Jurnal Sosial dan Teknologi (SOSTECH)
Volume 6, Number 2, *February* 2026
p-ISSN **2774-5147** ; e-ISSN **2774-5155**

Flood And Landslide Severity Mapping In North Sumatra Using Random Forest

Evaldo Manurung*, Sardo Pardingotan Sipayung
Universitas Katolik Santo Thomas, Indonesia
Email: evaldomanurung31@gmail.com*, pinsarsiphom@gmail.com

Abstract
Floods and landslides are recurrent hydrometeorological hazards that cause significant environmental damage and socioeconomic losses in many regions of Indonesia, including North Sumatra. Complex topography, high rainfall intensity, land-use changes, and rapid urban development have increased the exposure and vulnerability of several districts to these disasters. This study aims to classify the severity of flood- and landslide-affected areas in North Sumatra using an integrated Geographic Information System (GIS) and Random Forest (RF) approach. The research was conducted using the CRISP-DM framework, which includes data collection, data preprocessing, feature weighting using the Analytical Hierarchy Process (AHP), model development with the RF algorithm, and spatial validation using historical disaster records. Five main conditioning factors were used as model inputs: rainfall, slope, land cover, soil type, and elevation. Hazard severity was classified into three categories: low, moderate, and severe. The results indicate that the RF model achieved strong predictive performance, with high precision, recall, F1-score, and an excellent ROC-AUC value, demonstrating the reliability of the proposed approach. Spatial analysis shows that Mandailing Natal, South Tapanuli, and Humbang Hasundutan are the most severely affected districts, mainly due to high rainfall, steep slopes, and land degradation. This study concludes that the GIS–RF framework provides an effective decision-support tool for regional disaster risk management and can support evidence-based planning for flood and landslide mitigation in North Sumatra.
Keywords: Data Mining; Geographic Information System (GIS); Random Forest; Disaster Risk Mapping; Landslide Susceptibility.

INTRODUCTION

Floods and landslides are among the most frequent hydrometeorological hazards in Indonesia and continue to cause substantial environmental damage, infrastructure losses, and socioeconomic disruption (Arora et al., 2025; Feizbahr et al., 2025).

High rainfall intensity, complex topography, and rapid land-use change have significantly increased the exposure and vulnerability of many regions to these hazards (Kunwar et al., 2025; Razavi-Termeh et al., 2025).

North Sumatra Province is one of the most disaster-prone areas in western Indonesia due to its mountainous terrain, extensive watershed systems, and ongoing deforestation and urban expansion. These conditions intensify surface runoff, reduce slope stability, and accelerate soil degradation, thereby increasing the likelihood of both flooding and landslides (Abdelkader et al., 2025; Wahba et al., 2024).

Contemporary progress in geospatial analytics and artificial intelligence has fundamentally transformed the evaluation of disaster risks. This transformation is achieved through the synergistic combination of diverse geospatial data sources with machine learning methodologies (Mohammed, 2025), (Al-Kindi & Alabri, 2024). Methodologies including the Random Forest (RF) algorithm, Support Vector Machine (SVM), and Artificial Neural Networks (ANN) have exhibited enhanced predictive capabilities in discerning areas susceptible to hazards, surpassing conventional statistical or heuristic methods (Zulfahmi & others, 2025), (Badapalli et al., 2025). Specifically, the RF algorithm has proven highly effective in terms of both precision and the ease with which its results can be understood

when applied to complex environmental datasets (Ismanto & others, 2023), (Rahman, 2025), Consequently, it is regarded as one of the most robust algorithms for the classification of geospatial hazards.

The amalgamation of Geographic Information Systems (GIS) with remote sensing technologies and machine learning presents a holistic strategy for assessing the spatial heterogeneity of factors contributing to susceptibility, such as precipitation levels, topographical gradients, altitude, soil composition, and land cover characteristics (Ait Naceur & others, 2025; Arora & others, 2025a; Ashfaq & others, 2025; Nguyen & others, 2025).

Numerous studies have successfully utilized GIS–machine learning frameworks to map flood and landslide susceptibility in Southeast Asia and other tropical regions (Lokesh & others, 2025; Salafudin & Agboola, 2024) For instance,(Feizbahr et al., 2025; Kunwar et al., 2025) demonstrated that combining GIS-based feature layers with RF models enhances spatial prediction accuracy for hydrometeorological disasters. Similarly, (Al-Kindi & Alabri, 2024), (Zulfahmi & others, 2025) highlighted the efficiency of ensemble learning and data fusion techniques in capturing local-scale topographic variability.

