Journal of the Civil Engineering Forum, May 2026, 12(2):267-280
DOI 10.22146/jcef.24796
Available Online at https://jurnal.ugm.ac.id/v3/jcef/issue/archive

Machine Learning Approaches to Soil Erosion Risk Mapping: A Comparison between Logistic Regression and Fast Large Margin

Muhammad Ramdhan Olii[1*], Marzieh Mokarram[2], Erwin Anshari[3], Rizky Selly Nazarina Olii[4], Ririn Pakaya[5]

[1]Department of Civil Engineering, Engineering Faculty, Universitas Gorontalo, Gorontalo, INDONESIA
[2]Department of Range and Watershed Management, College of Agriculture and Natural Resources of Darab, Shiraz University, IRAN
[3]Department of Mining Engineering, Faculty of Mathematics and Natural Sciences, Halu Oleo University, Southeast Sulawesi, INDONESIA
[4]Department of Architecture, Engineering Faculty, Universitas Gorontalo, Gorontalo, INDONESIA
[5]Department of Public Health, Public Health Faculty, Universitas Gorontalo, Gorontalo, INDONESIA
*Corresponding author: mr.olii@unigo.ac.id

SUBMITTED 23 September 2025 **REVISED** 13 November 2025 **ACCEPTED** 18 December 2025

ABSTRACT Soil erosion is a critical environmental issue that accelerates land degradation, reduces agricultural productivity, and increases sedimentation in water bodies. Despite its importance, spatial prediction of erosion risk remains a challenge due to the complex interaction of topographic and vegetation-related factors. Previous studies have often overlooked the integration of topographic and remote sensing indices into advanced predictive models, thereby limiting the accuracy of erosion risk mapping. This study aims to evaluate the spatial distribution of soil erosion risk in the Tamalate Watershed, Gorontalo Province, Indonesia, by integrating topographic and remote sensing conditioning factors into Logistic Regression (LR) and Fast Large Margin (FLM) models. Eight conditioning factors—Normalized Difference Moisture Index (NDMI), Terrain Ruggedness Index (TRI), Stream Power Index (STI), Soil Adjusted Vegetation Index (SAVI), Normalized Difference Tillage Index (NDTI), Topographic Wetness Index (TWI), Sediment Power Index (SPI), and Vegetation Condition Index (VCI)—were analyzed using multicollinearity diagnostics and weighted scoring to quantify their relative importance. The results revealed that NDMI (0.253 in LR; 0.258 in FLM) and TRI (0.193 in LR; 0.244 in FLM) were the most influential factors controlling erosion risk, followed by STI (0.186 in LR; 0.166 in FLM). Spatially, both models classified most of the watershed into moderate risk (44.26% in LR; 48.31% in FLM) and high risk (26.09% in LR; 22.35% in FLM) categories, while very high-risk areas were minimal (<0.2%), yet critically important for soil conservation. The findings confirm that integrating topographic and remote sensing indices enhances the precision of erosion risk assessment. This research contributes theoretically and practically by demonstrating the robustness of the FLM approach in soil erosion risk modeling and by providing spatial evidence to support land management and conservation strategies in tropical watershed environments.

KEYWORDS Soil erosion risk; Logistic regression; Fast large margin; Topographic indices; Remote sensing.

1 INTRODUCTION

Soil erosion is widely acknowledged as one of the most pervasive forms of land degradation worldwide, threatening agricultural productivity, ecosystem stability, and sustainable development (Olii, Olii, Olii, Djau, Mokoagow, Kironoto, Bachtiar, Olii and Pakaya, 2025; Olii et al., 2023; Olii, Kironoto, Olii, Pakaya and Olii, 2024). The Food and Agriculture Organization (FAO, 2022) estimates that approximately 24 billion tons of fertile soil are lost each year, primarily due to unsustainable land management practices and climate-induced extreme rainfall events. Beyond its direct impact on agricultural lands, soil erosion contributes to sedimentation in reservoirs, deterioration of water quality, and increased vulnerability of rural communities to flooding and landslides. The Intergovernmental Panel on Climate Change (IPCC, 2022) has further emphasized the urgent need to identify erosion-prone regions as climate variability intensifies and increases the frequency of hydrometeorological events. These global concerns underscore the importance of developing accurate, scalable, and spatially explicit risk mapping techniques that can guide effective soil conservation and land management strategies (Aven and Zio, 2021).

In recent decades, soil erosion risk assessment has evolved from empirical and process-based models, such as the Revised Universal Soil Loss Equation (RUSLE), to advanced geospatial and statistical approaches (Sujatha and Sridhar, 2018; Das et al., 2020; Mahala, 2018; Gaubi et al., 2017). Empirical models remain useful due to their simplicity, but they often rely on generalized assumptions and cannot easily capture the complexity of interactions among multiple environmental and anthropogenic factors (Hong et al., 2017). To overcome these limitations, the integration of remote sensing data, Geographic Information Sys-

tems (GIS), and machine learning has emerged as a promising direction (Olii, Nento, Doda, Olii, Djafar and Pakaya, 2025). Numerous studies have applied algorithms such as Random Forest (Ghosh and Maiti, 2021), Gradient Boosted Trees (Naceur et al., 2024), and Support Vector Machines (Olii, Nento, Doda, Olii, Djafar and Pakaya, 2025) for erosion risk mapping. These machine learning techniques have demonstrated superior predictive accuracy compared to traditional statistical models, owing to their ability to capture nonlinear relationships and complex interactions among conditioning factors such as topography, rainfall, soil properties, and land cover (Olii, Olii, Olii, Djau, Mokoagow, Kironoto, Bachtiar, Olii and Pakaya, 2025; Olii, Nento, Doda, Olii, Djafar and Pakaya, 2025).

Among statistical approaches, Logistic Regression (LR) has been widely used in natural hazard assessment, particularly in landslide risk studies, due to its interpretability, relatively low computational requirements, and probabilistic outputs (Lee and Pradhan, 2007). However, its application to soil erosion risk prediction remains limited and has not been systematically examined in an index-based framework. Likewise, large-margin classifiers such as Support Vector Machines have shown strong performance in environmental classification. Still, standard implementations are computationally intensive and scale poorly with large, high-dimensional geospatial datasets (Baiddah et al., 2023; Nguyen et al., 2023). Fast Large Margin (FLM) algorithms have been developed to overcome these computational constraints, yet to date, they have not been applied to soil erosion risk prediction. This gap in the literature restricts our understanding of how LR and FLM perform when applied to soil erosion risk mapping, particularly in terms of predictive accuracy, spatial generalization, and practical interpretability.

This study aims to develop an index-based soil erosion risk mapping framework that integrates topographic and remote sensing indicators, while simultaneously conducting a comparative evaluation of Logistic Regression and fast large margin classifiers. Through this approach, the study contributes to advancing methodological practices in soil erosion risk modeling. It provides practical insights for sustainable land and watershed management under increasing pressures from climate change and intensified land use. The research was conducted in the Tamalate Watershed (DAS Tamalate), located in Gorontalo Province, Indonesia. This watershed was selected as the study area due to its critical ecological and socio-economic importance, encompassing extensive agricultural lands and rapidly expanding settlements. In recent years, the Tamalate Watershed has experienced notable land use changes and increased erosion rates, largely driven by unsustainable agricultural practices and deforestation. These characteristics make it an ideal representative area for evaluating the applicability and performance of advanced erosion risk modeling techniques in tropical watershed environments.

