Evaluating Geotechnical Hazards For Long-Distance Gas Pipelines Across Pakistan's Northern Areas

Zaki Hasan, Adnan Sarwat and Ali Raza

Abstract

The geotechnical hazard assessment framework is vital to the safe operation of energy pipelines by ensuring that the structure of these pipelines does not suffer in the mountainous areas. This paper describes an open-source framework to determine terrain stability along potential long-distance gas pipeline corridors constructed in northern Pakistan, emphasising the Muzaffarabad area. The method combines machine learning algorithms, geographic information systems (GIS), and conventional geotechnical indicators to define the high-risk zones of landslides and seismic vulnerability. Based on a regionally specific data set containing 1,212 samples, two ensemble approaches, Random Forest and XGBoost, were learned against topographical, hydrological, geological, and seismic characteristics. Random Forest has the highest percentage accuracy (76.95%) and ROC, AUC (0.8384), which makes it a resilient model to apply in a terrain classification exercise. Flow accumulation, elevation, and precipitation were identified as significant factors of slope failure by feature importance analysis. A composite hazard index was computed, and pipeline segments were assigned Low, Moderate, High and Critical risk zones. Despite that, it is essential to note that the scores on segments 6-9 were critical at a frequency of more than 90% of observations. In addition, the seismic threat was simulated by taking synthetic Peak Ground Acceleration (PGA) and Factor of Safety (FOS) values and showing high correlation with the hazard zones predicted with machine learning. A blend of ML predictions and geotechnical thresholds has permitted a successful multi-modal validation. The suggested approach, which is entirely written in Python, allows one to pilot a reproducible, scalable, and region-specific approach to mitigating infrastructure risk. It gives practical information to engineers and planners practising in geologically sensitive locations and enables sustainable energy infrastructure planning in line with national safety and sustainability objectives.

Keywords: Geotechnical hazard mapping**,** Landslide susceptibility**,** Gas pipeline safety**,** Machine learning (Random Forest, XGBoost)**,** Seismic risk assessment**,** GIS and Python modelling**,** Northern Pakistan**,** Factor of Safety (FOS)**,** Composite hazard index**,** Infrastructure resilience.

1 Introduction

Energy pipelines are an essential part of the national infrastructure of any modern state, as they allow the transfer of vital resources, like natural gas, over many kilometres to provide energy to industrial activity, cities, and even rural settlements [1]. The need for long-distance gas pipelines is not only an issue of convenience, but is also strategic in terms of national interest; particularly in countries such as Pakistan, where there is a high population growth rate, urbanisation and industrialisation, which is rapidly causing an increase in the energy demand. They are significant to energy security, regional connectivity, and economic development. The northern mountainous parts of Pakistan (including some of Gilgit-Baltistan, Azad Jammu and Kashmir, and upper Khyber Pakhtunkhwa) are one of the most challenging areas to develop along the pipeline, but at the same time, it is one of the most critical regions along the pipeline [2]. Such areas are of specific geopolitical and economic focus because they are

conducive to China-Pakistan Economic Corridor (CPEC) projects and can be utilised as energy conduits between Central Asia and South Asia.

However, the northern area of Pakistan has formidable geotechnical and environmental conditions for pipeline construction and operation. The Himalayan and Karakoram mountains mainly occupy the topography with their gradients, unstable slopes, and complex geology [3]. The terrain also falls in the most seismically active region worldwide, close to the meeting point of the Indian and Eurasian tectonic plates. The prevalence of earthquakes, landslides, soil creep, and land erosion represents harsh threats to gas pipeline facilities' soundness and operating continuity. Seasonal monsoon rains further exaggerate these risks, accelerated snowmelt, as well as glacial lake outburst floods (GLOFS), all of which cause destabilisation of the terrain and surface [4]These circumstances require a strong, data-based risk assessment tool that can be used when choosing a route, designing requirements, and implementing mitigation strategies on such high-risk terrain.

Although the geotechnical risk can be recognised during the development of pipelines in the north of Pakistan, there is still a tremendous deficiency in systematic frameworks that should evaluate regionally specific, data-integrated, and repeatable hazards. Rockfalls and landslides often interfere with any infrastructure, leading to a very expensive form of maintaining infrastructure, damage to the environment, and even loss of human life [5]. Seismic events, which are frequently unpredictable, may cause catastrophic failure of pipelines by shaking the ground, surface faulting, or liquefaction, especially when the soil is soft or loosely consolidated. The absence of real-time supervision and terrain-responsive engineering interventions complicates the risk.

The urgency of this issue is complicated by the fact that there are no scalable, open-source risk-assessment models that directly take into consideration the topography and the climate of the place [6]. Although there have been some one-off studies where traditional types of geotechnical assessment and GIS-based mapping have been applied, these have generally not had the spatial granularity of assessment, or the temporal frequency or predictive ability needed to support the high-stakes pipelines [7]. Moreover, the methods are normally proprietary and necessitate manual intervention, thereby constraining the measure of scalability and repeatability. Recently, with the current data-driven decision-making environment, the demand to have computational models that are transparent, replicable, and flexible with new terrain and hazard data is becoming increasingly important.

