Jurnal Teknik Sumber Daya Air, Desember 2025, 5(2): **213-228**
DOI 10.56860/jtsda.v5i2.157
Tersedia pada https://jtsda.hathi.id/index.php/jtsda/issue/view/11

INTEGRATING MACHINE LEARNING WITH GEOGRAPHIC INFORMATION SYSTEMS AND REMOTE SENSING FOR EROSION RISK MAPPING IN THE TAMALATE WATERSHED

Muhammad Ramdhan Olii[1]*, Sartan Nento[2], Ririn Pakaya[3],
Moh. Isnaen Muhidin[4], dan Erwin Anshari[5]

[1,2]Civil Engineering Department, Engineering Faculty, Universitas Gorontalo, Gorontalo, Indonesia, 96214
[3]Public Health Department, Public Health Faculty, Universitas Gorontalo, Gorontalo, Indonesia, 96214
[4]Sulawesi II River Basin Center, Gorontalo, Indonesia, 96214
[5]Department of Mining Engineering, Faculty of Mathematics and Natural Sciences Halu Oleo University, Southeast Sulawesi, 93231

*kakaramdhanolii@gmail.com

Pemasukan: 8 November 2025 Perbaikan: 22 Desember 2025 Diterima: 24 Desember 2025

Intisari

Soil erosion poses a serious threat to environmental sustainability, particularly in tropical watersheds with complex topographic and hydrological conditions. Accurate and spatially reliable erosion risk mapping is therefore essential for effective land management. This study evaluates the performance of five machine learning models—Random Forest (RF), Gradient Boosting Tree (GBT), Decision Tree (DT), Generalized Linear Model (GLM), and Support Vector Machine (SVM)—for erosion risk prediction in the Tamalate Watershed, Indonesia, by integrating topographic and remote sensing–derived variables. Erosion and non-erosion ground-truth samples (553 and 793 points, respectively) were obtained through visual interpretation of temporally consistent high-resolution Google Earth imagery aligned with Landsat-9 acquisition, ensuring data validity. Eight environmental predictors were derived at a consistent spatial resolution and screened for multicollinearity (VIF < 3). Model performance was assessed using spatially explicit validation based on accuracy, AUC, precision, recall, sensitivity, specificity, and F-measure. Results show that RF achieved the best overall performance (accuracy = 0.727; AUC = 0.772), comparable to recent erosion modeling studies in similar tropical environments. Topographic Wetness Index (TWI) and Normalized Difference Moisture Index (NDMI) were identified as the most influential predictors. While high recall and sensitivity indicate strong capability to detect erosion-prone areas, relatively low specificity—particularly in GLM and DT—suggests a tendency to overestimate erosion risk, with implications for management prioritization. Ensemble-based models produced more stable and realistic spatial risk patterns. This study provides a transferable machine learning framework for erosion risk mapping to support sustainable watershed management in data-limited tropical regions.

Keywords : Erosion risk, machine learning, remote sensing, topographic indices, Tamalate watershed

Introduction

Soil erosion remains one of the most pressing and widespread forms of land degradation, particularly in tropical watersheds that experience intense rainfall, steep terrain, and anthropogenic pressure (Panagos et al., 2015; Poesen, 2017). In many developing regions, including Indonesia, watershed ecosystems are increasingly threatened by unsustainable land-use practices such as deforestation, shifting cultivation, and poorly managed agriculture (Eekhout & de Vente, 2022). The Tamalate Watershed in Gorontalo Province exemplifies this vulnerability, where topographic variability, unregulated slope farming, and vegetation clearance have significantly accelerated erosion processes. These conditions not only lead to the loss of fertile topsoil (Woldemariam et al., 2018) but also contribute to sedimentation in downstream water bodies, reducing the effectiveness of irrigation infrastructure and increasing the risk of flash floods (Gaubi et al., 2017). Despite the severity of erosion-related impacts, erosion risk assessment in many Indonesian watersheds remains limited to generalized approaches, with insufficient attention to spatial prediction reliability and local process representation (Susanti et al., 2019).

Conventional erosion modeling approaches—such as the Universal Soil Loss Equation (USLE) or Revised USLE (RUSLE)—have been widely applied due to their simplicity and accessibility (El Jazouli et al., 2017; Issaka & Ashraf, 2017). However, these models rely on linear assumptions and often fail to capture the spatial heterogeneity and complex interactions among environmental factors (Borrelli et al., 2020). Furthermore, they are generally site-specific and not readily transferable to regions with different climatic or geomorphological conditions. In the Indonesian and broader tropical context, most erosion studies continue to rely on empirical or single-model approaches, while systematic comparisons of multiple machine learning algorithms using spatially validated datasets remain scarce (Dharmawan et al., 2023). As the availability of high-resolution remote sensing data and computing power grows, machine learning (ML) methods have emerged as a transformative tool for environmental modelling (Zhong et al., 2021). ML algorithms can identify intricate, non-linear patterns within large datasets and have demonstrated high performance in spatial prediction tasks, including erosion risk assessment (Arif et al., 2017). Nevertheless, the applicability and relative performance of different machine learning models in heterogeneous tropical watersheds have not been sufficiently explored, particularly at the watershed scale (Olii et al., 2025).

