Landslide Susceptibility Assessment in Western External Rif Chain using Machine Learning Methods

Marouane Benmakhlouf [1*] , Younes El Kharim [1] , Jesus Galindo-Zaldivar [2, 3] ,
Reda Sahrane [1]

[1] GERN, FS, Abdelmalek Essaadi University, 93000, Tétouan, Morocco.

[2] Instituto Andaluz de Ciencias de la Tierra, CSIC-University of Granada, 18071 Granada, Spain.

[3] Department of Geodynamics, University of Granada, 18071 Granada, Spain.

Abstract

Landslides are a major natural hazard in the mountainous Rif region of Northern Morocco. This study aims to create and compare landslide susceptibility maps in the Western External Rif Chain context using three advanced machine learning models: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and K-Nearest Neighbors (KNN). The landslide database, created by satellite imagery and field research, contains an inventory of 3528 cases of slope movements. A database of 12 conditioning factors was prepared, including elevation, slope, aspect, curvature, lithology, rainfall, topographic wetness index (TWI), stream power index (SPI), distance to streams, distance to faults, distance to roads, and land cover. The database was randomly divided into training and validation sets at a ratio of 70/30. The predictive capabilities of the models were evaluated using overall accuracy (Acc), area under the receiver operating characteristic curve (AUC), kappa index, and F score measures. The results indicated that RF was the most suitable model for this study area, demonstrating the highest predictive capability (AUC= 0.86) compared to the other models. This aligns with previous landslide studies, which found that ensemble methods like RF and XGBoost offer superior accuracy. The most important causal factors of landslides in the study area were identified as slope, rainfall, and elevation, while the influence rate of TWI and SPI was the minimum. By analyzing a larger inventory of landslides on a more extensive scale, this study aims to improve the accuracy and reliability of landslide predictions in a west Mediterranean morphoclimatic context that encompasses a wide variety of lithologies. The resulting maps can serve as a crucial resource for land use planning and disaster management planning.

Keywords: Landslides; XGBoost; K-nearest Neighbour; Random Forest; Susceptibility.

1. Introduction

Northern Morocco, specifically the mountainous region of the Rif cordillera, is frequently affected by landslides, including flows, rock falls, and complex movements [1–5]. These landslides are not only prevalent in the central region but also in the western external part, as will be demonstrated in this study (Figure 1). Therefore, it is crucial to compile comprehensive and accurate databases, including a geographic database that contains landslide inventories and condition factors. A detailed landslide inventory can enhance the prediction of these events and help manage the associated geohazards more effectively [6].

Various methods have been employed regionally for landslide susceptibility mapping to aid in hazard assessment and risk mitigation. Earlier studies used heuristic and statistical methods [7, 8], while recent years have seen the

emergence of machine learning algorithms as powerful tools for landslide modeling in northern Morocco. These algorithms include artificial neural networks, support vector machines, and others [9]. The integration of machine learning (ML) algorithms and Geographic Information Systems (GIS) has significantly advanced landslide susceptibility mapping and hazard assessment. This combination allows for the detection of complex patterns to predict the spatial probability of landslide occurrence, hazards, and risk management [10, 11].

Using a landslide inventory and various factors as training and validation data, machine learning models can achieve high accuracy in predicting landslide susceptibility and hazards. The predictions can then be visualized and analyzed in GIS to identify hazardous zones and help authorities develop proper risk mitigation strategies and land use planning policies. Internationally, previous studies have utilized machine learning algorithms such as K-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGBoost), and Random Forest (RF) for landslide susceptibility mapping in a South Asian morphoclimatic context [12, 13]. These studies highlight the need to complement these studies in different environmental settings and validate the results.

The aim of this contribution is to compare the results of landslide susceptibility maps in the Western External Rif (Morocco) calculated by three of the most efficient advanced machine learning models: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and K-Nearest Neighbors (KNN). We will discuss their accuracy and suitability. To facilitate this study, we have compiled a comprehensive landslide database that takes into account several parameters. By analyzing this larger inventory of landslides on a larger scale, we hope to improve the accuracy and reliability of landslide predictions in a west Mediterranean morphoclimatic context that includes a wide variety of lithologies.

2. Geological Setting

This research study is focused on the Western External Rif, located in the northern part of Morocco (Figure 1), and covers 5000 km². It is part of the Rif Mountain range that belongs to the heavily deformed and eroded alpine orogeny of the westernmost Mediterranean. The main mountain-building event that took place during the Late Cretaceous and since the Early Tertiary [14]. The Rif mountains (Figure 1) are located at the northern boundary of the African plate, mainly formed by the West African Craton, a stable, ancient landmass that has been relatively unchanged for over a billion years. This boundary region has been affected by the collision of the African and Eurasian plates, which has resulted in the uplift and folding of the Rif mountains [15, 16]. The Rif is subdivided into three structural domains: the internal Rif, the Flysch nappes, and the external Rif.
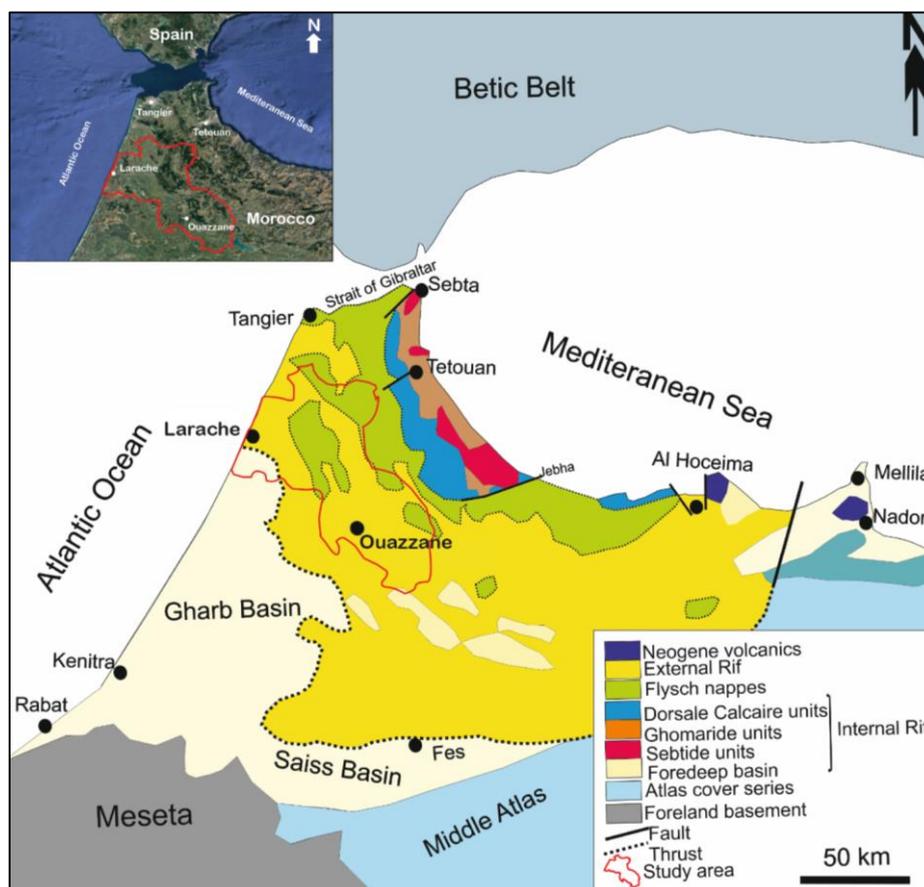

Figure 1. Geological map of the Rif (based on the structural map of the Rif 1/500000)

The Internal Zone, comprising the Sebtide and Ghomaride metamorphic complexes, underwent the main Alpine deformation during the Late Oligocene to Early Miocene [14, 17, 18]. The Flysch divides the External Zones from the Internal Zones, although rocks from the Late Miocene usually cover the contact between the two zones in the Rif Chain (Figure 1).

The flysch nappes that constitute the basement geology of the Rif external zones were deposited as marine sediments during the Cretaceous period, between 145-65 million years ago [19]. They comprise alternating beds of sandstone, shales, and schists accumulated along the northern margin of Africa [20]. Intense compression during the Alpine orogeny caused the thrusting and transport of these sedimentary sequences for hundreds of kilometers, stacking them into nappe complexes [21]. In the study area, the predominant flysch lithologies are phylitic shales and schists intercalated with sandstone [22].

From a geomorphological point of view, the geological structures impart a topography of range-and-valley landscapes to the Western Rif external zones. The unstable flysch units weather to form gentler external dip slopes, along with steeper internal fold-and-thrust belt reliefs [22]. Ongoing tectonic activity maintains this elevated topography through crustal shortening processes and influences seismic activity in the region [21]. The undulating terrain is dissected by a dense drainage network incised into the weak flysch lithologies.

Overall, the Western External Rif region has a warm temperate Mediterranean type climate with mild winters and hot, dry summers. However, the region can experience extreme weather events, such as heavy rainfall and thunderstorms, particularly during the winter months [23].

