

Modelling Susceptibility to Water Erosion in the Moroccan High Atlas Using Machine Learning Model: The Case of the Upstream Tassaoute Watershed

Oussama Nait-taleb^{1}, Maryem Ismaili¹, Sana Elomari¹, Insaf Ouchkir¹, Jaouad El Atiq², Fatima Ezzahra El Kamouni¹, Mohamed El Haou¹, Samira Krimissa¹, Mustapha Namous¹, and Abdenbi Elaloui¹*

¹ Data Science for Sustainable Earth Laboratory (Data 4 Sustainable Earth), Faculty of Sciences and Techniques, Sultan Moulay Slimane University, Beni Mellal, Morocco.

² Geomatics, Georesources and Environment Laboratory, Faculty of Sciences and Techniques, Sultan Moulay Slimane University, Beni Mellal, Morocco.

Abstract. Water erosion is one of the most widespread land degradation processes in arid and semi-arid mountainous regions, causing significant soil loss and severely impacting natural resources. This study aims to assess water erosion susceptibility in the Upper Tassaoute watershed (High Atlas, Morocco) using two machine learning models: Random Forest (RF) and Support Vector Machine (SVM). An inventory of approximately 200 eroded sites, established through the integration of field observations and satellite imagery, was used for model training (70%) and validation (30%). Twenty environmental conditioning factors were selected, encompassing topographic, geological, climatic, soil, and vegetation variables. The performance of both models was evaluated using the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC), showing satisfactory predictive accuracy for both RF and SVM. The analysis of variable importance revealed that NDVI, slope, curvature, soil properties, and lithology were among the most influential factors. The results confirm the effectiveness of machine learning approaches for mapping water erosion vulnerability and provide a robust scientific basis to support sustainable land management strategies in sensitive mountainous environments.

Keywords: Water erosion; Machine learning; Random Forest; Support Vector Machine; Upper Tassaoute watershed; Morocco.

*Corresponding author: oussama.nait-taleb@usms.ma

1. Introduction

Soil erosion, although a natural process, is accelerating at an alarming rate as a result of anthropogenic pressures. This phenomenon leads to a loss of fertility in agricultural soils, an alteration in the quality of water resources and the degradation of ecosystems. It is estimated to be the main cause of land degradation, accounting for around 85% of cases worldwide [1]. This worrying dynamic is not just an abstract concept; it has a crying resonance in Morocco, where almost 40% of land is affected, with annual soil losses estimated at between 23 and 55 tonnes per hectare, reaching as much as 524 tonnes per hectare in the most vulnerable areas [13]. In the face of this devastating scale, erosion - and water erosion in particular - is attracting growing interest among researchers and land managers. This urgent attention is motivated by a progressive loss that is increasing, compromising the sustainability of agricultural systems and even threatening the safety of human infrastructures [2]. Faced with this multifaceted threat to agriculture and infrastructure, the search for sustainable solutions is imperative. To anticipate and limit these impacts, it is essential to understand the factors that govern erosion, such as rainfall distribution, soil type and land-use patterns. This complexity explains why the study of erosion mobilizes a wide range of disciplines - from physical geography to economics - reflecting the resolutely interdisciplinary nature of this environmental issue. It is in this context of interdisciplinary research that machine learning techniques have emerged as a powerful tool for assessing susceptibility to water erosion. Developed by numerous researchers, these data-driven models are generally characterized by high predictive accuracy and improved performance in mapping areas at risk [13]. Recently, this interest in predictive modeling has taken shape in the work of researchers such as Garossi et al. Their application enables sensitivity maps to be produced using a wide spectrum of approaches, from traditional statistical methods to the most recent machine learning algorithms. Among the latter, techniques such as Random Forest (RF), Neural Networks (ANN) and Support Vector Machines (SVM) stand out for their performance, confirming the field's transition towards purely data-driven models [3].

The objective of this study is to propose and validate machine learning approaches for modeling susceptibility to water erosion in the upper Tassaoute watershed. Using methods such as random forest (RF) and support vector machines (SVM), it aims to produce an operational map of vulnerability classes to guide the prioritization of soil conservation and erosion control measures at the watershed level. These tools thus contribute to strengthening decision support for the planning and implementation of sustainable water and soil resource management strategies.

2. Materials and methods

2.1 Study area

The Tassaoute drains the land upstream of Lake Moulay Youssef in the central High Atlas region, southeast of Oum Er-Rbia, and extends over 1,418 km², from 31°33'56" N to 31°64'47" N latitude and 6°48'40" W to 7°33'40" W longitude. It lies in the northern sub-atlasic zone known as the Demnate Atlas, to the east of the Haouz plain (Figure 1). Some 35 km from the town of Demnate and 90 km from Marrakech, it is crossed by the Tassaoute river, which rises at an altitude of 3,978 meters. Flowing from northeast to northwest, it is the main tributary of the Oum Er-Rbia River, which has a significant hydraulic load here. The terrain consists of Pre-Permian rocks, with Jurassic rocks to the east of Azilal. This location was chosen for its characteristics: a semi-dry climate marked by torrential rains, uneven terrain, worn-out plants and numerous friable rocks. All this accelerates soil degradation, depriving it of its qualities, and seriously damaging lives and nature in ways that could be irreparable [4].

