ORIGINAL ARTICLE

Landslide susceptibility mapping using GIS-based statistical and machine learning modeling in the city of Sidi Abdellah, Northern Algeria

Bourenane Hamid[1] · **Braham Massinissa**[1] · **Guessoum Nabila**[1]

Abstract

Landslides are one of the most common and damaging geological hazards that constrain the urban planning and development of many cities in northern Algeria. Therefore, landslide susceptibility maps (LSMs) constitute an essential tool for effective hazard management and long-term development planning in landslide-prone areas. The aim of this work is to prepare and compare the LSMs by applying GIS‑based statistical and machine learning models for the new city of Sidi Abdellah and surrounding areas (Northern Algeria). We implemented the statistical models of the frequency ratio (FR), statistical index (SI), and weights of evidence (WoE) models, and the machine learning models represented by a logistic regression (LR) model for landslide susceptibility prediction. An historical landslide inventory map was produced using the interpretation of Google Earth satellite images, available historical records, and geological field investigations. The obtained landslides were randomly divided into the training (70%) and validation (30%) processes. Furthermore, 12 influencing factors for landslide occurrence (including precipitation, slope, elevation, distance to drainage, aspect, land use, density of streams, distance to road, lithology, distance to fault, seismicity, and density of roads) were selected to prepare thematic maps and were considered for susceptibility analysis. Subsequently, landslide susceptibility assessment and mapping are performed by considering the inventoried landslide events and their related predisposing factors using LR, SI, WoE, and FR models in GIS. The accuracy of the four models was verified, validated, and compared using the area under curve (AUC) value of the Receiver Operating Characteristics Curves (ROC) method. The validation results showed that all the used statistical models provided a good accuracy in predicting landslide susceptibility than the machine learning models, with the SI model having a higher AUC value of 80.1% than the WoE (78.2%), FR (78.2%), and LR (64.2%) models. Based on these results, we conclude that the established maps can be used as useful tools for land use planning and risk reduction in the urban area of Sidi Abdellah.

Keywords Landslide · Validation · Susceptibility map · Statistical model · Machine learning model · GIS · Algeria

Introduction

Landslides are a major concern in northern Algeria due to the damage they cause to properties and infrastructure, as well as the loss of human lives. They affect many urban areas, constituting serious threats to the population and presenting a significant constraint to land use planning and development. Large casualties and huge economic losses from devastating landslides have been reported during recent decades in many Algerian cities, such as Constantine, Azazga, Ain El Hammam, Tigzirt, Bejaia, and Algiers (Hadji et al. 2013; Guirous et al. 2014; Laribi et al. 2014; Bourenane et al. 2016; Djerbal et al. 2017; Hallal et al. 2017). The impact of these landslides has been significantly increased and exacerbated by the following : (i) uncontrolled development of the built environment in landslide-prone areas; (ii) inappropriate planning and management; (iii) a lack of policy instruments; and (iv) insufficient understanding about landslide hazards.

Actually, the natural hazards and risks resulting from landslide‑prone areas remain unknown throughout the Algerian territory. Moreover, there seems to be too little consideration given to possible problems arising from bad

✉ Bourenane Hamid
  b.hamid725@gmail.com

[1] Division Microzonage Sismique, Centre National de Recherche Appliquée en Génie Parasismique (CGS), Algiers, Algeria

planning of land use and slope management. There are no strategic disaster risk reduction plans serving to manage and prevent landslide occurrences. The available urban planning instruments represented by both the Master Plan for Urban Planning and Development (PDAU) and the Land Use Plans (POS) did not consider and integrate natural hazards. As a result, many urban settlements grew in naturally hazardous areas. Consequently, it is essential to develop an accurate landslide susceptibility map (LSM) for disaster prediction and management in such a landslide-prone area.

This article attempts to consider the challenges of landslide hazards in land use planning to initiate durable policies and legislation for mitigation and prevention purposes. The LSM contributes to the risk mitigation and management policy for sustainable urban planning and territory development in areas prone to slides. Thus, identifying landslide susceptibility in the urban zone is the first step in any approach intended to reduce the landslides risk. The LSM gives an essential location of where future mass movements will probably occur based on the identification of zones of past landslide occurrences and areas where comparable physical properties exist. This procedure, known as "landslide susceptibility zoning" plays a vital role in the regulation, management, and measures for the reduction of the risks related to identified and potential future landslides.