3. Research Methodology

The methodology for analyzing landslide occurrence began with an examination of a landslide database, which included twelve causal factors derived from various sources. These factors were elevation, slope, aspect, curvature, lithology, rainfall, topographic wetness index (TWI), stream power index (SPI), distance to stream, distance to fault, distance to roads, and landcover. The data resolutions were not uniform across all factors. The ASTER GDEM and Landsat images had a resolution of 30 m, and thus, the factors derived from these data shared the same resolution. However, the geology, rainfall, distance from the road, distance from the stream, and fault map maps had a higher resolution. For the susceptibility map to be produced, all causal factors need to have the same resolution. Given that 12 out of 10 factors had a resolution of 30 m, this resolution was selected for the susceptibility maps. This decision was also supported by Rabby et al. (2020) [24], who noted that a resolution of 30 m led to the highest accuracy for landslide susceptibility mapping. The landslide susceptibility mapping process is complex and requires careful consideration of various factors and their resolutions. The resolution of the data can significantly impact the accuracy of the susceptibility map. For instance, higher-resolution data does not necessarily lead to more accurate susceptibility maps, especially in areas of high relative relief [25]. The flowchart of the research methodology that was used to achieve the study's aims is shown in Figure 2.
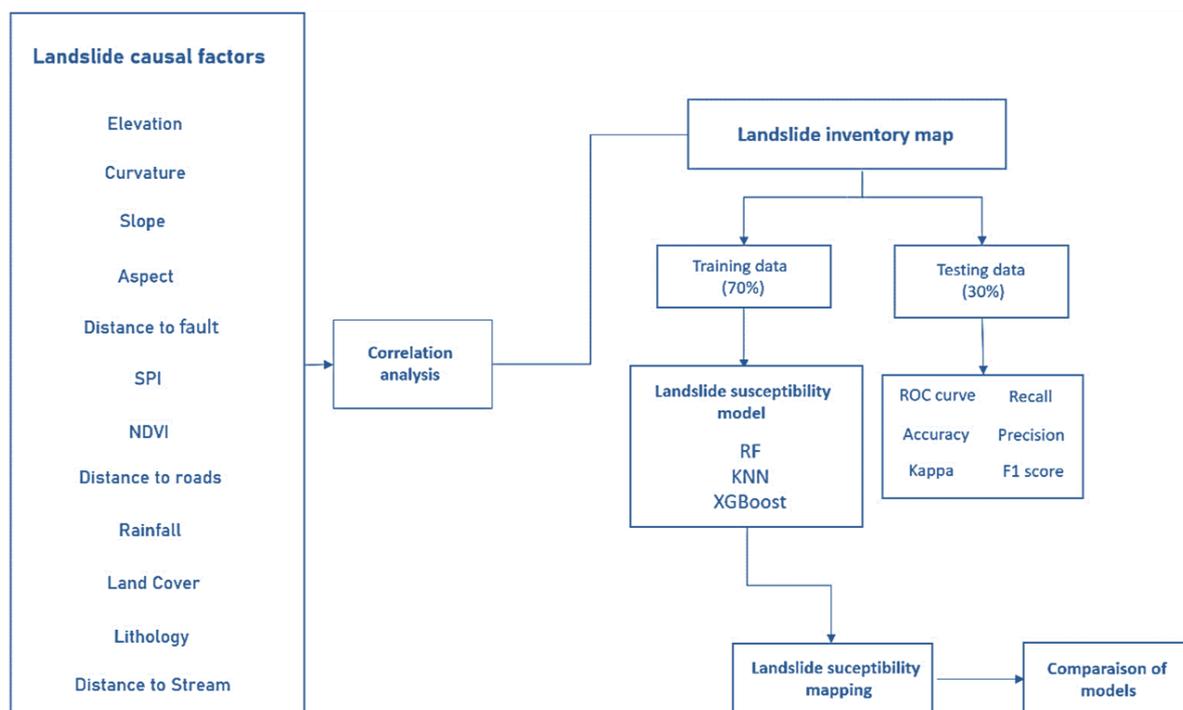

Figure 2. The workflow of the study

Landslide locations were divided into training and validation data sets for the purpose of landslide susceptibility mapping [26]. The total of 3528 landslides were randomly split into two subsets with a 70:30 ratio. Machine learning techniques were used for the mapping, which required dichotomous data for the training and validation data sets (1: presence; 0: absence). Consequently, an equal number of non-landslide locations (absence) were randomly produced for the landslide-free area [27, 28]. These locations without landslides were also divided in the same 70:30 proportion. Landslide sites were assigned a value of "1", and non-landslide locations a value of "0". Three models were used for this process: Random Forest (RF), K-nearest Neighbor (KNN), and XGBoost, and their results were compared.

3.1. Random Forest

Random Forest (RF), an ensemble learning technique, was introduced by Breiman in 2001. This method is highly flexible and can be used for both regression and classification tasks. It employs a strategy known as bagging (bootstrap aggregating) to mitigate overfitting and enhance the model's stability [29]. In the RF approach, each decision tree is constructed using a bootstrap sample from the training set. This means that a random subset of the training data is utilized for each tree's growth. To further prevent overfitting and enhance stability, only a subset of the features is considered at each node when determining the optimal split. This randomization process helps to decrease the correlation among individual trees, leading to a more diverse and robust ensemble. The RF models have several hyperparameters, including the number of trees to be grown, the maximum number of terminal nodes that the trees in the forest can have, and the node size. The RF model was implemented using the "randomForest" package in the R 3.6 statistical software [30].

3.2. K-nearest Neighbor (KNN)

The K-Nearest Neighbors (KNN) algorithm is a fundamental yet powerful tool used for both classification and regression problems [31]. This algorithm is nonparametric, meaning it does not make any assumptions about the underlying data distribution. Instead of learning a discriminative function from the training data set, it memorizes the data, earning it the label of a 'lazy learning' algorithm [32].

In the KNN classifier, the parameter 'k' is a positive integer. When predicting the class of unseen data, the algorithm computes the distances between the new data and all samples in the training set. This is typically done using the Euclidean distance. The 'k' nearest neighbors of the new data are then selected. The optimal 'k' value is determined through hyperparameter optimization, which calculates the model's accuracy for different 'k' values. In this study, a grid search (ranging from 1 to 100 in increments of 1) was performed using the 'Caret' Package in the R environment to select the most suitable 'k' value [33].

3.3. XGBoost

XGBoost, an abbreviation for Extreme Gradient Boosting, is a highly efficient library that has been optimized for gradient boosting. It was created by Chen & Guestrin [34]. This library is applicable to both classification and regression problems. A standout feature of XGBoost is its ability to mitigate overfitting. Overfitting is a prevalent issue in machine learning where a model is excessively complex and fits the training data too closely, resulting in subpar performance when applied to new, unseen data. XGBoost addresses this issue by managing the growth of decision trees using methods such as setting a maximum tree depth and pruning. The library grows the tree to a certain depth and then prunes it backward until the improvement in the loss function is below a certain threshold [35]. Several model parameters need to be selected when using XGBoost models. These include nrounds (the maximum number of iterations), colsample_bytree (the ratio of columns to consider when constructing each tree), subsample (the ratio of training instances), and max_depth (the depth of a tree for effective optimization).

3.4. Model Validation and Comparison

The landslide susceptibility maps produced from KNN, RF, and XGBoost models were evaluated for their predictive capabilities using Area Under Curve (AUCs) and various statistical measures. These statistical measures included the true positive rate (sensitivity) Equation 1, true negative rate (specificity) Equation 2, total accuracy, and the Kappa index (Equation 3). The true positive rate, also known as sensitivity, is the percentage of landslide pixels that are correctly identified as landslide occurrences. On the other hand, the false-positive rate is the percentage of landslide pixels that are incorrectly identified as non-landslide occurrences [26, 36]. The overall accuracy is calculated as the ratio of correctly identified landslide and non-landslide pixels to the total number of landslide and non-landslide pixels.

The Kappa index is a reliable method for evaluating the performance of these models as it takes into account the agreement caused by the rater's guessing [37]. It is often used to assess the agreement between two raters and can also be used to assess the performance of a classification model. Unlike overall accuracy, the Kappa index takes into account the imbalance in class distribution, making it more complex to interpret but potentially more accurate in cases of class imbalance.

$$Sensitivity = TP/(TP + FN) \tag{1}$$

$$Specificity = TN/(FP + TN) \tag{2}$$

$$K = (Po - Pe)/(1 - Pe) \tag{3}$$

Where; Po= Observed Agreement, Pe= Expected Agreement.

The Area Under the Curve (AUC) is another measure used to evaluate the performance of the models. The AUC ranges from 0.5-1.0 (or 50-100%), [27, 38] with the validation data set used for the prediction rate curve. A higher AUC value indicates better performance of the model.

