Journal of the Civil Engineering Forum, May 2025, 11(2):217-232
DOI 10.22146/jcef.19581
Available Online at https://jurnal.ugm.ac.id/v3/jcef/issue/archive

Transformation of Geospatial Modelling of Soil Erosion Susceptibility Using Machine Learning

Muhammad Ramdhan Olii[1,*], Sartan Nento[1], Nurhayati Doda[1], Rizky Selly Nazarina Olii[2], Haris Djafar[3], Ririn Pakaya[4]
[1]Department of Civil Engineering, Engineering Faculty, Universitas Gorontalo, Gorontalo, INDONESIA
[2]Department of Architecture, Engineering Faculty, Universitas Gorontalo, Gorontalo, INDONESIA
[3]Sulawesi II River Basin Center, Gorontalo, INDONESIA
[4]Department of Public Health, Public Health Faculty, Universitas Gorontalo, Gorontalo, INDONESIA
*Corresponding author: mr.olii@unigo.ac.id

SUBMITTED 27 January 2025 **REVISED** 15 February 2025 **ACCEPTED** 20 February 2025

ABSTRACT Soil erosion presents substantial environmental and economic challenges, especially in areas prone to land degradation. This study assesses the use of Machine Learning (ML) methods—Support Vector Machines (SVM) and Generalized Linear Models (GLM)—to model Soil Erosion Susceptibility (SES) in the Saddang Watershed, Indonesia. It incorporates environmental, hydrological, and topographical factors to improve prediction accuracy. The approach includes multi-source geospatial data collection, erosion inventory mapping, and relevant factor selection. SVM and GLM were applied to classify SES, with performance evaluated using accuracy, AUC, and precision metrics. Results show SVM classified 40.59% of the area as moderately susceptible and 38.50% as low susceptibility. GLM identified 24.55% as very low and 38.59% as low susceptibility. Both models demonstrated high accuracy (SVM: 87.4%, GLM: 87.2%) and strong AUC values (SVM: 0.916, GLM: 0.939), though GLM showed better specificity and recall. Feature importance analysis highlights that GLM favors hydrological factors like river proximity and drainage density, while SVM balances across various environmental inputs. These findings affirm the value of ML-based geospatial modeling for SES assessment, supporting interventions such as reforestation and erosion control. SVM is suitable for localized planning, whereas GLM offers strategic-level insights. This research contributes to advancing environmental modeling by embedding domain knowledge into ML frameworks, and suggests future work integrate real-time remote sensing and more sophisticated models for broader SES prediction.

KEYWORDS Soil Erosion Susceptibility (SES); Geospatial Modelling; Machine Learning (ML); Support Vector Machines (SVM); Generalized Linear Models (GLM).

1 INTRODUCTION

Soil erosion remains a critical environmental challenge that impacts ecosystems, agriculture, and infrastructure (Olii et al., 2023). The degradation of fertile topsoil, the sedimentation of waterways, and the loss of vegetation cover are among the detrimental effects caused by soil erosion, which can lead to long-term ecological damage (Arabameri et al., 2019) and economic losses (Almouctar et al., 2021). Therefore, predicting and managing soil erosion susceptibility (SES) is essential for sustainable land use and environmental conservation (Kucuker and Giraldo, 2022), requiring accurate and reliable models to assess erosion risk across different areas. Traditionally, SES modeling has relied on empirical methods that use historical data and simple statistical relationships to forecast erosion patterns (Saini et al., 2015). However, these methods often fail to capture the complex interactions among various environmental variables that influence soil erosion processes, such as rainfall intensity, soil type, land use, and topography (Olii, Olii, Olii, Pakaya and Kironoto, 2024). Consequently, there is a growing need for more sophisticated modeling approaches that can better account for these complexities and provide more accurate predictions (Kucuker and Giraldo, 2022; Golijanin et al., 2022).

When combined with Machine Learning (ML) techniques, geospatial modeling significantly enhances SES analysis and prediction. ML models such as Random Forest (RF), Decision Tree (DT),

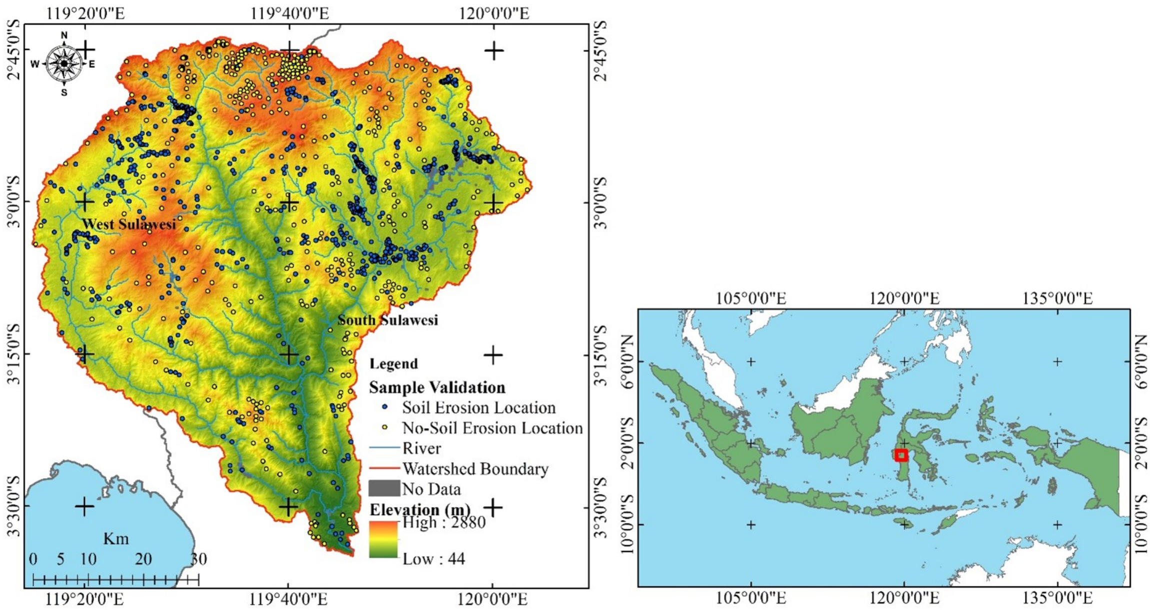


Figure 1 Location of the study area

Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Generalized Linear Models (GLM) can handle complex, non-linear relationships between environmental factors that influence SES, which might not be fully captured by traditional geospatial models (Gayen et al., 2019; Al-Bawi et al., 2021). Integrating these models with Geographic Information System (GIS) and remote sensing data enables more accurate mapping and prediction of SES spatial patterns (Olii, Kironoto, Olii, Pakaya and Olii, 2024). This approach supports a data-driven framework in which models can learn from large datasets, adapt to diverse geographical contexts, and improve prediction accuracy by incorporating classified and weighted factors tailored to local environmental conditions. The synergy between geospatial modeling and ML thus provides powerful tools for more effective land management and soil erosion prevention strategies.

