Geospatial Machine Learning Models for Integrated Landslide Risk and Asset Exposure Assessment

Swastika Vinayak Pawar[1], Aravinth R[2] *

[1,2] Bharati Vidyapeeth Deemed University, Institute of Environment Education and Research, Pune, 411043, INDIA,
*Corresponding author

Keywords: Landslide Susceptibility, Risk Assessment, Asset Loss, Machine Learning, Google Earth Engine, Google Colab

Abstract

Landslides increasingly threaten communities, infrastructure, and ecosystems, especially in areas with steep gradients, heavy precipitation, and unregulated land utilization. This research employs a hybrid geospatial and machine learning (ML) methodology to evaluate landslide susceptibility and asset exposure in Western Maharashtra, an area noted for its significant vulnerability. Three machine learning models—Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Back Propagation Neural Network (BPNN)—were trained utilizing topographic, hydrological, land use, and soil variables. XGBoost attained the best accuracy at 91.17%, generating a susceptibility map divided into five risk groups. To assess exposure, the susceptibility outputs were combined with building footprint data, indicating significant threats to residential areas, infrastructure systems, and agricultural land. Google Earth Engine (GEE) facilitated satellite-based analysis, whilst Google Colab enabled model training and validation. The results indicate robust model performance; however, limitations include reliance on static input data and the lack of real-time environmental monitoring. Future endeavors intend to integrate dynamic datasets, advanced deep learning architectures, and IoT-based early warning systems. The research highlights the need of combining geospatial analysis with machine learning methods for sustainable disaster risk mitigation and informed spatial planning.

1. Introduction

Landslides are among the most frequent and destructive natural hazards, involving the downslope movement of soil, rock, or debris under the influence of gravity. These events are typically triggered by a range of factors such as intense or prolonged rainfall, seismic activity, and anthropogenic drivers including deforestation, slope modification, and unplanned construction activities. Their impacts are often severe, resulting in disruptions to transportation networks, destruction of property, environmental degradation, and significant loss of human life, particularly in mountainous and monsoon-affected regions of the world.(1)

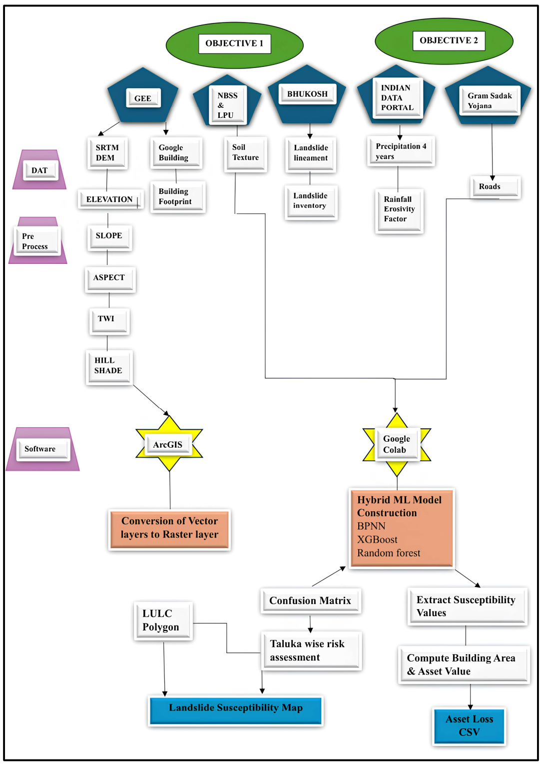
The Western Ghats and Konkan belt of Maharashtra, India, represent regions of particularly high landslide susceptibility due to their steep terrain, heavy seasonal rainfall, and increasing anthropogenic pressures. Historical events, such as the 2023 Irshalwadi landslide in Raigad district and the widespread monsoon-triggered landslides of 2021, underscore the vulnerability of this region. The Geological Survey of India (GSI) has demarcated over 0.09 million km² across six districts of Maharashtra as landslide-prone, with more than 500 fatalities reported in the past decade alone (2).

To mitigate these risks, landslide susceptibility mapping (LSM) has been widely applied as a tool to identify and delineate areas prone to future failures. Such mapping commonly integrates terrain, geology, hydrology, and land-use parameters. Traditional approaches to LSM have relied on heuristic or statistical methods. While these methods have contributed valuable insights, they often suffer from limited adaptability and insufficient capacity to capture the complex, nonlinear interactions among conditioning factors, thereby restricting their predictive reliability in dynamic environments (3).

In recent years, machine learning (ML) techniques have gained prominence in susceptibility modelling due to their ability to handle large spatial datasets, capture nonlinear relationships, and provide improved predictive accuracy compared to conventional approaches. However, much of the existing literature remains focused primarily on hazard prediction, with limited attention given to risk evaluation—particularly the exposure of physical assets and infrastructure. Furthermore, many studies employ single ML algorithms, whereas ensemble-based approaches, which combine the strengths of multiple models, have the potential to enhance prediction robustness but remain underexplored in this domain. Another critical gap lies in the scale of application: while most research addresses regional or national-level susceptibility, there is limited work at sub-regional (taluka) scales, where actionable planning and disaster mitigation strategies are typically implemented.

Addressing these gaps, the present study develops an integrated framework for landslide risk assessment in Western Maharashtra. The framework advances beyond conventional susceptibility mapping by incorporating exposure and asset-loss analysis. Using ensemble machine learning models—specifically Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Backpropagation Neural Networks (BPNN)—this research generates high-resolution susceptibility maps. These outputs are subsequently integrated with land-use and infrastructure datasets to evaluate risk zones and quantify potential asset losses. By providing a comprehensive and spatially explicit risk assessment, the study aims to support evidence-based decision-making for localized disaster preparedness and mitigation in the Western Ghats region.

2. Datasets and Methodology



Methodology Flowchart

This investigation utilized a machine learning-driven geospatial framework to evaluate landslide susceptibility and asset exposure. A total of thirteen causative factors were identified: slope, elevation, curvature, aspect, topographic wetness index (TWI), precipitation, rainfall erosivity, NDVI, sand, silt, clay, road, and landslide lineament. The conversion of vector layers, specifically road and lineament, to raster format was accomplished through the application of the Euclidean Distance tool. This tool effectively assigns each cell the shortest straight-line distance to the nearest feature.