A number of methods have been developed and applied in the literature for the spatial prediction of landslide susceptibility. Heuristic, deterministic, statistical models, and machine learning are the four primary categories of these methods (Guzzetti et al. 1999; Lee and Min 2001; Lee and Pradhan 2007; Yalcin 2008; Yilmaz 2009; Pradhan and Lee 2010a, 2010b; Pradhan and Youssef 2010; Tien Bui et al. 2011; Ozdemir and Altural 2012; Pourghasemi et al. 2013; Bourenane et al. 2016; Xiao et al. 2019; Merghadi et al. 2020; Huang et al. 2020; Zhou et al. 2021; Huang et al. 2022).

Deterministic approaches appropriate at a small scale, such as versant or catchment, are specifically used to provide early warning of imminent slope failure and deal with mathematical modeling (Goetz et al. 2015). The used physical-based models (such as the infinite slope model) need large quantities of detailed data on the slope failure at site-specific locations to give reliable results.

The heuristic method is a direct subjective approach based on expert knowledge (subjective decision rules) to perform a qualitative LSM (Guzzetti et al. 1999; Thiery et al. 2007). The Multicriteria, Fuzzy Logic, and Boolean Logic evaluation models for landslide mapping correspond to this research type (Zhou et al. 2003).

The statistical method is a quantitative method that is based on statistical relationships between landslide-controlling factors and the landslide distribution. They have been developed to reduce the subjectivity in qualitative

expert analysis. The central idea underlying the quantitative approaches is that the causative factors of future landslides are the same as those imposed in the past (Guzzetti et al. 1999). During the past decades, statistical techniques have yielded entirely satisfactory results and are, consequently, regarded as more objective and more appropriate for landslide susceptibility mapping at medium (1:50,000, 1:25,000) and large scales (1:10,000) because of their ability to reduce errors initiated by expert subjectivity (Bonham-Carter et al. 1989; Van Westen et al. 1997; Lee and Min 2001; Thiery et al. 2007; Lee and Pradhan 2007; Yalcin, 2008; Yilmaz 2009; Pradhan and Lee 2010a, 2010b; Pradhan and Youssef 2010; Tien Bui et al. 2011; Park et al. 2013; Ozdemir and Altural 2012; Pourghasemi et al. 2013; Regmi et al. 2014; Nourani et al. 2014; Goetz et al. 2015; Bourenane et al. 2016; Xiao et al. 2019; Merghadi et al. 2020; Huang et al. 2020; Zhou et al. 2021; Huang et al. 2022). The statistical methods include bivariate (frequency ratio, statistical index, and weights of evidence) and multivariate (logistic regression and discriminant analysis) approaches which were developed for landslide susceptibility mapping.

Machine learning models have been demonstrated to be a distinct elucidation for dealing with large-data spatial analysis when statistical rules are unreliable and the variety of hypothesized understandings of a problem is incomplete (Merghadi et al. 2020). They treat a large range of variable-scale input data without any obligation to pre-existing data structures (e.g., variable transformation or normal distribution). Several machine learning models, including logistic regression, artificial neural networks, fuzzy inference systems, and decision trees, have been developed for landslide susceptibility modeling (Yesilnacar and Topal 2005; Yilmaz 2009; Pradhan and Lee 2010b; Nourani et al. 2014; Goetz et al. 2015; Merghadi et al. 2020; Xiao et al. 2019; Merghadi et al. 2020; Huang et al. 2020; Zhou et al. 2021; Huang et al. 2020).