4. Landslide Database

To conduct a comprehensive inventory of landslides in the study area, the survey was based on the identification of gravitational processes through the interpretation of aerial and ortho satellite photos, as well as topographic maps. Concurrently, field missions were executed to verify the data and gather additional information, such as the morphological characteristics of the relief, states and activity indices, factors, impacts, and so on. The landslide inventory (Figure 3) across the studied sector, which covers an areaof 5000 km², enabled the identification of 3528 instances of mass movements. These movements include slides, earth/debris flows, creep, rockfalls, and complex movements. This represents an average density of less than 0.75 movements per km². However, the total area affected is approximately 80.43 km², which is nearly 1.65% of the region's territory. Interestingly, two-thirds of these slope movements identified (2777 cases) are of medium or small size (less than 10,000 m² in area), totaling only 6.51 km², or 0.13% of the study region's area.

In terms of the type of movement, slides are the most frequent, followed by flows, which often take the form of mudslides, debris, and solifluctions (Figure 4).

4.1. Landslide Causal Factors

A total of twelve tested causal factors were divided into four groups:

- Geological factors;
- Topographic factors;
- Hydrogeological factors;
- Environmental and anthropogenic factors.

All these twelve factors were generated for whole the study area in raster format in GIS with a resolution of 30m and the same projection system Merchich.
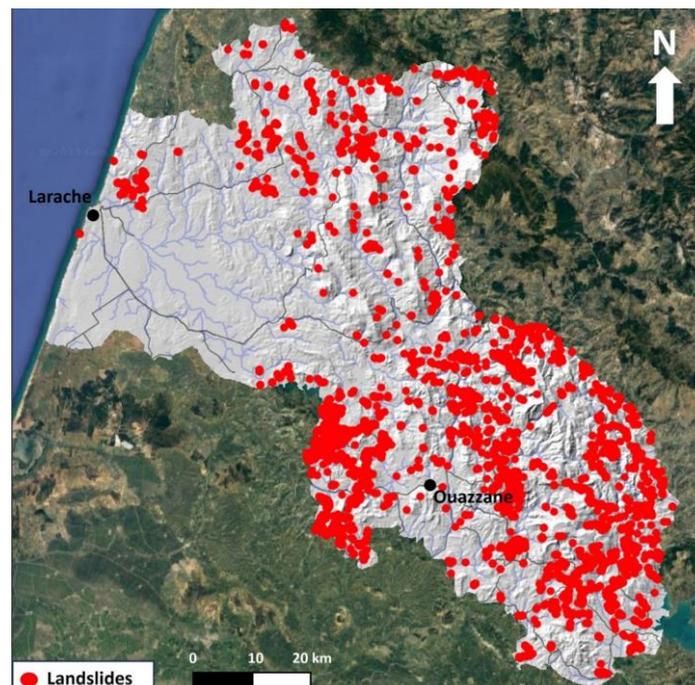

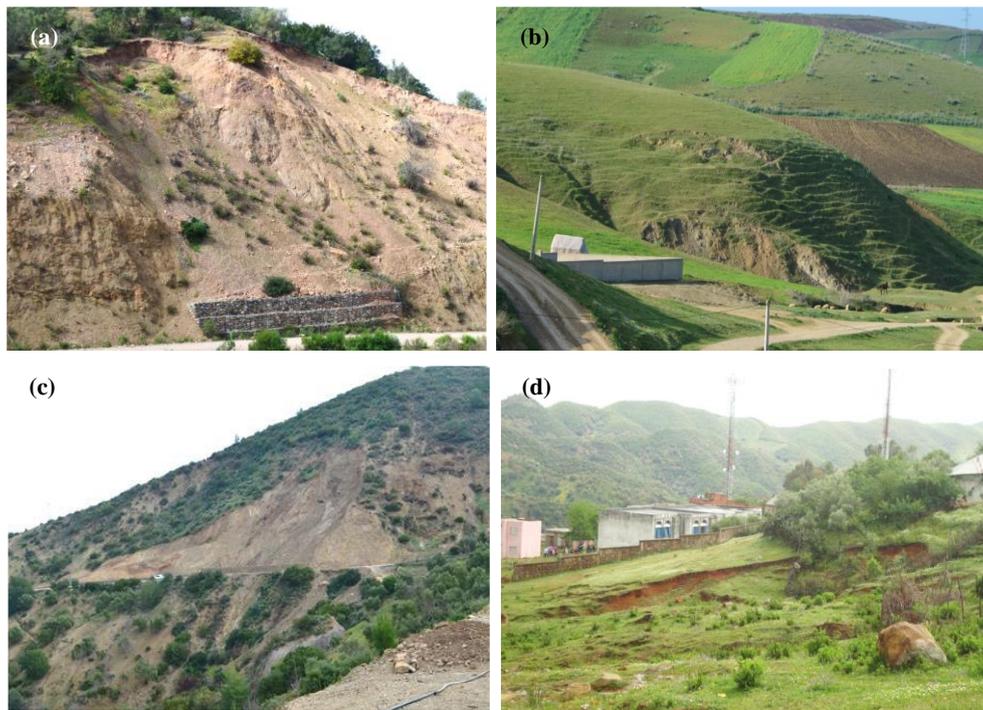**Figure 3. Spatial distribution of landslides in the study area**

Figure 4. Examples of landslides in the field: a). Debris flow at the upstream slope of a track; b). Active landslide; c). Ancient reactivated by the layout and development of the road; d). Reactivated ancient rotational landslide

4.1.1. Geological factors

Two casual factors: lithology and distance from fault. Lithology refers to the physical characteristics of rock units, such as their mineral content, grain size, and texture [39]. It plays a significant role in landslide susceptibility as different rock types have varying degrees of resistance to weathering and erosion, which can influence the stability of slopes. In our study, we identified three dominant lithological classes: pelite, sandstone and marl alternation (28.93%), marl-sandstone alternation (16.99%), and pelite-schists (14.81%). These classes reflect the geology of the study area and are crucial in understanding the geomechanical characteristics of the land. We produced a lithological map of the region by reclassifying and grouping geological formations. This map helps us understand the geomechanical characteristics of the land that outcrops in the territory of the study area (Figure 5).
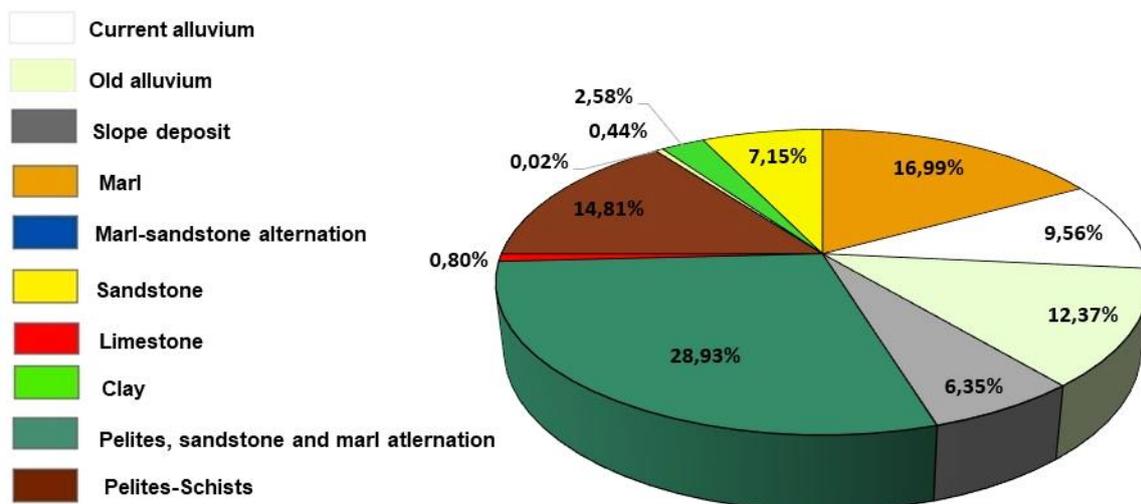