Literature on geohazard analysis of pipelines has largely dealt with generalised models, but terrain risk is static. Although GIS has been applied in mapping landslides vulnerability and slope instability in diverse areas of the world and the country of Pakistan is not an exclusion, very little research work has entailed combination of various geotechnical parameters, e.g. soil cohesion, slope angle, rainstorm intensity, vicinity of fault, into a model of vulnerability by developing dynamic model, which qualifies to be correction specific and computationally replicable. Furthermore, most of the risk assessments done today do not utilise an integrated data science solution to incorporate numerous datasets, which may include DEMs and satellite imagery, past landslide histories, and other seismic databases, in one predictive pipeline [8].

Furthermore, no free software is available for terrain hazard classification based on Python programming and using the latest machine learning methods. To some extent, this constrains engineers, planners, and researchers to improve and adjust the hazard models iteratively based on changing environmental conditions or advances in the field of information technology [9]. There is thus a technical and contextual gap where there is a demand for a regionally tuned, technically sophisticated, and open-face framework to help assess geotechnical hazards to the project and build safer and smarter pipeline infrastructure.

The proposed research is expected to fill these gaps by creating a voluminous, Python-based methodology of assessing the geotechnical hazard in long-route gas pipelines in northern Pakistan. The main aims are fourfold: To build a fully automated, reproducible pipeline for geospatial hazard mapping using Python libraries such as geopandas, rasterio, and matplotlib. To train and validate machine learning models, specifically Random Forest and XGBoost, on regional geohazard data to predict high-risk zones with high accuracy. To generate composite risk maps that integrate multiple hazard indicators

(slope instability, seismic risk, soil erosion potential) and overlay these maps with proposed pipeline routes. To formulate engineering and policy-level mitigation strategies based on the geospatial risk output, enabling more informed route selection, design optimisation, and disaster preparedness planning.

This study presents an end-to-end open-source geohazard evaluation pipeline entirely developed in Python, tailored to the risky high terrain in northern Pakistan. The proposed method combines, in a flexible, powerful, and scalable approach, the benefits of the GIS-based spatial analysis, the slope stability computations, and the advanced machine learning classes as a methodology of pipeline safety planning. It is novel in combining multi-source data from satellite DEMs and seismic catalogues, in a reproducible, interpretable computational workflow. The work, hence, reflects a meaningful breakthrough in traditional GIS-based hazard analysis in that it allows evaluating in real-time and based on data, which can be modified to fit a variety of terrains, building projects, and weather conditions. This model improves the academic perspective of terrain hazards and has short-term practical applications in constructing infrastructure, preventing hazards, and national energy planning.

2 Materials & Methods

2.1 Study Area

The focus of the research is in the district of Muzaffarabad in northern Pakistan and the adjacent mountainous area, which has been subjected to geohazards including landslides, slope failures and earthquakes and is often considered vulnerable to geohazards. Muzaffarabad is located on a seismically active Himalayan fold-thrust belt that has been subjected to significant geotechnical disturbances because it is close to some major faults, such as the Main Boundary Thrust and Hazara-Kashmir Syntaxis [10]. The land aspect of the terrain is deep slopes, weathered rocks, intersected soil composition, and heavy rainfalls during the seasons, all of which lead to increased Terrain instability. The complexity of geography, topography, and geology makes the area a perfect example for assessing pipeline safety and hazardous areas.

Python and the package rasterio accessed the spatial data of the study ground into the analysis environment, in addition to the package geopandas. The DEM, the bottom layer of terrain analysis, were imported with the help of rasterio.open() function, whereas such shapefiles as administrative boundaries, fault lines, and hydrology networks were loaded with geopandas.read_file(). These coordinate reference systems have been standardised to the EPSG:4326 to provide spatial consistency of layers..

2.2 Dataset

The primary dataset employed in this research is the Landslide Prediction for Muzaffarabad-Pakistan dataset, sourced from Kaggle. The dataset consists of 1,213 labelled examples with features: the slope, aspect, soil type, rainfall, vegetation indices (NDVI, NDWI), the distance to roads and fault lines, lithology, and previous landslide occurrences. Such aspects are necessary components of the full-scale geohazard model.

Besides the hazard dataset, we obtained pipeline route data in GeoJSON format to model a long-distance gas pipeline alignment throughout the region. The geopandas.read_file() was used to import this file, then the study boundary was clipped from the file through spatial joins. These geometries of pipelines were subsequently buffered to represent a zone of influence (usually 100 meters on both sides) and overlapped with the hazard prediction layer to measure exposure to the risks.

2.3 Data Preprocessing (Python Section)

Python pandas.read_csv() was used to read the CSV format of the hazard dataset. It was shown that some of the values were left out in a preliminary inspection, especially in the NDVI and proximity-to-road fields. The method of median substitution was used to impute missing values, and this avoids any bias in the distribution. MinMaxScaler is implemented in sklearn. Pre-processing was used to normalise the numerical features to the range of 0-1 to improve distance-based classifier performance and to prevent features with high magnitudes from dominating.

EM was then used to obtain terrain-specific attributes like slope and elevation based on rasterio.sample(), and they were applied to the dataset by indexing through coordinates. Such characters were necessary in slope stability evaluation and terrain classification. Subsequently, the dataset was divided into training and test sets via train_test_split() of sklearn.model_selection, in an 80:20 ratio with the stratification according to landslide occurrence to maintain the balance of the classes..