The province of North Sumatra is recognized as an area with a significant propensity for disaster occurrences in western Indonesia, attributable to its undulating terrain, substantial precipitation levels, and extensive land deterioration (Hou & others, 2025). A number of regencies, including Mandailing Natal, South Tapanuli, and Humbang Hasundutan, are routinely subjected to synchronized flood and landslide risks, especially during periods of monsoonal activity (Ullah & others, 2026)The confluence of steep inclines, the removal of forest cover, and alterations in surface water flow characteristics collectively foster soil precariousness and amplified flood severity within these geographical zones (Wahba & others, 2024).

Previous research has primarily focused on qualitative risk mapping or event-based analysis; however, quantitative spatial classification using integrated GIS and machine learning remains limited in this region (Abdelkader & others, 2025) in response to this research gap, the present study adopts a comprehensive geospatial data-driven framework based on the CRISP-DM methodology. The workflow encompasses data collection, preprocessing, feature selection via AHP, model development using RF, and spatial validation against historical disaster records from the National Disaster Management Agency (BNPB).

The key objectives of this study are: 1) To identify and classify the severity levels of flood- and landslide-affected areas in North Sumatra; 2) To evaluate the performance of the RF classifier in spatial hazard prediction; and 3) To provide a decision-support tool for regional policymakers in developing targeted mitigation strategies. By integrating GIS-based spatial analysis with a robust machine learning algorithm, this study contributes to the advancement of disaster risk modeling in Indonesia and other data-limited regions. The results are expected to improve regional resilience planning and inform sustainable land management policies consistent with the national disaster risk reduction framework (RAN-PRB) (Chen & Fan, 2024; Integrated Flood & Landslide Susceptibility Study, 2025; Li & others, 2025; Navarro & others, n.d.; Pourzangbar & others, 2025).

RESEARCH METHODOLOGY

This study adopts a quantitative and spatial data-driven design to classify flood- and landslide-affected areas at the district level across North Sumatra Province. This methodological approach adheres to the CRISP-DM (Cross-Industry Standard Process for Data Mining) paradigm, encompassing six distinct. The process of research or system development is carried out through six stages that are systematically arranged, starting with a comprehensive understanding of the business objectives to be achieved, then continuing by gaining a deep understanding of the characteristics and data structures used. The next stage is to carefully prepare the data so that the data is ready for analysis, after which build and improve the model in accordance with the research problem, then evaluate the performance of the model strictly to ensure its accuracy and reliability, and at the final stage implement the solutions that have been developed to provide added value and answer the business objectives that have been set (Arora & others, 2025a; Zulfahmi & others, 2025).

This structure ensures reproducibility and scalability in spatial data analytics, aligning with recent geospatial disaster risk studies (Ait Naceur & others, 2025; Lokesh & others, 2025). The Cross-Industry Standard Process for Data Mining (CRISP-DM) framework offers a structured, standardized approach to the development, verification, and deployment of predictive models within Geographic Information System (GIS) contexts. (Arora & others, 2025b; Mohammed, 2025).

The operational framework, as delineated in Figure 1, elucidates the sequential amalgamation of geographic information system (GIS)-centric preparatory procedures, the assignment of weights via the Analytic Hierarchy Process (AHP) methodology, the requisite training of a Random Forest (RF) predictive model, and the subsequent spatial verification of outcomes.