Most studies utilizing ML for SES modeling rely on raw or normalized continuous data without prior classification into discrete classes or the assignment of weights based on expert judgment (Phinzi and Szabó, 2024; Huang et al., 2023; Golkarian et al., 2023). While this approach facilitates automation, it can result in less interpretable models, as the continuous nature of the data may obscure important distinctions between categories of environmental factors. Moreover, the absence of expert-informed weighting can lead to underestimation or overestimation of variable significance, potentially reducing model accuracy and robustness.

To address this issue, this study introduces a novel approach that integrates traditional classification and weighting of environmental factors with advanced ML models, such as SVM and GLM, for SES mapping. In contrast to previous studies that apply these models in isolation, this study emphasizes the innovative combination of pre-classifying variables into discrete categories and assigning expert-derived weights. This methodology enhances model interpretability, making results more accessible to practitioners and decision-makers while also mitigating common ML challenges, such as data complexity, overfitting, and multicollinearity. Furthermore, the flexible classification framework allows the model to be customized to diverse geographic settings, thereby increasing its applicability and robustness. Overall, this study offers a significant advancement over conventional approaches by improving both prediction accuracy and model stability, particularly in regions with distinct environmental characteristics.

2 MATERIALS AND METHODS

2.1 Study Area

The Saddang Watershed, located in southwestern Sulawesi, Indonesia, covers an area of 4,909 km^2

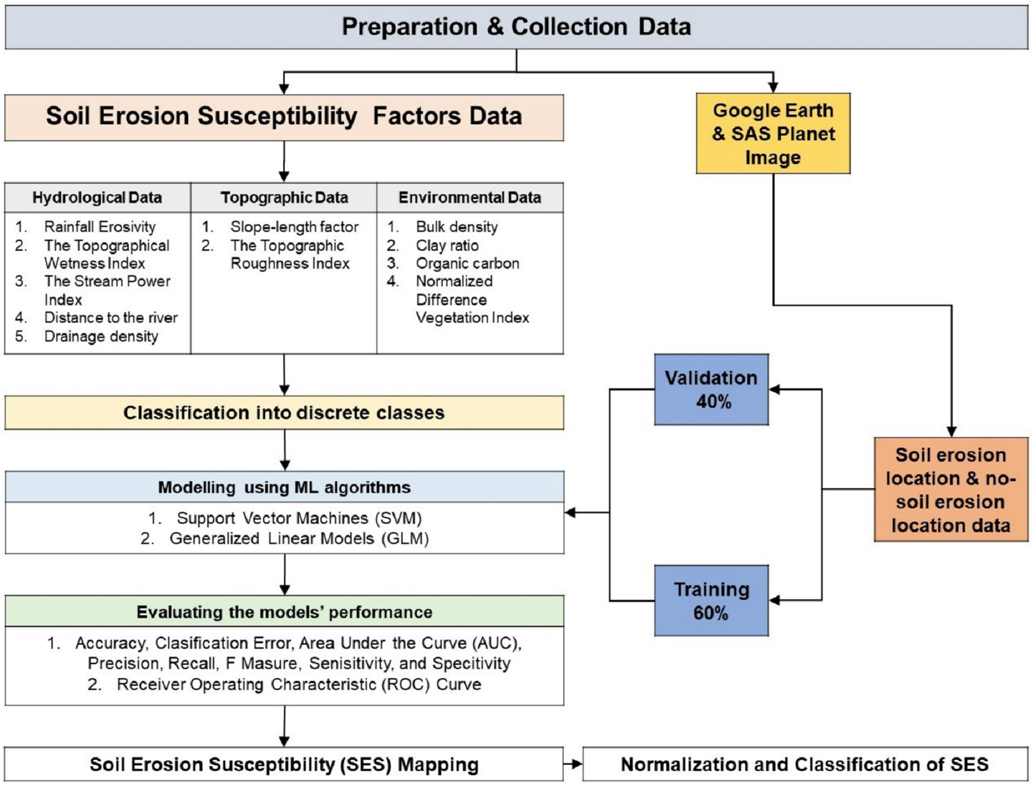


Figure 2 Research flowchart

across South and West Sulawesi provinces, with coordinates of 2°43'–3°34'S and 119°14'–120°3'E (Figure 1). This location is traversed by the Saddang River, which flows through the districts of Enrekang, Tana Toraja, North Toraja, and Polewali before discharging into the Makassar Strait via the Barbana and Paria Figure 1 stuaries. This watershed is crucial for irrigation and energy production. The Benteng Dam supports irrigation across more than 94,000 hectares of farmland, while the Bakaru Hydroelectric Power Plant generes 128 MW of electricity. Furthermore, its groundwater potential, estimated at 1.354 million m^2year^{-1}, contributes significantly to local water supply and agriculture.

The region features a diverse geomorphology, including karst hills, volcanic mountains, and deeply eroded folded terrain. Land use within the watershed is equally varied, comprising settlements, rice fields, plantations, forests, and mixed dryland agriculture, all of which support vital local economic activities. The topography is highly varied, with elevations ranging from 44 m to 2,880 m and an average altitude of 1,277 m. The watershed experiences a tropical climate with an average annual temperature of 23°C. October is typically the warmest month (26°C), while June is the coolest (22°C). Average annual rainfall is approximately 2,500 mm, peaking in May (387 mm) and reaching a minimum in September (68 mm). These characteristics underscore the watershed's ecological and economic significance, highlighting its critical role in regional development and environmental management.

2.2 Overview of Methodological Framework

The methodological framework of this study comprised several key steps for assessing SES, as illustrated in Figure 2.

2.2.1 Data Collection

This study utilized a range of geospatial and environmental datasets from various platforms with distinct spatial resolutions. Shuttle Radar Topography Mission (SRTM) data at a 30×30 m^2 resolution provided detailed elevation information for topographic analysis. Landsat 9 OLI/TIRS imagery, also at 30×30 m^2 resolution, was used to assess vegetation greenness; both datasets were accessed via the USGS Earth Explorer. SoilGrids contributed soil property maps at a 250×250 m^2 resolution, including information on soil texture, organic carbon content, and bulk density, which were essential for land use analysis. Rainfall data, with a spatial resolution of 0.25°×0.25°, was obtained from the NASA POWER Data Access Viewer and was used to analyze precipitation patterns. High-resolution imagery from SAS Planet and Google Earth provided detailed visual monitoring of land surface changes, enabling the identification of soil erosion features. Administrative boundary data in shapefile format from GADM supported geographic analysis. These datasets provided comprehensive insights into the study area, facilitating analyses of elevation, land cover, soil properties, and precipitation patterns for environmental research.