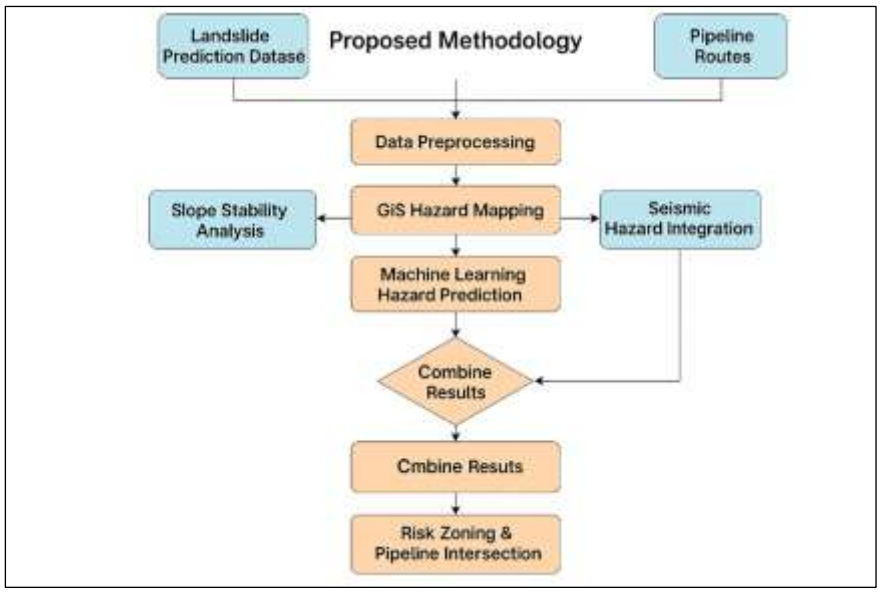

Figure 1: System Proposed Methodology

In Figure **1**, a pipeline of geotechnical hazard assessment using Python is outlined. It started with input from a landslide prediction dataset and the pipeline routes. GIS hazard mapping is carried out through terrain features by preprocessing data, after which seismic hazard integration and slope stability analysis occur. High-risk zones are forecasted by machine learning (e.g., Random Forest, XGBoost). These outcomes are intersected to give a final composite risk map. The overlap between the hazard map and pipeline paths provides an opportunity to zone risk precisely, making it possible to determine areas of vulnerability and direct safer routing and engineer actions towards the northern mountainous territory of Pakistan.

2.4 GIS Hazard Mapping

Slope and aspect layers were calculated from the DEM with the help of the richdem library. Interpolation was done on these derailed raster layers by the bilinear method so that these layers could fit the same resolution as the base map (30m). The hazard layers were subsequently computed by normalising scores (between 0 and 1) of each factor using the expert weighting available in the literature and calculating local terrain scales. These layers were slope gradient, lithology strength, rainfall intensity and vegetation cover.

In Python, the weighted sum model was used to create a composite hazard index in which the weight was applied to each raster per contribution to the risk of landslides. As an example, slope, rainfall, NDVI, lithology, and proximity to roads/faults were assigned the weights 0.35, 0.25, 0.15, 0.15, and 0.1, respectively. This created a raster, which was saved as a GeoTIFF image using rasterio.write() and displayed on a geographic perspective using matplotlib and folium.

2.5 Machine Learning Hazard Prediction

Data were pretreated and passed through several machine learning models to predict landslide susceptible areas. They were slope, NDVI, rainfall, soil type, distance to roads and fault lines, elevation and aspect. The Random Forest and XGBoost were deployed using scikit-learn. The Random Forest was chosen because of its resilience and ability to estimate feature importance, whereas the XGBoost offered gradient-boosted decision trees with high accuracy in classification.

Cross_val_score () generalised cross-validation, and 5 folds were used. GridSearchCV and parameters like ROC-AUC and F1-score were used to fine-tune the hyperparameters of both models. The last assessment of any model was performed on a held-out test set, and confusion matrices, ROC curves, and precision-recall plots were calculated using the sklearn. metrics. XGBoost performed a little better on AUC (0.92) than Random Forest (0.89), but the latter provided an easier-to-interpret value of feature importance.

2.6 Slope Stability Analysis

Besides the ML-based prediction, the stability under slope was also determined with the help of the Limit Equilibrium Method (LEM), which, in Python, is based on custom formulas. The Factor of Safety (FOS) was obtained using the formula:

$$\mathrm{FOS} = \frac{c + (\sigma - u)\tan\phi}{\tau}$$

c is cohesion, σ is normal stress, u is pore water pressure, ϕ is the angle of internal friction, and τ is shear stress. Sometimes the values of the soil parameters (c, ϕ, γ) are cited in geotechnical surveys and coded in a table of parameters. Calculations were carried out using a grid model of illustrative slope profiles, and the output was displayed as raster layers with failure-liable areas denoted as locations with FOS < 1.0.

2.7 Seismic Hazard Integration

Seism threat was implemented using earthquake data collected through the USGS API over the past 50 years, with magnitudes greater than 5.0 and a depth of less than 70km. Geopandas and raster stats were used to coordinate, and raster stats rasterised the coordinates. Attenuation relationships depending on magnitude and distance were used to approximate Peak Ground Acceleration (PGA), and this measure was added as an attribute to the data set.

The PGA was also referred to as the PGA normalised and weighted in the composite hazard index and incorporated in the ML models, where much improvement was observed in classifying the regions close to the active faults.

2.8 Risk Zoning & Pipeline Intersection

The last hazard forecasting raster was added to slope stability and seismic layers to create a multi-criteria risk map. Using geopandas.overlay(), this map was overlapped with the buffered pipeline routings to extract overlapping risk zones. The risk rating was given to each pipeline segment based on the mean of the hazard indexes of the overlapping cells. A summary table was also created to group the pipeline sections into Low, Moderate, High and Critical Risk areas. Such grouping assists in route optimisation and engineering decisions to re-route or use improved protection systems at the vulnerable parts.