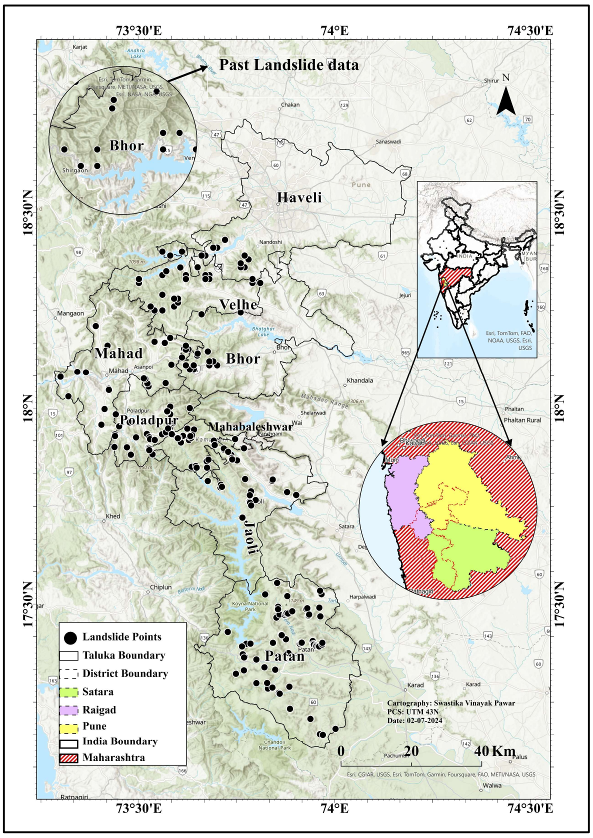
Three machine learning models were developed: Back Propagation Neural Network (BPNN), Extreme Gradient Boosting (XGBoost), and Random Forest (RF). The BPNN was employed due to its capacity to model intricate relationships using a feed-forward architecture that is trained through error backpropagation. XGBoost is an advanced boosting algorithm that enhances prediction accuracy by systematically addressing errors through the construction of decision trees based on residuals. Random Forest, a collection of decision trees trained on various subsets of data, effectively minimized overfitting and highlighted the significance of different variables.

Land Use Land Cover (LULC) data were combined with susceptibility maps to identify risk zones and building footprint data were utilized to assess asset exposure and potential loss at the taluka level. The findings were archived in CSV format. The training and processing of the model were carried out utilizing Google Colab, employing raster inputs at a 30m resolution derived from 11 geospatial layers, which encompass DEM, slope, aspect, hillshade, NDVI, rainfall, R-factor, curvature, TWI, road, and lineament.

The landslide inventory data were organized into five susceptibility categories according to feature dimensions: Very Low (<20m), Low (20–40m), Moderate (40–60m), High (60–80m), and Very High (>80m), with 69 points allocated to each category, totaling 345 points. The data underwent conversion to shapefiles, followed by uploading to Google Earth Engine, and subsequently merged with raster values utilizing Geopandas, Rasterio, and Pandas. Features underwent label encoding and were standardized through the application of StandardScaler.

In the model training process, the random forest was set up with 300 trees and balanced class weights. The XGBoost model was optimized using "mlogloss" as the loss function, while the backpropagation neural network incorporated two hidden layers consisting of 64 and 32 neurons, utilized the Adam optimizer, and produced a softmax output. SMOTE was utilized to tackle the issue of class imbalance. A total of 482 landslide points were identified, with 345 utilized in the model (90% for training, 5% for evaluation, and 5% for testing). The evaluation of model performance involved metrics such as accuracy, precision, recall, and F1-score, with XGBoost demonstrating the highest overall accuracy.

3. Study Area



Study Area

This study selected eight talukas across three districts in Western Maharashtra. The locations encompass Bhor, Velhe & Haveli, Mahabaleshwar, Patan and Jaoli, Mahad, and Poladpur. The selection of these talukas is based on the fact that a significant number of landslides in western Maharashtra take place in these regions. The talukas in Western Maharashtra exhibit a significant vulnerability to landslides, positioning them as a vital area of investigation for this study. This study addresses a gap in the existing literature, as prior investigations have not thoroughly explored landslide risk at the taluka level, underscoring the importance of this work. This study employs machine learning techniques to deliver a comprehensive evaluation of risk and asset loss for each taluka on an individual basis. This method is groundbreaking as it fills a void in current studies, providing a detailed and localized examination of landslide risks in the area.

Performing comparative analyses of various machine learning models is crucial for identifying the most effective strategy for rugged, mountainous terrains. The combination of machine learning and Geographic Information System (GIS) technologies has revolutionized landslide risk assessment by incorporating a spatial aspect to the data. The use of GIS facilitates the development of intricate maps that pinpoint high-risk zones and assess potential asset loss, thereby improving regional-scale modeling and early warning systems. This study employs hybrid machine learning models, representing a notable progression in the assessment of landslide risk. By concentrating on particular talukas in Western Maharashtra, it offers a detailed examination that can act as a model for other mountainous areas encountering comparable issues. Furthermore, the comparative analysis of various machine learning models aids in pinpointing the most efficient methods for forecasting landslides and evaluating asset loss in intricate landscapes.

4. Results and Discussion

4.1 Landslide risk assessment

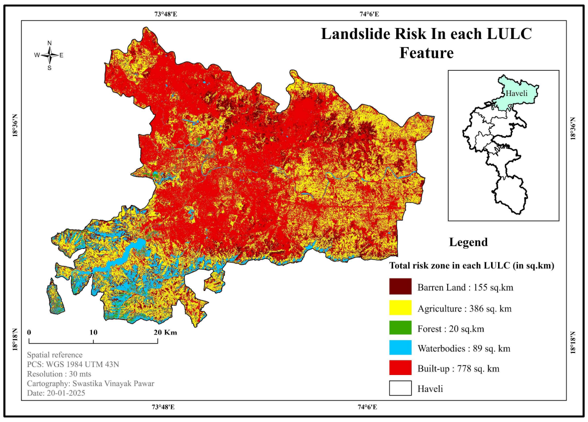


Figure 1 Risk Assessment Map of Haveli Taluk

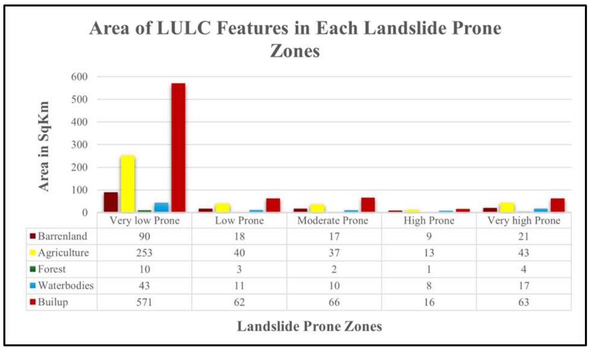


Figure 2 Graph representing LULC feature in each Landslide zone

The examination of Land Use and Land Cover (LULC) characteristics across several landslide susceptibility zones uncovers significant spatial patterns and exposure hazards. Built-up areas are predominantly centered in very low-prone zones (571 sq. km), although a significant growth into very high-prone zones (63 sq. km) signifies increasing urban encroachment into hazardous regions.