Figure 5. Types of lithologies in the study area

The distance from the fault is another important factor in landslide susceptibility. Faults are fractures in the Earth's crust where significant displacement has occurred [40]. Areas closer to fault lines are generally more prone to landslides due to the increased seismic activity and ground shaking. In our study, we created a map using satellite imagery and on-site observations to determine the proximity to the fault by buffering the fault map (Figure 6).
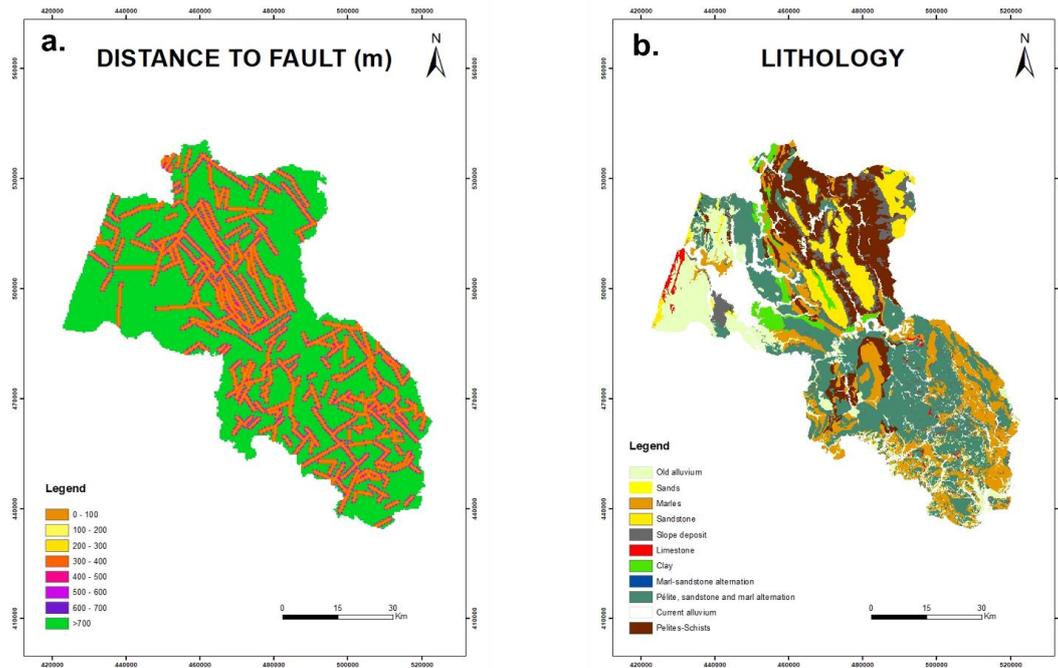
Figure 6. Geological factors: a). Distance to Fault; b). Lithology

4.1.2. Topographic Factors

Topographic variables play a significant role in landslide susceptibility models (LSM), which express the likelihood of a landslide event occurring in a given area based on local terrain conditions. Four topographic factors (Figure 7): Elevation, slope, aspect and curvature which were derived from the ASTER DEM (Digital Elevation Model). Elevation refers to the height of a point on the Earth's surface above a reference point, usually sea level. In the study area, the highest elevation is 1691 meters. Slope is the measure of the steepness of a terrain and is an important conditioning factor for landslides. The highest slope is 63°. Aspect refers to the orientation of a slope, measured clockwise in degrees from 0° to 360°, where 0° is north-facing, 90° is east-facing, 180° is south-facing, and 270° is west-facing.

Curvature is a measure of the rate of change in slope direction and combines both profile curvature (affecting the acceleration and deceleration of flow, which affects erosion and deposition) and planform curvature (affecting the convergence and divergence of the flow). High curvature values can indicate areas where water accumulates or flows rapidly, potentially increasing the risk of landslides.
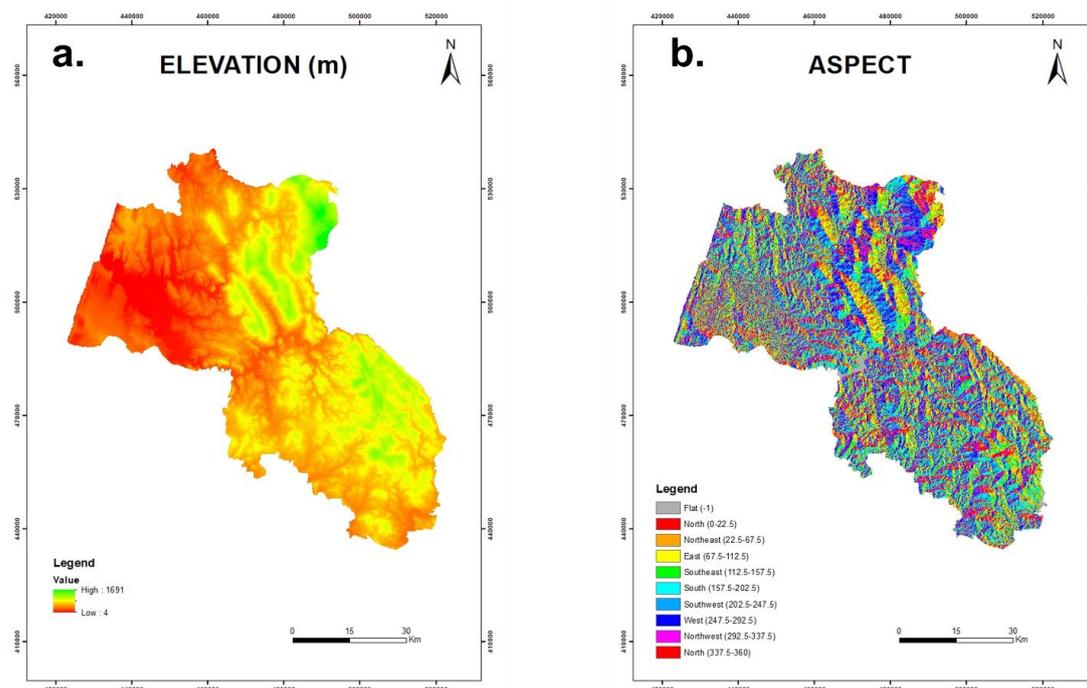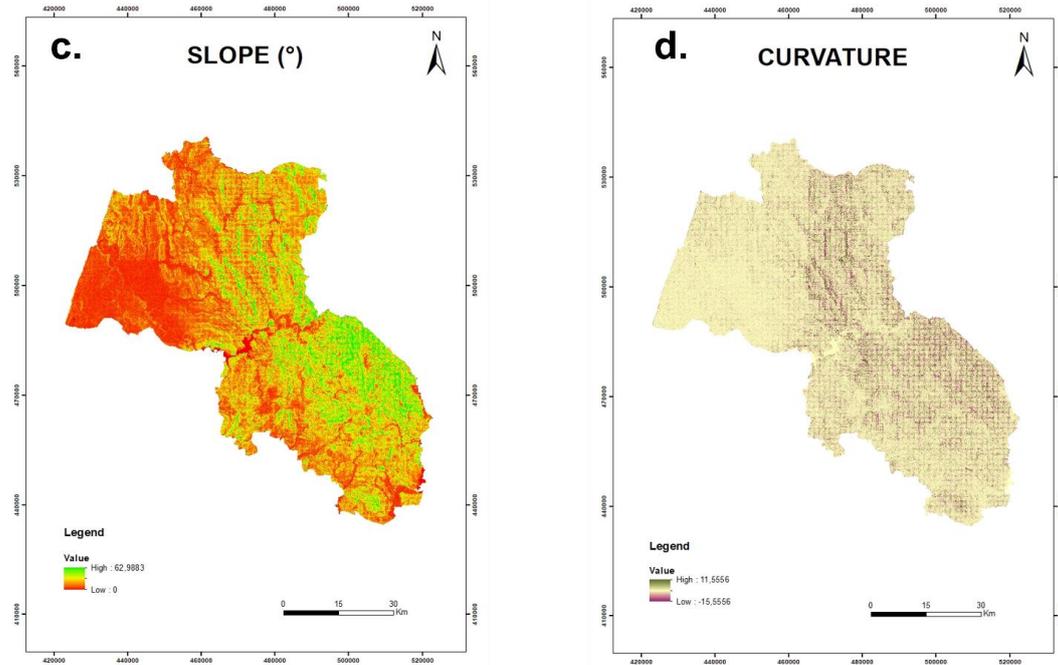
Figure 7. Topographic Factors: a). Elevation; b). Aspect; c). Slope; d). curvature
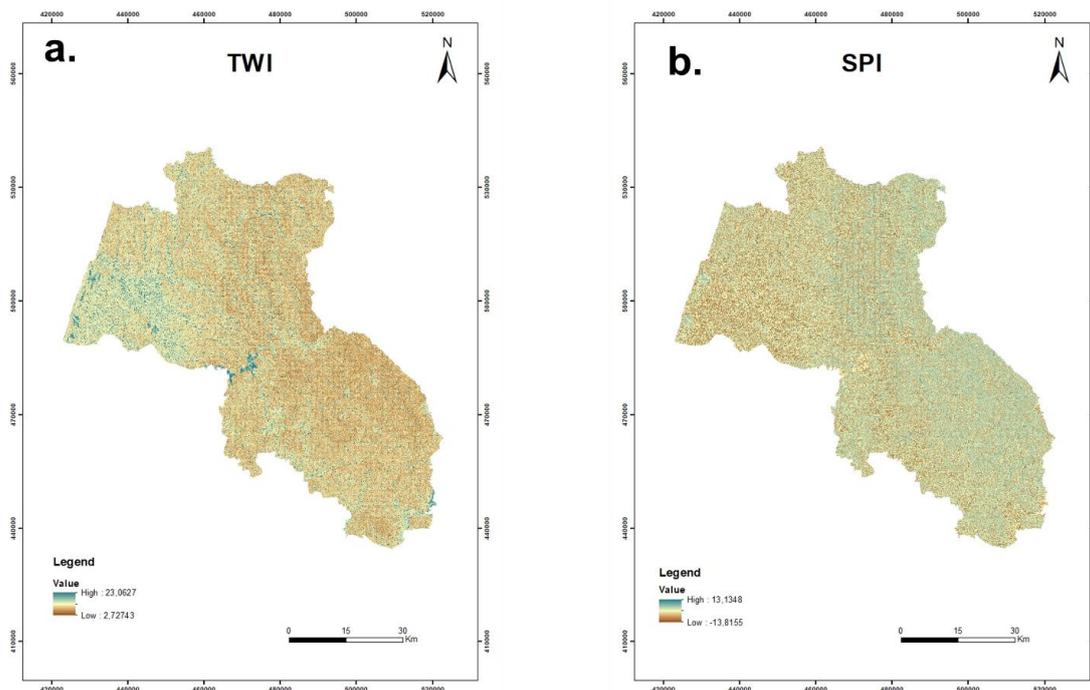
4.1.3. Hydrogeological Factors