In this study, a data-driven modeling approach is employed to map erosion risks in the Tamalate Watershed by integrating five machine learning algorithms: Random Forest (RF), Gradient Boosted Trees (GBT), Decision Tree (DT), Generalized Linear Model (GLM), and Support Vector Machine (SVM). These models are trained using a combination of terrain attributes, topographic and satellite-derived indices derived from satellite imagery and GIS-based spatial analysis. The performance of each algorithm is compared using rigorous classification metrics to identify the most suitable method for erosion prediction in this context. The novelty of this research lies in the systematic comparison of multiple machine learning models for erosion risk mapping in a tropical watershed that has not been

extensively studied, coupled with the use of diverse topographic and satellite-derived indices predictors integrated from GIS and remote sensing. By providing spatially explicit erosion risk maps and model-based performance insights, this study offers practical support for watershed management, conservation prioritization, and land-use planning in data-limited tropical regions. Moreover, this study contributes a transferable framework for erosion risk modeling that can inform spatial planning and conservation efforts in similar vulnerable watersheds across Indonesia.

Methodology

Data

The Digital Elevation Model (DEM) utilized in this research was obtained from the Earth Explorer platform (https://earthexplorer.usgs.gov/) , based on the Landsat 9 dataset (LC08_L1TP_113060_20241230_20250104_02_T1), captured on December 30, 2024. This dataset was chosen due to its suitable spatial resolution and coverage of the study area. All spatial datasets used in this study were processed and standardized to a raster spatial resolution of 30 × 30 m² to ensure consistency across analyses. Administrative boundaries were retrieved from the GADM database (https://gadm.org/), known for its comprehensive and regularly updated geopolitical data. To delineate erosion and non-erosion areas, high-resolution imagery from Google Earth was interpreted visually, with image capture dates intentionally aligned with the Landsat 9 acquisition date. This temporal alignment between satellite imagery and ground-truth observations ensures that the input data reflects consistent environmental conditions. As a result, the erosion hazard models developed in this study benefit from improved temporal accuracy, thereby increasing the robustness and validity of the spatial analysis and prediction outputs.

Method

This study was designed to develop an advanced erosion risk model for the Tamalate Watershed (Figure 1) by integrating topographic indices and satellite-derived indices into a machine learning framework. The overall methodology comprised four major phases: **(1) selection and analysis of erosion risk factors, (2) erosion sample selection and dataset preparation, (3) model evaluation using multiple performance metrics (4) erosion risk modeling with machine learning and mapping**. Each step was carefully structured to ensure the methodological rigor necessary for spatial predictive modeling in a highly dynamic watershed environment.

1. Selection and Analysis of Erosion Risk Factors

This study began by selecting key factors associated with erosion, incorporating both topographic and remote sensing indices, which represent the dominant physical and surface processes controlling soil detachment and transport in tropical watersheds. Specifically, four topographic factors—Sediment Transport Index (STI), Topographic Wetness Index (TWI), Terrain Ruggedness Index (TRI), and

Stream Power Index (SPI)—were extracted from a high-resolution Digital Elevation Model (DEM), as these indices describe runoff concentration, flow energy, terrain instability, and erosion potential driven by topographic controls. Concurrently, remote sensing-derived indices, including the Normalized Difference Tillage Index (NDTI), Normalized Difference Moisture Index (NDMI), Soil Adjusted Vegetation Index (SAVI), and Vegetation Condition Index (VCI), were generated from Landsat-9 imagery to characterize land surface disturbance, soil moisture conditions, vegetation cover, and vegetation health, which have been widely validated as key indicators influencing erosion processes in humid tropical environments.. All indices were processed and standardized using ArcGIS 10.8 (Figures 2 and 3) to ensure spatial alignment and resolution uniformity across the dataset (Table 1).