Agricultural land exhibits comparable exposure, encompassing 253 sq. km in low-risk areas and 43 sq. km in very high-risk zones, underscoring its susceptibility. Barren terrain, while predominantly situated in low-risk zones, also encompasses 21 sq. km in high-susceptibility regions. Forest cover is restricted yet regularly observable, whereas water bodies are more uniformly dispersed throughout regions.

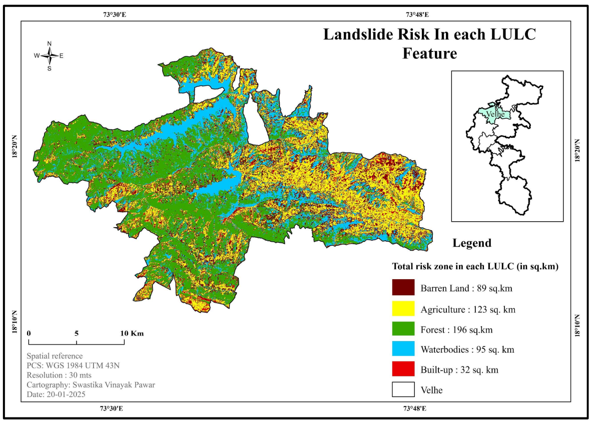


Figure 3 Risk Assessment Map of Velhe Taluk

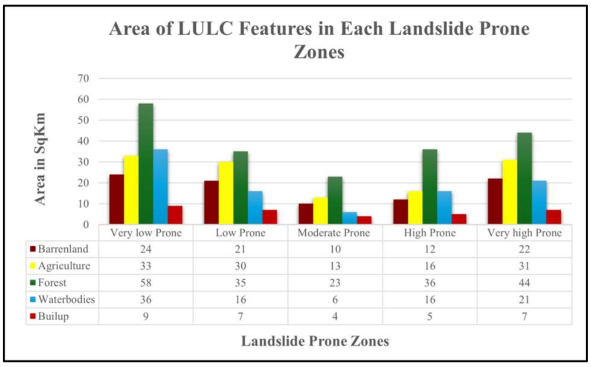


Figure 4 Graph representing LULC feature in each Landslide zone

The research indicates that wooded regions prevail in the landscape, encompassing 58 sq. km in the very low-prone zone and 44 sq. km in the very high-prone zones, signifying vegetation presence on both stable and susceptible slopes. Agricultural land is equitably allocated, comprising 33 sq. km in areas of very low susceptibility and 31 sq. km in regions of very high susceptibility. Developed regions are restricted (maximum of 9 sq. km in low-risk areas), although their emergence in high-risk zones is alarming. Barren land and water bodies occupy 22 sq. km and 21 sq. km, respectively, in areas of very high susceptibility, highlighting the mixed-use exposure of the region.

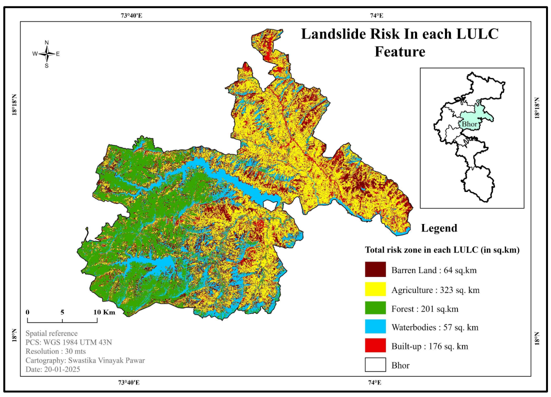


Figure 5 Risk Assessment Map of Bhor Taluk

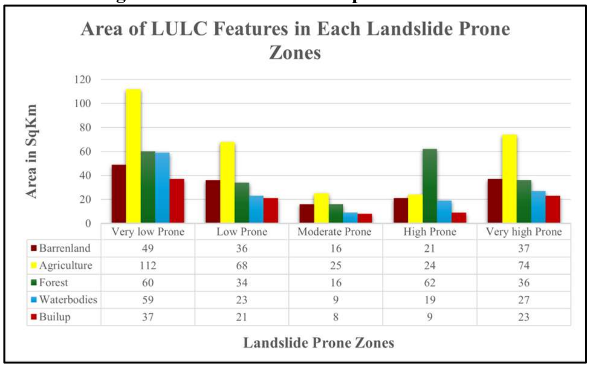


Figure 6 Graph representing LULC feature in each Landslide zone

Agriculture constitutes the predominant land use and land cover (LULC) category, encompassing 112 sq. km in the very low-prone zone and 74 sq. km in the very high-prone zone, signifying substantial vulnerability of agricultural lands to landslide hazards. Forest coverage is significant (60 sq. km in low-risk areas and 36 sq. km in high-risk zones), but developed regions exhibit considerable encroachment into high-risk territories (23 sq. km). Waterbodies (59–27 sq. km) and desolate areas (49–37 sq. km) encompass both stable and susceptible zones, underscoring the necessity for controlled land development.

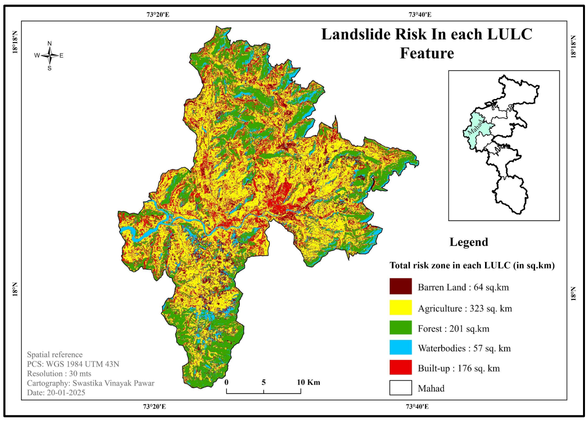


Figure 7 Risk Assessment Map of Mahad Taluk

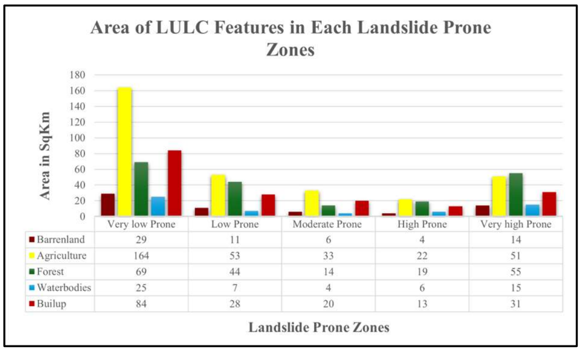


Figure 8 Graph representing LULC feature in each Landslide zone

Agricultural terrain predominates, with 164 sq. km in low-prone zones and 51 sq. km in very high-prone areas, highlighting its vulnerability to landslide hazards. Forest areas are substantial (69–55 sq. km), indicating their existence on both stable and unstable slopes. Developed areas (84 sq. km low; 31 sq. km high) exhibit concerning urban expansion into hazardous regions. Waterbodies and arid land are allocated (25–15 sq. km and 29–14 sq. km, respectively), indicating varied topographical features within the taluka.