In this study, four key conditioning factors were considered: Topographic Wetness Index (TWI), Stream Power Index (SPI), mean annual rainfall and distance to stream (Figure 8). Rainfall, which has a dominant influence on landslide occurrences, was analyzed using the mean annual rainfall data from 2011 to 2020 [41]. The proximity of a location to a stream, which can significantly impact the likelihood of a landslide, was assessed using the Euclidean distance tool in ArcGIS 10.7 applied to the drainage network map. The Topographic Wetness Index (TWI), a tool used to quantify the impact of topography on hydrological processes, was computed using a specific formula (Equation 4) in the ArcGIS 10.7 software. Similarly, the Stream Power Index (SPI), which assesses the erosive power of flowing water [42], was determined using a different specific formula (Equation 5) in the same software, ArcGIS 10.7.

$$TWI = \ln(A/\tan\beta) \tag{4}$$

$$SPI = A \times \tan\beta \tag{5}$$

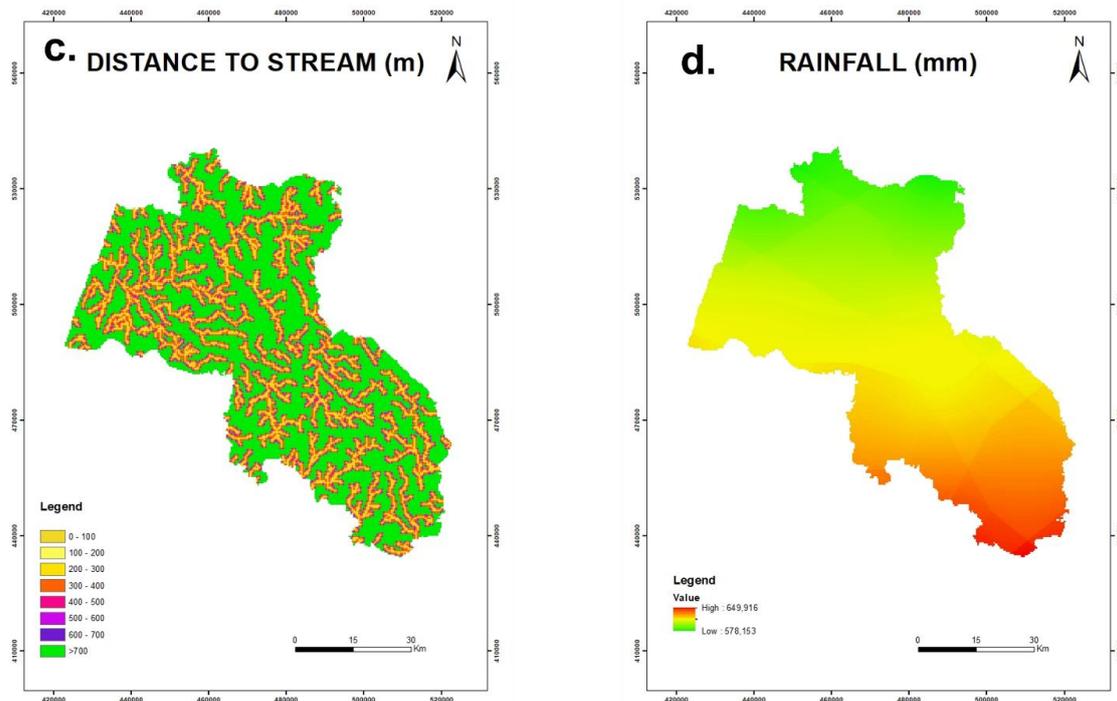Where: A= Area of a Specific Catchment, ß= Slope Gradient of the Specific Area.

Figure 8. Hydrogeological Factors: a). TWI; b). SPI; c). Distance to stream (m); d). Rainfall (mm/yr)

4.1.4. Environmental and Anthropogenic Factors

Two factors used: land cover and distance to roads (Figure 9). Land cover plays a crucial role in landslide susceptibility as it provides information about the type and distribution of vegetation, which can influence the stability of slopes [43]. The land cover map in this research was digitized from Landsat images and Google Earth. Distance to roads is another important factor in landslide susceptibility. Road construction and maintenance can directly or indirectly impact the stability of slopes, leading to an increased risk of landslides. In this study, the Euclidean distance tool in ArcGIS 10.7 was used to prepare the distance to roads map from the national road network.
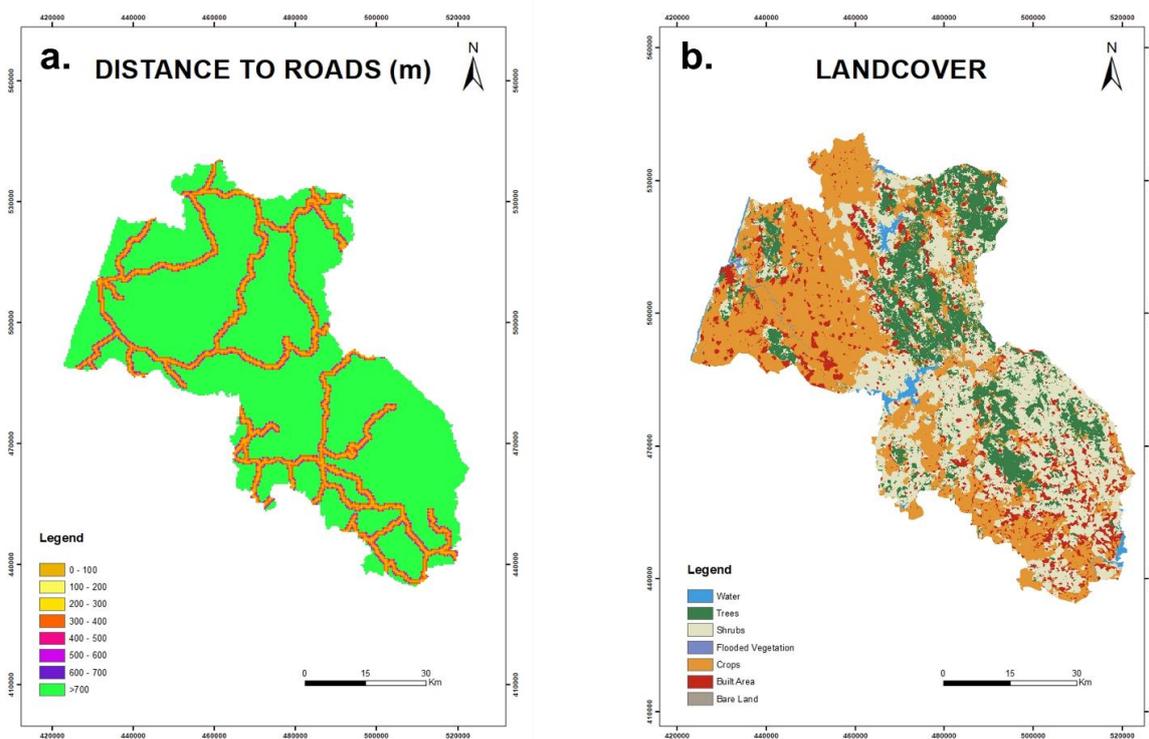

Figure 9. Environmental and anthropogenic Factors: a). Distance to roads (m); b). Land cover

5. Results and Discussion

5.1. KNN

The process of selecting the best K value for the model was carried out using a grid search, ranging from 1 to 100. The model performed optimally when K was set to 9, resulting in an overall accuracy of 0.74. The variable importance plot for the K-Nearest Neighbors (KNN) model indicated that the slope was the most significant factor, followed by elevation, rainfall, and landcover. The distance to roads, Topographic Wetness Index (TWI), and Stream Power Index (SPI) were not present in the plot. The KNN model was used to calculate landslide probabilities for each pixel in the study area. These probabilities were then reclassified into four categories: low, moderate, high, and very high, using the Jenks natural break method. This reclassification was presented as a landslide susceptibility map. In the landslide susceptibility map, 34.31% of the area was identified as having low susceptibility to landslides. On the other hand, 22.42% of the area was classified as having high or very high susceptibility. The model identified the northwestern part of the region as having high susceptibility in certain zones.