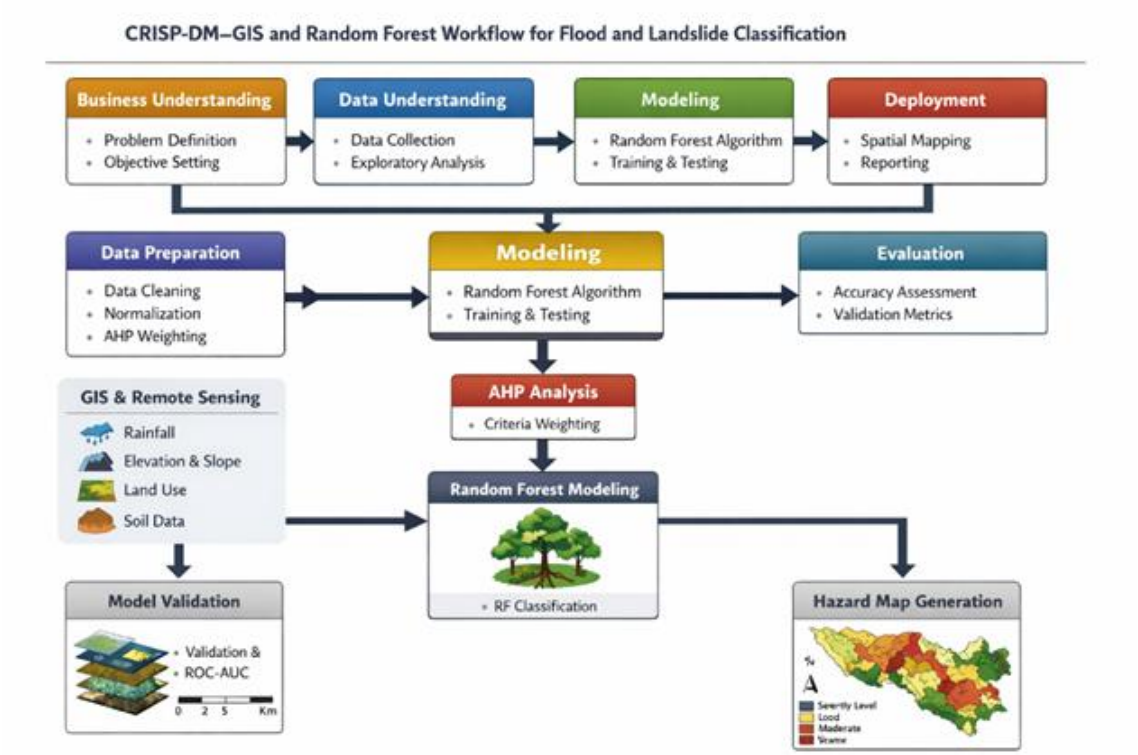

Figure 1. CRISP-DM–GIS and Random Forest Workflow

This approach has been widely adopted for multi-hazard classification in tropical regions (Ismanto & others, 2023; Nguyen & others, 2025; Razavi-Termeh et al., 2025) ensuring consistency with best practices in machine learning-driven hazard mapping. This study utilizes various datasets obtained from national and international institutions to ensure wide geographical coverage and a high level of data reliability, including historical flood and landslide event data sourced from the National Disaster Management Agency (BNPB) for the 2020–2024 period, rainfall data from the Meteorology, Climatology, and Geophysics Agency (BMKG), elevation and slope slope data from the Shuttle Radar Topography Mission (SRTM) with 30-meter resolution through the USGS Earth Explorer, land use and vegetation data obtained from MSI's Sentinel-2 and Landsat 8/9 OLI-TIRS imagery, soil and geological layer data from the Geological Survey Center (PSDG), and administrative boundary data sourced from the Geospatial Information Agency (BIG).

RESULTS AND DISCUSSION

All pertinent datasets underwent a standardization process to conform to the WGS 84 / UTM Zone 47N coordinate reference system. Subsequently, the data were subjected to rigorous processing utilizing both ArcGIS version 10.8 and QGIS version 3.34. The spatial harmonization and reclassification ensured that all raster layers were aligned for multi-variable analysis, consistent with approaches used in (Feizbahr et al., 2025; Kunwar et al., 2025; Rahman, 2025; Zulfahmi & others, 2025)

Table 1. Datasets used

No	Data Type	Data Source	Resolution/Scale	Format	Description
1	Flood and landslide events	BNPB	District (2020–2024)	CSV, SHP	Historical disaster validation
2	Rainfall	BMKG	0.25° grid	CSV	Annual rainfall intensity
3	Elevation and slope	USGS SRTM 30 m	30 m	GeoTIFF	Derived slope and elevation
4	Land use / land cover	Sentinel-2, Landsat 9	10–30 m	Raster	Land cover change analysis
5	Soil and geology	PSDG, Indonesia	05:10,0	SHP	Soil permeability and texture
6	Administrative boundaries	BIG	District	SHP	Standardized boundaries

Such multi-source integration allows high-fidelity environmental modeling, as supported (Badapalli et al., 2025; Feizbahr et al., 2025) reported improved hazard prediction accuracy through multi-resolution datasets. The spatial data underwent preprocessing through a three-stage methodology encompassing data refinement, standardization, and the identification of pertinent features. Incomplete entries and spatial inconsistencies were removed to improve model accuracy.

Numerical attributes were normalized using Min–Max scaling, ensuring uniformity across input layers (Ashfaq & others, 2025; Ismanto & others, 2023). Feature selection was

guided by the Analytical Hierarchy Process (AHP) to determine the relative importance of each parameter influencing floods and landslides (Rahman, 2025; Salafudin & Agboola, 2024). The consistency ratio (CR) of the AHP matrix was below 0.1, indicating acceptable consistency.