2 MATERIALS AND METHODS

2.1 Study Area and Data Sources

The Tamalate Watershed, situated in Gorontalo Province, Indonesia, is a sub-catchment that contributes significantly to the hydrological regime of Gorontalo City. Geographically, it lies between approximately 0°31’53.5”N–0°38’52.5” N latitude and 123°3’24.5”–123°13’10.5” E longitude, encompassing an area of about 105.44 km² (Figure 1). The watershed exhibits a tropical humid climate, with mean annual rainfall ranging from 909 to 2,877 mm, predominantly influenced by monsoonal circulation. The watershed exhibits a tropical humid climate, with mean annual rainfall ranging from 909 to 2,877 mm (NASA/POWER database for the period 1981–2024), predominantly influenced by monsoonal circulation. Its topography is highly variable, consisting of steep mountainous headwaters, rolling mid-slopes, and relatively flat lowland areas. This geomorphic gradient drives diverse hydrological and erosional processes. Land use within the Tamalate watershed is dominated by mixed agriculture, settlements, and fragmented forest patches. Unsustainable cultivation on steep slopes and rapid land conversion have intensified erosion, contributing to sedimentation in Lake Limboto. Given its strategic location and ecological importance, the Tamalate watershed represents a critical area for soil erosion risk analysis and sustainable watershed management planning.

The spatial datasets employed in this study were obtained from multiple reliable sources to ensure both accuracy and consistency. The Digital Elevation Model (DEM) was downloaded from the Earth Explorer platform (https://earthexplorer.usgs.gov/), derived from the Landsat 9 dataset (LC08_L1TP_113060_20241230_20250104_02_T1) acquired on December 30, 2024. This imagery was chosen due to its high spatial resolution and temporal relevance to the study area. Administrative boundaries were sourced from the Global Administrative Areas Database (GADM) database (https://gadm.org/), which provides detailed and regularly updated geographic boundary information suitable for scientific analysis. Furthermore, erosion and non-erosion sites were identified using high-resolution satellite imagery interpreted via Google Earth, with acquisition dates carefully aligned with the Landsat 9 imagery of December 30, 2024. Such synchronization across data

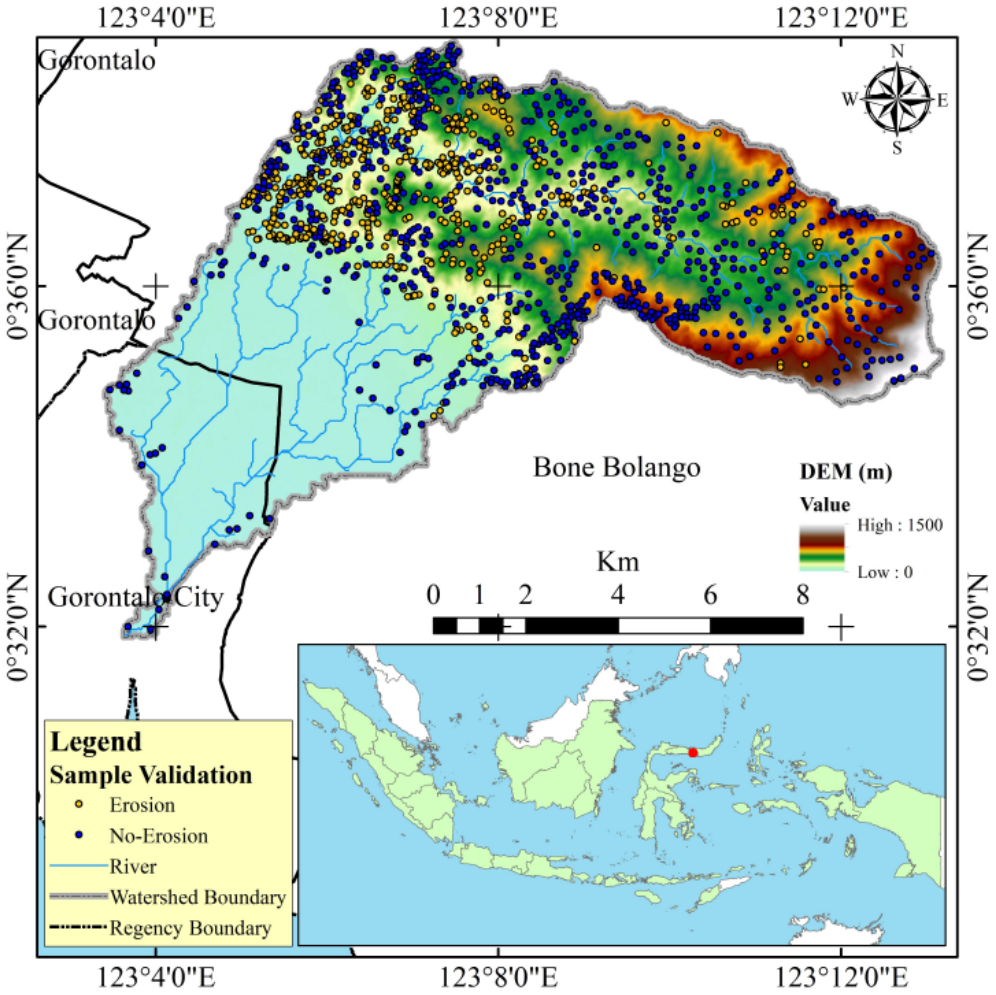


Figure 1. Map Showing The Study Area

sources ensured that environmental conditions were temporally consistent, thereby reducing uncertainties. By combining multi-source datasets with temporal alignment, the study strengthens the reliability of soil erosion risk modeling and enhances the robustness of predictive performance.

2.2 Selection of Conditioning Factors

Soil erosion risk was modeled using eight conditioning factors that represent both topographic and land surface dynamics. Four topographic indices—Sediment Transport Index (STI), Topographic Wetness Index (TWI), Terrain Ruggedness Index (TRI), and Stream Power Index (SPI)—were derived from the DEM using GIS-based terrain analysis tools. In parallel, four remote sensing-derived indices were extracted from Landsat-9 imagery: Normalized Difference Tillage Index (NDTI), Normalized Difference Moisture Index (NDMI), Soil Adjusted Vegetation Index (SAVI), and Vegetation Condition Index (VCI). These indices capture surface roughness, moisture distribution, vegetation cover, and tillage practices, all of which are critical drivers of erosion processes.

2.3 Data Preprocessing

Preprocessing steps in this study primarily focused on ensuring data uniformity and comparability across different spatial datasets. Landsat-9 imagery was resampled to a consistent spatial resolution of 30 m to match the DEM and other auxiliary datasets. All conditioning factors were subsequently normalized to a uniform ordinal scale ranging from 1 to 5, representing low to high risk levels. This transformation enabled a more interpretable comparison between topographic and remote sensing indices while preserving their relative importance. To address potential redundancy among predictors, collinearity analysis was performed using the Variance Inflation Factor (VIF), Tolerance (TOL), and coefficient of determination (Coefficient of Determination, R^2), where variables with VIF values greater than 10, TOL values lower than 0.1, or excessively high R^2 values indicating multicollinearity were excluded from the analysis (Miles, 2014). In order to reduce sampling bias, balanced datasets were generated through strati-

Table 1. Classification of Topographic and Remote Sensing Indices into Soil Erosion Risk Scores

Indices	Soil Erosion Risk Factor	Class	Score
Topographic	Sediment Transport Index (STI)	<5	1
		5 – 10	2
		10 -20	3
		20 – 40	4
		>40	5
	Topographic Wetness Index (TWI)	<4	1
		4 – 8	2
		8 – 12	3
		12 – 16	4
		>16	5
	Terrain Ruggedness Index (TRI)	<0.1	1
		0.1 – 0.2	2
		0.2 – 0.3	3
		0.3 – 0.4	4
		>0.4	5
	Stream Power Index (SPI)	<2	1
		2 – 4	2
		4 – 6	3
		6 – 8	4
		>8	5
Remote Sensing	Normalized Difference Tillage Index (NDTI)	<-0.4	1
		-0.4 – -0.2	2
		-0.2 – 0.0	3
		0.0 – 0.2	4
		>0.2	5
	Normalized Difference Moisture Index (NDMI)	>0.3	1
		0.1 – 0.3	2
		-0.1 – 0.1	3
		-0.3 – -0.1	4
		<-0.3	5
	Soil Adjusted Vegetation Index (SAVI)	>0.8	1
		0.6 – 0.8	2
		0.4 – 0.6	3
		0.2 – 0.4	4
		<0.2	5
	Vegetation Condition Index (VCI)	>80	1
		60 – 80	2
		40 – 60	3
		20 – 40	4
		<20	5

fied random sampling of soil erosion risk and non-risk areas, which were derived from expert-based soil erosion inventories.