5.2. RF

The optimal model chosen demonstrated the highest accuracy of 0.78, achieved with mtry set to 7, the number of trees set to 500, and the number of nodes set to 14 (Table 1). The variable importance plot (Figure 10) indicated that the slope was the most influential factor, followed by Rainfall, Distance to roads, and Elevation. The Random Forest (RF) model provided a probability of landslides for each pixel in the study area, similar to the K-Nearest Neighbors (KNN) model, and these probabilities were reclassified into four categories (Figure 11). In the landslide susceptibility map, nearly half of the area (47.12%) was classified as having low susceptibility. In contrast, 22.32% of the area was categorized as having high or very high susceptibility. The RF model identified the north-western part of the study area as having low to moderate susceptibility, a difference from the KNN model's classification. In the southern part of the area, areas of high susceptibility were observed near the streams.

Table 1. RF parameters

Parameters	Value
Mtry	7
Ntree	500
Node size	14

5.3. XGBoost

The performance of the XGBoost model is influenced by hyperparameter tuning. In this study, a grid search was conducted with various parameter combinations to identify the best model. The highest accuracy (0.78) was achieved with the following parameters (Table 2).

Table 2. XGBoost parameters

Parameters	Value
Nrounds	210
Colsample bytree	0.75
Subsample	1
Max_depth	6

The variable importance plot (Figure 10) for the XGBoost model revealed that slope was the most crucial factor, followed by Rainfall, Distance to roads, and Elevation. Similar to the KNN model, the probability of landslides was reclassified into four classes using the Jenks natural break method. In the landslide susceptibility map (Figure 11), 64.18% of the area was classified as low susceptibility zones, while 14.42% of the area was classified as either high or very high susceptibility zones. The XGBoost model demonstrated a greater ability to classify areas into high and low susceptibility zones compared to the previous two models. It classified only 21.40% of the regions as moderate susceptibility zones, whereas the other two models classified almost twice as many areas as moderate susceptibility zones. The XGBoost model identified the north-western part of the study area as having low susceptibility compared to the other models.

5.4. Model Validation and Comparison

The confusion matrix (Figure 11) serves as a tool to assess and contrast the performance of the three models. The models are performing well in terms of accuracy, with the RF model standing out as the most accurate of the three, indicating its potential superiority in accurately classifying examples. The performance and accuracy of the machine

learning model validation dataset provide an indication of the model's ability to predict future landslides. As per the confusion matrix (Figure 11), the precision rates for the KNN, XGBoost, and RF models are 77.54%, 79.15%, and 82% respectively. The overall accuracy of the three models is high, with the KNN model achieving an accuracy of 0.76, the XGBoost model 0.78, and the RF model 0.80.

The ROC curve and AUC (Figure 12) reveal that the KNN model had the lowest AUC value of 0.81, while the RF and XGBoost models achieved a higher AUC value of 0.86. Based on the analysis of performance metrics, the RF model emerges as the most predictive.
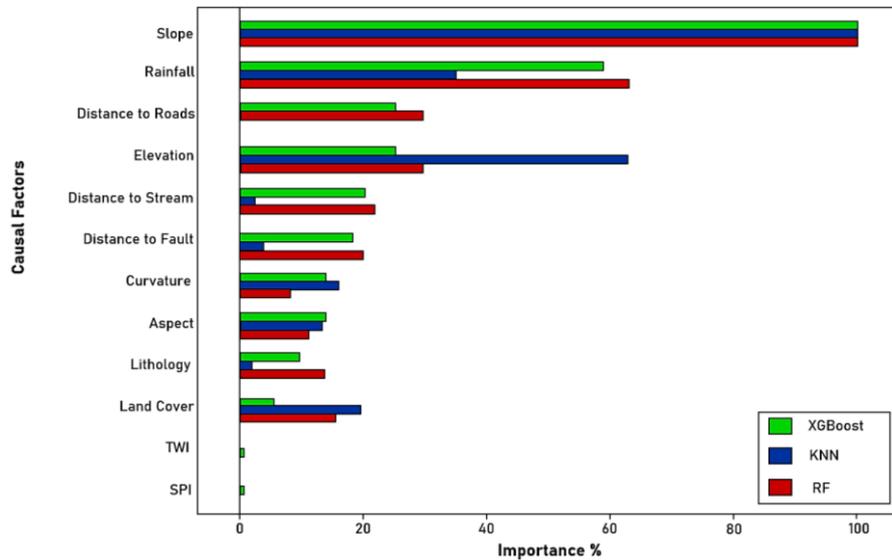
Figure 10. Variable Importance Plots of RF, KNN and XGBoost models

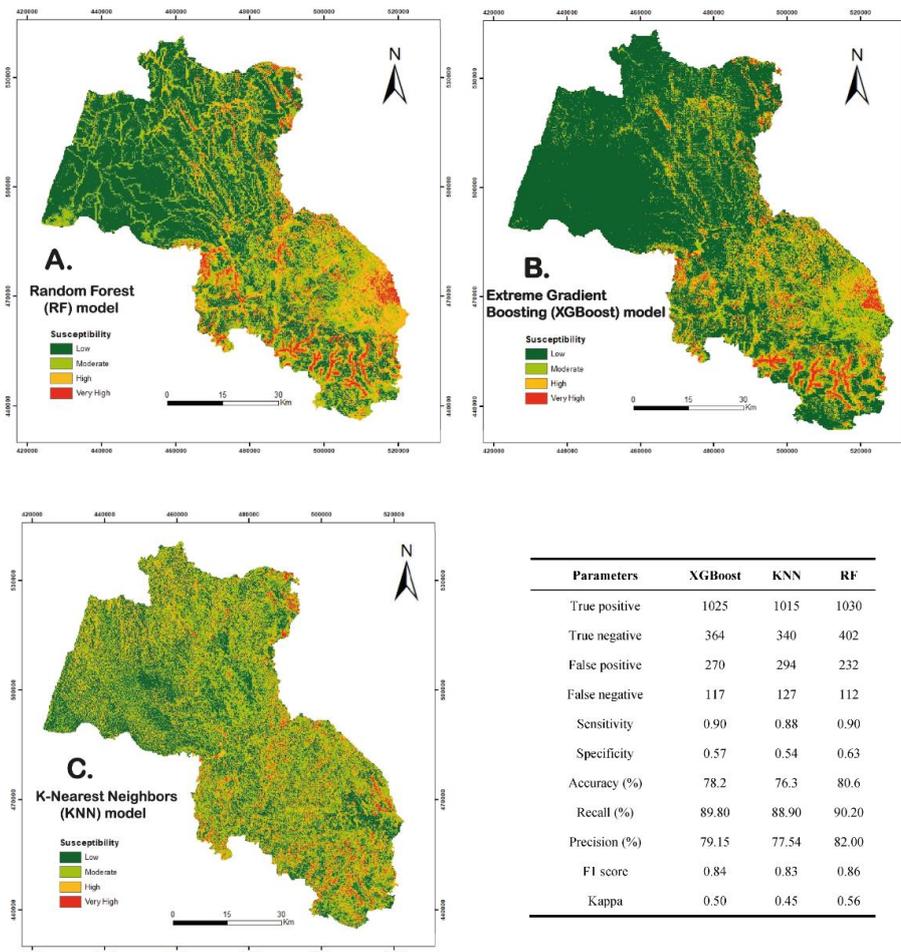| Parameters	XGBoost	KNN	RF
True positive	1025	1015	1030
True negative	364	340	402
False positive	270	294	232
False negative	117	127	112
Sensitivity	0.90	0.88	0.90
Specificity	0.57	0.54	0.63
Accuracy (%)	78.2	76.3	80.6
Recall (%)	89.80	88.90	90.20
Precision (%)	79.15	77.54	82.00
F1 score	0.84	0.83	0.86
Kappa	0.50	0.45	0.56