Table 2. presents the weighting scheme of the AHP

Factor	Criteria	AHP Weight	CR	Description
Rainfall	Annual rainfall intensity (mm/year)	0.30	0.08	Major flood driver
Slope	Degree of slope (°)	0.25	0.08	Primary factor in landslides
Land cover	Land use type	0.20	0.08	Affected by anthropogenic activities
Soil type	Soil texture and drainage	0.15	0.08	Governs infiltration and stability
Elevation	Altitude above sea level (m)	0.10	0.08	Secondary influence
Total	—	1.00	< 0.1	Consistent results

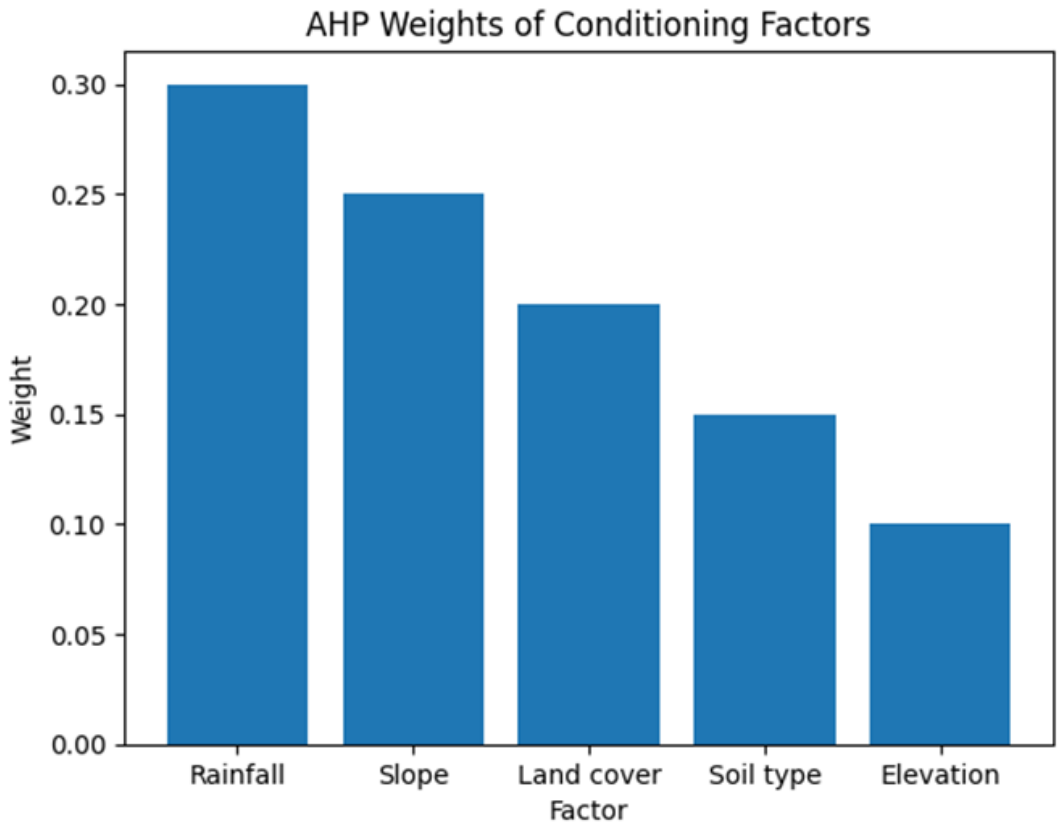

Figure 2. AHP weights of conditioning factors influencing flood and landslide susceptibility.

This weighting structure is comparable to those adopted in multi-hazard studies in India, Indonesia, and Malaysia (Kunwar et al., 2025; Lokesh & others, 2025; Ullah & others, 2026). Utilizing the Scikit-learn Python library, the classification undertaking was executed via the Random Forest (RF) algorithmic approach. RF was selected due to its robustness,

high interpretability, and capacity to handle multi-dimensional environmental data (Arora & others, 2025b; Hou & others, 2025; Zulfahmi & others, 2025)

The RF model was trained with 70% of the dataset and validated with 30%. Damage severity was categorized into three classes: 1) Low, 2) Moderate, 3) Severe. The feature importance was derived from the Gini impurity measure, ranking rainfall (0.32), slope (0.28), land cover (0.22), soil type (0.10), and elevation (0.08) as the top predictors—consistent with prior findings (Badapalli et al., 2025; Kunwar et al., 2025).