Furthermore, all spatial datasets were processed and integrated using ArcGIS 10.8 software, ensuring accurate spatial alignment and data management. For model development, AI Studio 2025.0.1 was employed to implement the machine learning framework, particularly for optimizing the weighting of conditioning factors in both LR and FLM models. These steps ensured robust and reliable inputs for subsequent modeling.

2.4 Machine Learning Models

Two classification algorithms were employed for soil erosion risk prediction:

a) Logistic Regression (LR): A statistical model estimating the probability of soil erosion occurrence as a function of conditioning factors. The model was fitted using a maximum likelihood approach, with ridge regularization applied to reduce overfitting and improve stability in the presence of correlated predictors (Hosmer et al., 2013).
b) Fast Large Margin (FLM): A scalable margin-based classifier optimized for high-dimensional data. Unlike conventional Support Vector Machines, FLM utilizes a stochastic gradient approximation to efficiently handle large geospatial datasets while maintaining high classification accuracy. Model hyperparameters, including margin penalty and kernel functions, were optimized via nested cross-validation (Özer et al., 2021).

2.5 Model Training and Validation

The dataset, consisting of 1,346 total locations—including 553 soil erosion risk sites and 793 non-soil erosion sites—was randomly divided into training (60%) and testing (40%) subsets, resulting in 808 locations for training and 538 locations for testing (Figure 1). Model performance was evaluated using accuracy, classification error, F-measure, precision, recall, sensitivity, and specificity to comprehensively assess the predictive capability of both LR and FLM models. In addition, the AUC–ROC curve was used to evaluate the overall discriminative ability of the models, while the confusion matrix provided detailed insight into classification performance through true and false prediction distributions.

2.6 Soil Erosion Risk Mapping

To evaluate and map the spatial distribution of soil erosion risk, a weighted linear combination approach was applied by integrating the classification scores of conditioning factors (Table 1) with their respective importance weights (Table 3).

$$SER = \sum_{i=1}^{n} (W_i \times Score_i) \quad (1)$$

where *SER* is the total soil erosion risk value, W_i is the weight of factor *i*, and $Score_i$ is the class score of conditioning factor *i* (Olii, Olii, Olii, Pakaya and Kironoto, 2024). The results of Eq. 1 generated from the LR and FLM models were rescaled to a continuous soil erosion risk index ranging from 0 to 1. For each grid cell, the total weights of conditioning factors were

Table 2. Multicollinearity Diagnostics of Conditioning Factors Based on R^2, Tolerance (TOL), and Variance Inflation Factor (VIF)

Factor	R^2	TOL	VIF
VCI	0.565	0.435	2.299
SAVI	0.507	0.493	2.030
NDMI	0.157	0.843	1.186
NDTI	0.184	0.816	1.225
TWI	0.556	0.444	2.250
SPI	0.413	0.587	1.705
TRI	0.128	0.872	1.146
STI	0.313	0.687	1.456

Table 3. Relative Importance Weights of Conditioning Factors Derived from LR and FLM Models

Conditioning Factors	Model	
	LR	FLM
NDMI	0.253	0.258
TRI	0.193	0.244
STI	0.186	0.166
SAVI	0.120	0.146
NDTI	0.101	0.096
TWI	0.099	0.094
SPI	0.073	0.092
VCI	0.043	0.077

normalized to ensure comparability across the study area. The normalized values were then classified into five risk categories using equal interval division, where each class represented a range of 0.2: very low (0.0–0.2), low (0.2–0.4), moderate (0.4–0.6), high (0.6–0.8), and very high (0.8–1.0) (Bui et al., 2020; Dinh et al., 2021). The use of equal-interval risk thresholds follows common practice in soil erosion and spatial risk mapping studies, ensuring consistency, comparability, and objective class boundaries. This method provides a simple and reproducible framework for comparing results across models and study areas (Olii, Olii, Olii, Djau, Mokoagow, Kironoto, Bachtiar, Olii and Pakaya, 2025; Baiddah et al., 2023). In contrast, alternative approaches such as natural breaks, quantile, or expert-based thresholds may lead to uneven class distributions or subjective class limits. Accordingly, this classification ensures a systematic interpretation of spatial soil erosion risk. The resulting risk maps were then validated using field-observed erosion points and expert-based inventories to assess both spatial accuracy and ecological plausibility of the model outputs.

3 RESULTS

The assessment of soil erosion risk requires the integration of multiple environmental indicators that capture both terrain characteristics and land surface conditions. Table 1 and Figure 2 provide a detailed classification of topographic and remote sensing indices into

Table 4. Performance Evaluation Metrics of LR and FLM Models

Statistical Metrics	Model	
	LR	FLM
AUC	0.763	0.749
Accuracy	0.706	0.693
Classification error	0.294	0.307
F-measure	0.785	0.780
Precision	0.690	0.673
Recall	0.911	0.929
Sensitivity	0.911	0.929
Specificity	0.411	0.354

soil erosion risk scores, serving as the basis for evaluating the spatial variability of soil erosion risk across the study area. Topographic factors such as the STI, TWI, TRI, and SPI were selected to represent hydrological flow dynamics, slope steepness, surface roughness, and the erosive power of flowing water. These indices are directly linked to soil detachment and transport processes, which are critical drivers of soil erosion in complex landscapes. In addition, remote sensing-based indices, including the NDTI, NDMI, SAVI, and VCI, were incorporated to provide information on vegetation cover, soil moisture conditions, and land management practices. By classifying each index into five classes with corresponding scores ranging from 1 (very low risk) to 5 (very high risk), the approach ensures a standardized evaluation of soil erosion risk factors from diverse data sources (Olii, Olii, Olii, Djau, Mokoagow, Kironoto, Bachtiar, Olii and Pakaya, 2025). This classification framework not only facilitates the comparison between different conditioning factors but also supports their integration into statistical and machine learning models for predicting soil erosion risk areas with higher accuracy.

Table 2 shows the results of multicollinearity diagnostics for the conditioning factors used in soil erosion risk modeling. The analysis employed R^2, TOL, and VIF to check the degree of correlation among variables. Factors such as VCI (R^2 = 0.565) and TWI (R^2 = 0.556) exhibit moderate correlation with others, while NDMI (0.157) and TRI (0.128) show lower relationships. TOL values for all factors are above 0.4, well above the critical threshold of 0.1, indicating that each factor provides a sufficient independent contribution. Similarly, VIF values range between 1.146 and 2.299, far below the acceptable limit of 10, confirming the absence of significant multicollinearity. These results suggest that all conditioning factors (VCI, SAVI, NDMI, NDTI, TWI, SPI, TRI, and STI) are statistically reliable and can be confidently used in further analysis without the risk of distortion due to multicollinearity.