2.2.2 Soil Erosion Inventory Mapping

The soil erosion inventory map is a key component in developing the SES model, serving as the dependent variable in this study. Mapping SES accurately in the Saddang watershed required identifying eroded and non-eroded areas based on geographic coordinates from 1,992 field survey locations—993 with erosion and 999 without—analyzed using SAS Planet and Google Earth. These data points were then used to construct a binary SES model, classifying each location based on the presence or absence of soil erosion. For model development, 1195 samples (60% of the total dataset) were randomly selected for training , while the remaining 797 samples (40%) were reserved for model validation (Figure 1). The types of soil erosion identified across the watershed included sheet, rill, gully, and mass movements.

2.2.3 Selection of the SES Factors

The selection of factors influencing SES for this study was guided by several criteria: (i) the availability and reliability of data, (ii) insights from previous studies, (iii) the connectivity and variability of the data, and (iv) the specific geo-environmental characteristics of the study area. Based on these considerations, a comprehensive set of 11 key factors influencing SES was identified and compiled. The spatial distribution of these SES factors is illustrated in Figure 3.

2.2.3.1 Rainfall Erosivity

The rainfall erosivity quantifies the potential of rainfall to cause soil erosion based on the intensity and kinetic energy of rainfall events. Higher rainfall erosivity values indicate a greater potential for soil detachment and transport. The most widely used method for estimating rainfall erosivity is through the R-factor in the Universal Soil Loss Equation (USLE), expressed as:

$$R = \sum_{i=1}^{12} 1.735 \times 10^{\left[1.5\log_{10}\left[\frac{Pm^2}{Pa}\right]-0.018188\right]} \tag{1}$$

where R is the rainfall erosivity factor (MJ mm ha^{-1} h^{-1} $year^{-1}$), P_m is the monthly rainfall (mm), and P_a is the annual rainfall (mm).

2.2.3.2 Topographic Wetness Index (TWI)

The Topographic Wetness Index reflects the susceptibility of an area to soil saturation and water accumulation, both of which can influence soil erosion by increasing soil moisture and reducing slope stability. TWI is calculated using the formula:

$$\mathrm{TWI} = \ln\left[\frac{A_s}{\tan\beta}\right] \tag{2}$$

where A_s is the upslope contributing area and β is the slope gradient (in radians).

2.2.3.3 Stream Power Index (SPI)

The Stream Power Index represents the erosive power of flowing water and its capacity to transport sediment. Higher SPI values indicate a greater potential for soil erosion due to the increased hydraulic force exerted by surface runoff. SPI is calculated using the following formula:

$$\mathrm{SPI} = \ln\left[A_s\tan\beta\right] \tag{3}$$

Where A_s is the upslope contributing area and β is the slope gradient (in radians).

2.2.3.4 Distance to River

The distance to the river significantly influences soil erosion by affecting the likelihood of sedi-

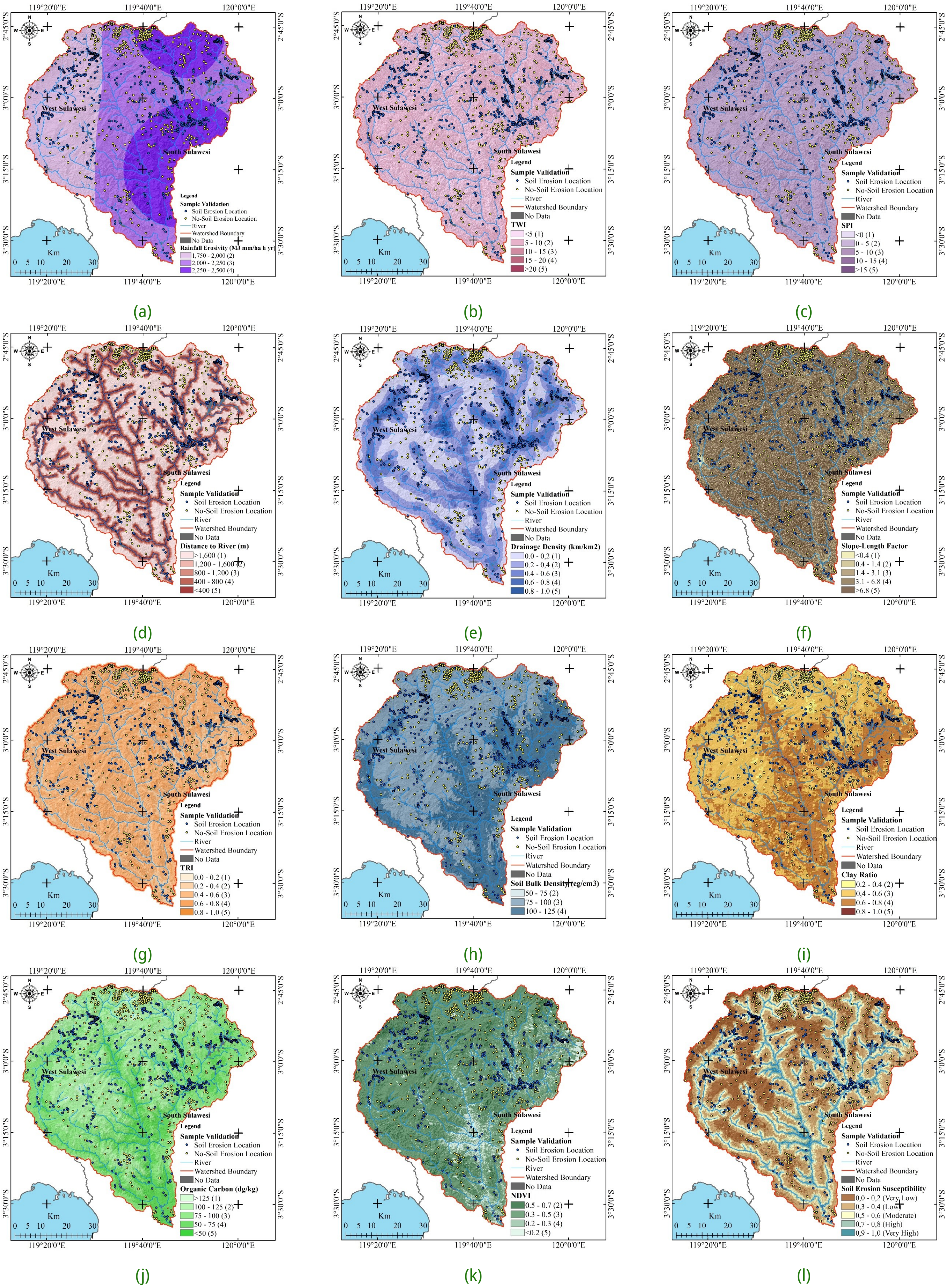


Figure 3 SES factors in Saddang Watershed

ment transport into water bodies. Areas located closer to rivers are generally more susceptible to soil erosion due to increased water flow velocity and higher potential for sediment detachment and movement. This proximity also contributes to greater riverbank instability and accelerates surface erosion processes. As the distance to the river decreases, SES tends to increase, there by elevating the risk of sedimentation in the river and intensifying erosion in these vulnerable areas.