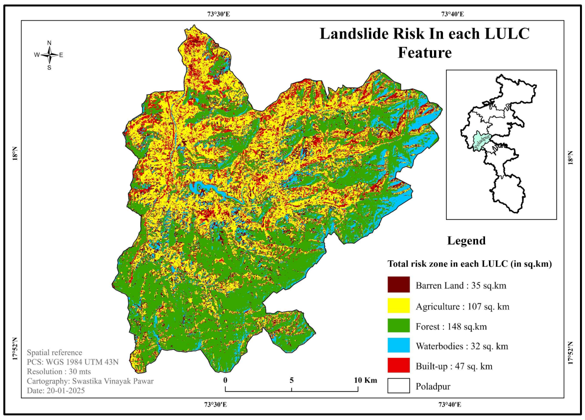


Figure 9 Risk Assessment Map of Poladpur Taluk

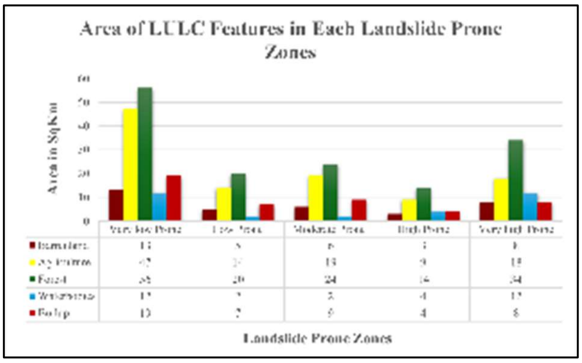


Figure 10 Graph representing LULC feature in each Landslide zone

Agricultural land is substantial, encompassing 47 sq. km in areas of very low susceptibility and 18 sq. km in regions of very high susceptibility. Forest cover, ranging from 56 to 34 square kilometres, varies between susceptibility levels. Built-up regions, however restricted in size (19–8 sq. km), are situated in high-risk zones, indicative of unregulated development. Waterbodies are uniformly distributed at both extremities (12 sq. km), although barren land encompasses 13 sq. km in low-prone zones and 8 sq. km in high-prone zones, highlighting the complex exposure pattern of this region.

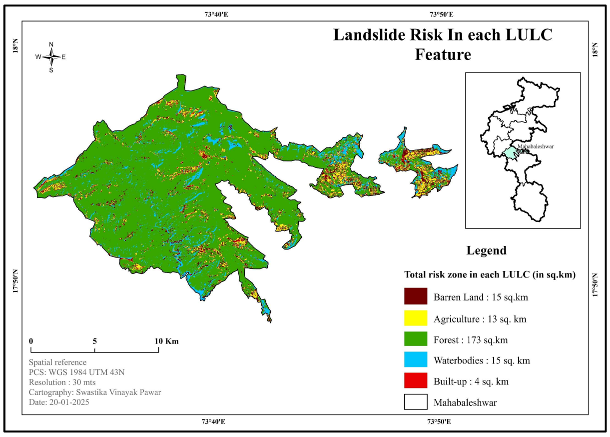


Figure 11 Risk Assessment Map of Mahabaleshwar Taluk

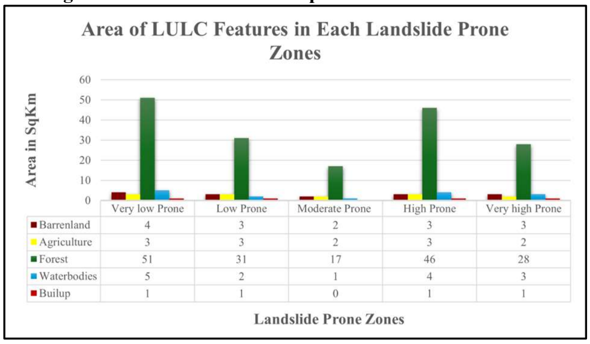


Figure 12 Graph representing LULC feature in each Landslide zone

Agriculture and forest are the principal land use and land cover types, each contributing 3 sq. km and 51–28 sq. km, respectively, across zones with very low and very high susceptibility. Built-up areas are limited (1 sq. km in both extremes), although their existence in high-risk zones indicates human habitation in perilous locations. Barren land and water bodies are uniformly dispersed, measuring 4–3 sq. km and 5–3 sq. km correspondingly, indicating a diminutive yet susceptible landscape need concentrated attention. Forests predominate the terrain, covering 96 sq. km in low-prone areas and 97 sq. km in high-prone zones, signifying ecological abundance and vulnerability.

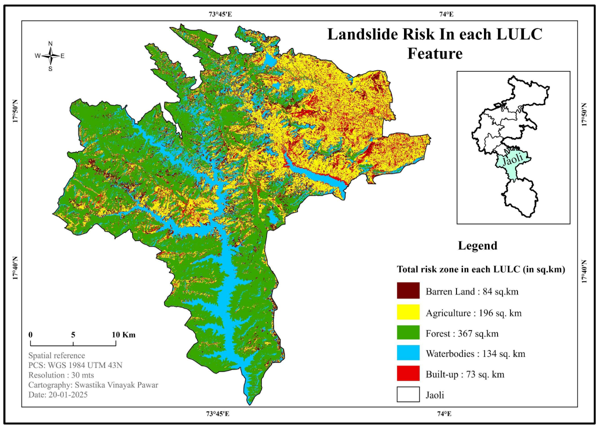


Figure 13 Risk Assessment Map of Jaoli Taluk

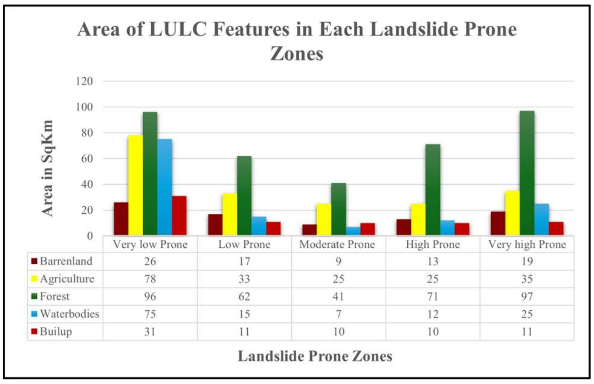


Figure 14 Graph representing LULC feature in each Landslide zone

Agricultural land exhibits considerable distribution, encompassing 78 sq. km in areas of very low susceptibility and 35 sq. km in regions of very high susceptibility. Built-up areas (31–11 sq. km), aquatic bodies (75–25 sq. km), and barren terrain (26–19 sq. km) are distributed throughout all zones, underscoring the region's extensive utilization and potential vulnerability. Patan exhibits the broadest land use and land cover exposure.