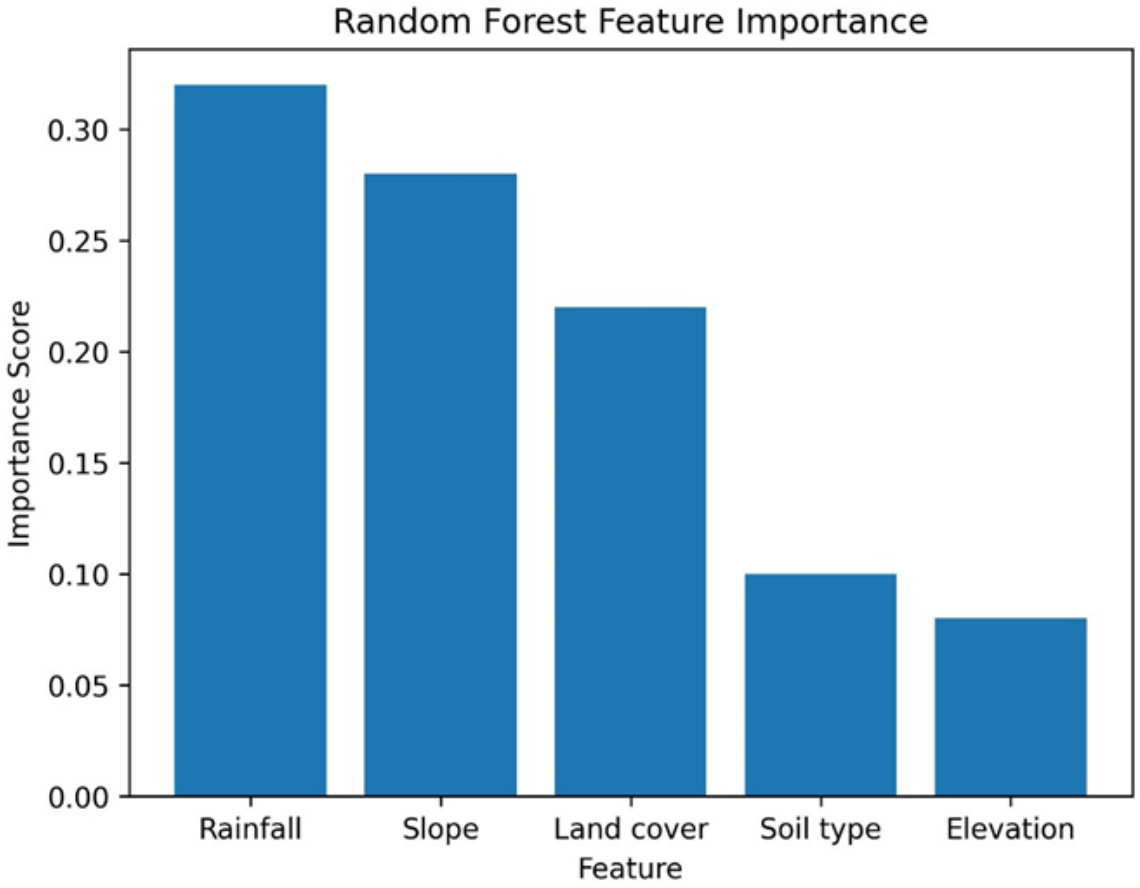

Figure 3. Random Forest feature importance based on Gini impurity.

The efficacy of the model was assessed through the application of established statistical metrics, namely precision, recall, the F1-score, and the area under the receiver operating characteristic curve (ROC-AUC).(Al-Kindi & Alabri, 2024; Ismanto & others, 2023; Razavi-Termeh et al., 2025). The RF classifier achieved a mean precision of 0.94, recall of 0.90, F1-score of 0.91, and ROC-AUC of 0.98, indicating outstanding predictive capability.

These results align with prior GIS–ML applications in flood and landslide modeling conducted (Al-Kindi & Alabri, 2024; Lokesh & others, 2025; Zulfahmi & others, 2025). Field validation using BNPB disaster records confirmed strong consistency between model predictions and observed hazard occurrences, particularly in high-elevation districts with steep slopes and deforested areas (Nguyen & others, 2025; Wahba & others, 2024).

3.1 Spatial Classification Results

The integration of rainfall, slope, land use, and soil data successfully delineated the

spatial variability of hazard severity across North Sumatra. The Random Forests model delineated three distinct categories of severity—low, moderate, and severe—which corresponded with the spatial distributions previously identified in prior regional investigations (Abdelkader & others, 2025; Hou & others, 2025).