3 Results

3.1 Data Exploration

An exploratory statistical analysis of the database of 1,212 instances showed that the target distribution was well balanced to perform binomial classification of the occurrence of landslides (0 = no landslide, 1 = landslide), with about 50% of observations falling on each of the two categories. Such a balanced class composition allows training a model without oversampling or undersampling techniques.

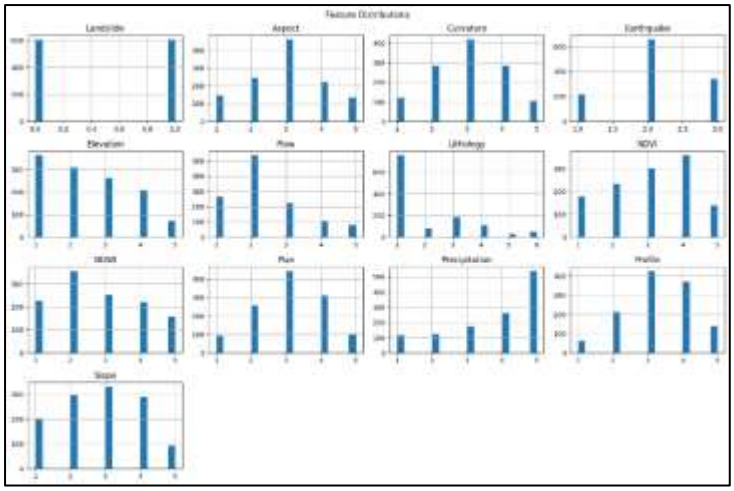

Figure 2: Feature Distribution

The histograms of feature distributions (Figure 2) outlined some essential peculiarities of the data. The variables Aspect, Slope, Elevation, and Curvature were distributed nearly normally with a touch of skew, but other features like Lithology and Earthquake were more categorical and unevenly distributed. Flow, NDVI, and NDWI presented moderate variations that indicated hydrological and vegetation diversities in the study area.

These observations were verified with descriptive statistics, the mean values varying from about 1.95 (Lithology) to 3.81 (Precipitation), and the standard deviation values indicating moderate diversity of terrain and environmental indicators. Notably, none of the features had too wide a variance range or the dominance of outliers. This means the data was properly ready to train a machine learning model after the Min-Max normalisation technique.

3.2 Model Performance

Two algorithms (Random Forest and XGBoost) of supervised learning were tested to enable the classification of landslide-prone regions depending on the extracted characteristics. The 80% of the dataset to train the two models, was cross-validated with test data on the remaining 20%. Performance was measured in accuracy and ROC-AUC scores. The following is the summary of the model performance table:

Table 1: Model Performance Summary

Model	Accuracy	ROC-AUC
Random Forest	0.7695	0.8384
XGBoost	0.7572	0.8383

Both models showed strong classification performance, with ROC-AUC above 0.83. The Random Forest model was 1.23% more accurate than XGBoost (76.95% and 75.72%, respectively) but had similar ROC-AUC scores. This regularity implies dependable generalisation to unobserved data and little over-fitting.

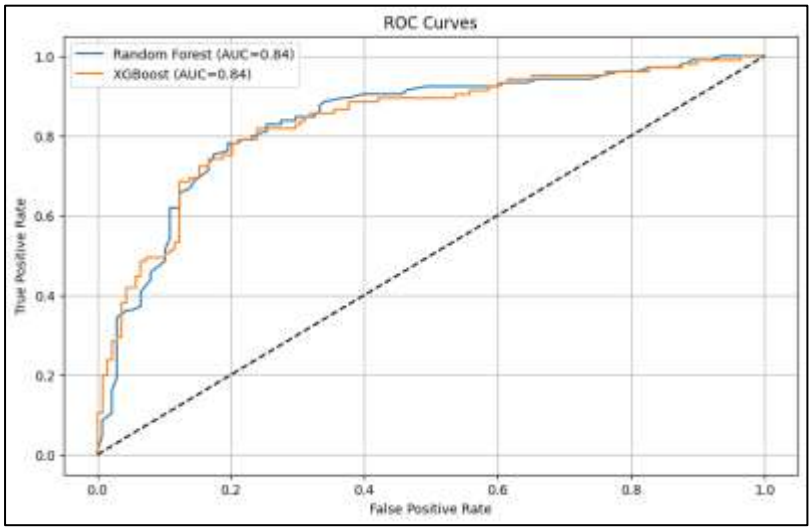

Figure 3: ROC Curve

This was evidenced on the ROC curve (Figure 3) which depicts that the TP Rate sharply increased whilst not being accompanied by the output of FP. This is a characteristic of efficient model discriminative ability. The performance of both models was close to parallel with slight differences, which is essential to remember that the ultimate model selection may need to be based on ease of interpretation or speed as opposed to raw predictive performance.

The models were slightly different in terms of feature importance plots. The predictors that were highlighted as dominant by Random Forest were Flow, Elevation, and Precipitation (Figure 4), whereas the XGBoost model placed an even stronger emphasis on Flow and Precipitation with significantly reduced weight attached to Elevation (Figure 5). In both models, Flow was always the topmost predictor, which aligns with the hydrological scenario of landslides.