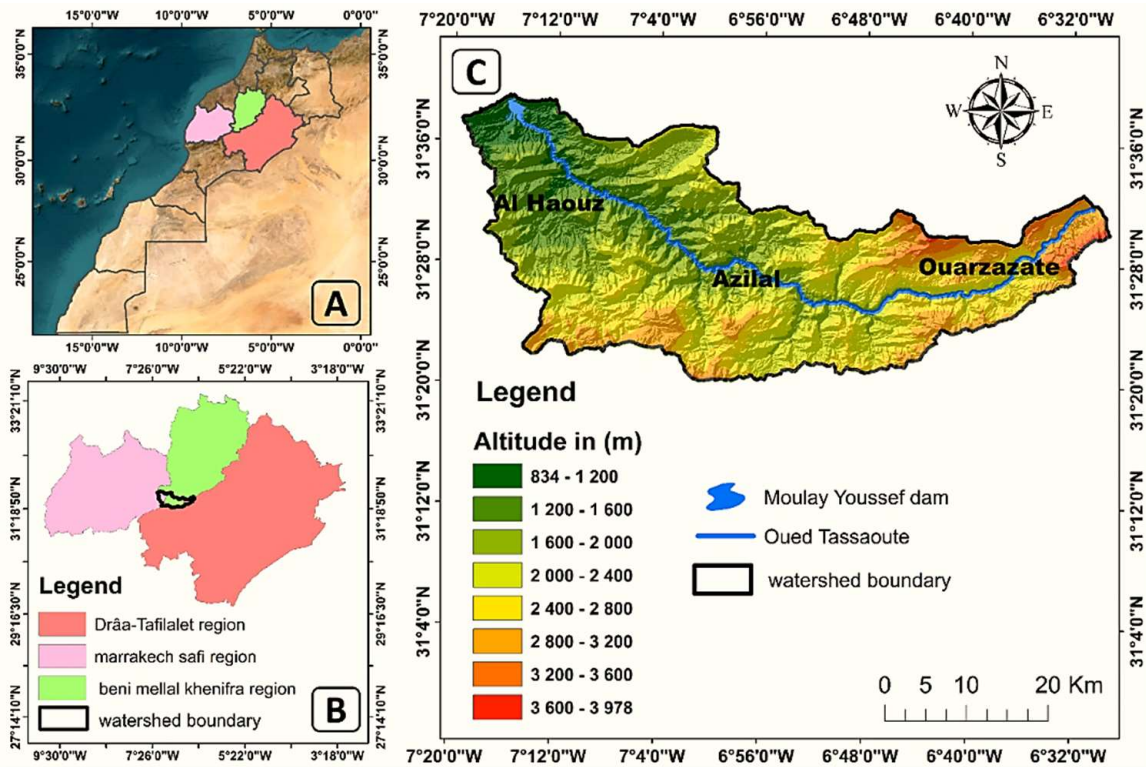


Fig. 1. Location map ((A): Morocco, (B): Regions, (C): Upstream Tassaoute watershed)

2.2 Methodology

The approach adopted in the present study is broken down into several essential stages, represented in the diagram in Figure 2. This figure also illustrates the approach developed for the probabilistic analysis of susceptibility to water erosion, mobilizing RF and SVM models to generate water erosion susceptibility maps with increased precision.