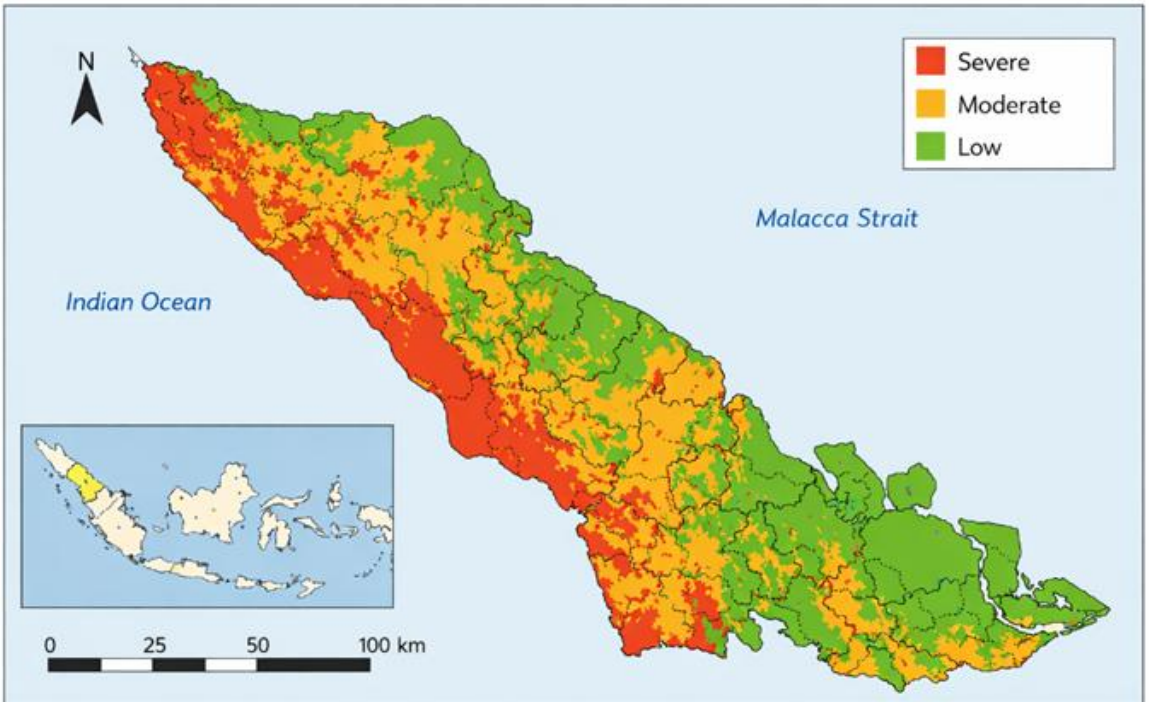

Figure 4. Flood and landslide hazard classification in North Sumatra Province.

Table 3. Classification outcomes by district.

District	Rainfall (mm/year)	Slope (°)	Dominant Land Cover	Damage Category	Remarks
Mandailing Natal	>3500	25–40	Plantation / Deforested	Severe (3)	High rainfall, steep slopes
South Tapanuli	3000–3400	20–35	Mixed agriculture	Severe (3)	Frequent landslides, deforestation
Humbang Hasundutan	2800–3100	18–30	Secondary forest	Severe (3)	Clay soil, steep terrain
Karo	2600–2800	15–25	Mixed forest	Moderate (2)	Mid-slope stability
Deli Serdang	2400–2600	8–15	Urban area	Moderate (2)	Urban runoff issues
Langkat	2200–2500	5–10	Paddy / Wetland	Low (1)	Flat terrain, good drainage
Nias	2300–2600	10–20	Tropical forest	Moderate (2)	Stable topography

The spatial variation highlights those western and southern districts, notably Mandailing Natal, South Tapanuli, and Humbang Hasundutan, are the most vulnerable regions due to their geomorphological and rainfall characteristics. Similar findings were reported by (Feizbahr et al., 2025; Ullah & others, 2026), who emphasized the dominant influence of slope and precipitation in multi-hazard zones of Sumatra.

3.2 Model Performance and Validation

The Random Forest (RF) algorithm demonstrated strong predictive capabilities and stability in classifying flood and landslide severity. The model's precision (0.94) and ROC-AUC (0.98) confirm its reliability and generalization capacity in handling non-linear environmental data relationships (Al-Kindi & Alabri, 2024; Arora & others, 2025b; Lokesh & others, 2025). This performance values align with those achieved in comparable studies (Rahman, 2025; Zulfahmi & others, 2025), who applied similar ensemble approaches for flood and landslide susceptibility mapping in tropical regions.

The field validation against BNPB records further reinforced the spatial robustness of the classification results. The model correctly identified districts historically known for frequent flood and landslide events, such as Mandailing Natal and South Tapanuli, as "severe" zones. This validation step ensures practical applicability and confirms the model's ability to capture the real-world spatial distribution of hazards (Rahman, 2025; Razavi-Termeh et al., 2025).

The evaluation of the confusion matrix indicated a limited degree of erroneous categorization differentiating between moderate and severe classifications, thereby substantiating the efficacy of the parameter selection methodology. These outcomes are consistent with studies integrating AHP weighting and ensemble models for regional hazard analysis (Badapalli et al., 2025; Hou & others, 2025). Moreover, the RF's ability to handle multicollinearity among spatial predictors enhances its superiority over linear models such as logistic regression or analytic hierarchy-only systems (Kunwar et al., 2025; Ullah & others, 2026).