The spatial prediction of landslide susceptibility using statistical approaches and machine learning supported by GIS has gained popularity and become a major topic of research in the last decade, particularly when dealing with the challenge of landslide susceptibility evaluation at large scales (1:10,000), in which sufficient geotechnical input data is provided. Most of the progress has been made on producing susceptibility maps at the regional- (1:100,000–1:50,000) and medium-scale (1:50,000–1:25,000). A limited number of studies on the effectiveness of these methods have been done at a scale of 1:10,000, which is the scale at which most regulatory landslide hazard and risk maps are produced. A limited number of studies on the effectiveness of these methods have been done at a scale of 1:10,000, which is the scale at which most regulatory landslide hazard and risk maps are produced.

The statistical index, the artificial neural network, the certainty factor, the frequency ratio, and the logistic regression models are the most efficient and reliable methods around the world when compared to physical ones, which require multiple simulations to prepare susceptibility outputs by finding specific geotechnical parameters (Yilmaz 2009; Pradhan and Lee 2010a, 2010b; Park et al. 2013; Nourani et al. 2014; Pradhan and Youssef 2010; Tien Bui et al. 2011; Ozdemir and Altural 2012; Nourani et al. 2014; Regmi et al. 2014; Bourenane et al. 2016). These methods, meanwhile, are expensive, site-specific, and largely rely on deep knowledge of geology and geomorphology. They do, though, have some limitations due to the difficulty in understanding the final output of the black box models and their prediction accuracy in the presence of limited training data sets.

The abovementioned study revealed numerous techniques for improving landslide susceptibility models, including data-related methods and others that focus on the model development and training process. This work aims to develop a reproducible methodology for validation and comparison of LSMs by applying GIS-based statistical and machine learning models in the case of the new city of Sidi Abdellah (Northern Algeria). This research is part of a larger thematic approach that focuses on a better understanding of landslide susceptibility as well as a technique for assessing and mapping landslide susceptibility at a large scale. This work completes prior research on the prediction of landslides at a large scale, allowing scientists to get a better knowledge of the spatial variation of landslide hazard in the urban area of Sidi Abdellah. The results can be used as guidelines for land use and development planning and provide useful guidance for reducing landslide hazards. The final objective of this work was to verify whether the statistical and machine learning approaches produce satisfactory performances that could be implemented in the Algerian context for landslide susceptibility mapping, and more commonly in other areas exposed to the same threat.

Study area characteristics

The new city of Sidi Abdellah, a western extension of the city of Algiers, the capital city of Algeria, is the new attractive urban pole of Algiers. The study area, defined by its geographical coordinates of 36° 37′ 39″ N to 36° 42′ 18″ N in latitude and 2° 48′ 49″ E to 2° 55′ 53″ E in longitude (WGS 1984 and UTM Zone 31 North), is located 25 km south-west of the city of Algiers in northern Algeria (Fig. 1a, b). The city is extended over the territories of five municipalities (Fig. 1c): Douera, Mahelma, Rahmania, Souidania, and Zeralda. It is planned to carry out a very large housing program of 30,000 housing units and the creation of an important concentration of investment as well as

government institutions (Master Plan 2003). The study area corresponds to the urban extent designated by the master city plan (PDAU), which covers a surface area of 31 km^2 (Fig. 1c).

The Sidi Abdellah province is susceptible to progressive landslides due to its geomorphic, geologic, and climatic characteristics and human activities. Landslides, in fact, pose a substantial impediment to the city's development and urban planning. The national and local governments are aware of the gravity of landslide prevention and management.