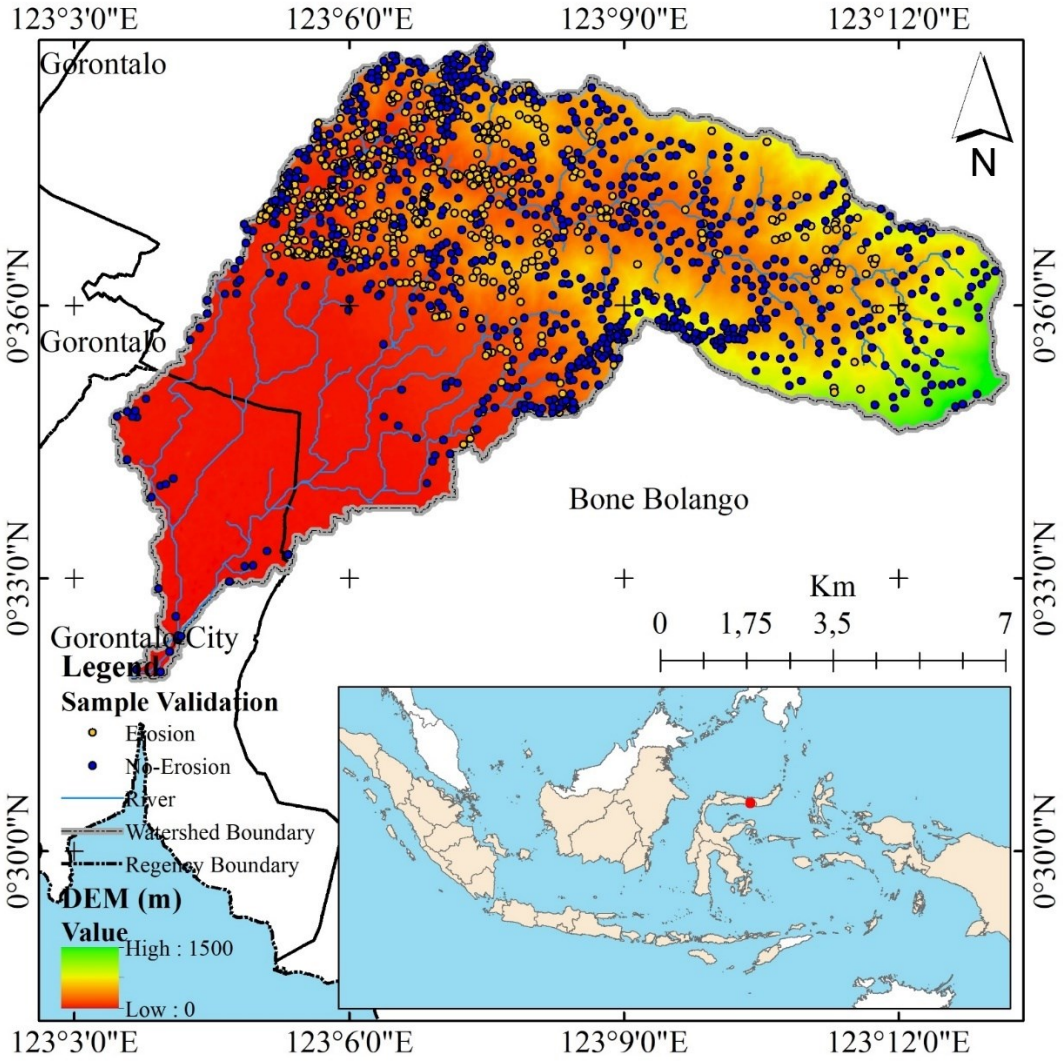

Figure 1. Study site

Table 1. Class n score of topographic and remote sensing indices

Indices	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

2. Erosion Sample Selection and Dataset Preparation

553 erosion and 793 non-erosion points were identified via high-resolution visual interpretation in Google Earth based on surface features (Figure 1) and used as ground-truth data for model training and validation. To minimize potential interpretation bias, samples were selected using consistent visual criteria and distributed across different topographic and land-cover conditions. Temporal consistency between Google Earth imagery and Landsat-9 acquisition was ensured to reduce misclassification. Eight environmental indices (STI, TWI, TRI, SPI, NDTI, NDMI, SAVI, BSI) were extracted for each point using ArcGIS 10.8 to form the predictor dataset, and multicollinearity was assessed using VIF and TOL metrics (Band et al., 2020; Ghorbanzadeh et al., 2020).

3. Model Evaluation Using Multiple Performance Metrics

Model evaluation using multiple performance metrics involves assessing the model's effectiveness through various measures, such as accuracy, classification error, AUC, precision, recall, F-measure, sensitivity, and specificity. These metrics provide a comprehensive understanding of how well the model predicts erosion risks, taking into account both the true positive and false positive rates. This evaluation process ensures that the model performs optimally and generalizes well to unseen data, providing reliable results for erosion risk assessment.