Table 3 presents the relative importance weights of conditioning factors in soil erosion risk modeling, de-

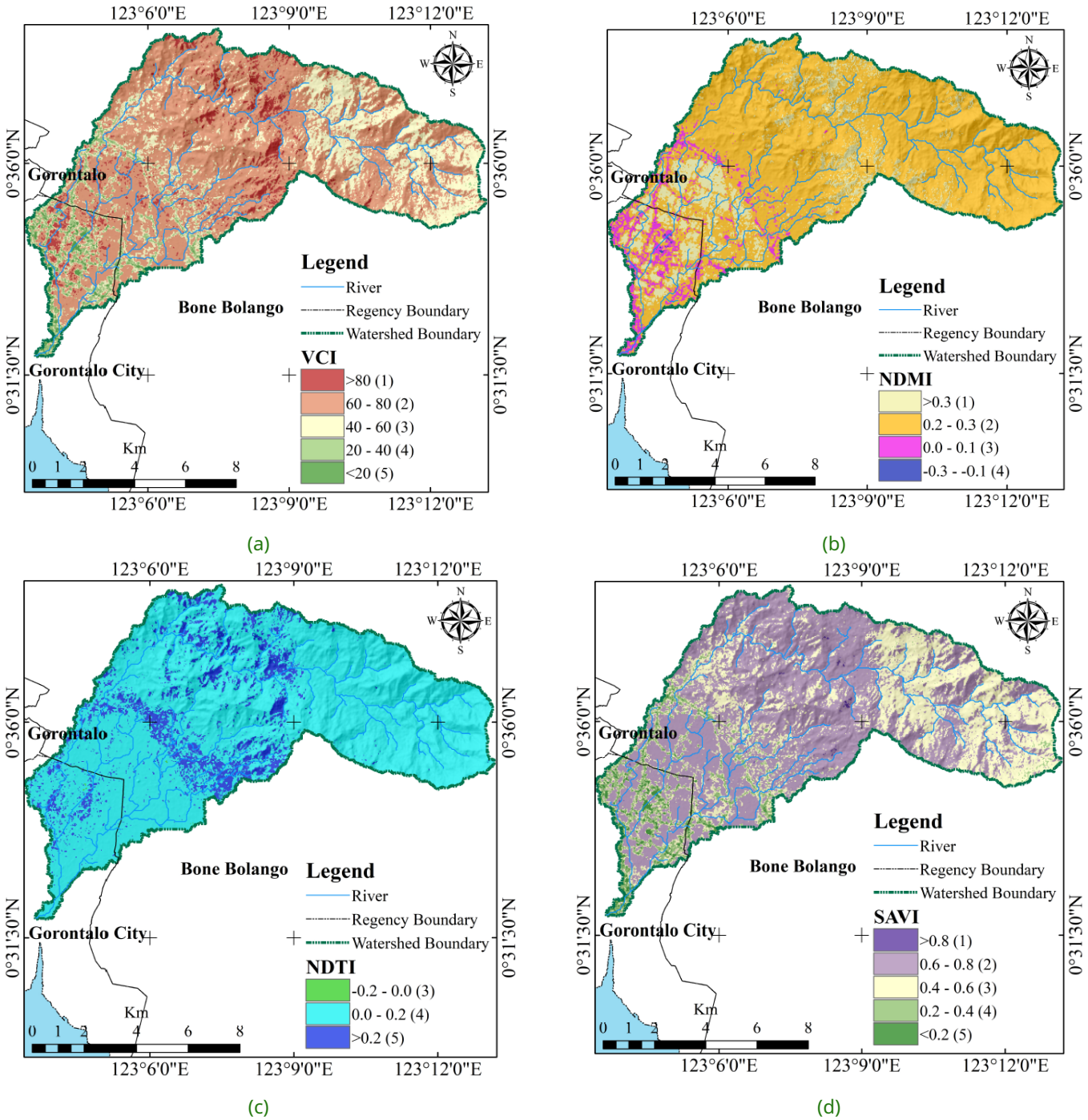
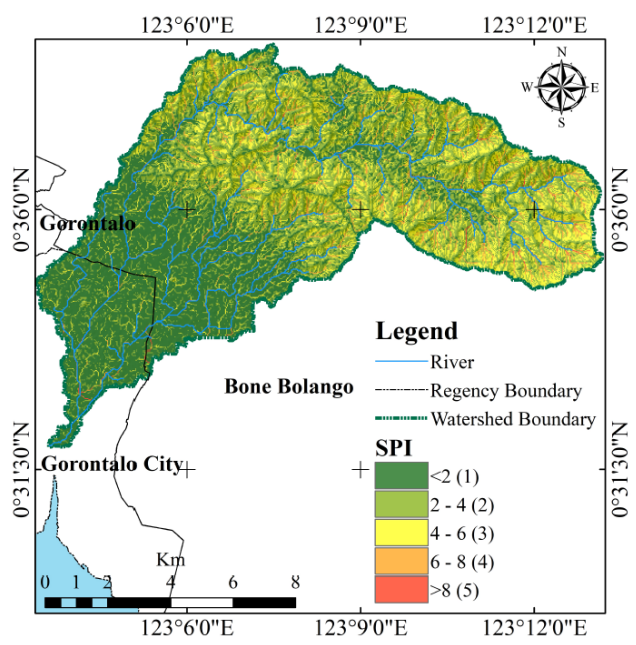
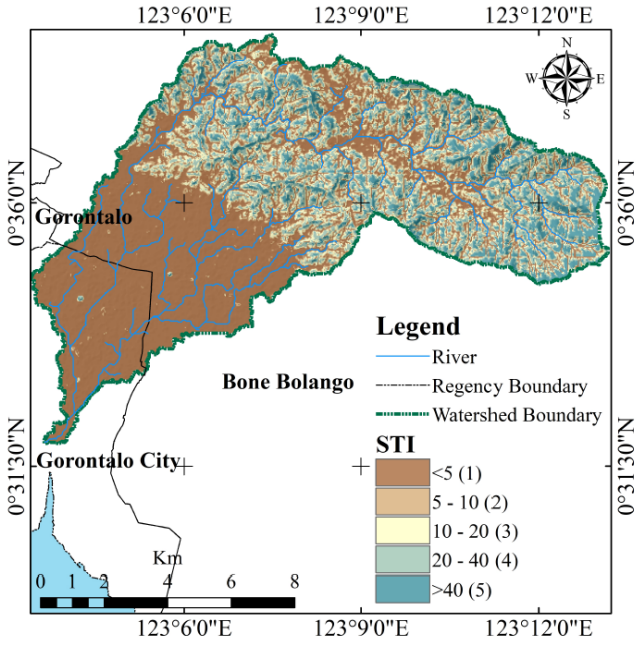


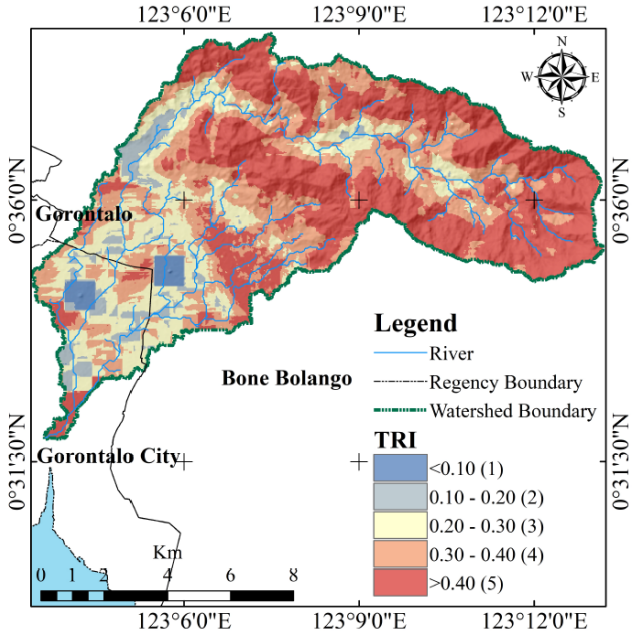
Figure 2. Spatial Distribution of Topographic and Remote Sensing Indices for Soil Erosion Risk