Figure 11. Landslide Susceptibility Maps (a. RF; b. XGBoost; c. KNN) and the confusion matrix of the three methods
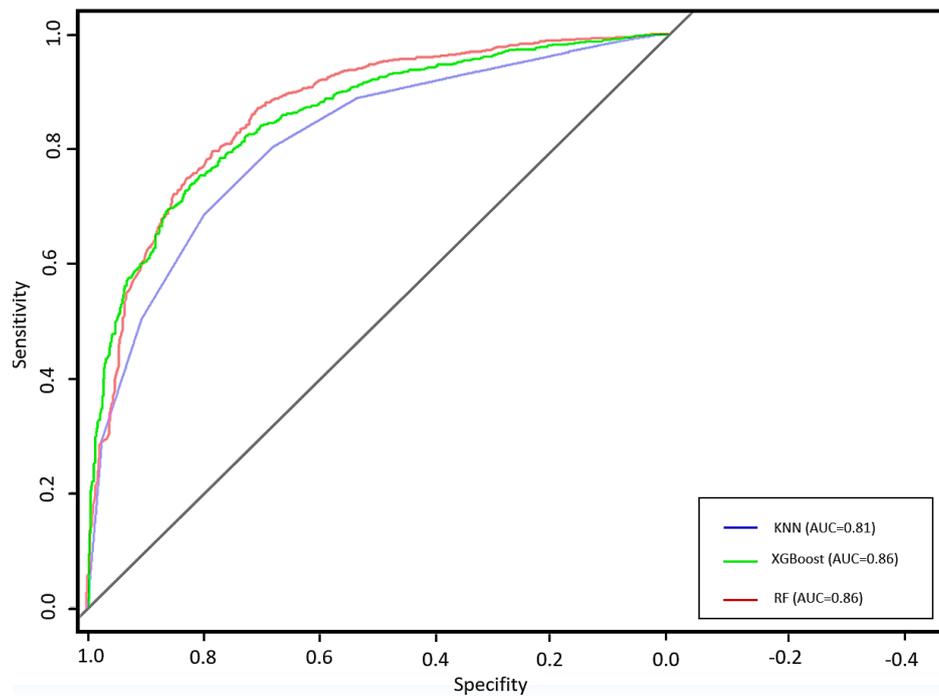
Figure 12. ROC plots of RF, KNN and XGBoost Models

The present study focused on the application of three machine learning models, K-Nearest Neighbor (KNN), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) for landslide susceptibility mapping in the Western External part of the Rif cordillera. The results indicated that these models have excellent precision for landslide susceptibility mapping, with some differences observed in detail. The landslide susceptibility map of XGBoost classified the north-west part as the lowest susceptibility zone in comparison to the KNN and RF model. In contrast, the south part showed some of the highest susceptibility near the streams. The success and prediction rates showed that Random Forest was the best model among the three. The most important causal factors of landslides in the study area were slope, rainfall, and elevation, while the influence rate of Topographic Wetness Index (TWI) and Stream Power Index (SPI) was the minimum.

The results of the study indicate that the Random Forest (RF) model outperforms the K-Nearest Neighbor (KNN) and Extreme Gradient Boosting (XGBoost) models in predicting future landslides in the Central-Western External Rif region of Morocco. The RF model achieved the highest accuracy and precision, as well as the highest Area Under the Curve (AUC) value, which is a measure of the model's ability to distinguish between positive and negative classes. The slope was identified as the most critical variable in all models, with the KNN model placing the greatest emphasis on elevation. These findings are consistent with previous studies conducted in other regions, such as the studies by Rabby et al. (2020) in three Upazilas of Rangamati Hill District, Bangladesh, and Hussain et al. (2022) in Muzaffarabad District in Pakistan [12, 13]. This previous research has also found RF and XGBoost to be effective for complex classification problems with high-dimensional data.

The study's findings also corroborate the results of a comparative analysis of machine learning techniques for landslide susceptibility mapping in Muzaffarabad district, which found that these techniques performed well in assessing landslide susceptibility [44]. Furthermore, a study in Bangladesh found that the RF and XGBoost models were effective in landslide susceptibility mapping [45].

In terms of variable importance, the study's results are in line with the findings of Pourghasemi & Kerle [46], who noted variations in the variable importance graphs for different models. The significance of variables in XGBoost and RF is somewhat influenced by the number of nodes and splits, while the KNN algorithm operates differently, with the number of neighbors determining the model.

One of the strengths of this study is the use of three different machine learning models, which allows for a comprehensive and robust analysis of landslide susceptibility. The study also benefits from a large inventory of landslides and a large study area, which may provide a more comprehensive understanding of landslide susceptibility in the Western External part of the Rif cordillera.

However, the study also has some limitations. For instance, the influence rate of TWI and SPI was found to be the minimum, which may suggest that these factors are not as significant in this particular region. Additionally, while the study found that the RF model was the best among the three, it's important to note that the performance of machine learning models can vary depending on the specific characteristics of the dataset and the study area.

6. Conclusion

This study has successfully applied three machine learning models: K-Nearest Neighbor (KNN), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) for landslide susceptibility mapping in the Western External part of the Rif Cordillera. The results indicate that these models have excellent precision for landslide susceptibility mapping, with the Random Forest model outperforming the other two in terms of accuracy, precision, and Area Under the Curve (AUC) value.

The landslide susceptibility map generated by the XGBoost model classified the north-west part as the lowest susceptibility zone, while the south part showed some of the highest susceptibility near the streams. The most important causal factors of landslides in the study area were identified as slope, rainfall, and elevation, while the influence rate of the Topographic Wetness Index (TWI) and Stream Power Index (SPI) was the minimum. These findings align with previous research conducted in other regions, such as the studies by Rabby et al. (2020) [13] in three Upazilas of Rangamati Hill District, Bangladesh, and Hussain et al. (2022) [12] in Muzaffarabad District, Pakistan. The study's results contribute to the growing body of evidence supporting the use of machine learning models, particularly the RF model, in predicting landslide susceptibility.

By analyzing a larger inventory of landslides on a more extensive scale, this study aims to improve the accuracy and reliability of landslide predictions in a west Mediterranean morphoclimatic context that encompasses a wide variety of lithologies. This comprehensive approach allows for a more accurate understanding of the factors contributing to landslide susceptibility and helps inform effective mitigation strategies in the region.

In conclusion, the study's findings underscore the potential of machine learning models to enhance our understanding of landslide susceptibility and inform effective mitigation strategies. Future research could further explore the application of these models in different geographical and climatic contexts and investigate other potential factors influencing landslide susceptibility.

7. Declarations

7.1. Author Contributions

Conceptualization, M.B.; methodology, M.B.; software, M.B.; validation, Y.E.K. and J.G.Z.; formal analysis, M.B.; investigation, M.B. and Y.E.K.; resources, M.B and Y.E.K.; data curation, M.B., Y.E.K. and J.G.Z.; writing—original draft preparation, M.B.; writing—review and editing, M.B, Y.E.K. and J.G.Z.; visualization, M.B.; supervision, Y.E.K. and J.G.Z.; project administration, Y.E.K. and J.G.Z.; funding acquisition, M.B. All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7.4. Conflicts of Interest

The authors declare no conflict of interest.

8. References

[1] Maurer, G. (1968). Les montagnes du Rif central: Etude géomorphologique. Université de Paris, Paris, France.

[2] Millies-Lacroix, A. (1968). Les glissements de terrains. Présentation d'une carte prévisionnelle des mouvements de masse dans le Rif (Maroc septentrional). Mines et Géologie, 27, 45–55.

[3] Rouai, M., & Jaaidi, E. B. (2003). Scaling properties of landslides in the Rif mountains of Morocco. Engineering Geology, 68(3-4), 353-359. doi:10.1016/S0013-7952(02)00237-5.

[4] Fonseca, A. (2014). Large deep-seated landslides in the northern Rif Mountains (Northern Morocco): inventory and analysis. Geological Society, London, Special Publications, 408, 195.

[5] El Kharim, Y. (2012). Geological features of the slope instability in Tetouan region (the northern Rif, Morocco). Boletín de la RSEHN. Sección Geológica, 106, 39–52.

[6] El Kharim, Y., Bounab, A., Ilias, O., Hilali, F., & Ahniche, M. (2021). Landslides in the urban and suburban perimeter of Chefchaouen (Rif, Northern Morocco): inventory and case study. Natural Hazards, 107(1), 355–373. doi:10.1007/s11069-021-04586-z.

[7] Sahrane, R., El Kharim, Y., & Bounab, A. (2022). Investigating the effects of landscape characteristics on landslide susceptibility and frequency-area distributions: the case of Taounate province, Northern Morocco. Geocarto International, 37(27), 17686–17712. doi:10.1080/10106049.2022.2134462.

[8] Bounab, A., Agharroud, K., El Kharim, Y., El Hamdouni, R., & Faghloumi, L. (2022). The importance of investigating causative factors and training data selection for accurate landslide susceptibility assessment: the case of Ain Lahcen commune (Tetouan, Northern Morocco). Geocarto International, 37(25), 9967–9997. doi:10.1080/10106049.2022.2028905.

[9] Sahrane, R., Bounab, A., & EL Kharim, Y. (2023). Investigating the effects of landslides inventory completeness on susceptibility mapping and frequency-area distributions: Case of Taounate province, Northern Morocco. Catena, 220, 106737. doi:10.1016/j.catena.2022.106737.

[10] Merghadi, A., Yunus, A. P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D. T., ... & Abderrahmane, B. (2020). Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. Earth-Science Reviews, 207, 103225. doi:10.1016/j.earscirev.2020.103225.

[11] Ado, M., Amitab, K., Maji, A. K., Jasińska, E., Gono, R., Leonowicz, Z., & Jasiński, M. (2022). Landslide Susceptibility Mapping Using Machine Learning: A Literature Survey. Remote Sensing, 14(13), 3029. doi:10.3390/rs14133029.