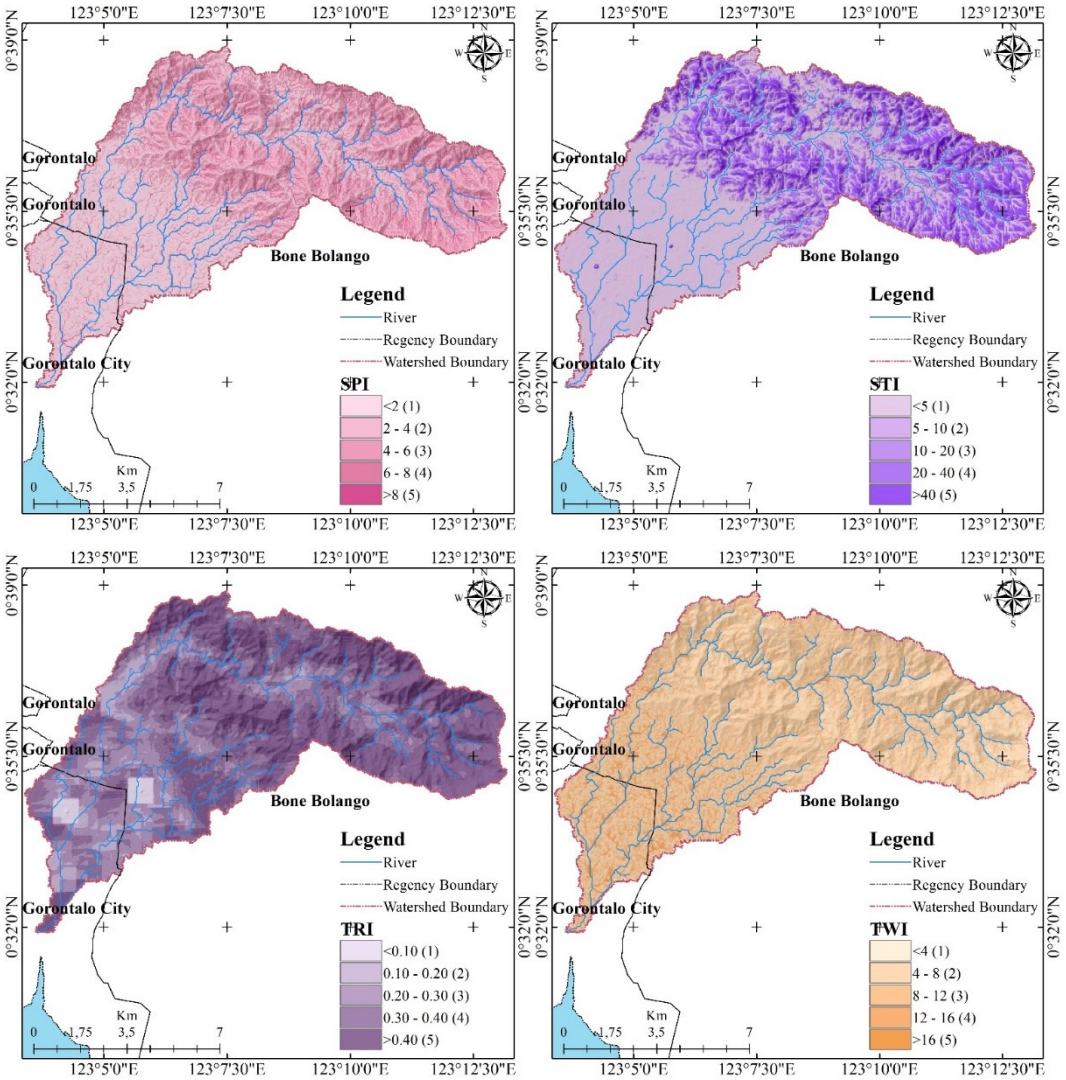

Figure 2. Topographic indices

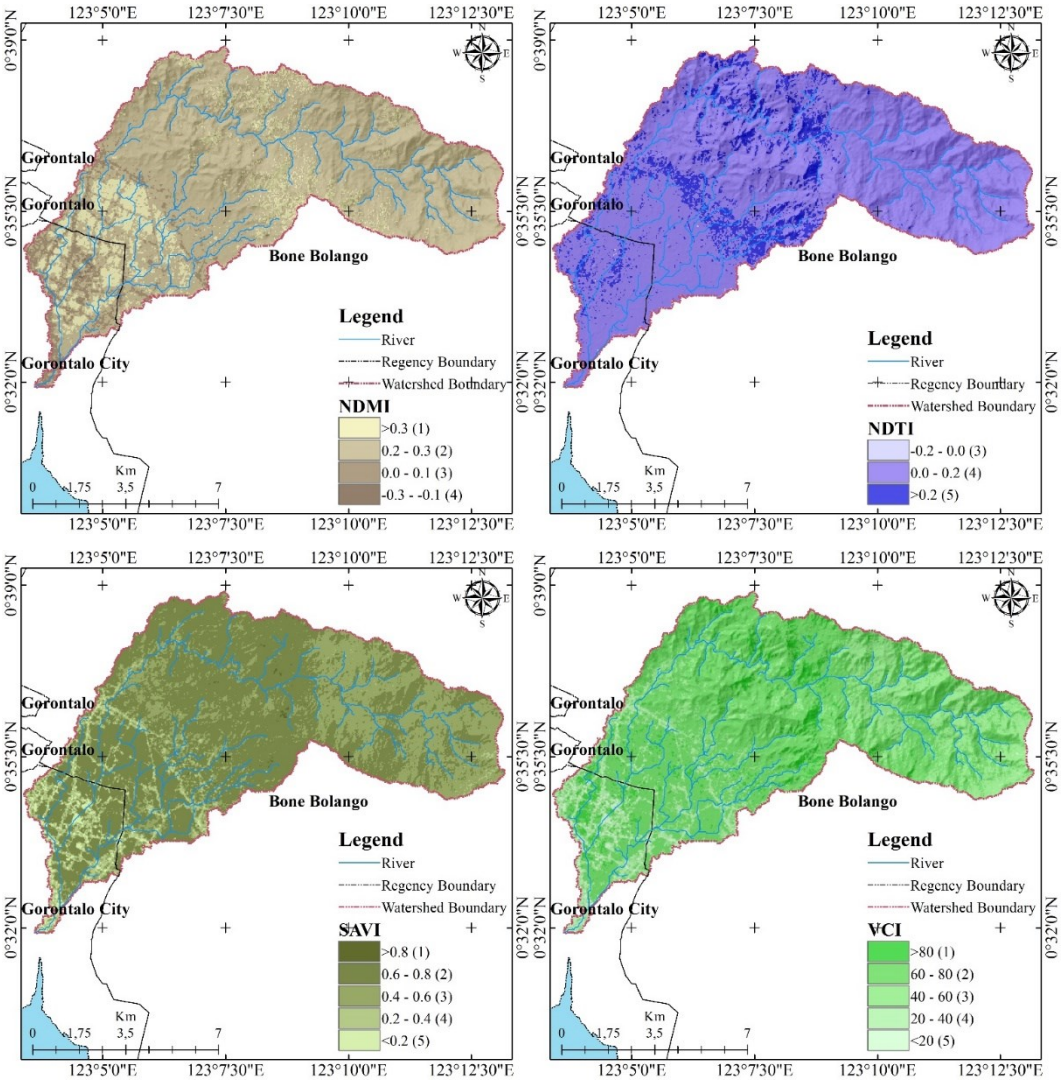

Figure 3. Remote sensing indices

4. Erosion Risk Modeling with Machine Learning and Mapping

Erosion risk was modeled using five machine learning algorithms: RF, GBT, DT, GLM, and SVM. Each model was trained to predict erosion risk based on the prepared dataset. To evaluate the contribution and reliability of each model, its predictive performance was assessed, and model-specific weights were derived. These weights were later used to integrate the outputs, enhancing the overall accuracy of erosion risk mapping, as expressed in the following equation:

$$\text{Erosion Risk} = \sum_{i=i}^{n} \left(W_i \times S_i \right) \quad (1)$$

where W_i *i*s the relative weight (importance) of factor i, and S_i is the classified score based on Table 1.

$$\text{Erosion Risk}_{norm} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \quad (2)$$

where X is the original value of erosion risk, X_{min} is the minimum value of erosion risk, and X_{max} is the maximum value of erosion risk. The normalized erosion risk values (ranging from 0 to 1) were classified into five categories—Very Low, Low, Moderate, High, and Very High—using equal interval classification, with each class spanning a range of 0.2.