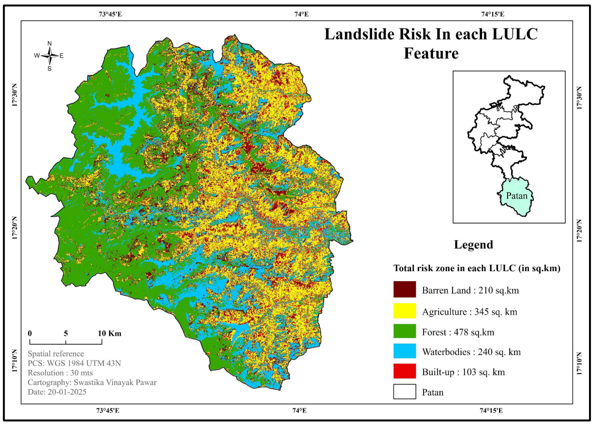


Figure 15 Risk Assessment Map of Patan Taluk

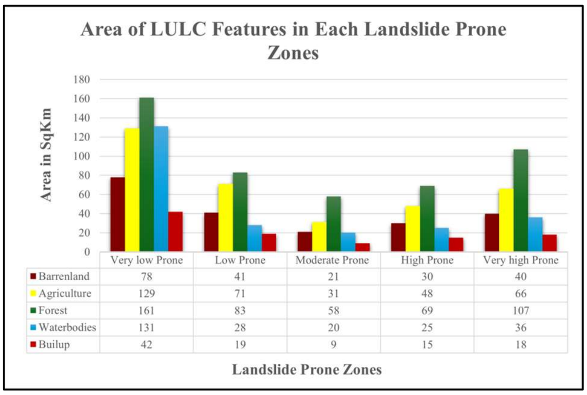


Figure 16 Graph representing LULC feature in each Landslide zone

Agriculture (129–66 sq. km), woodland (161–107 sq. km), and barren terrain (78–40 sq. km) are prevalent in both stable and highly vulnerable areas. Developed regions encompass 42 sq. km in areas of very low risk and 18 sq. km in areas of very high risk, signifying significant human presence in perilous zones. Waterbodies exhibit extensive coverage, encompassing 131 sq. km in low-prone areas and 36 sq. km in high-prone zones, highlighting the intricacy of risk throughout this taluka.

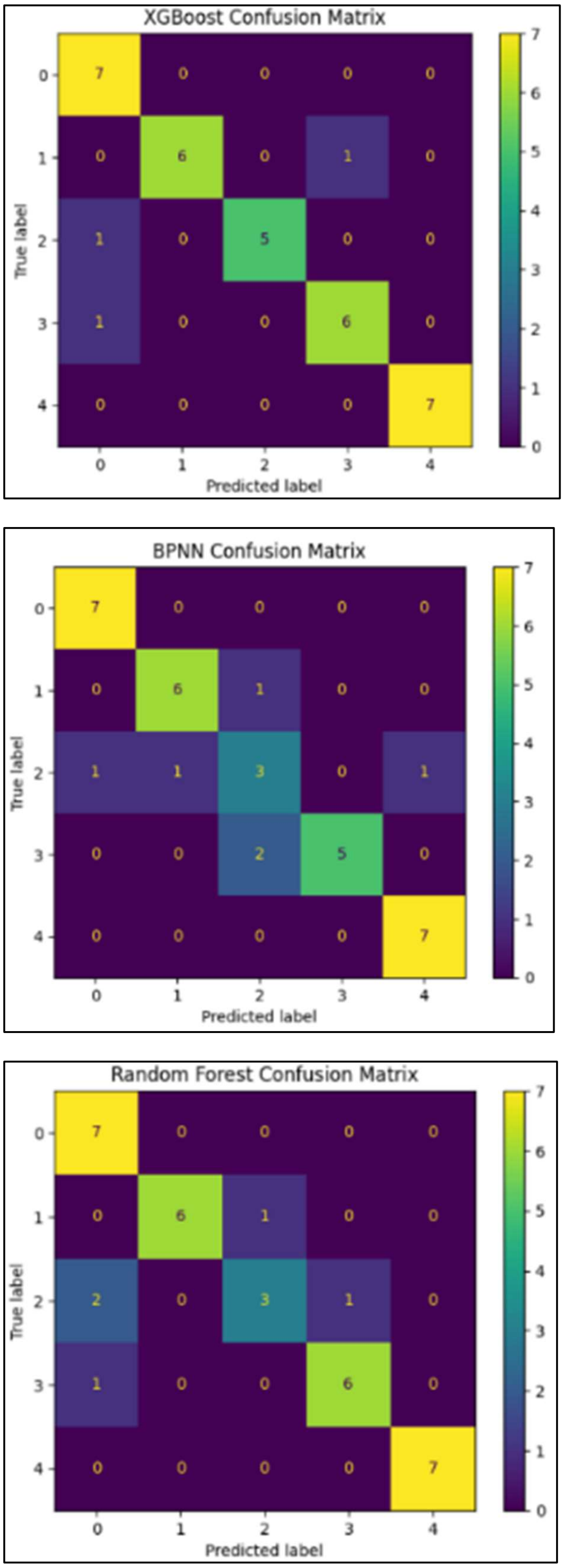


Figure 17 Confusion Matrix of XGB, BPNN & RF

The efficacy of the machine learning models—Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Back Propagation Neural Network (BPNN)—was evaluated utilizing essential criteria such as precision, recall, and F1-score. XGBoost attained the greatest classification accuracy of 91.17% among the three models, followed by Random Forest at 85.29% and BPNN at 82.35%. The confusion matrices produced for each model demonstrated their proficiency in accurately classifying landslide susceptibility across the five established risk categories. XGBoost demonstrated the best balanced and consistent performance, especially in successfully recognizing the mid-risk (class 2) group, which was frequently misclassified by the other models. XGBoost adept management of class imbalances and exceptional overall metrics render it the most dependable and efficient model for landslide susceptibility mapping in this research.