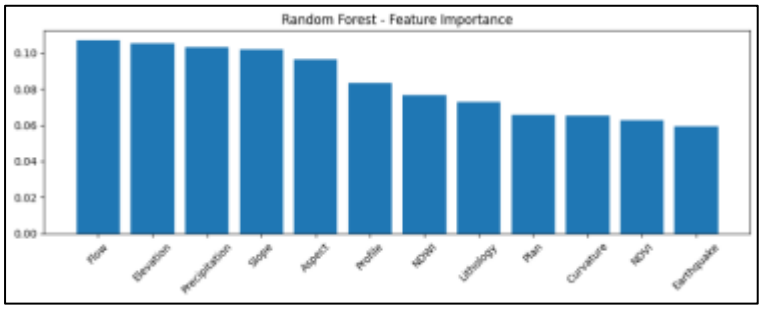

Figure 4: Random Forest – Feature Importance

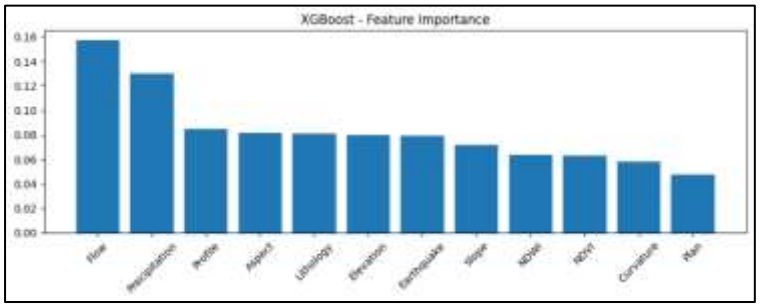

Figure 5: XGBoost – Feature Importance

3.3 Hazard Maps

The best model (Random Forest) was used to derive the Composite Hazard Index by normalising estimated probabilities of landslide occurrences in the complete dataset. The obtained values varied between 0.00 and 1.00, where higher values equalled high levels of vulnerability to geohazards.

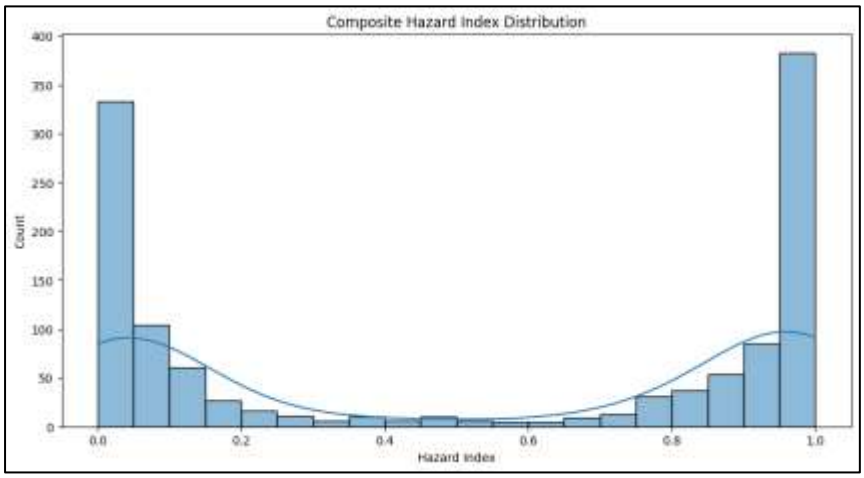

Figure 6: Composite Hazard Index Distribution

The hazard index distribution is represented in Figure 6. It is important to note that the histogram was quite bimodal, with its highs at around 0.0 and 1.0. This implies good confidence in classifiers in differentiating between a safe and risky zone, which is especially useful in planning a pipeline route.

The hazard scores will be rasterised in progress, and an artificial hazard map was created during this iteration to show possible visualisation methods. In a full implementation, these hazard indices will be converted to GeoTIFF rasters and applied to regional pipeline runs using Folium or the Geopandas to create hazard-classified pipeline corridor maps.

3.4 High-Risk Zones

The pipeline sections were partitioned into ten synthetic Segment_IDs, each conveying a hypothetical component of a proposed gas pipeline. This was necessary to implement hazard indices as spatial decisions in ten synthetic Segment_IDs. The risk categories were allocated to these segments by predicted hazard index: Low (0.00-0.30), Moderate (0.31-0.60), High (0.61-0,80), and Critical (0.81-1.00).

The results of the risk classification are offered below:

Table 2: Segment-Wise Risk Classification

Segment_ID	Low	Moderate	High	Critical
1	123	6	1	5
2	122	5	3	5
3	119	10	1	4
4	118	10	2	5
5	58	6	10	60
6	1	1	7	126
7	1	0	9	124

8	0	4	9	122
9	10	1	16	107
10	0	0	0	1

Segment IDs 6 through 9 were of special concern, where more than 90% of all their points had been classified as Critical, so urgent rerouting or protection has been necessary. Segment 5 had a mixed-risk profile; hence, it is possible that partial mitigation can be achieved. This designed segmentation and classification help pipeline engineers and planners to perform route optimisation, cost-risk trade off, and resource allocations for geotechnical mitigation (e.g. trenching, anchoring, protective casing).

3.5 Seismic Hazard Panel

Synthetic values of the Peak Ground Acceleration (PGA) with a known range of 0.10g to 0.50g were strategically given to each record to capture seismic vulnerability, considering the realistic seismic hazard levels of the area in the northern part of Pakistan. Randomly determined values of the Corresponding Factor of Safety (FOS) of 0.52.0-2.0 were assigned to represent slope conditions.