2.2.3.5 Drainage Density

Drainage density reflects the total length of streams and rivers per unit area in a watershed. A higher drainage density indicates a more dissected landscape, which can enhance surface runoff and increase the potential of soil erosion. Drainage density is expressed as follows:

$$D_d = \frac{l_s}{W_A} \tag{4}$$

Where l_s is the total length of the river (km) and W_A is a watershed area (km^2).

2.2.3.6 Slope Length Factor

The Slope Length Factor (LS) quantifies the combined effect of slope length and steepness on soil erosion. In general, longer and steeper slopes lead to increased surface runoff velocity and volume, thereby elevating soil erosion risk. The LS factor is a key component of the USLE and is typically calculated using the following expressions:

$$\mathrm{LS} = \left[\frac{\lambda}{22.13}\right]^m 10.8\mathrm{sin}\beta + 0.03 \text{ if } \mathrm{tan}\beta 0.09 \tag{5}$$

$$\mathrm{LS} = \left[\frac{\lambda}{22.13}\right]^m 16.8\mathrm{sin}\beta - 0.5 \text{ if } \mathrm{tan}\beta \geq 0.09 \tag{6}$$

$$m = \frac{F}{1+F} \tag{7}$$

$$F = \frac{\mathrm{sin}\beta/0.0896}{3(\mathrm{sin}\beta)^{0.8} + 0.56} \tag{8}$$

Where λ is slope length (m), β is the slope gradient (in radians), m is the slope length exponent, and F is the ratio between rill soil erosion and inter-rill soil erosion.

2.2.3.7 Topographic Roughness Index (TRI)

The Topographic Roughness Index quantifies the variability of terrain elevation, with higher values indicating steeper and more rugged landscapes. Such terrains are typically more prone to soil erosion due to enhanced surface runoff and reduced vegetation cover. TRI is commonly calculated as the standard deviation of elevation values within a specified window or grid cell, using the following equation (Riley et al., 1999):

$$\mathrm{TRI} = Y\left(\sum (x_{ij} - x_{00})^2\right)^{0.5} \tag{9}$$

In the context of the Topographic Roughness Index (TRI), the grid cell at (0,0) refers to the central cell within a defined moving window (commonly 3×3). The values of x_ij correspond to the elevations of the surrounding eight neighboring cells. TRI is calculated by measuring the elevation differences between the central cell and its neighbors, providing a quantitative measure of surface roughness or terrain variability. A higher TRI value indicates greater topographic irregularity, which can influence hydrological flow paths and soil erosion potential.

2.2.3.8 Bulk Density

The bulk density reflects the soil's compactness and porosity, influencing water infiltration and root penetration. High bulk density indicates compacted soil with reduced pore space, resulting in lower infiltration rates and increased surface runoff, which in turn elevates the risk of soil erosion.

2.2.3.9 Clay Ratio

The clay ratio represents the proportion of clay particles in the soil, which influences soil structure and its SES. Soils with higher clay content tend to be more cohesive, which can reduce erosion. However, under certain conditions, such as intense rainfall or poor drainage, these soils may become prone to surface crusting and erosion. The clay ratio is calculated using the following formula:

$$C_L = \frac{\%\text{clay}}{\%\text{sand+}\%\text{silt}} \tag{10}$$

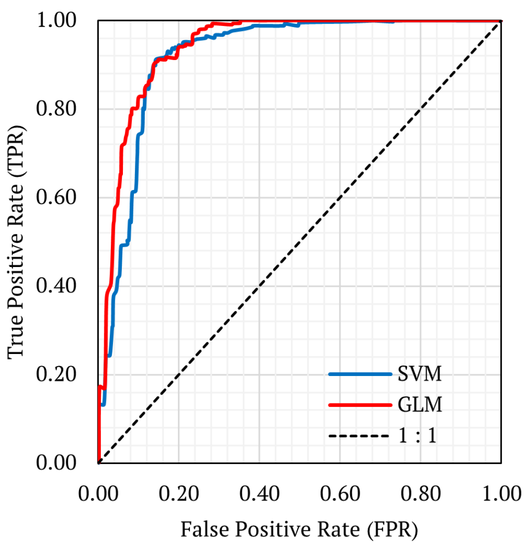


Figure 4 AUC-ROC Graph

where %clay, %sand, and %silt refer to the respective proportions of soil particle sizes.

2.2.3.10 Organic Carbon

Soil organic carbon content influences soil structure, stability, and resistance to soil erosion. Higher levels of organic carbon enhance soil aggregation, which improves infiltration capacity and reduces surface runoff, there by decreasing the risk of soil erosion. Soil organic carbon is typically measured in units of decigrams per kilogram (dg/kg) of soil.

2.2.3.11 Normalized Difference Vegetation Index (NDVI)

The Normalized Difference Vegetation Index (NDVI) is an indicator of vegetation cover and health, where higher NDVI values represent denser and more vigorous vegetation. Dense vegetation mitigates soil erosion by reducing the impact of raindrops and slowing down surface runoff. NDVI is derived from satellite imagery using the following formula:

$$\mathrm{NDVI} = \frac{\mathrm{NIR}_{\mathrm{band}} - \mathrm{Red}_{\mathrm{band}}}{\mathrm{NIR}_{\mathrm{band}} + \mathrm{Red}_{\mathrm{band}}} \tag{11}$$

where NIR is the reflectance in the near-infrared spectrum, and RED is the reflectance in the red portion of the spectrum.

2.2.4 Soil Erosion Modeling Using Machine Learning Models

Soil erosion modeling using ML involves predicting the likelihood and extent of soil erosion based on various environmental factors. ML models can capture complex relationships among these factors and the underlying soil erosion processes, offering a flexible and data-driven approach to SES assessment.

2.2.4.1 Support Vector Machines (SVM)

The Support Vector Machines are applied in soil erosion modeling to classify and predict areas with varying SES based on hydrological, environmental, and topographical factors. SVM identifies the optimal hyperplane that maximizes the margin between different SES classes, effectively distinguishing between high-susceptibility and low-susceptibility zones. To address complex, nonlinear relationships among input variables, SVM uses kernel functions that project the data into a higher-dimensional feature space where linear separation becomes feasible. A commonly used SVM classification function is:

$$f(x) = \mathrm{sign}(w\phi(x) + b) \tag{12}$$

where w is the weight vector, $\phi(x)$ represents the transformation of the input data into a higher-dimensional space, b is the bias term, and sign (.) determines the class label (+1 or -1). The function $f(x)$ outputs the predicted class based on the sign of the result. By training on datasets that include both eroded and non-eroded locations, SVM can effectively generalize and predict SES in previously untested areas, making it a powerful tool for soil conservation and land management planning.