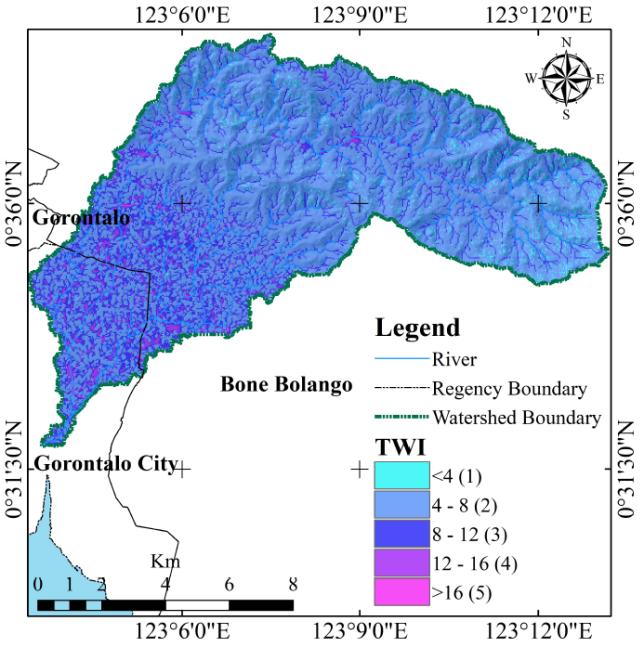
(e)



(f)



(g)



(h)

Figure 2. Spatial Distribution of Topographic and Remote Sensing Indices for Soil Erosion Risk (continued)

Table 5. Confusion Matrix of LR

	True - Soil Erosion Sites	True - Non-Soil Erosion Sites	Class Precision
Prediction of Soil Erosion Sites	65	20	76.47%
Prediction of Non-Soil Erosion Sites	93	206	68.90%
Class Recall	41.14%	91.15%	

Table 6. Confusion Matrix of FLM

	True - Soil Erosion Sites	True - Non-Soil Erosion Sites	Class Precision
Prediction of Soil Erosion Sites	56	16	77.78%
Prediction of Non-Soil Erosion Sites	102	210	67.31%
Class Recall	35.44%	92.92%	

rived from LR and FLM. Among the indices, NDMI consistently shows the highest contribution in both LR (0.253) and FLM (0.258), highlighting its strong influence on soil erosion risk. This is followed by TRI with notable weights (0.193 in LR and 0.244 in FLM), suggesting that terrain ruggedness is also a critical factor. STI ranks third in LR (0.186), while its importance slightly decreases in FLM (0.166). Vegetation-related indices, such as SAVI and VCI, show lower but meaningful contributions, particularly in FLM, where VCI increases to 0.077 compared to 0.043 in LR. Meanwhile, NDTI, TWI, and SPI exhibit moderate weights across both models. Overall, the table indicates that moisture and topographic factors play dominant roles, while vegetation indices gain relatively higher importance under the FLM approach.

Table 4 presents the performance evaluation of the LR and FLM models using six statistical metrics. The AUC values indicate that both models achieved good discriminatory power, with LR (0.763) slightly outperforming FLM (0.749) (Figure 3). In terms of overall accuracy, LR also performed better (0.706) compared to FLM (0.693), while classification error was slightly lower in LR (0.294) than in FLM (0.307). The F-measure values were high for both models (0.785 for LR; 0.780 for FLM), confirming a strong balance between precision and recall. Precision, however, was somewhat higher in LR (0.690) than in FLM (0.673). Interestingly, recall values show that FLM (0.929) identified positive cases more effectively than LR (0.911), highlighting its sensitivity to soil erosion risk areas. This trend is further supported by the confusion matrices (Tables 5 and 6), where both models show high recall for non-soil erosion sites (>91%), while LR demonstrates slightly higher recall and precision for soil erosion prediction compared to FLM. Overall, LR demonstrated marginally higher accuracy and precision, whereas FLM showed superior recall, suggesting its strength in detecting soil erosion risk even at the expense of slightly

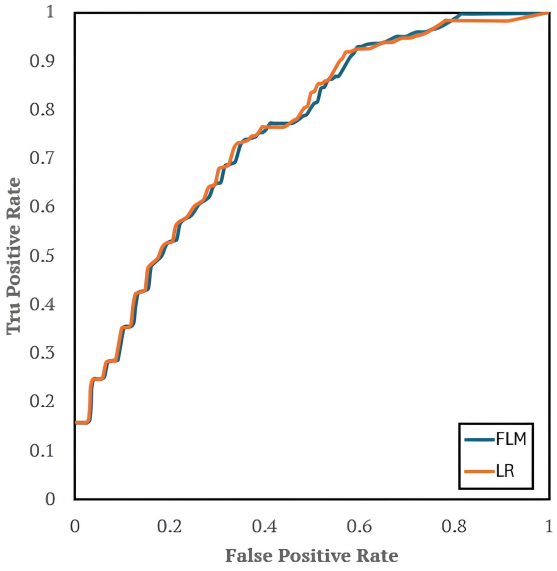


Figure 3. AUC Comparison Graph

lower precision.

Table 7 presents the spatial distribution of soil erosion risk classes derived from the LR and FLM models. The results indicate that both approaches yield comparable patterns, with some notable differences in the allocation of risk classes. In the LR model, the moderate risk class dominates, covering 50,991 grids (44.26% of the total area). Similarly, the FLM model also identifies the moderate class as the largest, with a slightly higher proportion of 55,653 grids (48.31%). This consistency demonstrates that moderate soil erosion risk is the prevailing condition in the study area according to both models.

The low-risk class is estimated at 24.31% (28,009 grids) in LR and 25.04% (28,844 grids) in FLM, representing relatively stable zones with limited vulnerability. The high-risk class, however, shows variation: LR identifies 26.09% (30,058 grids), while FLM estimates a lower share of 22.35% (25,747 grids). This reduction in the FLM output suggests that the model reallocates a portion of highly susceptible areas into the moderate category, providing a more balanced classification. At the extremes, the very low-risk class remains minor, at 5.17% in LR and 4.20% in FLM, while the very high-risk class is negligible in both models, below 0.2%. The total number of grids (115,211) is consistent across both approaches, ensuring data reliability. In summary, both LR and FLM highlight the dominance of moderate soil erosion risk. Still, the FLM model refines the spatial distribution by slightly lowering the high-risk extent and increasing moderate and low classes, offering a more nuanced perspective on soil erosion risk.