[12] Hussain, M. A., Chen, Z., Wang, R., Shah, S. U., Shoaib, M., Ali, N., Xu, D., & Ma, C. (2022). Landslide Susceptibility Mapping using Machine Learning Algorithm. Civil Engineering Journal (Iran), 8(2), 209–224. doi:10.28991/CEJ-2022-08-02-02.

[13] Rabby, Y. W., Hossain, M. B., & Abedin, J. (2022). Landslide susceptibility mapping in three Upazilas of Rangamati hill district Bangladesh: application and comparison of GIS-based machine learning methods. Geocarto International, 37(12), 3371–3396. doi:10.1080/10106049.2020.1864026.

[14] Chalouan, A., Michard, A., Feinberg, H., Montigny, R., & Saddiqi, O. (2001). The Rif mountain building (Morocco); A new tectonic scenario. Bulletin de La Société Géologique de France, 172(5), 603–616. doi:10.2113/172.5.603.

[15] Suter, G. (1980a). Carte géologique de la Chaine Rifaine, échelle 1:500.000. Ministère de l'Energie et des Mines du Maroc, Direction de la Gélogie, Rabat. Notes et Mem. du Service Geol. du Maroc, 245a.

[16] Suter, G. (1980b). Carte structurale du Maroc, échelle 1:500.000. Ministère de l'Energie et des Mines du Maroc, Direction de la Géologie, Rabat. Notes et Mem. du Service Geol. du Maroc, 245b.

[17] Chalouan, A., & Michard, A. (2004). The Alpine Rif belt (Morocco): A case of mountain building in a subduction-subduction-transform fault triple junction. Pure and Applied Geophysics, 161(3), 489–519. doi:10.1007/s00024-003-2460-7.

[18] Kornprobst, J. (1971). Contribution a l'étude pétrographique et structurale de la zone interne du Rif. Notes et Mémoires Du Service Géologique, Maroc, 251, 1–376.

[19] De Lamotte, D. F., Bezar, B. Saint, Bracène, R., & Mercier, E. (2000). The two main steps of the atlas building and geodynamics of the Western Mediterranean. Tectonics, 19(4), 740–761. doi:10.1029/2000TC900003.

[20] Michard, A. (1976). Eléments de Géologie Marocaine. Notes et Mémoires du Service Géologique du Maroc, 252.

[21] Bargach, K., Galindo-Zaldívar, J., Galdeano, C., & Akil, M. (2023). Tectonic interpretation of late Cenozoic faulting and stress field evolution in the Rif Mountains, Morocco. Tectonophysics, 340(3–4), 151–169. doi:10.1016/S0040-1951(01)00151-0.

[22] Mhammdi, N. A. (1991). Evolution tectonique Néogène et Quaternaire des zones externes de la région du Rif septentrional (Rif Nord-Oriental, Maroc). Ph.D. Thesis, Université Mohamed V, Rabat, Morocco.

[23] ABHL. (2023). Agence Tétouan du Bassin Hydraulique du Loukkos, Tétouan, Morocco. Available online: https://abhl.ma/ (accessed on July 2023).

[24] Rabby, Y. W., Ishtiaque, A., & Rahman, M. S. (2020). Evaluating the effects of digital elevation models in landslide susceptibility mapping in rangamati district, Bangladesh. Remote Sensing, 12(17), 2718. doi:10.3390/RS12172718.

[25] Gaidzik, K., & Ramírez-Herrera, M. T. (2021). The importance of input data on landslide susceptibility mapping. Scientific Reports, 11(1). doi:10.1038/s41598-021-98830-y.

[26] Chen, C. W., Chen, H., Wei, L. W., Lin, G. W., Iida, T., & Yamada, R. (2017). Evaluating the susceptibility of landslide landforms in Japan using slope stability analysis: a case study of the 2016 Kumamoto earthquake. Landslides, 14(5), 1793–1801. doi:10.1007/s10346-017-0872-1.

[27] Althuwaynee, O. F., Pradhan, B., Park, H. J., & Lee, J. H. (2014). A novel ensemble bivariate statistical evidential belief function with knowledge-based analytical hierarchy process and multivariate statistical logistic regression for landslide susceptibility mapping. Catena, 114, 21–36. doi:10.1016/j.catena.2013.10.011.

[28] Chen, W., Xie, X., Peng, J., Shahabi, H., Hong, H., Bui, D. T., ... & Zhu, A. X. (2018). GIS-based landslide susceptibility evaluation using a novel hybrid integration approach of bivariate statistical based random forest method. Catena, 164, 135-149. doi:10.1016/j.catena.2018.01.012.

[29] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. doi:10.1023/A:1010933404324.

[30] Liaw, A., & Wiener, M. (2002). Classification and Regression by Random Forest. R News, 2(3), 18-22.

[31] Nouh, R. M., Lee, H. H., Lee, W. J., & Lee, J. D. (2019). A smart recommender based on hybrid learning methods for personal well-being services. Sensors (Switzerland), 19(2), 431. doi:10.3390/s19020431.

[32] Marjanović, M., Bajat, B., & Kovačević, M. (2009). Landslide susceptibility assessment with machine learning algorithms. International Conference on Intelligent Networking and Collaborative Systems, 273–278. doi:10.1109/INCOS.2009.25.

[33] Kuhn, M. A. (2013). Short Introduction to the caret Package. The Comprehensive R Archive Network, 1-10. Available online: https://cran.r-project.org/web/packages/caret/vignettes/caret.html (accessed online July 2023).

[34] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016, 785–794. doi:10.1145/2939672.2939785.

[35] Cao, J., Zhang, Z., Du, J., Zhang, L., Song, Y., & Sun, G. (2020). Multi-geohazards susceptibility mapping based on machine learning—a case study in Jiuzhaigou, China. Natural Hazards, 102(3), 851–871. doi:10.1007/s11069-020-03927-8.

[36] Van Westen, C. J., Rengers, N., & Soeters, R. (2003). Use of geomorphological information in indirect landslide susceptibility assessment. Natural Hazards, 30, 399-419. doi:10.1023/B:NHAZ.0000007097.42735.9ͤ.

[37] Pham, B. T., Prakash, I., Dou, J., Singh, S. K., Trinh, P. T., Tran, H. T., Le, T. M., Van Phong, T., Khoi, D. K., Shirzadi, A., & Bui, D. T. (2020). A novel hybrid approach of landslide susceptibility modelling using rotation forest ensemble and different base classifiers. Geocarto International, 35(12), 1267–1292. doi:10.1080/10106049.2018.1559885.

[38] Bui, D. T., Lofman, O., Revhaug, I., & Dick, O. (2011). Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. Natural hazards, 59, 1413-1444. doi:10.1007/s11069-011-9844-2.

[39] Yu, X., Zhang, K., Song, Y., Jiang, W., & Zhou, J. (2021). Study on landslide susceptibility mapping based on rock–soil characteristic factors. Scientific Reports, 11(1), 15476. doi:10.1038/s41598-021-94936-5.

[40] Li, Y., Zhang, H., Huang, L., Li, H., & Wu, X. (2023). The formation mechanism of landslides in typical fault zones and protective countermeasures: A case study of the Nanpeng River fault zone. Frontiers in Earth Science, 10, 1-11. doi:10.3389/feart.2022.1092662.

[41] Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. Scientific Data, 7(1), 109. doi:10.1038/s41597-020-0453-3.

[42] Dai, X., Zhu, Y., Sun, K., Zou, Q., Zhao, S., Li, W., Hu, L., & Wang, S. (2023). Examining the Spatially Varying Relationships between Landslide Susceptibility and Conditioning Factors Using a Geographical Random Forest Approach: A Case Study in Liangshan, China. Remote Sensing, 15(6), 1513. doi:10.3390/rs15061513.

[43] Rabby, Y. W., Li, Y., Abedin, J., & Sabrina, S. (2022). Impact of Land Use/Land Cover Change on Landslide Susceptibility in Rangamati Municipality of Rangamati District, Bangladesh. ISPRS International Journal of Geo-Information, 11(2), 89. doi:10.3390/ijgi11020089.

[44] Chowdhury, M. S. (2023). A review on landslide susceptibility mapping research in Bangladesh. Heliyon, 9(7), e17972. doi:10.1016/j.heliyon.2023.e17972.

[45] Khalil, U., Imtiaz, I., Aslam, B., Ullah, I., Tariq, A., & Qin, S. (2022). Comparative analysis of machine learning and multi-criteria decision making techniques for landslide susceptibility mapping of Muzaffarabad district. Frontiers in Environmental Science, 10, 1028373. doi:10.3389/fenvs.2022.1028373.

[46] Pourghasemi, H. R., & Kerle, N. (2016). Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran. Environmental Earth Sciences, 75(3), 1–17. doi:10.1007/s12665-015-4950-1.