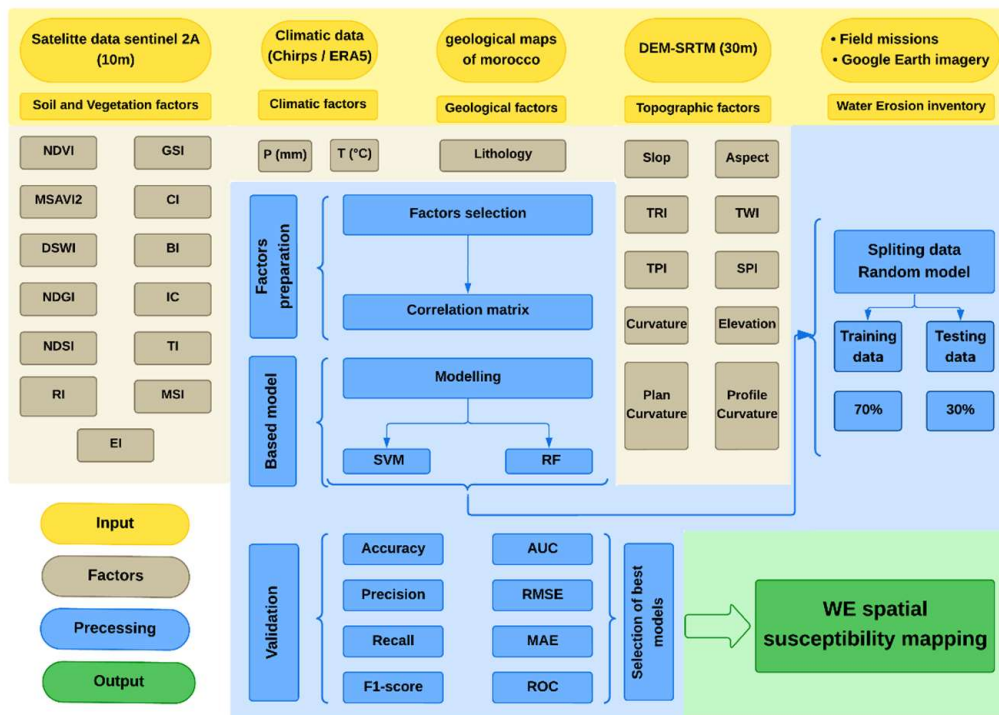


Fig. 2. Methodological flow chart for this study.

2.3 Mapping the water erosion inventory

A water erosion inventory map provides essential information on how water erosion is distributed in space and helps to better understand the links between different factors and water erosion.

In this research, the water erosion inventory map included 220 sites with water erosion and the same number of sites without water erosion (220 pixels). Sites with water erosion were located and represented as points using field observations and high-quality Google Earth images (Figures 3 and 4).

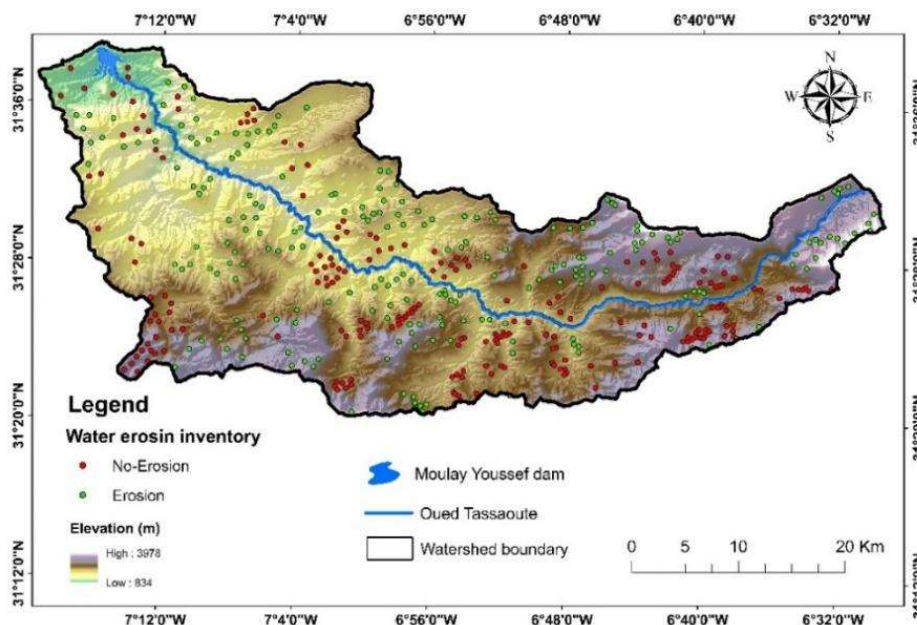


Fig. 3. Water erosion inventory map.



Fig. 4. Recent field photographs of water erosion in the study area.

2.4 Preparing data sets for spatial modelling

To assess susceptibility to water erosion, the selection of conditioning factors is a decisive step, as it directly influences the accuracy and reliability of predictive models. In the present study, and in line with the recommendations of the literature review, 26 factors deemed relevant to soil erosion analysis were collected and prepared. These factors cover various categories, including topographical, pedological, vegetation, climatic and geological

parameters, as presented in Table 1. A brief description of each of these predictive factors is then provided to clarify their role in the modelling process.

Table 1: Factors considered for mapping susceptibility to water erosion.