As shown in Figure 4, areas with low NDMI—indicating reduced vegetation moisture—combined with high TRI

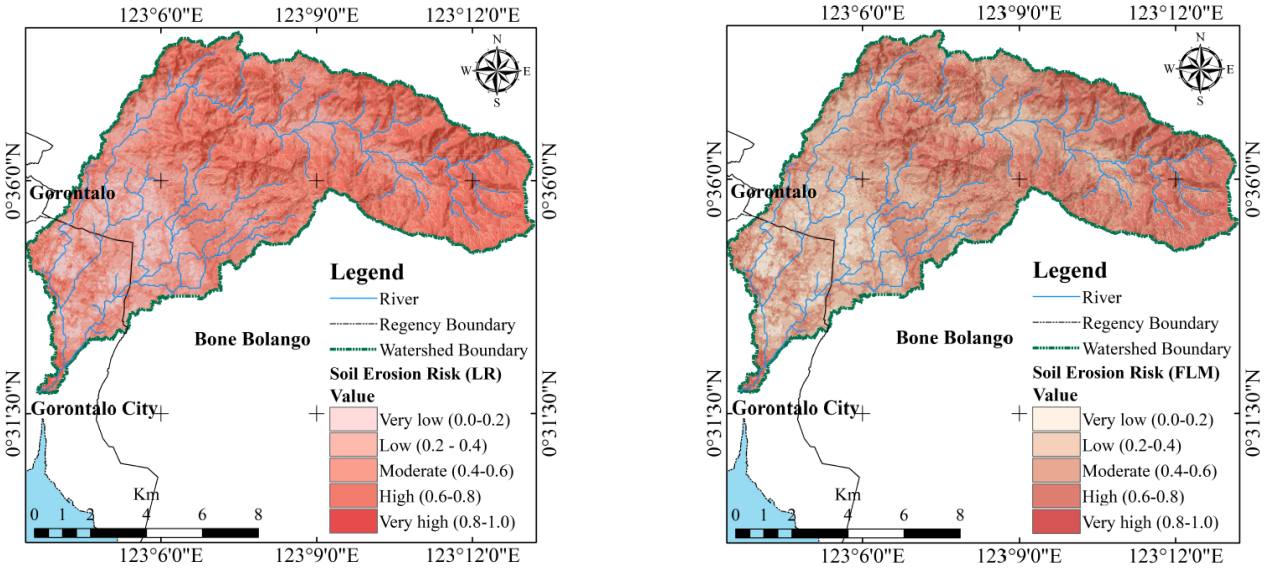


Figure 4. Comparison of Soil Erosion Risk Classification between LR and FLM Models

and STI values—representing steep terrain and strong flow transport capacity—correspond to zones of high soil erosion risk. NDMI exhibited the highest relative weight, underscoring the critical role of vegetation cover and soil moisture in maintaining slope stability. The decline in NDMI values is largely attributed to land cover changes from dense natural vegetation to dryland agriculture, particularly corn cultivation on sloping lands exceeding 15%, which substantially decreases canopy density, soil shading, and moisture retention capacity. This conversion exposes the soil surface to raindrop impact and accelerates surface runoff, thereby increasing erosion vulnerability. The upper part of the Tamalate Watershed, where steep slopes coincide with extensive agricultural expansion, shows the combined conditions of low NDMI, high TRI, and high STI, making it the most soil erosion-risk area.

In contrast, the downstream areas exhibit low soil erosion risk, even though vegetation density is relatively sparse. This is primarily because the slope gradient becomes gentler toward the lower reaches, which diminishes runoff energy and sediment transport capacity. The sparse vegetation in these downstream zones is mainly due to the dominance of residential and paddy field areas, where land surfaces are either built-up or irrigated, thus limiting erosion occurrence. These results confirm that corn cultivation on slopes greater than 15%, coupled with reduced vegetation, markedly accelerates soil erosion, highlighting the urgent need for soil conservation, contour farming, and reforestation strategies in the upland zones to maintain watershed stability.

4 DISCUSSIONS

The present study underscores the critical role of hydrological and topographic indices in shaping soil erosion risk within the study area. Among the conditioning factors, the NDMI, TRI, and STI consistently emerged as the most influential predictors in both LR and FLM models. This outcome reflects the fundamental processes driving soil erosion, where the interaction between terrain steepness, runoff concentration, and soil moisture plays a dominant role. NDMI indicates the balance between vegetation cover and soil moisture, which controls infiltration capacity and resistance to detachment (Golkarian et al., 2023). TRI highlights the degree of slope irregularity, which regulates water flow velocity and erosion pathways, while STI captures the erosive energy of accumulated runoff as it interacts with slope gradients (Olii et al., 2023; Jaafari et al., 2014). By contrast, vegetation indices such as SAVI and VCI showed lower predictive power, likely due to their localized influence and seasonal variability (Senanayake et al., 2020; Musasa et al., 2024; Amanollahi et al., 2025). These findings confirm that soil erosion risk is more strongly tied to long-term geomorphological and hydrological dynamics than to short-term vegetation fluctuations. Such patterns demonstrate that soil erosion is not evenly distributed across the landscape, but is concentrated in areas where hydrological processes are amplified by steep and rugged terrain, leading to disproportionate sediment mobilization. The identification of these dominant factors validates their robustness as transferable indicators for soil erosion modeling, particularly in tropical landscapes where climatic and terrain interactions intensify soil degradation processes (Amanollahi et al., 2025).

Table 7. Spatial Distribution of Soil Erosion Risk Classes Derived from LR and FLM Models

Soil Erosion Risk	LR			FLM		
	Grid	Area (km^2)	Area (%)	Grid	Area (km^2)	Area (%)
Very low	5961	5.36	5.17	4835	4.35	4.20
Low	28009	25.21	24.31	28844	25.96	25.04
Moderate	50991	45.89	44.26	55653	50.09	48.31
High	30058	27.05	26.09	25747	23.17	22.35
Very high	192	0.17	0.17	132	0.12	0.11
Total	115211	103.69	100	115211	103.69	100

A notable feature of the results lies in the convergence of predictor weightings across LR and FLM, suggesting that the significance of conditioning factors is driven by underlying natural processes rather than model choice. The stability of NDMI, TRI, and STI across both approaches indicates that these predictors are fundamental drivers of soil erosion and, therefore, likely to retain importance in diverse settings. This convergence enhances the credibility of the findings, as consistency across distinct algorithms minimizes the risk of model-specific artifacts. Previous studies lend strong support to this interpretation. For instance, Olii, Olii, Olii, Djau, Mokoagow, Kironoto, Bachtiar, Olii and Pakaya (2025) in the Saddang watershed of Sulawesi also found hydrological and topographic indices to outweigh vegetation factors in predictive performance (Olii, Olii, Olii, Djau, Mokoagow, Kironoto, Bachtiar, Olii and Pakaya, 2025; Olii, Nento, Doda, Olii, Djafar and Pakaya, 2025), while Chen and Zhang (2022) in the northwest of Yan'an City, Shaanxi Province, similarly reported the primacy of slope- and runoff-related indices in shaping gully erosion (Chen and Zhang, 2022). Furthermore, Sahour et al. (2021) demonstrated that multiple machine learning models converged on similar variable importance rankings, with slope and hydrological factors emerging as dominant drivers (Sahour et al., 2021). This methodological robustness highlights that the soil erosion process is consistently controlled by terrain–hydrology interactions across regions and modeling frameworks. From a practical standpoint, the stability of factor importance across models simplifies the decision-making process for practitioners, as it suggests that investing in accurate terrain and hydrological datasets may be more crucial than pursuing algorithmic complexity. This also indicates that relatively simple models can perform comparably to more advanced classifiers, provided that critical conditioning factors are properly represented in the dataset.