In terms of geomorphology, the Sidi Abdellah region is a part of the Sahel, an active faulted anticline structure limiting the quaternary Mitidja basin from the north. It is formed by plio-quaternary deposits extending along the Algiers coast (Meghraoui 1988). Sidi Abdellah city is a hilly area located on the foothills of the Sahel ridge, where the altitude ranges between 100 and 400 m. It is crossed by an expected branch of the Sahel active fault (Harbi et al. 2004; Meghraoui 1988; Moulouel et al. 2020).
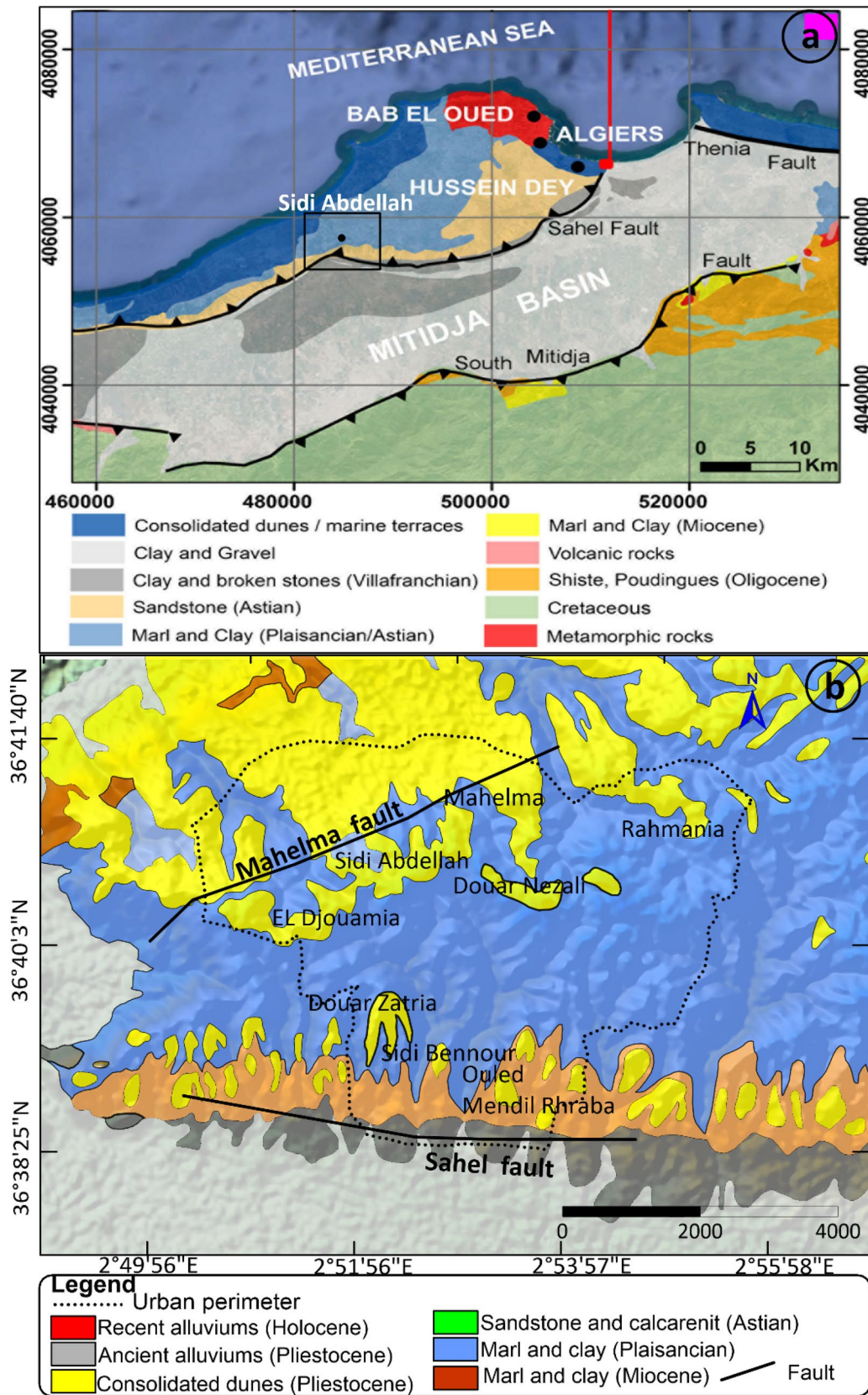
Geologically, the Sidi Abdellah region exhibits metamorphic rock outcrops surrounded by Neogene and Quaternary deposits of the Sahel anticline and the Mitidja basin (Fig. 2a). In terms of structural geology, the Sidi Abdellah region belongs to the internal zones of the Maghrebian chain. The lithology consists of two geological formations (Fig. 2b) as follows: (i) the Pliocene deposits formed by Plaisancian marls, Astien limestone, and sandstone; and (ii) the Quaternary deposits formed by the consolidated dunes and alluvial terraces. The Plaisancian marl formations, covering a large surface of the urban area, are very sensitive to the presence of water and have average-to-high plasticity, which favors landslide occurrence.

