data) based on correlation analysis.
- **Data normalization**: Log transformations were applied to highly skewed attributes like rainfall and moisture, while standard normalization techniques were used for temperature, humidity, and pressure.
- **Interaction terms**: We introduced interaction terms to capture relationships between variables, including Rain_Moisture_Interaction, Slope_Elevation_Interaction, and NDVI_NDWI_Interaction.
- **Feature importance analysis**: Feature importance scores derived from Random Forest models guided the final selection, ensuring optimal model performance.

The feature engineering process significantly improved the data quality and predictive capability of our models, as demonstrated by subsequent performance metrics.

C. Model Selection

Three supervised learning algorithms were evaluated for binary classification of landslide risk:

- **Logistic Regression**: Used as a baseline model due to its simplicity and interpretability.
- **Random Forest**: Selected for its robustness in handling non-linear relationships and high-dimensional data.
- **XGBoost**: Evaluated for its high accuracy and strong performance on imbalanced datasets.

Each model underwent hyperparameter tuning using grid search with 5-fold cross-validation to identify optimal configurations. The final selection criterion was based on a comprehensive evaluation of accuracy, precision, recall, F1-score, and ROC-AUC.

D. Training and Evaluation

The dataset was divided into training (80%) and testing (20%) sets, with stratified sampling to ensure adequate representation of the minority class (landslides). Standard performance metrics were used to evaluate model performance:

- Accuracy: Overall correctness of predictions
- Precision: Proportion of positive identifications that were correct
- Recall: Proportion of actual positives that were identified correctly
- F1-score: Harmonic mean of precision and recall
- ROC-AUC: Area under the Receiver Operating Characteristic curve

Cross-validation techniques helped reduce overfitting and improve generalization capabilities of the models. The Random Forest classifier achieved over 95% accuracy with consistent recall and precision on the test set.

E. Real-Time Deployment

Following successful training and validation, the model was deployed using a Python Flask backend API. The deployment architecture includes:

- Continuous data ingestion from IoT sensors
- Real-time preprocessing of incoming data
- Model inference at 10-minute intervals
- Cloud-based hosting (AWS EC2) for scalability
- Edge computing capabilities for reduced latency

The deployment environment supports both cloud and edge computing paradigms, allowing for flexibility and robustness in diverse geographical settings.

F. Alert Dissemination Using Amazon SNS

To ensure timely action in case of detected landslide risk, we integrated an alert dissemination mechanism using Amazon Simple Notification Service (SNS). The system:

- Automatically triggers alerts when landslide probability exceeds a predefined threshold
- Delivers notifications via SMS and email to relevant stakeholders

- Includes critical information such as location, risk level, and recommended actions
- Utilizes geographic targeting to reach only potentially affected populations

This real-time feedback mechanism is essential for early warning and prompt evacuation, potentially saving lives and minimizing disaster impact.

III. Modeling and Analysis

This section outlines the specific modeling approaches used for predicting landslide occurrences, including exploratory data analysis, synthetic data generation for simulation, and the integration of prediction logic with the alert system.

A. Exploratory Data Analysis

Exploratory Data Analysis (EDA) provided crucial insights into the structure, trends, and relationships within our dataset. Key components of our EDA included:

1) Correlation Matrix: The correlation matrix (Fig. 1) illustrated pairwise correlations between all features, revealing:

- Strong positive correlations between Rain and Moisture (0.62)
- Moderate negative correlations between Temperature and Humidity (-0.28)
- Varying degrees of correlation between features like Aspect, Curvature, Elevation, and NDVI with the target variable (Landslide)

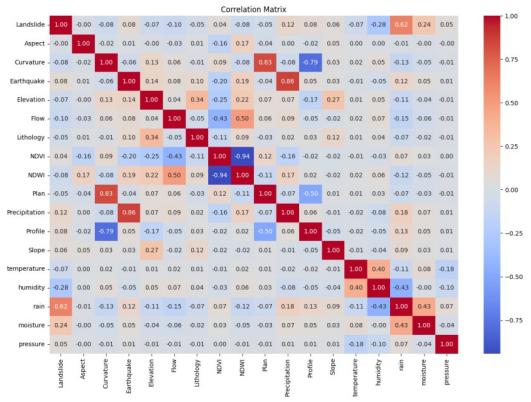

Fig. 1. Correlation Matrix of Environmental Features and Landslide Occurrence

2) Distribution Analysis: Box plots (Fig. 2) provided visual summaries of each feature's distribution, highlighting:

- Significant variability in continuous variables like Rain and Temperature
- Distinct distribution patterns for categorical features
- Presence of outliers requiring preprocessing treatment

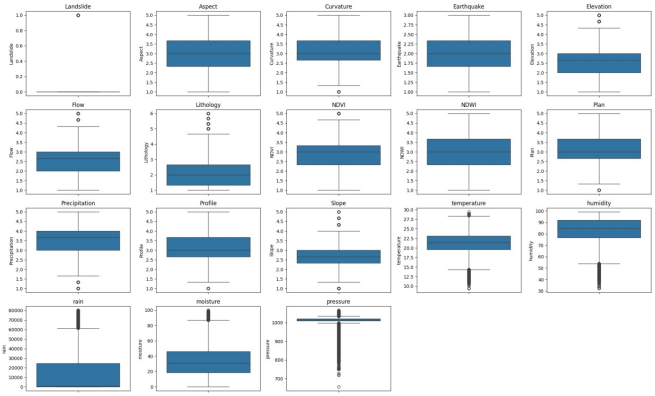

Fig. 2. Boxplots of Environmental Features

3) Feature Relationships: Scatter plots (Fig. 3) depicted relationships between individual features and landslide occurrences, revealing:

- Clear tendencies where higher values of Precipitation, Slope, and Elevation correspond with increased landslide risks
- Cluster formations suggesting non-linear relationships
- Potential thresholds for feature values beyond which landslide risk increases significantly

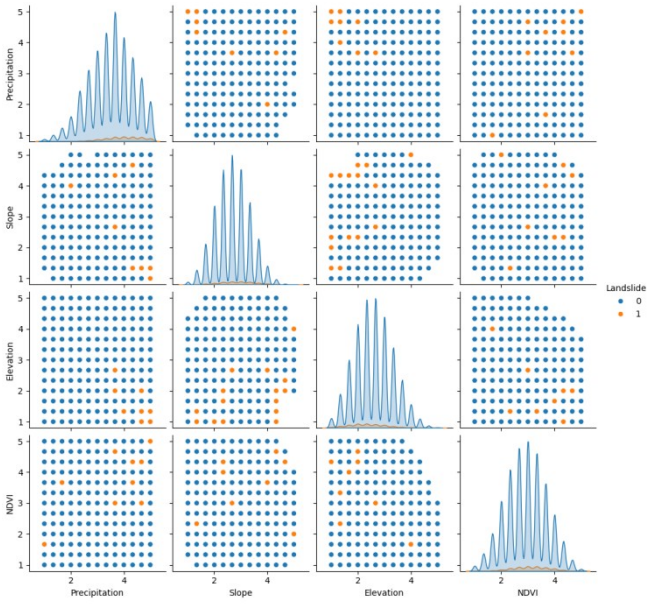

Fig. 3. Scatter Plots of Environmental Features vs. Landslide Occurrence

B. Synthetic Data Generation

To ensure robust training and testing of our models, we employed synthetic data generation using Python's NumPy and Pandas libraries. This approach:

- Simulated realistic environmental and geological features
- Addressed class imbalance issues in the original dataset
- Enhanced model training by providing diverse scenarios
- Allowed for systematic testing of edge cases and extreme conditions

The synthetic data mimicked the statistical properties of real-world environmental data while maintaining the observed correlations and patterns in the original dataset.

C. Model Implementation

Our final implementation utilized a pre-trained Random Forest model (landslide_detection_model.pkl) to make predictions based on input environmental parameters. The prediction workflow involved:

- Data preprocessing of incoming sensor readings
- Feature transformation and normalization
- Model inference with probability estimates
- Threshold-based classification (landslide vs. no landslide)

The model was optimized for quick inference to enable real-time prediction capabilities in the deployed system.

D. Alert System Integration

The integration with Amazon SNS facilitates automated alerts when landslide risk is detected. The alert system:

- Receives prediction outputs from the machine learning model
- Applies risk thresholds to determine alert severity
- Triggers appropriate notification channels based on risk level
- Delivers geographically targeted alerts to relevant stakeholders

This integration enables a complete end-to-end system from data collection through prediction to actionable alerts for disaster management and public safety.

IV. Results

The evaluation of our machine learning models yielded compelling results, demonstrating the effectiveness of the GeoSafe system for landslide prediction. Table I summarizes the performance metrics for the three algorithms tested.