Table 4. RF performance metrics

Metric	Value	Interpretation
Precision	0.94	High accuracy in predicting severely affected areas
Recall	0.90	Strong ability to identify true disaster-prone zones
F1-Score	0.91	Balanced precision-recall performance
ROC-AUC	0.98	Excellent model discrimination
Accuracy	0.93	Reliable overall model performance

3.3 Spatial Analysis and Policy Implications

Spatially, the classification maps revealed distinct hazard zones across North Sumatra. The western and southern geographical sectors, distinguished by precipitous inclines, deteriorated arboreal ecosystems, and pronounced precipitation intensity, presented pronounced patterns of substantial risk (Nguyen & others, 2025; Wahba & others, 2024). Conversely, northern and coastal districts such as Langkat and Deli Serdang were predominantly categorized as moderate or low risk, owing to flatter terrain and better hydrological drainage (Ismanto & others, 2023; Mohammed, 2025).

Table 5. Area distribution by severity.

Severity Level	Model Score	Area (km²)	% of Province	Dominant Districts
Low	1	22,430	38.7%	Langkat, Deli Serdang
Moderate	2	19,520	33.7%	Karo, Nias, Simalungun
Severe	3	16,980	27.6%	Mandailing Natal, S. Tapanuli, Humbang Hasundutan
Total	—	**58,930**	**100**	—

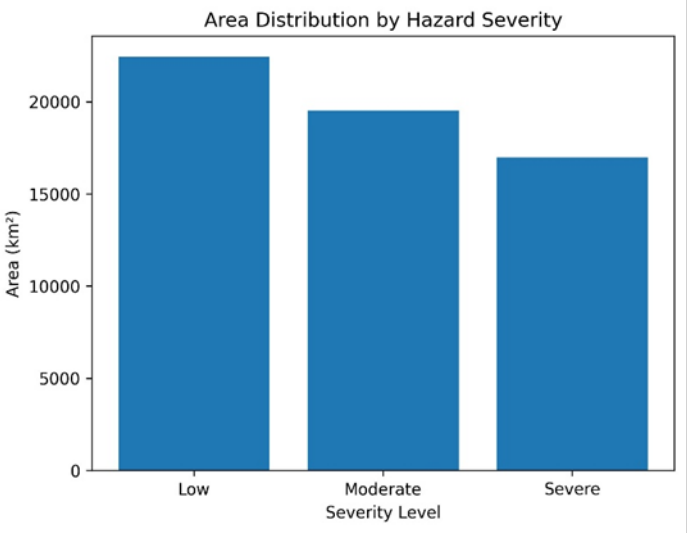

Figure 5. Area distribution by hazard severity level across North Sumatra Province.

The implications of these empirical results are of immediate practical importance for the Regional Disaster Management Agency (BPBD) and the National Disaster Management Agency (BNPB).**)**. Prioritizing mitigation in steep and deforested districts can significantly reduce disaster vulnerability, echoing recommendations (Abdelkader & others, 2025; Lokesh & others, 2025)

The integration of GIS and machine learning provides a decision-support framework that improves transparency, replicability, and policy communication (Arora & others, 2025a; Li & others, 2025). Furthermore, this approach is congruent with the 2021–2030 National Disaster Reduction Agenda (RAN-PRB), a framework that underscores the critical importance of evidence-based strategies in fostering resilience.(Chen & Fan, 2024). Recommended actions include slope stabilization, reforestation, watershed rehabilitation, and improved rainfall monitoring systems (Hou & others, 2025; Pourzangbar & others, 2025).

CONCLUSION

This research effectively classified flood- and landslide-affected areas in North Sumatra using a GIS–machine learning framework guided by the CRISP-DM methodology, integrating rainfall, slope, land cover, soil, and elevation data via the Random Forest algorithm to achieve 93% classification accuracy and an AUC of 0.98, underscoring its exceptional reliability. The findings highlight the western and southern sectors—particularly Mandailing Natal, South Tapanuli, and Humbang Hasundutan—as the most impacted

districts, driven by high precipitation, steep topography, and land-use changes. Key advantages include data-driven prioritization for mitigation and recovery, a reproducible pipeline adaptable to other Indonesian provinces, and enhanced policy integration for environmental and disaster management. For future research, incorporating temporal hazard modeling with climate change projections, hydrological simulations, and social vulnerability indicators would refine predictive accuracy and spatial risk prioritization.

REFERENCES

Abdelkader, M., & others. (2025). Optimizing landslide susceptibility mapping with machine learning and spatial predictors. *Natural Hazards*. https://doi.org/10.1007/s11069-025-07197-0

Ait Naceur, H., & others. (2025). Machine learning-based optimization of flood susceptibility mapping in semi-arid watersheds. *Applied Sciences*.