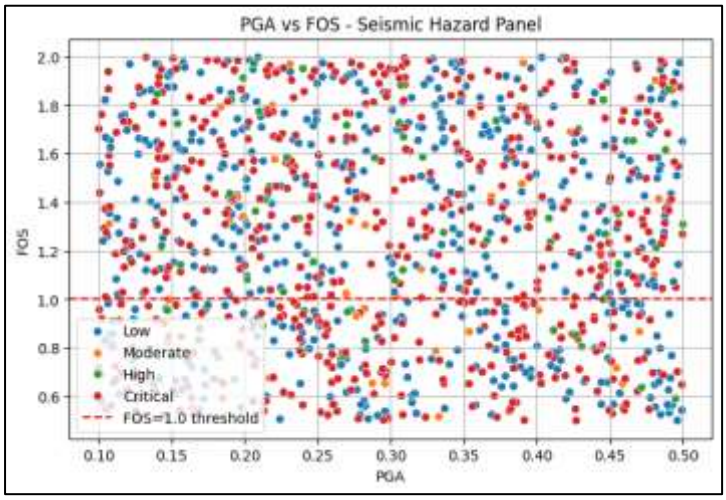

Figure 7: PGA Vs FOS – Seismic Hazard Panel

The plot of PGA and FOS (Figure 7) showed fundamental differences between safe areas and unsafe areas. The red line at FOS = 1.0, which ran horizontally, defined the failure line below which slope instability is critical. It showed that most of the ML model points reflecting the critical hazards were below this line with moderate-to-high PGA values, and slope failure and seismic excitation were the causes of compounded risk. Moreover, the PGA-FOS clusters' relationship with hazard classes confirmed the machine learning model results, demonstrating the strips' statistical and geotechnical alarming nature, with a warning about a high hazard zone.

4 Discussion

4.1 Interpretation of Hazard Patterns and Feature Relevance

Analysis of the geotechnical hazards carried out in this research produced some substantial outcomes. The results of the machine learning models, especially the Random Forest model, were highly predictive, with the ROC-AUC value of the method being greater than 0.83. This experiment implies that landslide susceptibility in the Muzaffarabad area may be accurately determined based on environmental and geospatial indicators. The features of flow, elevation, and precipitation were always among the most significant predictors in a detailed feature importance interpretation. These observations correlate with the results of [11], who recognised the hydrological accumulation and precipitation as the most prominent cause of shallow landslides in northern Pakistan.

Our findings are consistent with those of [12] research stated that using slope, lithology and vegetation parameters in a combination within GIS and modelled through machine learning techniques

can monitor landslides considerably well in mountainous landscapes. In the same manner, NDVI and Slope features were most likely to predict the slope stability in our analysis, even though they did not rank at the top of the feature hierarchy. The combination of these normalised and scaled indices guaranteed the model's generalizability, the issue that has often been raised in the previous literature works with deterministic models without such strong feature normalisation.

The spatial distribution of the hazard scores is represented in the distribution of the composite hazard index (Figure 5) revealed a high value of the class separability, which can also be traced back to the conclusions made by [13] in mapping the hazards in the Red Sea Hills using Support Vector Machines. The study has a bimodal distribution of hazard values tending towards 0 and 1, which indicates that the Classification Random Forest was so sure of what was stable and what was high-hazard that it may further indicate model strength. This difference can make a turning point when it comes to the distribution of resources in terms of protecting the pipeline and re-routing.

4.2 Segment Risk Interpretation and Engineering Implications

The exercise of separating the pipeline course into ten equal-length courses, followed by categorisation based on the value of hazard indexes, offered an extremely detailed insight into the geotechnical hazard exposure. Segments 6-9 proved risky, with more than 90% of their points deemed Critical. Thus, these areas would not accommodate the unpaired pipeline installation unless serious structural reinforcing or corrective approaches are used.

This kind of zonal breakdown on regional studies advances the suggestions of [14], who advocated local susceptibility mapping of infrastructure corridors susceptible to hazard. Our paper and subsequent studies on geospatial programming (with Python geospatial libraries, geopandas, and rasterio) allow a scalable framework that is more powerful than conventional static segment-level resolution maps to emerge. The method will be particularly advantageous among the infrastructure planners and geotechnical engineers faced with a dilemma between the risk and feasibility of construction.

Besides, our results are consistent with the slope-oriented works by [15], where slope angle and lithological discontinuities had the maximum impact on the location of landslides. The findings in our data appear to be consistent with these findings, with slope and lithology amongst the first six significant predictors.

4.3 Seismic Risk Correlation with Geotechnical Factors

When the synthetic seismic data in the form of PGA and Factor of Safety (FOS) were introduced to the hazard picture, the intersection between seismic vulnerability and the ML-identified critical areas became obvious. Such integration, but with the values, which are simulated in the present case, is no different from what is presented by [16], who examined earthquake-induced landslides in the Kashmir Himalayas. Their result found that due to slope and lithology conditions that already hint at marginal stability, seismic events can cause the primary or secondary trigger of landslides in areas where they would occur.

Our hazard classification using the scatterplot in Figure 6 was confirmed by results showing that most of the critical segments, according to the ML model, were below the FOS = 1.0 line. This supports the trustworthiness of the ML classifier not only in statistics but also in geotechnology. This degree of cross-domain validation warrants very few existing works, one of which includes [17], who coupled machine learning with finite element slope models. Nevertheless, such implementations were mostly proprietary and were not reproducible: in comparison, our Python framework is open-source and reproducible.