The hydrographic network is dense with high slopes (> 20%), represented to the north by the Bou Hayek, Erreba, and Sidi Bennour waterways, and Sidi Harrache, Larhat, El-Aggar, Eddalia, and Mahelma waterways to the south. These main waterways have a semi-permanent flow that are associated with affluent waterways having a temporary flow (Fig. 1c).

The Sidi Abdellah area belongs to the Mediterranean climate type, with a dry period from June to September and a rainy season from October to April. The intensity and frequency of precipitation are concentrated in a short period during the rainy season (November to February), which represents about 50–60% of the yearly precipitation. High rates of rainfall (600 to 800 mm/yr) and heavy storms during the winter and autumn seasons lead to the occurrence of landslides.
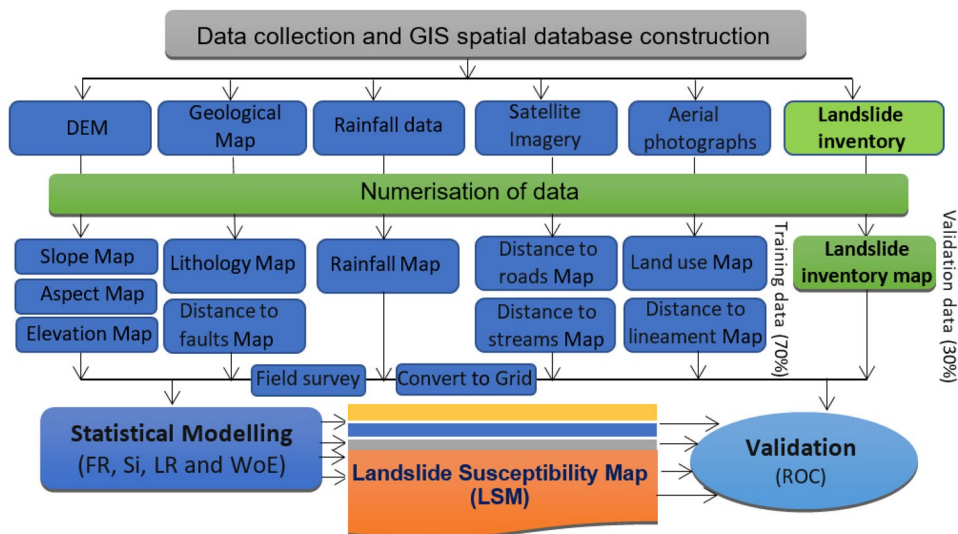
Sidi Abdellah city is characterized by high human activity and intense economic, scientific, and social infrastructure, as well as a high population density. Human activities evolved as a result of historical settlement on the Sidi Abdellah slopes, resulting in significant morphological change

Fig. 1 Geographical localization of Sidi Abdellah city: **a** geographical location in North Algeria; **b** administrative limit of the Sidi Abdellah municipality; **c** the Sidi Abdellah urban zone perimeter

and modification of soil stability conditions (deforestation, extensive clear-cut logging, and vegetation removal) without taking geological and geomorphological constraints into account. The continuous development of the urban area in the northern and southern parts of the city with inappropriate land-use practices is the main factor contributing to the increasing frequency of landslides.

Methodology

For the purpose of this research, the adopted methodology includes five steps that can be adapted to work with any modeling for landslide susceptibility studies (Fig. 3): (1) data collection and development of a spatial database based on GIS; (2) landslide inventory mapping (3) identification and mapping of landslide conditioning factors; (4) landslide susceptibility modeling and mapping using three statistical models and one machine learning model based on Geographic Information System (GIS); and (5) validation and comparison of the four used models after verification of the obtained LSMs using ROC Curves and statistical rules.

Data acquisition and spatial database construction

The first and main step in landslide susceptibility mapping is the data gathering and construction of the spatial database where the pertinent landslides and causative factors have been considered. The mapping of landslide susceptibility depends both on event landslide data and event-controlling factor information. The quality of LSM depends on the amount and quality of the used data.