Result and Discussion

The results of this study provide a comprehensive evaluation of erosion risk modeling using multiple machine learning algorithms, supported by robust statistical assessments and spatial analysis. Table 2 shows all factors in the analysis have VIF values below 3, indicating that there is no harmful multicollinearity that could negatively impact the regression model. While a few variables—specifically VCI, TWI, and SAVI—demonstrate moderate correlation with other predictors, their VIF values remain within acceptable limits and do not pose a serious concern. Therefore, it is appropriate to retain all variables for subsequent regression analysis without the need for exclusion or adjustment.

Table 2. Multicollinearity Assessment of Erosion Risk Factors

Erosion risk Factors	R^2	TOL	VIF	Interpretation
VCI	0.57	0.43	2.30	Moderate multicollinearity, still acceptable
SAVI	0.51	0.49	2.03	Moderate multicollinearity, still acceptable
NDMI	0.16	0.84	1.19	No multicollinearity
NDTI	0.18	0.82	1.22	No multicollinearity
TWI	0.56	0.44	2.25	Moderate multicollinearity, still acceptable
SPI	0.41	0.59	1.70	No serious multicollinearity
TRI	0.13	0.87	1.15	No multicollinearity
STI	0.31	0.69	1.46	No multicollinearity

Table 3 illustrates the relative importance of various environmental factors in predicting erosion risk using five machine learning models (RF, GBT, DT, GLM, SVM). Among the variables, TWI consistently holds the highest weight—especially in the RF (0.326) and SVM (0.238) models—indicating its critical role in identifying areas prone to moisture accumulation and surface runoff, which are key drivers of erosion. NDMI and STI also show substantial importance across models, reflecting the influence of soil moisture and runoff on erosion processes. In contrast, indices such as VCI and NDTI have relatively lower weights, suggesting a more indirect or minor role in erosion prediction. Moderate contributions are observed for terrain-related indices, such as TRI and SPI, which influence how water moves across landscapes. Overall, the results suggest that topography and hydrological indicators are the most influential factors in erosion modeling, whereas vegetation and land management indices contribute less significantly.

Table 3. Weight of each machine learning model

Erosion risk Factors	Machine Learning				
	RF	**GBT**	**DT**	**GLM**	**SVM**
VCI	0.037	0.047	0.031	0.045	0.087
SAVI	0.045	0.049	0.034	0.120	0.086
NDMI	0.095	0.103	**0.118**	**0.253**	0.169
NDTI	0.016	0.015	0.039	0.101	0.098
TWI	**0.326**	**0.249**	0.050	0.098	**0.238**
SPI	0.038	0.060	0.023	0.072	0.149
TRI	0.054	0.118	0.092	0.194	0.128
STI	0.086	0.119	0.106	0.187	0.128

The performance metrics of the five machine learning models (RF, GBT, DT, GLM, and SVM) in Table 4 align with the variable importance trends observed earlier, confirming the predictive power of topographic and moisture-related indices in erosion risk modeling. Random Forest (RF) demonstrates the highest overall performance with the highest accuracy (0.727), AUC (0.772), and F-measure (0.798), likely due to its strong handling of complex interactions among dominant predictors like TWI and NDMI. RF, DT, and GLM share the highest recall and sensitivity (0.916), indicating their strong ability to correctly identify erosion-prone areas, which is crucial in risk prevention. However, specificity remains relatively low across all models, particularly in the GLM (0.411), suggesting challenges in correctly identifying non-erosion areas, which may be due to class imbalance or overlapping feature distributions. SVM shows the highest precision (0.734) and specificity (0.558), suggesting it is more conservative in its predictions, leading to fewer false positives. Overall, RF offers the most balanced performance, effectively leveraging the most influential factors such as TWI and STI, thereby reinforcing the previous conclusion that terrain and hydrological indicators are critical for accurate erosion risk prediction.