Parameters	Formula	Parameter name	Reference
NDVI	$NDVI = (NIR - R) / (NIR + R)$	Normalized Difference Vegetation Index	[13].
MSAVI2	$MSAVI2 = \left(2NIR + 1 - \sqrt{(2NIR + 1)^2 - 8(NIR - R)} \right) / 2$	Modified Soil-Adjusted Vegetation Index 2	[14].
DSWI	$DSWI = (NIR + G) / (SWIR + R)$	Disease Water Stress Index	[5]
NDGI	$NDGI = (GR) / (G + R)$	Normalized Difference Greenness Index	[6]
MSI	$MSI = SWIR1/NIR$	Moisture Stress Index	[7]
TI	$TI = (SWIR1 - SWIR2) / (SWIR1 + SWIR2)$	Texture Index	[8]
IC	$CI = 3 * GR - 100$	Crust Index	[9].
BI	$BI = \sqrt{PIR^2 + R^2}$	Brightness Index	[9].
CI	$IC = (RV) / (R + V)$	Color Index	[9]
GSI	$GSI = (RB) / (R + V + B)$	Grain Size Index	[10]
EI	$EI = (R - B) / (R + B)$	Encrustation Index	[11]
RI	$RI = R^2/B * G3$	Redness Index	[12]
NDSI	$NDSI = (R - NIR) / (R + NIR)$	Normalized Difference Salinity Index	[14]
SPI	$SPI = As * \tan\beta$	sediment power index	[13]
TWI	$TWI = \ln\left(\frac{AS}{\tan\beta}\right)$	topographic wetness index	[13]
TRI	$TRI = \sqrt{\sum_k \epsilon N_8 (C_k - C_x)^2}$	Topographic Roughness Index	[13]
TPI	$TPI = z_0 - \frac{1}{N} \sum_{i=1}^n z_i$	Topographic Position Index	[13]

2.4.1 Topographical parameters

The set of DTM-derived topographic parameters used to assess susceptibility to water erosion in the upstream Tassaoute watershed is illustrated below. They include morphometric factors (altitude, slope, exposure, curvatures, TRI, TPI) as well as derived hydrological indices (SPI and TWI), recognized for their influence on surface runoff and the intensity of erosive processes.

2.4.2 Soil and vegetation parameters

In addition to topographical parameters, the analysis of susceptibility to water erosion considered soil and vegetation factors, derived from spectral indices. Soil indices include MSI, TI, IC, BI, CI, GSI, EI, RI and NDSI, which provide information on soil reflectance and susceptibility to erosion. Regarding vegetation: NDVI, MSAVI2, DSWI and NDGI were mobilized to characterize the density, vigor and water status of the vegetation cover, which plays a crucial role in protecting against runoff and soil degradation.

2.4.3 Climatic parameters

Climatic factors play a decisive role in water erosion processes, influencing both runoff intensity and vegetation dynamics. In this study, two major climatic variables were considered: mean annual precipitation and mean annual temperature. These parameters make it possible to assess the impact of climatic conditions on erosion susceptibility in the upstream Tassaoute watershed.

2.4.4 Geological parameters

Lithology is a fundamental factor in the study of susceptibility to water erosion, as it determines the resistance of geological formations to disintegration, runoff and transport processes. The lithological map of the upstream Tassaoute watershed highlights the diversity of geological formations. This lithological variability directly influences the spatial distribution of erosion and the dynamics of landscape evolution.

2.5 Multicollinearity analysis

In order to guarantee the model's robustness and avoid multicollinearity problems that could bias parameter estimation, a pre-selection of explanatory variables was carried out. This preliminary step was based on analysis of the Pearson linear correlation matrix. All pairs of variables with a correlation coefficient above a threshold of $|0.90|$ were identified. This systematic procedure ensured that the machine learning algorithms were presented with a set of predictor variables that were both informative and free from excessive linear redundancies, thus optimizing the generalizing capacity of the models.

2.6 Machine Learning Models

2.6.1 Random forest (RF)

Random Forest (RF), proposed by Breiman in 2001, is a supervised learning method that aggregates multiple decision trees. Each tree is trained on a bootstrap (resampled) subset of the data, and at every split the algorithm considers a randomly chosen subset of predictor variables to determine the optimal node division [14]. The final predictions are obtained by aggregating (averaging or majority voting) the results of the different trees. This approach, which transforms weak classifiers into a robust model, has rapidly aroused great interest in the scientific community due to its high accuracy, its superiority over other methods and its excellent overall performance [13].

2.6.2 Support Vector Machine (SVM)

SVM is a supervised learning approach rooted in statistical learning theory. Originally developed as a discriminant classifier, it can be used for both classification and regression tasks. Its principle is based on the idea of "support vectors": a small number of representative data points is sufficient to define the separating hyperplane and ensure good accuracy, even with a limited number of training samples. SVM stands out for its robustness, learning speed, self-adaptation and low dependence on sample size. It is particularly reliable for processing complex data, especially in remote sensing, as it can efficiently model non-linear relationships. However, this method has certain limitations: it is often difficult to interpret, consumes a lot of memory and computing power, does not directly provide probability estimates and remains sensitive to outliers [13].