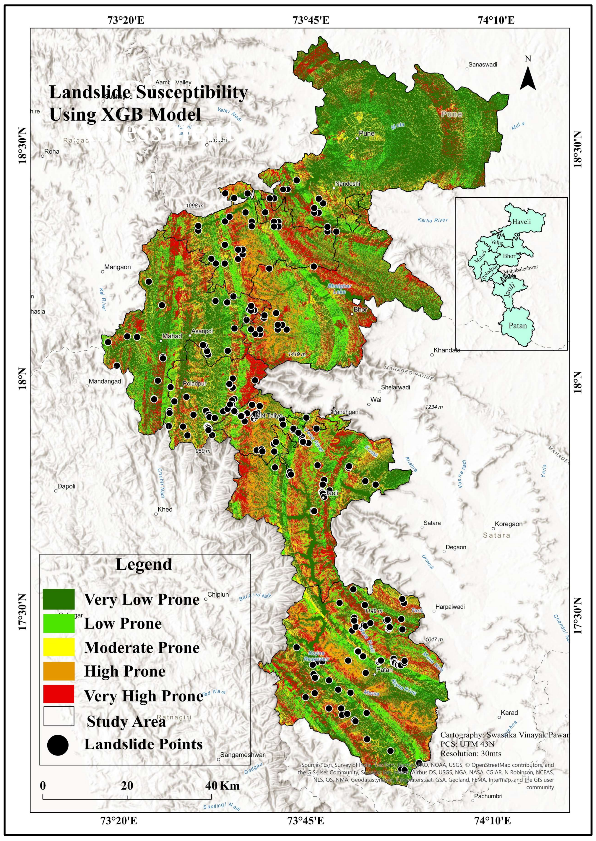


Figure 18 Landslide Susceptibility Map of XGB Model

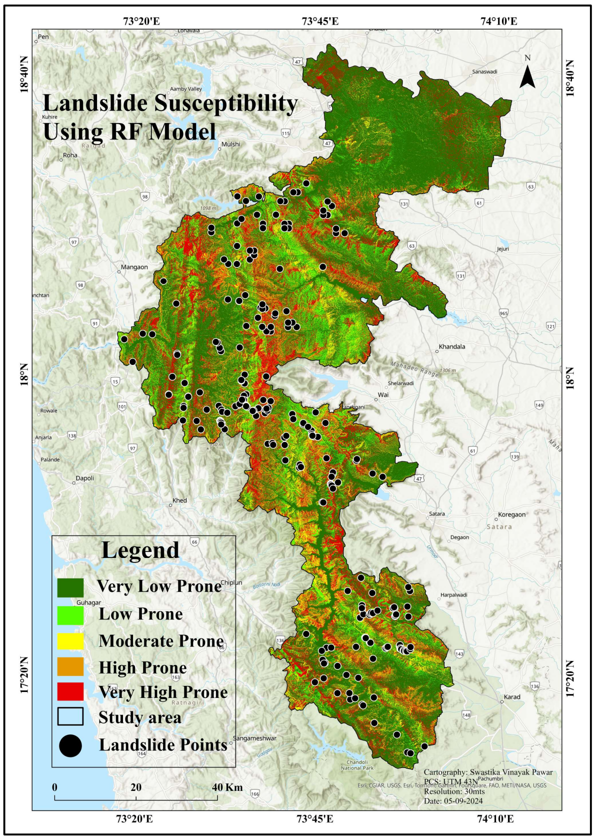


Figure 19 Landslide Susceptibility Map of RF Model

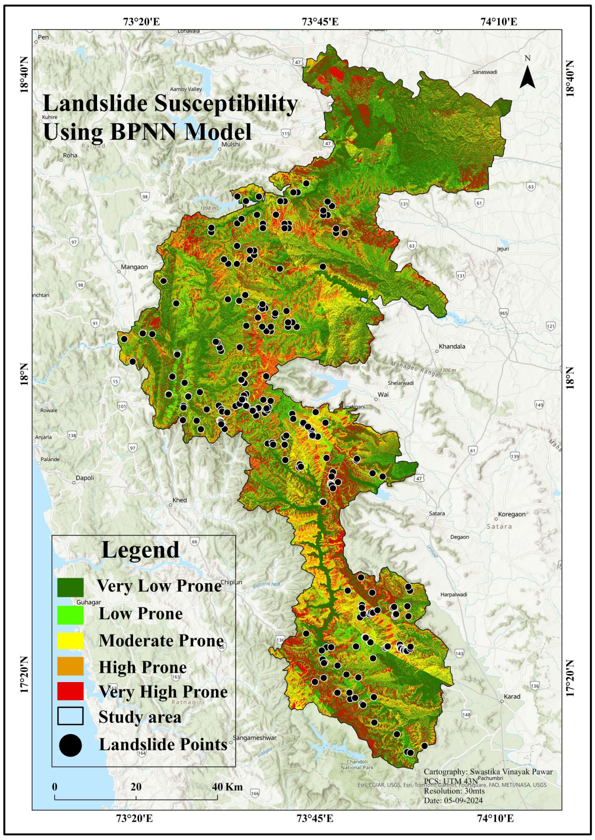


Figure 20 Landslide Susceptibility Map of BPNN Model

The landslide susceptibility map generated by the Extreme Gradient Boosting (XGB) model has superior forecast accuracy compared to the other two models—Random Forest (RF) and Backpropagation Neural Network (BPNN). The robust spatial link between the projected high-risk zones (shown in orange and red) and the actual landslide occurrence spots (depicted as black dots) is visible. The XGB model proficiently delineates places susceptible to landslides, notably in the southern and western sectors of the research area, including Patan, Mahabaleshwar, Poladpur, and Mahad, where terrain, slope, and other contributing factors are more pronounced. In comparison to the Random Forest (RF) and Back Propagation Neural Network (BPNN) models, the XGB model offers a more nuanced categorization, mitigates the overestimation of risk zones, and preserves a significant differentiation among various susceptibility levels.