Table 4. Performance metrics of each machine learning model

Performance Metrics	Machine Learning				
	RF	**GBT**	**DT**	**GLM**	**SVM**
Accuracy	**0.727**	0.716	0.716	0.708	0.721
Classification Error	0.273	0.284	0.284	**0.292**	0.279
AUC	**0.772**	0.765	0.736	0.763	0.742
Precision	0.707	0.713	0.697	0.691	**0.734**
Recall	**0.916**	0.872	**0.916**	**0.916**	0.836
F-measure	**0.798**	0.784	0.792	0.787	0.780
Sensitivity	**0.916**	0.872	**0.916**	**0.916**	0.836
Specificity	0.456	0.495	0.431	0.411	**0.558**

The erosion risk classification Table 5 and Figure 4 reveal notable differences in how each machine learning model categorizes the study area across five risk levels, which directly reflects their internal mechanisms and the relative weight they assign to various environmental factors. The variable importance analysis further indicates that topographic controls, such as TWI and SPI, together with vegetation-related

indicators such as NDMI, play a dominant role in shaping erosion susceptibility within the watershed. These variables are closely associated with known erosion processes, where moisture accumulation, flow concentration, and reduced vegetation cover on steep slopes intensify soil detachment and sediment transport, leading to spatial clustering of high-risk zones. RF, which places strong importance on TWI and NDMI, classifies the majority of the area as low (50.0%) and very low (22.0%) risk. This conservative risk estimation is consistent with RF's ensemble nature, which reduces overfitting and captures complex, non-linear interactions among dominant variables, particularly in landscapes characterized by heterogeneous terrain and variable vegetation conditions**.** In contrast, the Decision Tree (DT) model, which has a simpler structure and tends to overfit, assigns a much larger area to high (34.9%) and very high (4.5%) risk. This likely results from DT's sensitivity to strong local patterns in a few highly weighted variables, such as NDMI and TRI, without sufficient generalization, causing steep and sparsely vegetated areas to be disproportionately classified as high-risk. GBT and SVM, which strike a balance between model complexity and generalization, classify the majority of the area as moderate risk (53.3% and 54.8%, respectively), indicating a cautious but balanced risk assessment. GLM, which is linear and less capable of capturing complex variable interactions, also shows a higher percentage in moderate (42.1%) and high (25.2%) categories, possibly due to its limitations in modeling the non-linear behavior of erosion processes. These variations highlight how model architecture, learning strategies, and sensitivity to feature importance significantly impact erosion risk predictions, particularly when dominant factors such as terrain and moisture dynamics are involved.