2.7 Validation and accuracy assessment

In the context of water erosion risk modeling, the joint use of validation metrics confirms the reliability of the results. Accuracy, F1-score and AUC-ROC provide a good overall representation of model performance, while Kappa and MCC reinforce the robustness of the assessment to spatial imbalances between eroded and non-eroded areas. Finally, RMSE and MAE quantify prediction deviations, guaranteeing an accurate and consistent estimate of erosion risk.

3. Results

3.1 Data selection and analysis

Pearson's correlation matrix was applied to assess collinearity between all explanatory factors (Figure 5). The results show strong correlations ($r > 0.90$) between several variables, notably the spectral indices NDSI, NDVI, NDGI, IC, RI and GSI. These parameters, redundant in information and likely to introduce a bias in the modeling, were eliminated from the analysis. The other factors have moderate to low correlations, so their contribution to the modelling process is retained.

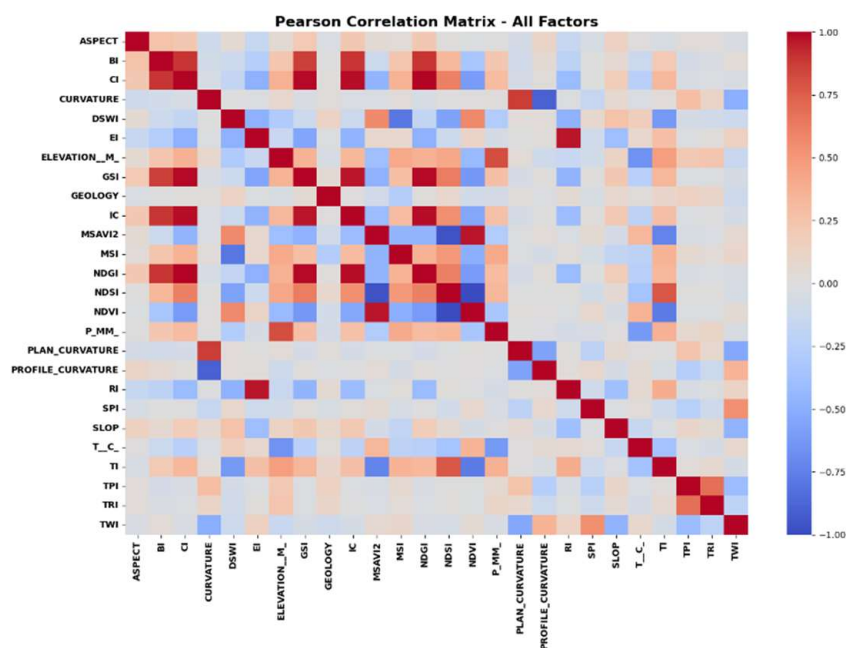


Fig. 5. The pearson correlation matrix of conditioning factors.

3.2 Importance factors using the RF and SVM model

To estimate the relative importance of the conditioning factors, two methods were applied. In the case of Random Forest, the Mean Decrease in Impurity was used to assess the contribution of each parameter to improving model accuracy. In addition, a linear SVM was used to extract the standardized coefficients of the variables, reflecting their relative weight in the classification. The results obtained indicate that certain parameters, such as EI, CI, TI and others, play a dominant role in predicting water erosion susceptibility, while other variables are of marginal importance (Figure 6). Overall, this combination of methods offers a robust and complementary view of the contribution of factors to modelling.