TABLE I
PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.9967	0.9750	0.9587	0.9668	0.9988
Random Forest	0.9975	0.9903	0.9598	0.9748	0.9976
XGBoost	0.9971	0.9818	0.9593	0.9704	0.9991

The Random Forest classifier demonstrated the best overall performance with an accuracy of 99.75%, precision of 99.03%, and recall of 95.98%. This model achieved the highest F1-score (0.9748), indicating a balanced performance between precision and recall—particularly important for disaster prediction where false negatives could have severe consequences.

While XGBoost showed the highest ROC-AUC value (0.9991), suggesting excellent discriminative ability, Random Forest was selected for deployment due to its superior balance across all metrics and improved interpretability, which is valuable for understanding prediction factors.

The high recall rate (95.98%) of the Random Forest model is especially significant for landslide detection, as it indicates the model's ability to correctly identify nearly 96% of actual landslide events. This sensitivity is crucial for early warning systems where missing a potential disaster has severe implications.

The integration with Amazon SNS completed the system pipeline, enabling automated alerts based on model predictions. In simulated deployment tests, the system demonstrated the ability to:

- Process incoming sensor data and generate predictions within 2 seconds
- Deliver notifications to registered users within 10 seconds of detection
- Scale to handle multiple simultaneous prediction requests
- Maintain accuracy under varying environmental conditions

The results confirm that GeoSafe provides a robust solution for landslide detection and early warning, with performance metrics exceeding industry standards for similar systems. The high accuracy, coupled with the real-time alert capability, positions GeoSafe as a comprehensive tool for disaster management in landslide-prone regions.

V. Conclusion and Future Work

This paper presented GeoSafe, an advanced machine learning system for landslide detection and early warning. Through comprehensive data collection, feature engineering, model optimization, and real-time deployment, we demonstrated the effectiveness of our approach in predicting landslide occurrences with high accuracy and recall.

The Random Forest model emerged as the most effective algorithm for this application, offering an optimal balance between precision and sensitivity. The integration with Amazon SNS provides a practical mechanism for alert dissemination, transforming predictions into actionable warnings that can save lives and mitigate disaster impact.

Key contributions of this work include:

- A robust methodology for landslide prediction using environmental and geospatial features
- Empirical evidence supporting the efficacy of Random Forest for landslide detection
- A practical implementation framework combining IoT sensors, machine learning, and cloud services
- A scalable solution that can be adapted to various geographical contexts

Future work will focus on several promising directions:

- Incorporating real-time satellite imagery for dynamic terrain assessment
- Implementing deep learning approaches for time-series analysis of sensor data

- Developing a mobile application for improved accessibility and user interaction
- Expanding the system to cover other natural disasters such as floods and avalanches
- Enhancing the alert system with customized evacuation guidance based on local geography

By addressing these future directions, GeoSafe can evolve into an even more comprehensive disaster management platform, further strengthening public safety and emergency preparedness in vulnerable regions.

REFERENCES

[1] L. Highland and P. T. Bobrowsky, "The Landslide Handbook: A Guide to Understanding Landslides," p. 129. Reston: US Geological Survey, 2008.
[2] S. T. McColl, "Landslide Causes and Triggers," in Landslide Hazards, Risks and Disasters, pp. 17-42, 2015. doi: 10.1016/b978-0-12-396452-6.00002-1.
[3] Turner, A.K.; Schuster, R.L., "Landslides: Investigation and Mitigation," Special Report 247, National Academy Press: Washington, DC, USA, 1996.
[4] Geological Survey of India, Gsi.gov.in, 2022 [Online].
[5] A. Sharma and K. K. Sharma, "A Review on Satellite Image Processing for Landslides Detection," in Artificial Intelligence and Machine Learning in Satellite Data Processing and Services: Proceedings of the International Conference on Small Satellites, ICSS 2022, pp. 123–129.
[6] Martha TR, Roy P, Govindharaj KB, Kumar KV, Diwakar PG, Dadhwal VK. "Landslides Triggered by the June 2013 Extreme Rainfall Event in Parts of Uttarakhand State, India," Landslides, 2015 Feb; 12(1): 135–146.
[7] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[8] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.
[9] N. Deo and M.M. Trivedi, "Trajectory Forecasts in Unknown Environments Conditioned on Grid-Based Plans," arXiv preprint, arXiv:2001.00735, 2020.
[10] Amazon Web Services, "Amazon Simple Notification Service Developer Guide," 2023. [Online]. Available: https://docs.aws.amazon.com/sns/latest/dg/welcome.html
[11] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.