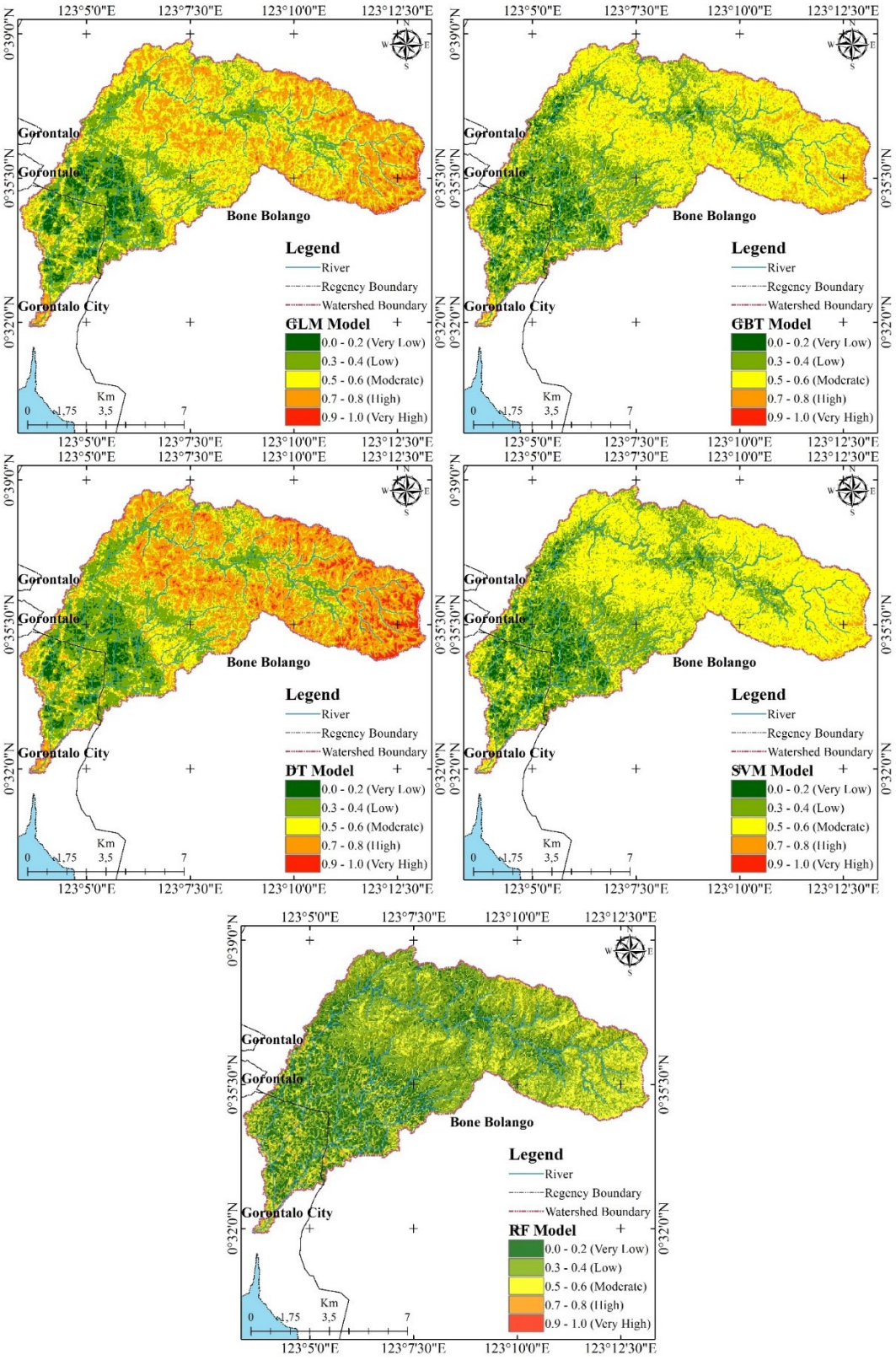

Figure 4. Spatial distribution of erosion risk

Table 5. Distribution of classes of erosion risk for each machine learning

Classes of Erosion Risk	RF		GBT		DT		GLM		SVM	
	Km^2	%	Km^2	%	Km^2	%	Km^2	%	Km^2	%
Very Low	22.8	22.0	6.2	5.9	4.3	4.1	5.3	5.1	7.3	7.0
Low	51.8	50.0	33.5	32.3	30.1	29.0	27.5	26.5	35.1	33.9
Moderate	25.4	24.5	55.2	53.3	28.4	27.4	43.6	42.1	56.8	54.8
High	3.5	3.3	8.6	8.3	36.2	34.9	26.1	25.2	4.4	4.2
Very High	0.2	0.2	0.2	0.2	4.7	4.5	1.2	1.2	0.1	0.1
Total	**103.7**	**100.0**	**103.7**	**100.0**	**103.7**	**100.0**	**103.7**	**100.0**	**103.7**	**100.0**

Discussion

The results of this study present several critical findings that merit in-depth discussion in the context of erosion risk modeling. The initial multicollinearity assessment confirms the statistical validity of the selected predictors, with all variables demonstrating VIF values below the commonly accepted threshold of 3 (O'brien, 2007). This ensures that the regression-based interpretations and machine learning models are not undermined by redundant information, thereby supporting the robustness of the variable selection process. However, the comparative analysis of variable importance across different algorithms reveals significant differences in model sensitivity to environmental parameters. Notably, RF and SVM consistently assign greater weight to topographic and hydrological indices—particularly TWI and NDMI—which is consistent with their superior predictive accuracy and balanced classification performance (Cutler et al., 2007; Fernández et al., 2023). The superior performance of ensemble-based models such as RF and GBT can be attributed to their ability to aggregate multiple decision rules and reduce variance, enabling them to better capture terrain variability and spectral heterogeneity characteristic of the Tamalate Watershed, where steep slopes, variable moisture conditions, and heterogeneous vegetation cover coexist. In contrast, DT model tends to overfit by emphasizing fewer dominant variables, leading to an overestimation of high and very high-risk zones (Kotsiantis, 2013). This overestimation is spatially evident in areas characterized by steep slopes and sparse vegetation cover, where localized patterns dominate decision rules and reduce the model's capacity to generalize across diverse landscape conditions. This limitation underscores the model's reduced generalizability in heterogeneous landscapes. Furthermore, although several models show high sensitivity and recall—crucial for identifying erosion-prone areas—they generally exhibit low specificity, particularly GLM and DT, indicating challenges in accurately detecting areas of minimal risk (Chawla et al., 2002). Such discrepancies may stem from imbalanced class distributions or overlapping feature spaces. The spatial distribution of high-risk zones further reflects the combined influence of topography and vegetation dynamics, where areas with high flow accumulation, elevated stream power, and reduced vegetation density consistently correspond to higher predicted erosion risk. Lastly, the spatial patterns derived from the classification outputs demonstrate that model architecture and complexity play a decisive role in the risk stratification of the landscape. Ensemble methods such as RF and GBT offer more nuanced and stable classifications, suggesting their suitability for applications requiring spatially explicit erosion risk assessments (Breiman, 2001; Friedman, 2001). Although a

formal uncertainty analysis was not conducted, potential sources of uncertainty related to input data quality, visual sample interpretation, and model structure have been acknowledged, and future work may incorporate uncertainty quantification approaches such as ensemble agreement analysis or probabilistic modeling to further strengthen predictive confidence. Overall, these findings emphasize the importance of selecting algorithms that not only perform well statistically but also reflect realistic spatial processes, and they advocate for the integration of diverse environmental indicators to enhance the reliability of erosion modeling.

Conclusions

This study demonstrates the effectiveness of machine learning algorithms in modeling erosion risk by integrating topographic, hydrological, and vegetation-related indicators. The absence of harmful multicollinearity among the predictors ensures the robustness of the input variables and supports their combined use in predictive modeling. Among the models tested, RF consistently outperforms others in terms of accuracy, AUC, and F-measure, largely due to its ability to capture complex, nonlinear interactions, especially among key factors such as TWI and NDMI. In contrast, simpler models like DT exhibit tendencies toward overfitting, leading to inflated high-risk classifications that may lack generalizability. The spatial distribution of erosion risk classes further confirms that model architecture significantly influences classification outcomes, with ensemble approaches like RF and GBT offering more balanced and realistic risk estimates. While high recall and sensitivity across models indicate their strong capacity to detect erosion-prone areas, the generally low specificity—particularly in GLM and DT—highlights ongoing challenges in distinguishing non-risk zones and increases the likelihood of false-positive predictions, partly due to class imbalance in the training data. **It should be noted that the conclusions drawn from this study are based on a single watershed and may not be directly transferable to regions with different geomorphological, climatic, or land-use characteristics.** These findings underscore the importance of selecting appropriate machine learning models and considering multiple environmental variables for more accurate and actionable erosion risk assessments. Future research should explore the integration of additional physical and socio-environmental factors and address class imbalance to further enhance model specificity and generalizability.