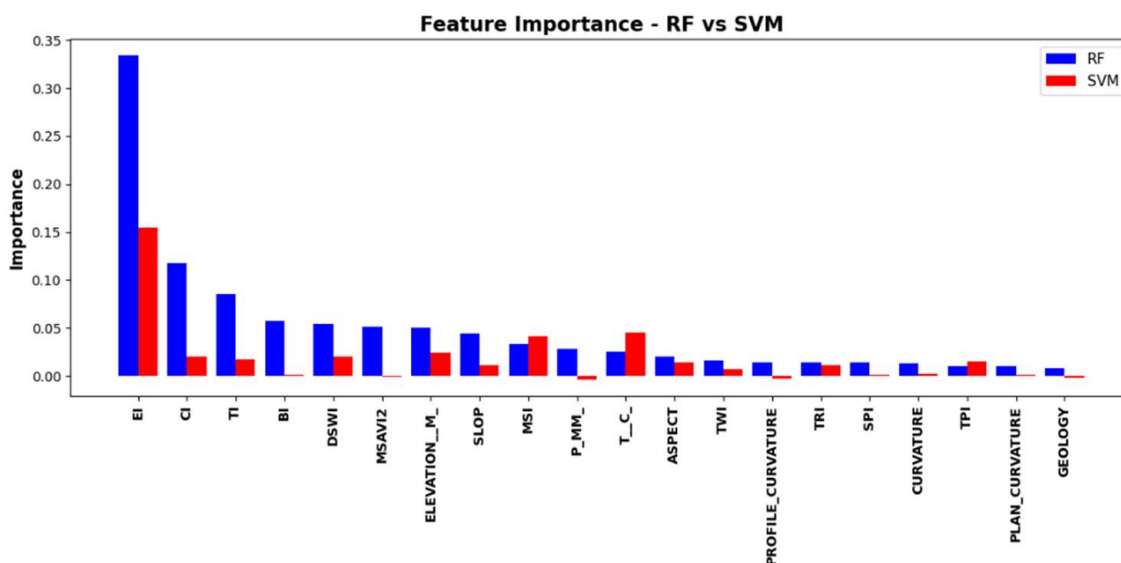


Fig. 6. Evaluation of the importance of conditioning factors using the RF and SVM algorithm.

3.3 Water Erosion Susceptibility Prediction

Water erosion susceptibility maps were generated using two machine learning algorithms, Support Vector Machine (SVM) and Random Forest (RF), by applying the partition scenario to the training and validation data (70%/30%) (Figure 7). The predictions of each model were made pixel by pixel over the entire upstream Tassaoute catchment, with probabilistic values ranging from 0 to 1, corresponding respectively to low and very high sensitivity to erosion. The resulting maps were then reclassified into four risk levels: low, medium, high and very high, using the ArcGIS 10.3 classification method.

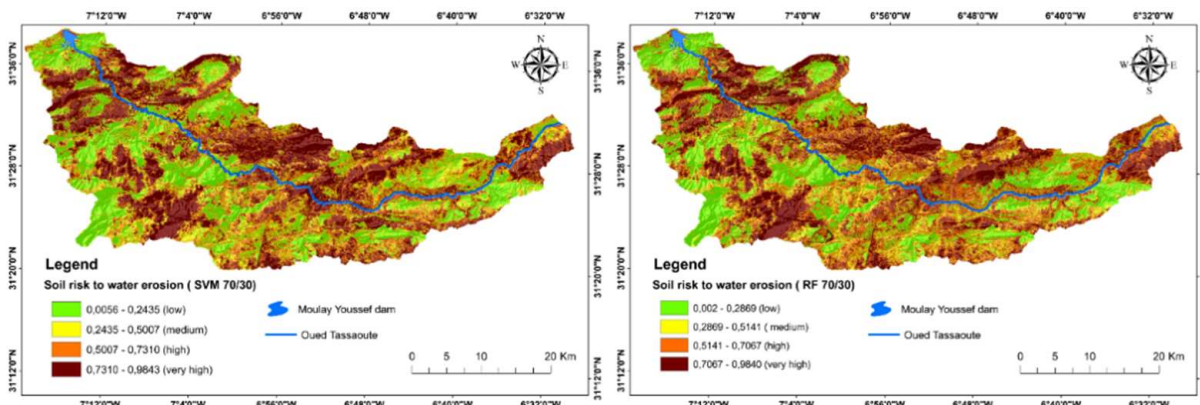


Fig. 7. Water erosion sensitivity maps using RF and SVM models.

Visual analysis of the two maps generated shows strong agreement between the two models in terms of the spatial distribution of vulnerable areas. Areas of low susceptibility mainly cover plains and areas with moderate topography, located in the downstream and central part of the basin. Medium-susceptibility zones are fragmented, generally in transition between flat areas and slopes. Areas of high and very high susceptibility are concentrated in mountainous and steeply sloping areas, particularly in the upstream part of the basin and along river systems, reflecting increased susceptibility to erosion.

3.4 Model accuracy and validation results

Evaluating the validation and accuracy of predictive performance is an essential step in optimally analyzing results. In this study, performance was assessed using several statistical indicators, including accuracy, F1 score, precision, sensitivity, specificity, Kappa coefficient, MCC, RMSE, MAE and area under the ROC (Receiver Operating Characteristic) curve (figure 8) (table 2).

Table 2: Validation results for the two prediction models (RF/SVM) using training and test data.

Model	Accuracy	Precision	Recall	F1	Kappa	MCC	AUC	RMSE	MAE
RF_Train	0,931818	0,90303	0,967532	0,934169	0,863636	0,865848	0,992452	0,234474	0,178468
RF_Test	0,848485	0,859375	0,833333	0,846154	0,69697	0,69729	0,942608	0,320185	0,245598
SVM_Train	0,915584	0,876471	0,967532	0,919753	0,831169	0,835691	0,976893	0,238733	0,147331
SVM_Test	0,843636	0,863636	0,863636	0,863636	0,727273	0,727273	0,932048	0,323414	0,209739