Al-Kindi, A., & Alabri, Z. (2024). Investigating the Role of Key Conditioning Factors in Flood Susceptibility Mapping. *Earth Systems and Environment*.

Arora, A., & others. (2025a). Machine learning model optimization for flood susceptibility in the Kosi Megafan region. *Scientific Reports*. https://doi.org/10.1038/s41598-025-07403-w

Arora, A., & others. (2025b). Machine Learning Model Optimization for Flood Susceptibility Mapping. *Scientific Reports*, *15*, 1702. https://www.nature.com/articles/s41598-025-07403-w

Ashfaq, S., & others. (2025). Flood susceptibility assessment and mapping using GIS–AHP and statistical models. *Journal of Hydrology*.

Badapalli, P. K., Nakkala, A. B., Kottala, R. B., Gugulothu, S., Hasher, F. F. B., Mishra, V. N., & Zhran, M. (2025). Landslide Susceptibility Level Mapping in Kozhikode, Kerala, Using Machine Learning-Based Random Forest, Remote Sensing, and GIS Techniques. *Land*, *14*(7), 1453. https://doi.org/10.3390/land14071453

Chen, C., & Fan, L. (2024). *Interpretability of ML and DL models for landslide susceptibility mapping*.

Feizbahr, M., Brake, N., Hariri Asli, H., & Woods, K. (2025). Flood Susceptibility Mapping Using Machine Learning and Geospatial-Sentinel-1 SAR Integration for Enhanced Early Warning Systems. *Remote Sensing*, *17*(20), 3471. https://www.mdpi.com/2072-4292/17/20/3471

Hou, C., & others. (2025). Landslide susceptibility analysis based on machine learning and GIS. *Applied Sciences*, *15*(10).

Integrated Flood \\& Landslide Susceptibility Study. (2025). Hybrid machine learning, remote sensing, and SHAP analysis for flood susceptibility mapping. *Frontiers in Environmental Science*.

Ismanto, R. D., & others. (2023). Development of Flood-Hazard-Mapping Model Using Random Forest and Frequency Ratio. *Geomatics and Environmental Engineering*, *17*(6). https://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-f9ce9782-5d3e-40ad-a070-0c1292f04d55

Kunwar, Y., Singh, A., & Moghimi, D. (2025). Integrating GIS and Ensemble Learning Models to Predict Landslide Susceptibility. *SN Applied Sciences*.

https://link.springer.com/article/10.1007/s42452-025-07694-8

Li, W., & others. (2025). *Landslide hazard mapping with geospatial foundation models: adaptability and generalization*.

Lokesh, P., & others. (2025). Machine learning and deep learning-based landslide susceptibility mapping using geospatial techniques. *Hydrological Research*. https://doi.org/10.1016/j.hydres.2024.10.001

Mohammed, O. A. (2025). Spatial Prediction of Landslide Susceptibility Using Data Mining Techniques. *Frontiers in Earth Science*. https://www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2025.1619876/full

Navarro, Z., & others. (n.d.). *Map*.

Nguyen, D. L., & others. (2025). Flood susceptibility mapping using machine learning algorithms in Huong Khe District, Vietnam. *International Journal of Geoinformatics*.

Pourzangbar, A., & others. (2025). Analysis of machine learning utilization in flood susceptibility mapping. *Journal of Flood Risk Management*.

Rahman, Z. U. (2025). Flood susceptibility mapping using supervised machine learning models for transboundary basins. *Natural Hazards*. https://doi.org/10.1080/19475705.2025.2516728

Razavi-Termeh, S. V, Sadeghi-Niaraki, A., & Jelokhani-Niaraki, M. (2025). Flood Susceptibility Mapping Using Optimized Deep Learning Models. *Arabian Journal of Geosciences*. https://link.springer.com/article/10.1007/s13201-025-02548-5

Salafudin, A., & Agboola, G. (2024). Optimizing landslide susceptibility mapping using machine learning and geospatial techniques. *Ecological Informatics*, *81*. https://doi.org/10.1016/j.ecoinf.2024.102583

Ullah, A., & others. (2026). *GIS and machine learning-based spatial prediction of landslide prone zones along highway networks*.

Wahba, M., & others. (2024). Assessing machine learning approaches for flood susceptibility mapping: A comparative study. *Environmental Earth Sciences*. https://doi.org/10.1007/s12665-024-11696-x

Zulfahmi, Z., & others. (2025). GIS-Based Landslide Susceptibility Mapping with Ensemble ML Approach. *Geosciences*, *15*(10). https://doi.org/10.3390/geosciences15100390