2.2.4.2 Generalized Linear Models (GLMs)

Generalized Linear Models are used in soil erosion studies to analyze the relationship between soil erosion factors and observed outcomes—such as the presence or absence of erosion—by extending traditional linear regression to accommodate non-normal response variable distributions, including binary and count data. GLMs connect predictor variables to the response variable through a specified link function, enabling the modeling of

Table 1. Weights, Classes, and Scores of SES Factors

No.	SES Factors	Categories	Weight SVM	Weight GLM	Classes of Factors	Area (km^2)	Area (%)	Scores
1	Rainfall erosivity (MJ mm ha^{-1} h^{-1} $year^{-1}$)	Hydrological Data	0.076	0.085	<1,750	-	-	1
					1,750-2,000	1,430	29.1	2
					2,000-2,250	1,951	39.7	3
					2,250-2,500	1,528	31.1	4
					>2,500	-	-	5
2	Topographical Wetness Index (TWI)		0.093	0.063	<5	1,960	39.9	1
					5-10	2,692	54.8	2
					10-15	227	4.6	3
					15-20	28	0.6	4
					>20	2	0.0	5
3	Stream Power Index (SPI)		0.038	0.032	<0	19	0.4	1
					0-5	4,014	81.8	2
					5-10	830	16.9	3
					10-15	41	0.8	4
					>15	5	0.1	5
4	Distance to River (m)		0.107	0.376	>1,600	2,777	56.6	1
					1,200-1,600	454	9.3	2
					800-1,200	492	10.0	3
					400-800	540	11.0	4
					<400	645	13.1	5
5	Drainage Density (km/km^2)		0.134	0.282	0.0-0.2	2,011	41.0	1
					0.2-0.4	1,815	37.0	2
					0.4-0.6	925	18.8	3
					0.6-0.8	150	3.1	4
					0.8-1.0	8	0.2	5
6	Slope Length Factor	Topographic Data	0.121	0.052	<0.4	949	19.3	1
					0.4-1.4	95	1.9	2
					1.4-3.1	210	4.3	3
					3.1-6.8	614	12.5	4
					>6.8	3,041	61.9	5
7	Topographic Roughness Index (TRI)		0.064	0.052	0.0-0.2	949	19.3	1
					0.2-0.4	95	1.9	2
					0.4-0.6	210	4.3	3
					0.6-0.8	614	12.5	4
					0.8-1.0	3,041	61.9	5
8	Bulk Density (cg/cm^3)	Environmental Data	0.022	0.001	<50	-	-	1
					50-75	18	0.4	2
					75-100	2,898	59.0	3
					100-125	1,993	40.6	4
					>125	-	-	5
9	Clay Ratio		0.099	0.143	0.0-0.2	-	-	1
					0.2-0.4	129	2.6	2
					0.5-0.6	2,837	57.8	3
					0.7-0.8	1,915	39.0	4
					0.8-1.0	28	0.6	5
10	Carbon Organic (dg/kg)		0.088	0.098	>125	62	1.3	1
					100-125	1,865	38.0	2
					75-100	2,175	44.3	3
					50-75	788	16.0	4
					<50	19	0.4	5
11	Normalized Difference Vegetation Index (NDVI)		0.108	0.283	>0.7	-	-	1
					0.5-0.7	313	6.4	2
					0.3-0.5	3,889	79.2	3
					0.2-0.3	305	6.2	4
					<0.2	402	8.2	5

Table 2. Accuracy metric of the SES Model

Accuracy Metrics	Unit	ML Algorithm	
		SVM	GLM
Accuracy	%	87.4	87.2
Classification Error	%	12.6	12.8
AUC		0.916	0.939
Precision	%	87.6	86.1
Recall	%	86.5	89.4
F Measure	%	87.0	87.7
Sensitivity	%	86.5	89.4
Specificity	%	85.0	88.3

complex and potentially non-linear relationships. This flexibility makes GLMs particularly suitable for predicting SES under varying hydrological, environmental, and topographical conditions. The general form of a GLM is expressed as:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \tag{13}$$

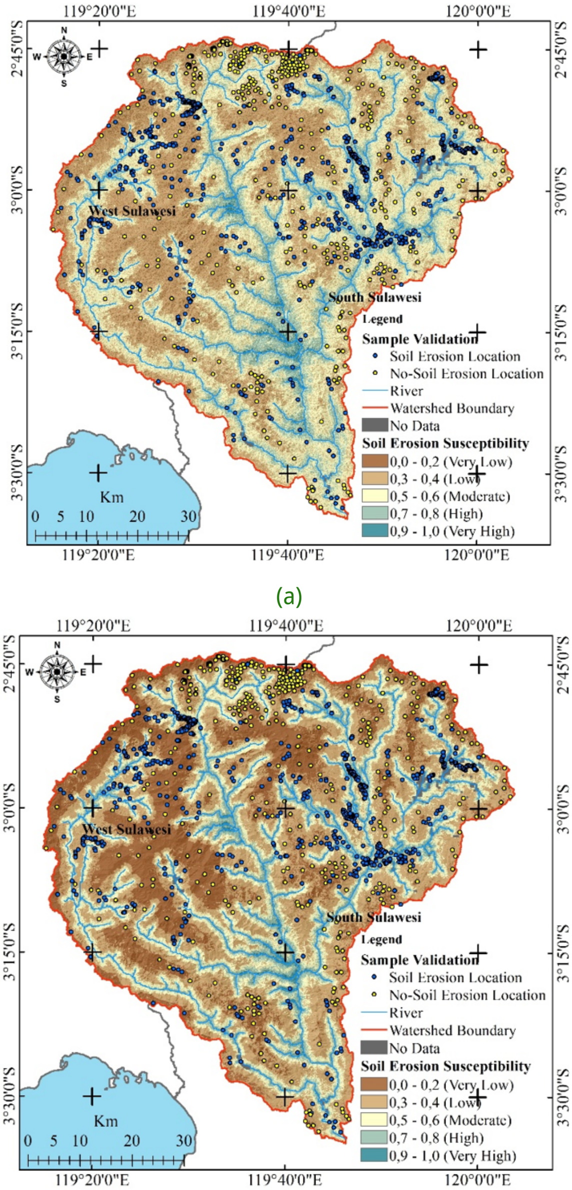
where $g(\mu)$ is the link function that relates the mean of the response variable μ to the linear predictors, while $\beta_0, \beta_1, \ldots, \beta_\mathrm{p}$ and $X_1, X_2, \ldots, X_\mathrm{p}$ are the independent variables (SES factors). GLMs are valuable tools for understanding the probabilistic nature of soil erosion and predicting its occurrence under various scenarios.