While the conditioning factors demonstrated stability, the comparative performance of LR and FLM revealed meaningful differences that highlight complementary strengths of the two models. LR achieved slightly higher overall performance in terms of AUC (0.763), accuracy (0.706), and precision (0.690), which indicates its strength in producing balanced classifications and minimizing false positives. Conversely, FLM attained a higher recall (0.929), demonstrating superior sensitivity in detecting true soil erosion risk areas, albeit at the expense of a slightly higher classification error (0.307). This trade-off reflects a broader pattern reported in the literature: regression-based models often deliver consistent and balanced performance, whereas margin-based classifiers such as FLM or SVM excel at identifying minority or high-risk classes (Mustafa et al., 2018; Fernández et al., 2023). From a management perspective, these differences are significant. LR may be more suitable for applications requiring balanced zoning of soil erosion risk for long-term land-use planning, while FLM is advantageous in early warning systems where failing to identify high-risk zones could lead to severe environmental or economic costs. The spatial risk distribution supports this interpretation: both models classified over 70% of the area as moderate to high risk, but the concentration of very high-risk zones along steep and hydrologically active slopes illustrates the FLM's advantage in capturing extreme cases.

From a broader perspective, the findings carry important scientific and practical implications. Scientifically, the study demonstrates that algorithmic complexity does not necessarily guarantee superior predictive performance, as simpler models can yield insights comparable to advanced classifiers when key conditioning factors are included. This echoes the argument by Mohammed et al. (2025), who stressed that model input quality often outweighs algorithm sophistication in predictive geomorphology (Mohammed et al., 2025). Practically, the results suggest that watershed managers can prioritize the acquisition and integration of high-quality hydrological and terrain data to build reliable soil erosion risk models, irrespective of the chosen algorithm. Nevertheless, several limitations must be acknowledged. First, critical variables such as rainfall erosivity, soil type, and land management practices were not incorporated, which may constrain predictive accuracy. Second, this study relied on spatially cross-validated data but did not account for temporal dynamics, thereby overlooking interannual variability in rainfall intensity and vegetation cover. Finally, while LR

and FLM provided robust predictions, ensemble and deep learning approaches, including Random Forest, Gradient Boosting, or Convolutional Neural Networks, could potentially enhance accuracy, as recently shown in several studies (Sahour et al., 2021; Band et al., 2020; Khosravi et al., 2023). Future research should therefore integrate temporal datasets, include additional environmental predictors, and apply ensemble-based models to strengthen predictive performance and generalizability.

5 CONCLUSIONS

This study integrated topographic and remote sensing indices through LR and FLM models to generate a spatial distribution of soil erosion risk. Results revealed that NDMI, TRI, and STI were the most influential factors shaping soil erosion risk, with vegetation-related indices such as SAVI also playing a substantial role. Specifically, the relative importance weights showed that NDMI contributed the most (0.253 in LR and 0.258 in FLM), followed by TRI (0.193 in LR and 0.244 in FLM) and STI (0.186 in LR and 0.166 in FLM), highlighting the dominant influence of moisture and topographic variability on erosion susceptibility. The spatial distribution analysis further indicated that moderate to high soil erosion risk classes dominate more than 70% of the watershed area, underscoring the urgency of targeted soil conservation efforts. In particular, moderate risk covered 44.26% in LR and 48.31% in FLM, while high risk accounted for 26.09% and 22.35%, respectively, indicating the predominance of erosion-prone zones across the watershed.

The novelty of this research lies in combining statistical learning with geospatial indices to produce reliable, transferable soil erosion risk maps, while addressing multicollinearity and enhancing prediction accuracy. Model performance metrics also demonstrated consistent reliability, with AUC values of 0.763 for LR and 0.749 for FLM, and accuracy rates above 0.69, confirming the robustness of both approaches. Although FLM exhibited slightly higher classification error (0.307), it achieved higher sensitivity and recall (0.929) compared to LR (0.911), suggesting better detection of erosion-prone areas. Beyond methodological contributions, the findings provide a practical decision-support tool for policymakers and land managers in prioritizing interventions, especially in soil erosion risk tropical environments.

Future work should incorporate soil physicochemical properties, rainfall erosivity, and land management practices to refine model accuracy. Expanding validation across diverse geographic settings will also strengthen the applicability of this framework for sustainable land-use planning and soil erosion risk mitigation at larger scales.

DISCLAIMER

Data supporting this study are available from the corresponding author upon reasonable request.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Faculty of Engineering, Universitas Gorontalo, for providing financial support through its internal research grant program.

REFERENCES

Amanollahi, J., Gharibi, S. and Rastkhadiv, A. (2025), 'Prediction of soil erosion control ecosystem service using machine learning based on the ANN model in Asia', *Environmental and Sustainability Indicators* **26**, 100723.
URL: *https://doi.org/10.1016/j.indic.2025.100723*

Aven, T. and Zio, E. (2021), 'Globalization and global risk: How risk analysis needs to be enhanced to be effective in confronting current threats', *Reliability Engineering and System Safety* **205**, 107270.
URL: *https://doi.org/10.1016/j.ress.2020.107270*

Baiddah, A., Krimissa, S., Hajji, S., Ismaili, M., Abdelrahman, K., El Bouzekraoui, M., Eloudi, H., Elaloui, A., Khouz, A., Badreldin, N. and Namous, M. (2023), 'Head-cut gully erosion susceptibility mapping in semi-arid region using machine learning methods: insight from the High Atlas, Morocco', *Frontiers in Earth Science* **11**, 1–19.
URL: *https://doi.org/10.3389/feart.2023.1184038*

Band, S. S., Janizadeh, S., Saha, S., Mukherjee, K., Khosrobeigi Bozchaloei, S., Cerdà, A., Shokri, M. and Mosavi, A. (2020), 'Evaluating the efficiency of different regression, decision tree, and Bayesian machine learning algorithms in spatial piping erosion susceptibility using ALOS/PALSAR data', *Land* **9**(10), 1–22.
URL: *https://doi.org/10.3390/land9100346*

Bui, D. T., Tsangaratos, P., Nguyen, V. T., Van Liem, N. and Trinh, P. T. (2020), 'Comparing the prediction performance of a deep learning neural network model with conventional machine learning models in landslide susceptibility assessment', *Catena* **188**, 104426.
URL: *https://doi.org/10.1016/j.catena.2019.104426*

Chen, B. and Zhang, X. (2022), 'Effects of slope vegetation patterns on erosion sediment yield and hydraulic parameters in slope-gully system', *Ecological Indicators* **145**, 109723.
URL: *https://doi.org/10.1016/j.ecolind.2022.109723*

Das, B., Bordoloi, R., Thungon, L. T., Paul, A., Pandey, P. K., Mishra, M. and Tripathi, O. P. (2020), 'An integrated approach of GIS, RUSLE and AHP to model

soil erosion in West Kameng watershed, Arunachal Pradesh', *Journal of Earth System Science* **129**(1), 1–18.
URL: *https://doi.org/10.1007/s12040-020-1356-6*

Dinh, T. V., Hoang, N. D. and Tran, X. L. (2021), 'Evaluation of different machine learning models for predicting soil erosion in tropical sloping lands of Northeast Vietnam', *Applied and Environmental Soil Science* **2021**, 1–14.
URL: *https://doi.org/10.1155/2021/6665485*

FAO (2022), *FAO Statistical Yearbook – World Food and Agriculture*, FAO, Rome, Italy.
URL: *https://doi.org/10.1016/S0140-6736(59)91820-3*

Fernández, D., Adermann, E., Pizzolato, M., Pechenkin, R., Rodríguez, C. G. and Taravat, A. (2023), 'Comparative analysis of machine learning algorithms for soil erosion modelling based on remotely sensed data', *Remote Sensing* **15**(2).
URL: *https://doi.org/10.3390/rs15020482*

Gaubi, I., Chaabani, A., Ben Mammou, A. and Hamza, M. H. (2017), 'A GIS-based soil erosion prediction using the Revised Universal Soil Loss Equation (RUSLE) (Lebna watershed, Cap Bon, Tunisia)', *Natural Hazards* **86**(1), 219–239.
URL: *https://doi.org/10.1007/s11069-016-2684-3*