In the present work, the data were gathered from various sources and have been used to construct thematic layers. The type and source of data used in this work are presented in detail in Table 1. Initially, the landslide inventory map is elaborated based on the exploration of Google Earth satellite images, which are confirmed and completed by field investigation. Furthermore, 12 landslide predisposing factors, including the slope, altitude, distance to drainage, aspect, land use, distance to road, lithology, precipitation, distance to fault, density of roads, seismicity, and density of streams, have been extracted from satellite images, geologic maps, the Digital Elevation Model (DEM), and a precipitation map. ArcGIS (v10.2) software tools were used

Fig. 2 Geological setting of the Sidi Abdellah locality: **a** Geological framework of the Mitidja basin with active faults affecting the Mitidja basins (Aymé 1954); **b** the locations of the Mahelma and Sahel active faults

Fig. 3 Methodological flow chart for landslide susceptibility mapping

Table 1 Database used in the landslide susceptibility assessment

Data layers	Map	Source
Geomorphology	Landslide inventory map	Landslide inventory database, satellite images of Google Earth and extensive field surveys
Relief	Slope gradient map	Digital elevation model DEM (5-m resolution) obtained by interpolation of elevation points of topographic map at scale of 1:2000 from VNSA (ville nouvelle de Sidi- Abdelah 2019)
	Aspect Map	
	Altitude map	
Geology	Lithological map	Digitization of the published geological map at a scale of 1:50,000 by the National Office of Geological and Mining Research (ORGM, 1997) and geological field investigation
	Distance to faults map	
Hydrology	Precipitation map	Precipitation database from local meteorological stations of the National Agency of Meteorology and Hydrology (ANRH 2005) located in the vicinity of the study area has been used to create a mean annual precipitation map by using Kriging interpolation in Arc-GIS (10.2) software tools
	Distance to stream map	Published available documents, Google Earth satellite Imagery, topographic map (1:20,000), and extensive field surveys. The distance to stream map and stream density were prepared from the drainage map using line density and buffer analysis in ArcGIS software environment. The distance to road map and road density were prepared from the road map using line density and buffer analysis in ArcGIS software environment
	Stream density map	
Land use	Land use map	
Anthropy	Distance to road map	
	Road density map	
Seism	Acceleration map	Seismic hazard study from the National Center of Research in Earthquake Engineering (CGS 2009)

to georeference layers, apply coordinate systems and data, visualize, extract, and geoprocess raster datasets. The data were all georeferenced using Algeria's national projection system (UTM Zone 31 North and WGS 1984).

Landslide inventory mapping

The historical landslide inventory constitutes an imperative basic step in landslide susceptibility assessment, principally when a probability modeling approach is adopted. The landslide inventory map of the Sidi Abdellah urban area was elaborated from a combination of the following steps (Table 1): (i) the analysis and interpretation of Google Earth Pro® satellite images with a spatial resolution of 15 m from 2003 to 2018; (ii) available historical records

(landslide reports, newspaper records, thesis, master plans) verified, validated, and completed by (iii) geological fieldwork investigations (between 2015 and 2020). The verification procedure not only provides clear evidence of landslide occurrence but also evidence of the landslide characteristics, as well as estimates of the triggering mechanism, landslide depth, type classification, and identification of conditioning factors.

The landslides were defined by the tension fractures, headscarf, bulges, grab ends, undrained depressions, and lobes. Figure 4 depicts cases of recent observations of various types of landslides observed in the Sidi Abdellah urban area. Figure 5 shows the spatial distribution of landslides as well as their characteristics, such as size (area, perimeter, and failure depth), geological discontinuities and tension

cracks, involved lithology, degree of development, human activity, average slope angle, land cover, and geotechnical features. The landslide perimeter covers approximately an area of 0.5229 km^2 (522.9 ha), which represents 2% of the total perimeter of the urban area. The mapped landslides are defined by different surfaces ranging from 800 m^2 to approximately 58,800 m^2. A diverse variety of failure types, movement rates, and triggering factors are observed across the locale visits.