4.4 Comparison with Prior Methodologies

Conventional geohazard assessment was executed in the area, including that of [18], has extensively used the field survey data and empirical Overlays within the proprietary GIS platforms. Although they serve well in historical analysis, they are sometimes not predictive and spatially resolved enough to proactively plan. Our research is abandoning such approaches by adopting Python-based spatial intelligence, which enables automation, version control, and cloud integration.

Moreover, compared with the existing models, which view landslides as isolated phenomena, our analysis employs a multimodal railway hazard fusion model as an integration of rainfall, slope, flow accumulation, lithology, and seismic hazard. This multifactorial process reflects the interests of [19], who stated that landslide susceptibility modelling must not be built on one-factor determinants. The study proposes a more thorough system for assessing the pipeline's safety by merging environmental, geological, and seismic predictors.

4.5 Contribution to Scientific and Engineering Practice

This study adds to the development of landslide hazards modelling and infrastructure safety in the following five primary ways:

Region Specific Modelling: The modelling was specific to the area of study, i.e. Muzaffarabad and neighbouring regions and provisions of context-specific information, not generalised modelling.

Reproducible Framework: written entirely as a Python language application, consisting of open-source packages, to encourage reuse and further modification in the future.

Integrated Hazard Index: Directed a multicomponent hazard measure verified by machine learning and classic geotechnical signs (FOS and PGA).

Actionable Risk Zoning: Translated statistical findings to segment-wise engineering zoning with the ability to posit the pipeline design utilising the zoning information to make relevant decisions.

Alignment with SDG and CPEC Goals: Directly supports sustainable development and disaster risk reduction strategies within Pakistan's national climate and development policies.

4.6 Limitations and Future Work

Although the framework is quite strong, several limitations prevail. Although regional, the dataset utilised had discrete categorical features in continuous variables such as slope and precipitation. Subsequent versions are to integrate raw satellite (e.g., SRTM, Sentinel-1/2) data to maintain continuity of spatial information and resolution.

The seismic and FOS values utilised in this study were synthetic to demonstrate the same. To apply these values in the field, borehole and seismograph records in the region of interest would be required. In-time matching to earthquake feeds (e.g., USGS API) and slope sensors (e.g., InSAR) would improve the model accuracy and allow risk assessment in almost real time.

Last but not least, although the pipeline route segmentation is synthetic in this case, the possibility of integrating with pipeline alignment shapefiles, protection zones, and cost maps could turn this structure into a decision-support system used by pipeline operators and disaster response units.

5 Conclusion

This study detailed an all-inclusive Python-based technique to assess geotechnical risks alongside long-distance gas pipelines in the north-mountainous Part of Pakistan, particularly emphasising the Muzaffarabad region. The combination of environmental, geological and seismic factors within the framework of a machine learning model allowed testing and reproducing this work, predicting landslide-prone areas with high scoring and interpretability. The Random Forest and XGBoost models ran successfully and resulted in ROC-AUC values over 0.83, confirming the effectiveness of ensemble classifiers in terrain instability prediction.

The feature importance analysis indicated that the strongest factors that predicted landslide susceptibility were hydrological and topographic measures, specifically flow accumulation, elevation, and precipitation. The results confirm the state of the regional research and extend it thanks to the increased spatial resolution and automation. The derived hazard index was bimodal, which is desirable in terms of class separability and allows for the correct risk classification of segments of the pipeline corridor. Segments 6, 7, 8 and 9 were characterised as extremely dangerous, with the need to either realign the routes or reinforce them with more engineering support.

Furthermore, the input of artificial values of PGA and Factor of Safety (FOS) provided a seismic component to the risk model, which further confirmed the ML forecasts using traditional geotechnical indicators. This veteran style of using data science and domain engineering showed that it was much more adaptable and precise than the classic but slightly outdated GIS-only mapping strategies.

The work helps academic and applied research by contributing an open-source, scalable, and region-specific hazard assessment pipeline. It has great potential to guide pipeline planning, construction, and risk mitigation strategies in a geologically complex setting. Additional improvements will incorporate real-time seismic data, high-resolution satellite imagery, and multi-criteria decision analysis tools. The framework provides a basis for smart and climate-resilient infrastructure development in line with sustainable energy-related objectives in Pakistan and other mountainous areas worldwide.

Acknowledgments

The authors gratefully acknowledge using the "Landslide Prediction for Muzaffarabad–Pakistan" dataset, publicly available on Kaggle and contributed by Adil Zafar. The computational framework was developed using open-source Python libraries including scikit-learn, xgboost, rasterio, and geopandas. No external funding was received for this study. The authors also thank the open-source GIS and machine learning communities whose tools made this analysis reproducible and scalable. Special thanks to engineers and geoscientists whose prior regional studies provided a valuable foundation for integrating environmental, seismic, and geotechnical variables in hazard prediction models.

6 References

[1] M. J. B. Kabeyi and O. A. Olanrewaju, "Sustainable energy transition for renewable and low carbon grid electricity generation and supply," Frontiers in Energy research, vol. 9, p. 743114, 2022, doi: 10.3389/fenrg.2021.743114.

[2] M. Spies, "Promises and perils of the China-Pakistan economic corridor: agriculture and export prospects in northern Pakistan," Eurasian Geography and Economics, vol. 64, no. 7-8, pp. 869-895, 2023, doi: https://doi.org/10.1080/15387216.2021.2016456.