Suggestion

Future research should incorporate additional variables such as detailed land use, long-term rainfall data, and socio-economic factors. Exploring ensemble learning techniques and improving field validation are recommended, particularly through targeted field surveys and longitudinal monitoring to capture temporal erosion dynamics better and reduce classification uncertainty. Integrating models into decision-support GIS platforms can enhance their applicability for sustainable watershed and erosion risk management, allowing local agencies to support erosion control planning, prioritize conservation measures, and allocate resources more effectively at the regional scale.

Acknowledgements

The authors would like to express their sincere gratitude to the Faculty of Engineering, Universitas Gorontalo, for the support and facilities provided throughout the course of this research. The continuous institutional support from the faculty greatly contributed to the successful completion of this study, both in terms of research infrastructure and constructive academic encouragement.

References

Arif, N., Danoedoro, P., & Hartono. (2017). Analysis of Artificial Neural Network in Erosion Modeling: A Case Study of Serang Watershed. *IOP Conference Series: Earth and Environmental Science*, *98*(1). https://doi.org/10.1088/1755-1315/98/1/012027

Band, S. S., Janizadeh, S., Saha, S., Mukherjee, K., Bozchaloei, S. K., Cerdà, A., Shokri, M., & 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. https://doi.org/10.3390/land9100346

Borrelli, P., Robinson, D. A., Panagos, P., Lugato, E., Yang, J. E., Alewell, C., Wuepper, D., Montanarella, L., & Ballabio, C. (2020). Land use and climate change impacts on global soil erosion by water (2015-2070). *Proceedings of the National Academy of Sciences*, *117*(36), 21994–22001. https://doi.org/10.1073/pnas.2001403117

Breiman, L. (2001). Random Forest. *Machine Lerning*, *45*, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/https://doi.org/10.48550/arXiv.1106.1813

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forests for Classification in Ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Dharmawan, I. W. S., Pratiwi, Siregar, C. A., Narendra, B. H., Undaharta, N. K. E., Sitepu, B. S., Sukmana, A., Wiratmoko, M. D. E., Abywijaya, I. K., & Sari, N. (2023). Implementation of Soil and Water Conservation in Indonesia and Its Impacts on Biodiversity, Hydrology, Soil Erosion and Microclimate. *Applied Sciences (Switzerland)*, *13*(13). https://doi.org/10.3390/app13137648

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

El Jazouli, A., Barakat, A., Ghafiri, A., El Moutaki, S., Ettaqy, A., & Khellouk, R. (2017). Soil erosion modeled with USLE, GIS, and remote sensing: a case study of Ikkour watershed in Middle Atlas (Morocco). *Geoscience Letters*,

4(1). https://doi.org/10.1186/s40562-017-0091-6

Fernández, D., Adermann, E., Pizzolato, M., Pechenkin, R., Rodríguez, C. G., & Taravat, A. (2023). Comparative Analysis of Machine Learning Algorithms for Soil Erosion Modelling Based on Remotely Sensed Data. *Remote Sensing*, *15*(2). https://doi.org/10.3390/rs15020482

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gaubi, I., Chaabani, A., Ben Mammou, A., & 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. https://doi.org/10.1007/s11069-016-2684-3

Ghorbanzadeh, O., Shahabi, H., Mirchooli, F., Valizadeh Kamran, K., Lim, S., Aryal, J., Jarihani, B., & Blaschke, T. (2020). Gully erosion susceptibility mapping (GESM) using machine learning methods optimized by the multi-collinearity analysis and K-fold cross-validation. *Geomatics, Natural Hazards and Risk*, *11*(1), 1653–1678. https://doi.org/10.1080/19475705.2020.1810138

Issaka, S., & Ashraf, M. A. (2017). Impact of soil erosion and degradation on water quality: a review. *Geology, Ecology, and Landscapes*, *1*(1), 1–11. https://doi.org/10.1080/24749508.2017.1301053

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, *41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Olii, M. R., Olii, A. K. Z., Olii, A., Djau, A. R., Mokoagow, M. A., Kironoto, B. A., Bachtiar, Olii, R. S. N., & 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. https://doi.org/10.2166/wcc.2025.603

Panagos, P., Borrelli, P., Meusburger, K., Alewell, C., Lugato, E., & Montanarella, L. (2015). Estimating the soil erosion cover-management factor at the European scale. *Land Use Policy*, *48*, 38–50. https://doi.org/10.1016/j.landusepol.2015.05.021

Poesen, J. (2017). Soil erosion in the Anthropocene: research needs. *Earth Surface Processes and Landforms*, *43*(11). https://doi.org/https://doi.org/10.1002/esp.4250

Susanti, Y., Syafrudin, S., & Helmi, M. (2019). *Soil Erosion Modelling at Watershed Level in Indonesia : a Review 3 Soil Erosion Modelling at Watershed Skale in Indonesia*. *8*(201 9).

Woldemariam, G. W., Iguala, A. D., Tekalign, S., & Reddy, R. (2018). Spatial Modeling of Soil Erosion Risk and Its Implication for Conservation Planning: the Case of the Gobele Watershed, East Hararghe Zone, Ethiopia. *Land*, *7*, 1–25. https://doi.org/10.3390/LAND7010025

Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., Ma, X., Marrone, B. L., Ren, Z. J., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B. M., Xiao, X., Yu, X., Zhu, J.-J., & Zhang, H. (2021). Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environmental Science & Technology*, *55*(19), 12741–12754. https://doi.org/10.1021/acs.est.1c01339