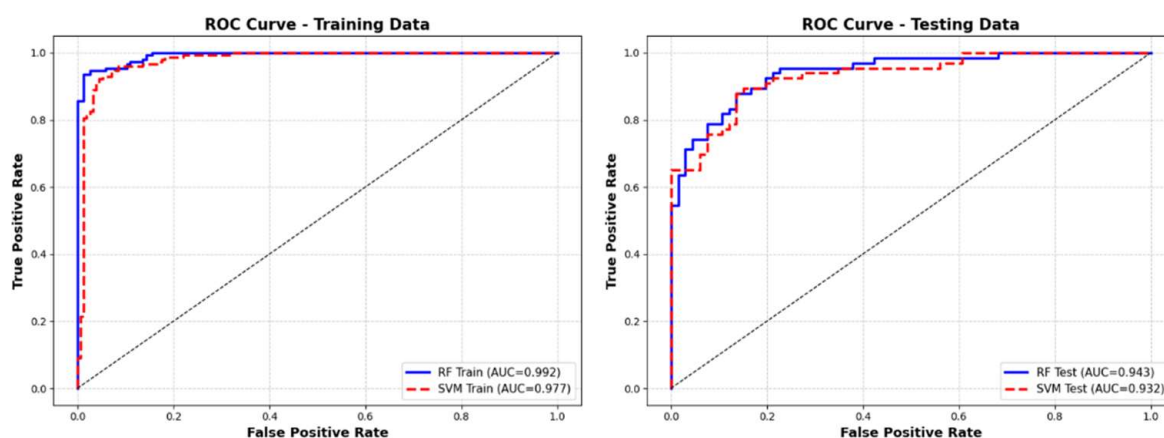


Fig. 8. Receiver operating characteristic (ROC) curves: success rate (training data) and predictive rate (test data).

ROC curves and AUC values highlight the excellent performance of both models, with a notable advantage for Random Forest (RF). On training data, RF achieves an AUC of 0.992, slightly higher than that of SVM (0.977), demonstrating its ability to distinguish classes perfectly during the learning phase. This trend is confirmed on test data, where RF maintains an AUC of 0.943 versus 0.932 for SVM, indicating better generalization to new data. What's

more, the minimal deviation between training and test AUCs for both models reveals little overlearning, confirming their robustness. So, although both approaches show outstanding results, Random Forest stands out for its slightly higher discriminating power and stability, making it the most reliable choice for this classification task.

4. Discussion

In this study, the results obtained highlight the relevance of evaluating the performance of predictive models of gully susceptibility, using both discrimination and reliability criteria. The evaluation was carried out using the scenario (70/30%), and through several statistical indicators, notably the Kappa index, the area under the ROC curve (AUC), the root means square error (RMSE) and the mean absolute error (MAE). The results show that the RF model outperformed (AUC = 0.94), while the SVM model (AUC = 0.93).

These results are in line with many previous works that underline the superiority of Random Forest (RF) models in the prediction of complex environmental phenomena.

Indeed, the RF has several notable advantages: it is based on a robust machine learning algorithm, capable of efficiently handling large datasets without requiring prior variable reduction, while providing unbiased estimates of generalization error as the forest is built. In addition, it identifies the most decisive factors in classification, reliably handles missing data and retains its validity even in the presence of large gaps in the databases [15].

Thus, the use of predictive models such as RF appears particularly advantageous in terms of cost and resource mobilization, as it enables managers and decision-makers to target intervention priorities. By improving the effectiveness and relevance of strategic choices, these approaches contribute to better management of environmental risks and the implementation of appropriate, sustainable solutions.

5. Conclusion

This study assessed the performance of the machine-learning algorithms Support Vector Machine (SVM) and Random Forest (RF) for modeling water-erosion susceptibility in the upstream Tassaoute watershed. A total of 220 erosion sites and 220 no-erosion sites were identified and randomly split into training (70%) and testing (30%) datasets. Twenty-six conditioning factors were derived from multiple databases. The results indicate that both models achieve satisfactory overall predictive performance.

In the specific context of this watershed, subject to strong anthropic pressure and a semi-arid climate, the susceptibility maps produced are proving to be a valuable decision-making tool. They enable the most vulnerable areas to be delineated with fine spatial resolution, providing local players with an objective basis for prioritizing interventions and optimizing soil conservation strategies.

For future research, several avenues of improvement are conceivable. The integration of complementary variables, such as detailed soil indicators or high temporal resolution land use data, could refine predictions. In addition, the exploration of hybrid models or ensemble learning approaches would be a promising avenue for enhancing the stability and accuracy of the susceptibility maps generated.

Acknowledgements: The authors express their sincere gratitude to the researchers supporting the PRIMA RESILINK and GEANTech projects for their valuable support of this research. Special thanks are extended to the anonymous reviewers for their insightful and constructive comments.