2.2.5 Weighting and Scoring

Modeling SES using weighted forms of SVM and GLM involved incorporating the relative importance of each contributing factor. Weights were assigned to individual factors based on their influence on SES, as shown in Table 1. These weights were then multiplied by factor-specific scores for different factor classes in Table 1, ensuring that the contributions of each factor were accurately represented in the model's predictions.

2.2.6 Evaluating The Models' Performance and Normalization

The performance of SVM and GLM in predicting SES was evaluated using various metrics and methods to assess accuracy, reliability, and generalization. Common evaluation metrics included accuracy, precision, recall, and F-measure, which assessed the models' ability to correctly identify true positives while minimizing false positives and



(a)

(b)

Figure 5 Soil Erosion Susceptibility (SES) Maps Based on SVM [a] and GLM [b]

negatives. In addition, Receiver-Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) were used to evaluate the overall discriminative power of the models. These performance assessments are critical in determining the effectiveness of the models in identifying SES-prone areas and ensuring their predictive robustness on unseen data. These evaluation methods provide a comprehensive framework for refining and enhancing SVM and GLM performance in environmental modelling and SES analysis.

The normalization of SES results is an important post-processing step that involves scaling prediction outputs to a common range to facilitate consistent interpretation and comparison across different datasets or scenarios. This process transforms SES values to a standardized scale—typically between 0 and 1—thereby reducing biases introduced by differences in data magnitude and allowing for clearer visualization and ranking of erosion susceptibility. Normalization is commonly performed using the min-max scaling method, defined as:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (14)$$

where X is the original value, X_{min} is the minimum value in the dataset, and X_{max} is the maximum value in the dataset.

3 RESULTS

3.1 Model Performance Evaluation

Figure 4 shows that GLM outperformed SVM in terms of AUC, with values of 0.939 and 0.916, respectively. This result indicated that GLM had a stronger discriminative ability in distinguishing between susceptible and non-susceptible areas, which is essential for generating reliable SES maps. Furthermore, GLM's recall and sensitivity were higher than SVM's, 89.4% and 86.5%, respectively, indicating that GLM was more effective in identifying true positives, reducing the susceptibility to underestimating SES areas—a key factor in environmental conservation and land management (Rahmati et al., 2017). Additionally, GLM achieved higher specificity (88.3% vs. SVM's 85%), indicating better performance in correctly identifying non-susceptible areas and reducing false positives. This higher specificity enhances the model's utility in supporting targeted and cost-effective soil conservation strategies (Bui et al., 2020). While SVM exhibited slightly better precision (87.6% vs. GLM's 86.1%), reflecting a lower rate of false positives, the overall performance of GLM across multiple evaluation metrics suggest that it provides a more balanced and robust approach to SES modeling—particularly in contexts where minimizing both false negatives and false positives is critical.

3.2 Spatial Distribution of Soil Erosion Susceptibility (SES)

Table 3 compares SES across different classes using the SVM and GLM, revealing notable contrasts in their predictions. One of the most striking differences is seen in the "Very Low" susceptibility class: GLM assigned 24.55% of the area to this category, while SVM assigned only 4.37%. This result suggests that GLM tends to generalize and classify broader regions as having minimal susceptibility, possibly due to its smoothing tendencies over fine-scale variations. In contrast, SVM's sensitivity to subtle patterns in the data results in more conservative and spatially restricted identification of "Very Low" SES areas.

A similarly notable difference emerged in the "Moderate" class, with SVM assigning 40.59% of the area compared to only 21.84% by GLM. This result suggests that SVM is more precise in capturing gradual transitions in SES, thereby distributing the susceptibility values more evenly across mid-range categories. On the other hand, both models demonstrated strong agreement in the "Low" susceptibility class, with nearly identical area proportions—38.50% for SVM and 38.59% for GLM—indicating a shared ability to identify regions with consistently low erosion risk, likely due to more robust patterns in the underlying data. Differences resurface in the "High" and "Very High" classes. SVM identified a greater area in the "High" susceptibility class (15.56%) compared to GLM (13.32%), while GLM slightly surpassed SVM in the "Very High" category (1.23% vs. 0.98%). This pattern may reflect SVM's ability to detect more transitional zones from moderate to high susceptibility, whereas GLM may be better suited to identifying extreme cases with strong, generalized signals.

Figure 5 further illustrates that areas categorized as "High" and "Very High" susceptibility by both models are predominantly located along river channels and tributaries. This pattern is consistent with the geomorphological reality that proximity to water bodies increases exposure to erosive forces. The presence of concentrated water flow, frequent saturation, and riverbank dynamics—such as meandering and high-discharge events—amplify soil detachment and transport in these zones, making them particularly vulnerable to erosion. Overall, the SES distribution patterns suggest that SVM offers a more nuanced

Table 3. The SES class area of each Machine Learning Algorithm

No.	Normalization Range	Classes of Soil Erosion Susceptibility	Support Vector Machines (SVM)			Generalized Linear Models (GLM)		
			Grid Total	Area (km^2)	%	Grid Total	Area (km^2)	%
1	0.0 - 0.2	Very Low	238,436	215	4.4	1,339,038	1,205	24.6
2	0.2 - 0.4	Low	2,099,817	1,890	38.5	2,104,967	1,894	38.6
3	0.4 - 0.6	Moderate	2,213,906	1,993	40.6	1,191,423	1,072	21.8
4	0.6 - 0.8	High	848,573	764	15.6	726,729	654	13.3
5	0.8 - 1.0	Very High	53,583	48	1.0	67,158	60	1.2
		Total	5,454,315	4,909	100	5,429,315	4,886	100

and detailed classification, making it advantageous for localized, site-specific erosion control and management. In contrast, GLM provides a broader, more generalized assessment, which is well-suited for strategic-level planning and prioritization of conservation resources. These differences highlight the importance of selecting a modeling approach that aligns with the objectives of the study—whether for precision targeting of erosion hotspots or for broader land-use planning and policy development.

3.3 Feature Importance and Interpretation

Table 1 highlights the weights assigned to various factors by each model, revealing insights into their relative significance in predicting SES. For SVM, the highest weights were assigned to distance to the river (0.107), drainage density (0.134), and slope length factor (0.121). This weighting pattern suggests that SVM places substantial emphasis on topographic and hydrological features. These priorities are consistent with findings in recent literature. For example, Band et al. (2020) emphasized the importance of slope length and drainage density in predicting soil erosion due to their direct influence on runoff and soil erosion processes. Slope length affects the velocity and energy of surface water flow, thereby impacting the degree of soil detachment, while drainage density influences the distribution and concentration of runoff across the landscape (Vu Dinh et al., 2021). The importance of distance to rivers also aligns with studies such as Pourghasemi et al. (2020), which found that proximity to water bodies plays a key role in sediment transport and the onset of erosion processes.