Ghosh, A. and Maiti, R. (2021), 'Soil erosion susceptibility assessment using logistic regression, decision tree and random forest: study on the Mayurakshi river basin of Eastern India', *Environmental Earth Sciences* **80**(8), 1–16.
URL: *https://doi.org/10.1007/s12665-021-09631-5*

Golkarian, A., Khosravi, K., Panahi, M. and Clague, J. J. (2023), 'Spatial variability of soil water erosion: Comparing empirical and intelligent techniques', *Geoscience Frontiers* **14**(1), 101456.
URL: *https://doi.org/10.1016/j.gsf.2022.101456*

Hong, E.-M., Pachepsky, Y. A., Whelan, G. and Nicholson, T. (2017), 'Simpler models in environmental studies and predictions', *Critical Reviews in Environmental Science and Technology* **47**(18), 1669–1712.
URL: *https://doi.org/10.1080/10643389.2017.1393264*

Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X. (2013), *Applied Logistic Regression*, John Wiley & Sons, Hoboken, NJ, USA.
URL: *https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387*

IPCC (2022), *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge.
URL: *https://doi.org/10.1017/9781009157926*

Jaafari, A., Najafi, A., Pourghasemi, H. R., Rezaeian, J. and Sattarian, A. (2014), 'GIS-based frequency ratio and index of entropy models for landslide susceptibility assessment in the Caspian forest, Northern Iran', *International Journal of Environmental Science and Technology* **11**(4), 909–926.
URL: *https://doi.org/10.1007/s13762-013-0464-0*

Khosravi, K., Rezaie, F., Cooper, J. R., Kalantari, Z., Abolfathi, S. and Hatamiafkoueieh, J. (2023), 'Soil water erosion susceptibility assessment using deep learning algorithms', *Journal of Hydrology* **618**.
URL: *https://doi.org/10.1016/j.jhydrol.2023.129229*

Lee, S. and Pradhan, B. (2007), 'Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models', *Landslides* **4**(1), 33–41.
URL: *https://doi.org/10.1007/s10346-006-0047-y*

Mahala, A. (2018), 'Soil erosion estimation using RUSLE and GIS techniques—a study of a plateau fringe region of tropical environment', *Arabian Journal of Geosciences* **11**(13).
URL: *https://doi.org/10.1007/s12517-018-3703-3*

Miles, J. (2014), 'Tolerance and Variance Inflation Factor', *Wiley StatsRef: Statistics Reference Online* **4**, 2055–2056.
URL: *https://doi.org/10.1002/9781118445112.stat06593*

Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F. and Harmouch, H. (2025), 'The effects of data quality on machine learning performance on tabular data', *Information Systems* **132**, 102549.
URL: *https://doi.org/10.1016/j.is.2025.102549*

Musasa, T., Dube, T. and Marambanyika, T. (2024), 'Landsat satellite programme potential for soil erosion assessment and monitoring in arid environments: A review of applications and challenges', *International Soil and Water Conservation Research* **12**(2), 267–278.
URL: *https://doi.org/10.1016/j.iswcr.2023.10.003*

Mustafa, M. R. U., Sholagberu, A. T., Yusof, K. W., Hashim, A. M., Khan, M. W. A. and Shahbaz, M. (2018), SVM-Based Geospatial Prediction of Soil Erosion under Static and Dynamic Conditioning Factors, *in* 'MATEC Web of Conferences', Vol. 203.
URL: *https://doi.org/10.1051/matecconf/201820304004*

Naceur, H. A., Abdo, H. G., Igmoullan, B., Namous, M., Alshehri, F. and Albanai, J. A. (2024), 'Implementation of random forest, adaptive boosting, and gradient boosting decision trees algorithms for gully erosion susceptibility mapping using remote sensing and GIS', *Environmental Earth Sciences* **83**(3), 1–21.
URL: *https://doi.org/10.1007/s12665-024-11424-5*

Nguyen, C. Q., Tran, T. T., Nguyen, T. T. T., Nguyen, T. H. T., Astarkhanova, T. S., Vu, L. V., Dau, K. T., Nguyen, H. N., Pham, G. H., Nguyen, D. D., Prakash, I. and Pham, B. (2023), 'Mapping of soil erosion susceptibility using advanced machine learning models at Nghe An, Vietnam', *Journal of Hydroinformatics* **26**(1), 1–16.
URL: *https://doi.org/10.2166/hydro.2023.327*

Olii, M. R., Kironoto, B. A., Olii, A., Pakaya, R. and Olii, A. K. Z. (2024), Advancing Soil Erosion Assessment: Application of Remote Sensing and Geospatial Techniques in Bulango Ulu Reservoir Basin, *in* 'E3S Web of Conferences', Vol. 476, pp. 1–15.
URL: *https://doi.org/10.1051/e3sconf/202447601041*

Olii, M. R., Nento, S., Doda, N., Olii, R. S. N., Djafar, H. and Pakaya, R. (2025), 'Transformation of geospatial modelling of soil erosion susceptibility using machine learning', *Journal of Civil Engineering Forum* **11**(2), 217–232.
URL: *https://doi.org/10.22146/jcef.19581*

Olii, M. R., Olii, A. K. Z., Olii, A., Djau, R. A., Mokoagow, M. A., Kironoto, B. A., Bachtiar, B., Olii, R. S. N. and Pakaya, R. (2025), 'Tree-based machine learning algorithms for soil erosion vulnerability (SEV) prediction in Saddang Watershed, South Sulawesi, Indonesia', *Journal of Water and Climate Change* **16**(4), 1459–1476.
URL: *https://doi.org/10.2166/wcc.2025.603*

Olii, M. R., Olii, A. K. Z., Olii, A., Pakaya, R. and Kironoto, B. A. (2024), 'Spatial modeling of soil erosion risk using a multi-criteria decision-making (MCDM) approach in the Paguyaman watershed, Gorontalo, Indonesia', *Arabian Journal of Geosciences* **17**(226), 1–13.
URL: *https://doi.org/10.1007/s12517-024-12032-0*

Olii, M. R., Olii, A., Pakaya, R. and Olii, M. Y. U. P. (2023), 'GIS-based analytic hierarchy process (AHP) for soil erosion-prone areas mapping in the Bone Watershed, Gorontalo, Indonesia', *Environmental Earth Sciences* **82**(9), 1–14.
URL: *https://doi.org/10.1007/s12665-023-10913-3*

Sahour, H., Gholami, V., Vazifedan, M. and Saeedi, S. (2021), 'Machine learning applications for water-induced soil erosion modeling and mapping', *Soil and Tillage Research* **211**, 1–12.
URL: *https://doi.org/10.1016/j.still.2021.105032*

Senanayake, S., Pradhan, B., Huete, A. and Brennan, J. (2020), 'A review on assessing and mapping soil erosion hazard using geo-informatics technology for farming system management', *Remote Sensing* **12**(24), 1–25.
URL: *https://doi.org/10.3390/rs12244063*

Sujatha, E. R. and Sridhar, V. (2018), 'Spatial prediction of erosion risk of a small mountainous watershed using RUSLE: A case-study of the Palar sub-watershed in Kodaikanal, South India', *Water* **10**(11), 1–17.
URL: *https://doi.org/10.3390/w10111608*

Özer, C., Çevik, T. and Gürhanlı, A. (2021), 'A machine learning-based framework for predicting game server load', *Multimedia Tools and Applications* **80**(6), 9527–9546.
URL: *https://doi.org/10.1007/s11042-020-10067-5*

[This page is intentionally left blank]