Declaration of generative AI: The authors declare that they have used generative AI for the creation of this manuscript. During the preparation of this work, they used GPT-4 to correct the grammar and structure of the English language, ensuring sentence clarity for the reader. After using this tool/service, the author revised and corrected the content as required and took full responsibility for the publication's content.

References

- 1 A. Vrieling, S. C. Rodrigues, H. Bartholomeus, et G. Sterk, « Automatic identification of erosion gullies with ASTER imagery in the Brazilian Cerrados », *International Journal of Remote Sensing*, vol. **28**, no 12, p. 2723-2738, juin 2007, doi: 10.1080/01431160600857469.
- 2 H. Mosaid, A. Barakat, V. Bustillo, et J. Rais, « Modeling and mapping of soil water erosion risks in the Srou Basin (Middle Atlas, Morocco) using the EPM model, GIS and magnetic susceptibility », *Journal of Landscape Ecology*, vol. **15**, n° 1, p. 126-147, 2022.
- 3 T. Svoray, E. Michailov, A. Cohen, L. Rokah, et A. Sturm, « Predicting gully initiation: Comparing data mining techniques, analytical hierarchy processes and the topographic threshold », *Earth Surf. Processes Landf.*, vol. **37**, n° 6, p. 607-619, 2012, doi: 10.1002/esp.2273.
- 4 A. Elaloui *et al.*, « Soil erosion under future climate change scenarios in a semi-arid region », *Water*, vol. **15**, n° 1, p. 146, 2023.
- 5 Z. Bochenek, K. Dąbrowska-Zielińska, R. Gurdak, F. Niro, M. Bartold, et P. Grzybowski, « Validation of the LAI biophysical product derived from Sentinel-2 and Proba-V images for winter wheat in western Poland », *Geoinformation Issues*, vol. **9**, n° 1, p. 15-26, 2017.
- 6 R. Nedkov, « NORMALIZED DIFFERENTIAL GREENNESS INDEX FOR VEGETATION DYNAMICS ASSESSMENT », 2017, 2017.
- 7 T. Benabdelouahab, R. Balaghi, R. Hadria, H. Lionboui, et B. Tychon, « Assessment of vegetation water content in wheat using near and shortwave infrared SPOT-5 Data in an irrigated area », *rseau*, vol. **29**, n° 2, p. 97-107, juin 2016, doi: 10.7202/1036542ar.
- 8 S. Li *et al.*, « Combining color indices and textures of UAV-based digital imagery for rice LAI estimation », *Remote Sensing*, vol. **11**, n° 15, p. 1763, 2019.
- 9 S. Maimouni, A. Bannari, A. El-Harti, et A. El-Ghmari, « Potentiels et limites des indices spectraux pour caractériser la dégradation des sols en milieu semi-aride », *Canadian Journal of Remote Sensing*, vol. **37**, n° 3, p. 285-301, juin 2011, doi: 10.5589/m11-038.
- 10 Laboratory of Natural Resources Management, Department of Geography, University of Yaoundé I, Cameroon *et al.*, « Assessment of Land Degradation Status and Its Impact in Arid and Semi-Arid Areas by Correlating Spectral and Principal Component Analysis Neo-Bands », *IJARSG*, vol. **5**, n° 1, p. 1539-1560, févr. 2016, doi: 10.23953/cloud.ijarsg.77.
- 11 K. J. Ambouta, « Etude des facteurs de formation d'une croute d'érosion et de ses relations avec les propriétés internes d'un sol sableux fin au Sahel. », 1996, Consulté le: 1 juin 2025. [En ligne]. Disponible sur: <https://elibrary.ru/item.asp?id=5601796>
- 12 Gadal S, Gbetkom P, and Mfondoum A. A new soil degradation method analysis by sentinel 2 images combining spectral indices and statistics analysis: application to the

- Cameroonians shores of lake Chad and its hinterland. In: Proceedings of the 7th international conference on geographical information systems theory, applications and management. AMU - Aix Marseille Université: SCITEPRESS - Science and Technology Publications (2021). p. 25–36. doi: 10.5220/0010521200250036
- 13 H. Eloudi *et al.*, « Robustness of optimized decision tree-based machine learning models to map gully erosion vulnerability », *Soil Systems*, vol. 7, n° 2, p. 50, 2023.
 - 14 A. Parmar, R. Katariya, et V. Patel, « A Review on Random Forest: An Ensemble Classifier », in *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, J. Hemanth, X. Fernando, P. Lafata, et Z. Baig, Éd., Cham: Springer International Publishing, 2019, p. 758-763. doi: 10.1007/978-3-030-03146-6_86.
 - 15 C. Jiang, W. Fan, N. Yu, et E. Liu, « Spatial modeling of gully head erosion on the Loess Plateau using a certainty factor and random forest model », *Sci. Total Environ.*, vol. 783, 2021, doi: 10.1016/j.scitotenv.2021.147040.