In contrast, GLM assigned the highest weights to the distance to the river (0.376), drainage density (0.282), and NDVI (0.283). This result suggests a greater sensitivity of GLM to both proximity to hydrological features and land cover conditions. The high weight given to distance to rivers and drainage density reinforces their critical roles in shaping erosion patterns, consistent with the findings of Band et al. (2020). Meanwhile, the elevated importance of NDVI in the GLM model highlights the relevance of vegetation cover in modulating SES. As Bui et al. (2020) observed, NDVI serves as a reliable indicator of vegetative health and density, both of which contribute to soil stabilization and reduced erosion risk.

4 DISCUSSION

4.1 Strengths and Weaknesses of Each Model

For SES modelling, the GLM offers several advantages, particularly in terms of flexibility and interpretability. GLM accommodates a variety of predictor variable types and can model different forms of SES outcomes, including binary and continuous responses. Its interpretability is especially valuable in understanding the influence of environmental factors on SES, thereby supporting evidence-based land management decisions. For instance, GLM assigned high weights to the distance to river (0.376) and NDVI (0.283), highlighting their significant roles in soil erosion processes. These factors affect sediment transport and vegetation cover, which are consistently identified in the literature as key drivers of erosion dynamics (Arabameri et al., 2019; Gayen et al., 2019; Rahmati et al., 2017; Igwe et al., 2020).

However, one limitation of GLM lies in its assumption of a linear relationship between predictors and the response variable, which may restrict its ability to capture the inherently complex and non-linear nature of soil erosion processes (Jiang et al., 2021). Additionally, GLM may underperform when dealing with high-dimensional datasets or intricate interactions among variables. While GLM demonstrated strong performance, as evidenced by an AUC of 0.939 (vs. SVM's 0.916), its capacity to fully capture non-linear relationships may be limited compared to more advanced machine learning techniques.

On the other hand, SVM is particularly well-suited for SES modelling due to its ability to model complex, non-linear relationships between predictors and response variables. SVM excels in handling high-dimensional data and capturing intricate interactions, making it a powerful tool for representing the multifaceted processes underlying soil erosion. The SVM model placed strong emphasis on drainage density (0.134) and slope length factor (0.121), highlighting its effectiveness in capturing topographic influences, in line with findings from previous studies (Olii et al., 2023).

SVM's flexibility in employing various kernel functions allows it to adapt to different data structures, thereby enhancing predictive accuracy in SES assessments (Mustafa et al., 2018). Nonetheless, SVM has some drawbacks, including its computational intensity and limited interpretability—challenges for practitioners who require a clear understanding of the factors driving erosion (Devos et al., 2009). Despite these drawbacks, SVM demonstrated strong predictive performance, with a precision of 87.6% and a recall of 86.5%, indicating its robustness and reliability in SES classification tasks.

4.2 Implications for Spatial Susceptibility Assessment and Environmental Management

The findings of SES assessments carry significant implications for land management practices, particularly when integrating ML models like SVM and GLM with geospatial data. As shown in Table 3, both models effectively delineate areas with varying degrees of susceptibility. Notably, GLM identified a substantially larger area as having very low susceptibility (24.55%) compared to SVM (4.37%), suggesting a broader generalization of low-risk zones. This ability to delineate high-susceptibility zones enables more targeted interventions—such as reforestation, terracing, and the construction of check dams—which are essential for preventing soil erosion and enhancing land productivity

Furthermore, the data suggests that the two models may emphasize different risk areas. For example, SVM assigned a higher percentage of the area to the moderate susceptibility class (40.59%) than GLM (21.84%), highlighting SVM's sensitivity to transitional zones. These insights support previous findings that targeted soil conservation measures can enhance agricultural yields and reduce environmental degradation, demonstrating the practical benefits of using ML models in SES assessments (Mosavi et al., 2020). The application of such data-driven insights in land management not only optimizes resource allocation but also aligns with sustainable development goals by mitigating the adverse impacts of soil erosion on ecosystems and human livelihoods (Rahmati et al., 2017; Arabameri et al., 2018).

The broader environmental implications of spatial SES assessments extend beyond immediate land management applications, particularly when considering the differing weight distributions and classification patterns identified by SVM and GLM models. For instance, SVM's higher sensitivity in identifying moderately susceptible areas (40.59%) suggests that it may be more effective in predicting and managing soil erosion in regions where SES is less apparent but still ecologically and economically significant. In contrast, GLM's emphasis on very low susceptibility areas (24.55%) can assist in prioritizing zones for preventive measures, ensuring that areas currently at low risk are managed to avoid future degradation.

Accurate SES mapping is crucial for watershed management (Farhan et al., 2013), as it supports the prediction of sediment loads in rivers and reservoirs, a key factor in maintaining water quality and preventing downstream flooding (Olii, Olii, Olii, Pakaya and Kironoto, 2024). Moreover, the adaptability of these models to various geographic and environmental contexts, as indicated by the correlation between different model outputs and environmental factors, highlights their utility in broader global environmental management efforts. This adaptability is particularly important for integration into climate change adaptation strategies, where SES models can be used to anticipate shifts in erosion patterns resulting from changing precipitation regimes and land use dynamics (Eekhout and de Vente, 2022).

Ultimately, SES models, with their ability to characterize nuanced differences in susceptibility levels, serve as valuable tools for addressing global environmental challenges such as desertification, biodiversity loss, and unsustainable land use. Their application supports long-term sustainability goals by providing a scientific basis for policymaking, conservation planning, and land degradation mitigation at multiple scales (Eekhout and de Vente, 2022).

5 CONCLUSION

The integration of ML techniques, such as SVM and GLM, in the geospatial modelling of SES significantly enhances predictive accuracy and reliability compared to traditional methods. The analysis revealed that the SVM model predominantly classified the study site as areas having moderate (40.59%) and low (38.50%) SES, whereas the GLM model identified a larger portion of areas with very low (24.55%) and low (38.59%) SES. Both models exhibited high accuracy metrics, with SVM achieving an accuracy of 87.4%, closely followed by GLM at 87.2%.

In terms of model discrimination ability, GLM recorded a slightly higher AUC of 0.939 compared to 0.916 for SVM, indicating its superior performance in distinguishing between different SES classes. An analysis of contributing factors further highlighted differences in model behavior: GLM placed greater emphasis on hydrological variables, such as distance to rivers and drainage density, while SVM distributed weight more evenly across a wider range of topographical and environmental factors. This finding suggested that GLM may offer a more hydrologically focused interpretation of soil erosion dynamics, whereas SVM provides a broader, more balanced perspective.

The robust performance of both models—validated through cross-validation techniques—underscores their potential for dynamic and reliable SES assessments. Such data-driven insights are essential for effective land management and soil conservation strategies, particularly in erosion-prone regions. Ultimately, these findings highlight the value of adopting innovative machine learning approaches in environmental modeling to address the complex and evolving challenges of land degradation, watershed protection, and sustainable development.