4.2 Landslide Asset loss model

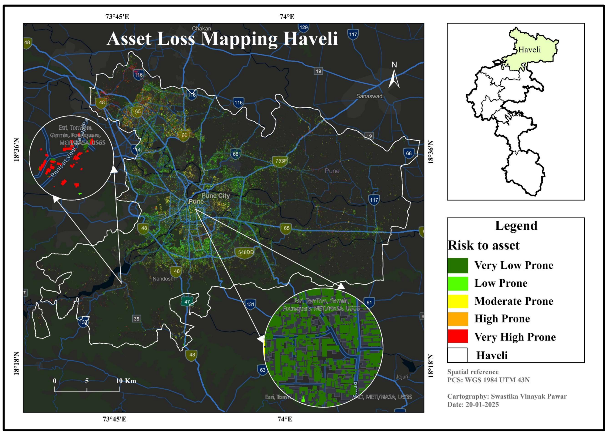


Figure 21 Asset Loss Map Haveli Taluk

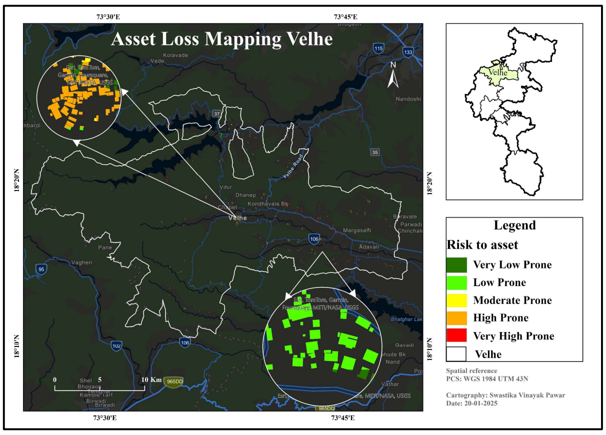


Figure 22 Asset Loss Map Velhe Taluk

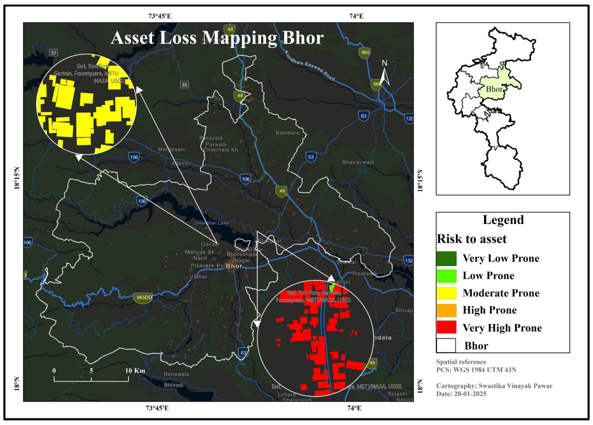


Figure 23 Asset Loss Map Bhor Taluk

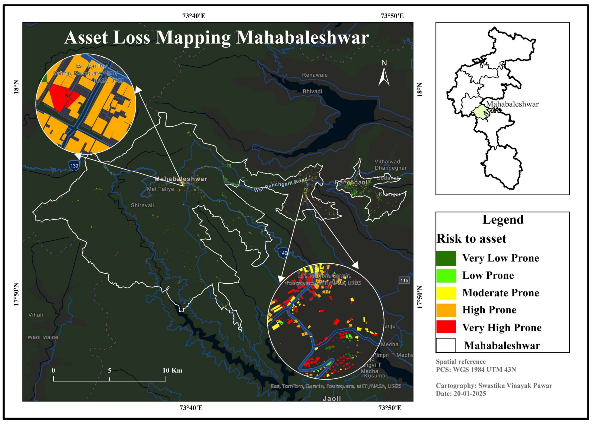


Figure 26 Asset Loss Map Mahabaleshwar Taluk

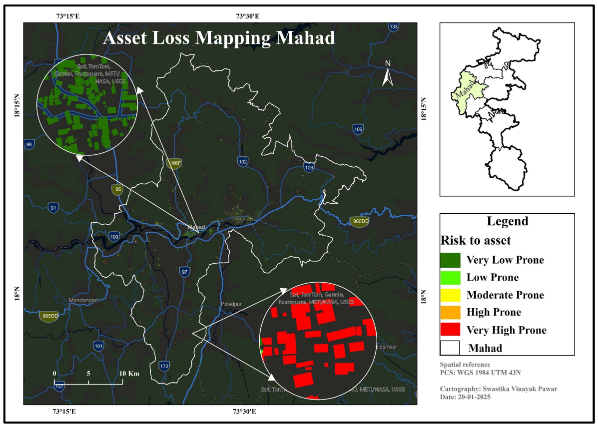


Figure 24 Asset Loss Map Mahad Taluk

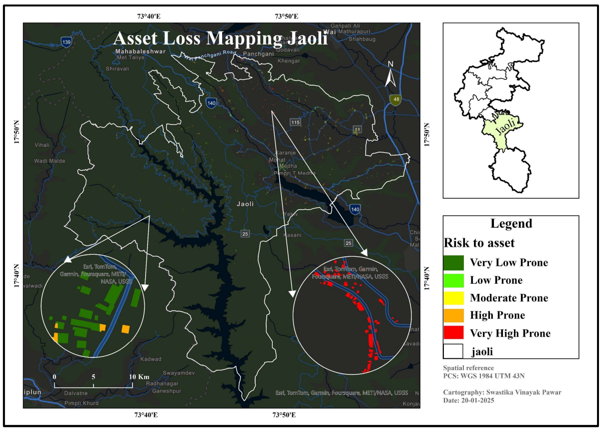


Figure 27 Asset Loss Map Jaoli Taluk

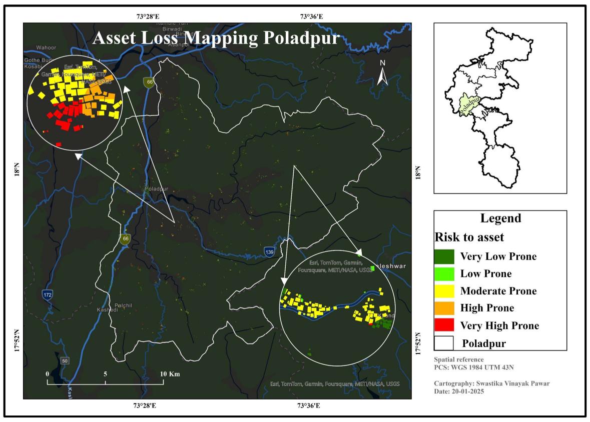


Figure 25 Asset Loss Map Poladpur Taluk

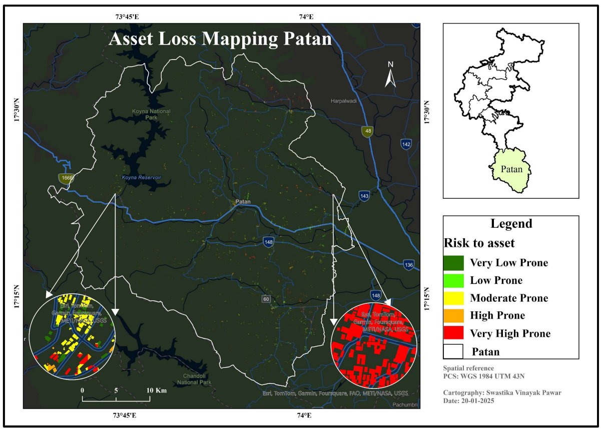


Figure 28 Asset Loss Map Patan Taluk

Taluk Name	Susceptibility	Building Count	Total Asset Value in Cr
Haveli	Very Low	769867	3,25,000
	Low	116521	30,000
	Moderate	101402	41,000
	High	22584	68,300
	Very High	85486	32,800