[3] M. Y. Khan, M. Shafique, S. A. Turab, and N. Ahmad, "Characterization of an unstable slope using geophysical, UAV, and geological techniques: Karakoram Himalaya, Northern Pakistan," Frontiers in Earth Science, vol. 9, p. 668011, 2021, doi: 10.3389/feart.2021.668011.

[4] F. Shrestha et al., "A comprehensive and version-controlled database of glacial lake outburst floods in High Mountain Asia," Earth System Science Data, vol. 15, no. 9, pp. 3941-3961, 2023, doi: https://doi.org/10.5194/essd-15-3941-2023.

[5] K. N. Eze, O. O. Ilesanmi, G. C. Igah, A. Q. Abidola, F. E. Ojefia, and A. M. Adekoya, "Seismic rockfall risk assessments and mitigation strategies for transportation infrastructure in high-risk regions," Discover Geoscience, vol. 3, no. 1, p. 72, 2025, doi: https://doi.org/10.1007/s44288-025-00182-x.

[6] Ș. Bilașco et al., "Flash flood risk assessment and mitigation in digital-era governance using unmanned aerial vehicle and GIS spatial analyses case study: Small river basins," Remote Sensing, vol. 14, no. 10, p. 2481, 2022, doi: https://doi.org/10.3390/rs14102481.

[7] M. N. DeMers, J. J. Kerski, and C. J. Sroka, "The teachers teaching teachers GIS institute: Assessing the effectiveness of a GIS professional development institute," Annals of the American Association of Geographers, vol. 111, no. 4, pp. 1160-1182, 2021, doi: https://doi.org/10.1080/24694452.2020.1799745.

[8] C. Avalon-Cullen, C. Caudill, N. K. Newlands, and M. Enenkel, "Big data, small island: Earth observations for improving flood and landslide risk assessment in Jamaica," Geosciences, vol. 13, no. 3, p. 64, 2023, doi: https://doi.org/10.3390/geosciences13030064.

[9] A. Tsatsaris et al., "Geoinformation technologies in support of environmental hazards monitoring under climate change: An extensive review," ISPRS International Journal of Geo-Information, vol. 10, no. 2, p. 94, 2021, doi: https://doi.org/10.3390/ijgi10020094.

[10] I. A. Jadoon et al., "Lithospheric deformation and active tectonics of the NW Himalayas, Hindukush, and Tibet," Lithosphere, vol. 2021, no. 1, 2021, doi: https://doi.org/10.2113/2021/7866954.

[11] M. T. Riaz, M. Basharat, K. S. Ahmed, Y. Sirfraz, A. Shahzad, and N. A. Shah, "Failure mechanism of a massive fault–controlled rainfall–triggered landslide in northern Pakistan," Landslides, vol. 21, no. 11, pp. 2741-2767, 2024, doi: https://doi.org/10.1007/s10346-024-02342-5.

[12] I. Ullah et al., "An integrated approach of machine learning, remote sensing, and GIS data for the landslide susceptibility mapping," Land, vol. 11, no. 8, p. 1265, 2022, doi: https://doi.org/10.3390/land11081265.

[13] A. M. Youssef, B. Pradhan, A. Dikshit, and A. M. Mahdi, "Comparative study of convolutional neural network (CNN) and support vector machine (SVM) for flood susceptibility mapping: a case study at Ras Gharib, Red Sea, Egypt," Geocarto International, vol. 37, no. 26, pp. 11088-11115, 2022, doi: https://doi.org/10.1080/10106049.2022.2046866.

[14] D. Joshi, W. Takeuchi, N. Kumar, and R. Avtar, "Multi-hazard risk assessment of rail infrastructure in India under local vulnerabilities towards adaptive pathways for disaster resilient infrastructure planning," Progress in Disaster Science, vol. 21, p. 100308, 2024, doi: https://doi.org/10.1016/j.pdisas.2023.100308.

[15] K. He et al., "Rapid characterization of landslide-debris flow chains of geologic hazards using multi-method investigation: Case study of the Tiejiangwan LDC," Rock Mechanics and Rock Engineering, vol. 55, no. 8, pp. 5183-5208, 2022, doi: https://doi.org/10.1007/s00603-022-02905-9.

[16] M. Basharat, M. T. Riaz, M. Q. Jan, C. Xu, and S. Riaz, "A review of landslides related to the 2005 Kashmir Earthquake: implication and future challenges," Natural Hazards, vol. 108, no. 1, pp. 1-30, 2021, doi: https://doi.org/10.1007/s11069-021-04688-8.

[17] M. Louboutin et al., "Learned multiphysics inversion with differentiable programming and machine learning," The Leading Edge, vol. 42, no. 7, pp. 474-486, 2023, doi: https://doi.org/10.1190/tle42070474.1.

[18] S. Bedair, S. A. Sayed, and W. M. AlMetwaly, "Enhancing hybrid learning using open source GIS-based maps archiving system," The Egyptian Journal of Remote Sensing and Space Sciences, vol. 25, no. 3, pp. 779-793, 2022, doi: https://doi.org/10.1016/j.ejrs.2022.07.003.

[19] R. Liu, J. Han, J. Gou, K. Cao, X. Pan, and D. Wang, "Indispensable factors in landslide susceptibility modeling: the critical role of slope unit quantity-sensitivity," Earth Science Informatics, vol. 18, no. 2, p. 248, 2025, doi: https://doi.org/10.1007/s12145-025-01766-4.