DISCLAIMER

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We sincerely thank the Faculty of Engineering at Universitas Gorontalo for their invaluable support and resources, which were crucial to the success of this research. We also appreciate the contributions of our colleagues, research assistants, and collaborators, whose expertise and dedication significantly enhanced the quality of this study.

REFERENCES

Al-Bawi, A. J., Al-Abadi, A. M., Pradhan, B. and Alamri, A. M. (2021), ‘Assessing gully erosion susceptibility using topographic derived attributes, multi-criteria decision-making, and machine learning classifiers’, *Geomatics, Natural Hazards and Risk* **12**(1), 3035–3062.
URL: *https://doi.org/10.1080/19475705.2021.1994024*

Almouctar, M. A. S., Wu, Y., Zhao, F. and Dossou, J. F. (2021), ‘Soil erosion assessment using the rusle model and geospatial techniques (remote sensing and gis) in south-central niger (maradi region)’, *Water (Switzerland)* **13**(24).
URL: *https://doi.org/10.3390/w13243511*

Arabameri, A., Cerda, A. and Tiefenbacher, J. P. (2019), ‘Spatial pattern analysis and prediction of gully erosion using novel hybrid model of entropy-weight of evidence’, *Water (Switzerland)* **11**(6), 1–23.
URL: *https://doi.org/10.3390/w11061129*

Arabameri, A., Pradhan, B., Pourghasemi, H. R. and Rezaei, K. (2018), ‘Identification of erosion-prone areas using different multi-criteria decision-making techniques and gis’, *Geomatics, Natural Hazards and Risk* **9**(1), 1129–1155.
URL: *https://doi.org/10.1080/19475705.2018.1513084*

Band, S. S. et al. (2020), 'Novel ensemble approach of deep learning neural network (dlnn) model and particle swarm optimization (pso) algorithm for prediction of gully erosion susceptibility', *Sensors (Switzerland)* **20**(19), 1–28.
URL: *https://doi.org/10.3390/s20195609*

Bui, D. T., Tsangaratos, P., Nguyen, V. T., Van, 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*

Devos, O., Ruckebusch, C., Durand, A., Duponchel, L. and Huvenne, J. P. (2009), 'Support vector machines (svm) in near infrared (nir) spectroscopy: Focus on parameters optimization and model interpretation', *Chemometrics and Intelligent Laboratory Systems* **96**(1), 27–33.
URL: *https://doi.org/10.1016/j.chemolab.2008.11.005*

Eekhout, J. P. C. and de Vente, J. (2022), 'Global impact of climate change on soil erosion and potential for adaptation through soil conservation', *Earth-Science Reviews* **226**, 103921.
URL: *https://doi.org/10.1016/j.earscirev.2022.103921*

Farhan, Y., Zregat, D. and Farhan, I. (2013), 'Spatial estimation of soil erosion risk using rusle approach, rs, and gis techniques: A case study of kufranja watershed, northern jordan', *Journal of Water Resource and Protection* **5**(12), 1247–1261.
URL: *https://doi.org/10.4236/jwarp.2013.512134*

Gayen, A., Pourghasemi, H. R., Saha, S., Keesstra, S. and Bai, S. (2019), 'Gully erosion susceptibility assessment and management of hazard-prone areas in india using different machine learning algorithms', *Science of the Total Environment* **668**, 124–138.
URL: *https://doi.org/10.1016/j.scitotenv.2019.02.436*

Golijanin, J. et al. (2022), 'Estimation of potential soil erosion reduction using gis-based rusle under different land cover management models: A case study of pale municipality, b&h', *Frontiers in Environmental Science* **10**(August), 1–13.
URL: *https://doi.org/10.3389/fenvs.2022.945789*

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*

Huang, D., Su, L., Zhou, L., Tian, Y. and Fan, H. (2023), 'Assessment of gully erosion susceptibility using different dem-derived topographic factors in the black soil region of northeast china', *International Soil and Water Conservation Research* **11**(1), 97–111.
URL: *https://doi.org/10.1016/j.iswcr.2022.04.001*

Igwe, O., John, U. I., Solomon, O. and Obinna, O. (2020), 'Gis-based gully erosion susceptibility modeling, adapting bivariate statistical method and ahp approach in gombe town and environs northeast nigeria', *Geoenvironmental Disasters* **7**(1).
URL: *https://doi.org/10.1186/s40677-020-00166-8*

Kucuker, D. M. and Giraldo, D. C. (2022), 'Assessment of soil erosion risk using an integrated approach of gis and analytic hierarchy process (ahp) in erzurum, turkiye', *Ecological Informatics* **71**, 101788.
URL: *https://doi.org/10.1016/j.ecoinf.2022.101788*

Mosavi, A., Sajedi-Hosseini, F., Choubin, B., Taromideh, F., Rahi, G. and Dineva, A. A. (2020), 'Susceptibility mapping of soil water erosion using machine learning models', *Water (Switzerland)* **12**(7).
URL: *https://doi.org/10.3390/w12071995*

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/mateconf/201820304004*

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., Olii, A. K. Z., Olii, A., Pakaya, R. and Kironoto, B. A. (2024), 'Spatial modeling of soil erosion risk: 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*

Phinzi, K. and Szabó, S. (2024), *Predictive machine learning for gully susceptibility modeling with geo-environmental covariates: main drivers, model performance, and computational efficiency*, Vol. 120, Springer Netherlands.
URL: *https://doi.org/10.1007/s11069-024-06481-9*

Pourghasemi, H. R., Sadhasivam, N., Kariminejad, N. and Collins, A. L. (2020), 'Gully erosion spatial modelling: Role of machine learning algorithms in selection of the best controlling factors and modelling process', *Geoscience Frontiers* **11**(6), 2207–2219.
URL: *https://doi.org/10.1016/j.gsf.2020.03.005*

Rahmati, O., Tahmasebipour, N., Haghizadeh, A., Pourghasemi, H. R. and Feizizadeh, B. (2017), 'Evaluating the influence of geo-environmental factors on gully erosion in a semi-arid region of iran: An integrated framework', *Science of the Total Environment* **579**, 913–927.
URL: *https://doi.org/10.1016/j.scitotenv.2016.10.176*

Riley, S. J., DeGloria, S. D. and Elliot, R. (1999), 'A terrain ruggedness index that quantifies topographic heterogeneity', *Intermountain Journal of Sciences* **5**(1–4), 23–27.

Saini, S. S., Jangra, R. and Kaushik, S. P. (2015), 'Vulnerability assessment of soil erosion using geospatial techniques-a pilot study of upper catchment of markanda river', *International Journal of Advanced Remote Sensing* **3**(1), 9–21.

Vu Dinh, T., 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* pp. 1–14.
URL: *https://doi.org/10.1155/2021/6665485*

[This page is intentionally left blank]