Velhe	Very Low	7568	12251.6
	Low	7312	15705.6
	Moderate	3194	5955.69
	High	3102	4273.1
	Very High	7732	13654.6
Bhor	Very Low	37889	63594.40
	Low	19566	24579.10
	Moderate	11276	11358.90
	High	5479	4861.51
	Very High	24337	28465.20
Mahad	Very Low	51695	14,651.8
	Low	11240	1,388.30
	Moderate	9682	1,592.70
	High	2282	225.7
	Very High	11313	1441.1
Poladpur	Very Low	14111	1,594.70
	Low	1975	175.7
	Moderate	4044	394.7
	High	1656	145.7
	Very High	3431	319.7
Mahabaleshwar	Very Low	5637	1050
	Low	8656	1150
	Moderate	3763	606
	High	4151	695
	Very High	3825	589
Jaoli	Very Low	25228	4182.2
	Low	11162	1724.3
	Moderate	9932	1361.5
	High	5414	720.1
	Very High	9045	1813.8
Patan	Very Low	54087	7487.2
	Low	23211	2675.1
	Moderate	15081	1753.5
	High	10545	1289.1
	Very High	26169	3114.7

Table 1 Asset loss value in Cr

The asset loss study demonstrates a distinct correlation between development and landslide susceptibility throughout the region. In all talukas, zones with very low susceptibility consistently exhibit the highest asset concentration, signifying a pronounced inclination for development in comparatively secure regions. These regions collectively encompass about ₹3.25 lakh crore in asset value, including ₹74.87 crore in Patan, ₹1,05,000 crore in Mahabaleshwar, and ₹6,35.94 crore in Bhor. These regions have the greatest concentration of structures and the most extensive built-up area, so increasing the inclination to develop in low-risk zones.

The research indicates a troubling concentration of assets in areas with very high, high, and moderate susceptibility, where infrastructure and communities face escalating vulnerability. For example, Velhe possesses assets valued at ₹1,36.54 crore in extremely high-risk sectors, but Mahabaleshwar and Bhor have assets of ₹58,900 crore and ₹2,84.65 crore, respectively. In Patan, a significant ₹31.14 crore of infrastructure is situated within extremely high-risk areas. Likewise, moderate and high-risk zones in talukas such as Jaoli (₹13.61 crore in moderate, ₹7.20 crore in high), Mahad (₹15.92 crore in moderate, ₹2.25 crore in high), and Poladpur (₹3.94 crore in moderate, ₹1.45 crore in high) indicate substantial advancement in hazard-prone areas.

This tendency indicates that urban expansion—prompted by land scarcity, population growth, or inadequate integrated risk-informed planning—has resulted in encroachment into vulnerable areas. The extensive vulnerability of valuable infrastructure to landslide hazards underscores the immediate necessity for proactive mitigating measures. These should encompass sustainable land use planning, rigorous enforcement of building restrictions in high-risk zones, the establishment of early warning systems, and community-oriented awareness initiatives.

The findings strongly support the incorporation of susceptibility-based risk assessments into developmental planning. With assets exceeding ₹43,563 lakh crore allocated across all vulnerability tiers in the region, the stakes are substantial. Investing in resilience and proactive planning can substantially mitigate possible economic losses and protect at-risk communities from the escalating threat of landslides.

5. Conclusion

This research combines hybrid machine learning models—Random Forest, XGBoost, and BPNN—for the purpose of landslide susceptibility mapping, exhibiting elevated spatial accuracy. Of the three machine learning models employed—Random Forest, XGBoost, and BPNN—XGBoost attained the best accuracy at 91.17%, successfully identifying high-risk areas. The analysis indicated that agricultural regions, urban developments, and transportation systems are especially susceptible due to topographical characteristics, precipitation patterns, and alterations in land use, including deforestation and haphazard Construction. The analysis estimated probable asset loss by merging susceptibility maps with building footprint data, emphasizing considerable infrastructure and economic vulnerability in high-risk areas. A significant disadvantage is the potential for overfitting, especially arising from the class imbalance present in landslide datasets. These models, despite their efficacy, frequently exhibit a lack of transparency, complicating the interpretation of individual elements such as slope or land cover, hence constraining their applicability for policy-level choices. A further disadvantage is the lack of temporal analysis. Landslide susceptibility is dynamic, shaped by causes including urbanization, climate change, and deforestation. Relying solely on static information limits the model's power to identify emerging threats, hence diminishing its prediction efficacy over time. Estimating asset loss by superimposing building footprints onto susceptibility zones provides a systematic risk evaluation. Nevertheless, edge instances adjacent to zone boundaries and the absence of data regarding structural robustness or building materials may impact accuracy. Furthermore, the dynamics of landslides, such as debris flow trajectories, are not simulated, constraining comprehension of cascading dangers. This study progresses the area by integrating machine learning with cloud-based platforms such as Google Earth Engine and Colab, facilitating efficient processing and replicability despite existing limits. The implementation of a multi-class susceptibility scale offers a more refined analysis of asset exposure than conventional binary classifications. Furthermore, taluka-level analysis facilitates targeted risk management methods. The results highlight the significance of machine learning in facilitating risk-informed spatial design and disaster mitigation. Future initiatives must integrate real-time data, deep learning algorithms, and IoT-based early warning systems to augment predictive accuracy and resilience methods. The methodology is applicable to various regions—such as the Himalayas, Western Ghats, or coastal areas—given suitable factor selection. Future research should incorporate time-series satellite data, real-time precipitation, and socio-economic indices to develop a more comprehensive risk model. Incorporating these models into early warning systems can facilitate proactive evacuation and planning, hence improving regional catastrophe resilience.

6. Acknowledgements

Acknowledgements of support for the project/paper/author are welcome. Note, however, that for the paper to be submitted for review all acknowledgements must be anonymized.

7. References

Kale, V., Nulkar, G., Upasani, D., Tamboli, A., Aslekar, U., and Kulkarni, H., 2022. Landslides of Western Maharashtra: Rapid appraisal for developing improved preparedness and early warning systems for landslides in Western Maharashtra and Konkan Regions, Maharashtra State. Unpublished report.

Ghaffarian, S., and Taghikhah, F.R., 2023. Explainable artificial intelligence in disaster risk management: Achievements and prospective futures. International Journal of Disaster Risk Reduction, 98, 104123. https://doi.org/10.1016/j.ijdrr.2023.104123

Ali, N., Chen, J., Fu, X., Ali, R., Hussain, M.A., Daud, H., et al., 2024. Integrating Machine Learning Ensembles for Landslide Susceptibility Mapping in Northern Pakistan. Remote Sensing, 16(6), 988. https://doi.org/10.3390/rs16060988

Song, Y., Yang, D., Wu, W., Zhang, X., Zhou, J., Tian, Z., et al., 2023. Evaluating Landslide Susceptibility Using Sampling Methodology and Multiple Machine Learning Models. ISPRS International Journal of Geo-Information, 12(5), 197. https://doi.org/10.3390/ijgi12050197

Sharma, Y., Tyagi, A., Lal Sharma, M., Sharma, P., and Aggarwal, A., 2023. Building Vulnerability Assessment using Artificial Intelligence for Landslide Susceptibility Zone in Champawat District, India. EGU General Assembly 2023, EGU23-1957. https://doi.org/10.5194/egusphere-egu23-1957