According to Varnes (1978), the mapped landslides can be divided into rotational and translational slides (99.5%) and falls (0.5%). For the purpose of determining landslide susceptibility, the landslide inventory map was divided at random into the following two portions: 30% for validation procedures and 70% for training or testing landslide models. It is suggested that the higher of the training and validation dataset ratio would improve and increase the accuracy of the testing accuracy.

Landslide predisposing factors

Landslides may manifest as a result of a combination of a number of factors that can be classified into the following two groups: (i) predisposing factors such as lithology, hydrology, land use, topography, and human activity (e.g., railway or road openings, excavation, etc.); and (ii) triggering factors such as earthquakes and rainfall. The assessment of landslide susceptibility is based on a comparison of the landslide-conditioning factor maps and the landslide inventory maps. The results are then extended across the entire investigated region, providing a final LSM output. Moreover, the landslide predisposing factors data must be selected based on landslide type, case study area characteristics, and dataset availability.

In this case study, 12
landslide predisposing factors (Fig. 6), including aspect, slope, landuse, altitude, lithology, seismicity, distance to a drainage network, precipitation, distance to a fault, distance to a road network, density of roads, and density of streams have been identified, analyzed, and considered for establishing LSMs using GIS statistical and machine learning-based models. Table 1 indicates the details of the source of the obtained and prepared versions of each landslide conditioning factor. Because of the non-uniform distribution of factors with dependent variables (e.g., altitude, slope angle, etc.), an autonomous reclassification method was used to produce classified data (i.e., data with class intervals and a number of classes). The classification of factors containing categorical and nominal data (e.g., lithology, stratigraphy, etc.) is the same as that supplied in the source data.

In order to facilitate easy raster calculation, the thematic layers have been sampled at a grid size of 10 m × 10 m using GIS technology. The selection of significant landslide conditioning factors is essential to evaluating the contributions of all factors to landslide occurrence. The feature selection of significant factors is performed based on the Spearman rank correlation coefficient (SRCC) method, which evaluates the contribution of factors by measuring Pearson's correlation between classes and factors.

In Spatial Analyst Tools from ArcGIS, a grid was used to compute the density in each class for each factor based on field data and the relationship of each factor related to each type of landslide. Figure 7 depicts the influence and density of landslides in each factor class.

The significance of landslide predisposing factors

One of the goal procedures in the landslide susceptibility assessment is the evaluation of the importance or influence of the predisposing factors as a result of the limitation of the mutual influence among those factors in developing the state of a landslide. Causal factor selection aims to reduce redundancy in predictor variables and save computation time when some of them are obtained through statistical analyses or reclassification on the same inputs. The redundancy in landslide susceptibility assessment can be caused by the existence of a linear correlation between some independent factors. This phenomenon, known as multicollinearity, can lead to false modeling by analyzing false datasets. Some factors that are insignificant to the occurrence of landslides should be removed to reduce noise and transition fitting issues, thereby improving the model's prediction accuracy. The Spearman rank correlation coefficient (SRCC) is frequently used to assess the contribution or influence factors to landslide occurrence with strong predictive ability in landslide susceptibility assessment to eliminate redundant features and reduce noises (Rodgers et al. 1988). Increasing the SRCC values indicates that the causal factor has a significant impact on the landslide model, and vice versa. In this study, we perform a correlation matrix based on SRCC, allowing us to quantify and detect multicollinearity in order to reduce it and optimize the results. The SRCC is defined as Eq. 1:

$$\text{SRCC}X,Y = \frac{\text{cov}(R(X), R(Y))}{\sigma R(X)\sigma R(Y)} = \frac{E((R(X) - \mu R(X))((R(Y) - \mu R(Y))}{\sigma R(X)\sigma R(Y)}, \tag{1}$$

where $\text{cov}(R(X), R(Y))$ is the covariance of the two variables, $the\sigma R(X) and \sigma R(Y)$, the standard deviations of the two variables, whereas $\mu R(X)$ and $\mu R(Y)$ are the mean values of the two variables. The absolute value of SRCC ranges from 0 to 1, whereas 0 corresponds to a weak linear correlation (complete absence of multicollinearity (presence of a problematic multicollinearity) and 1 corresponds to a strong linear correlation (complete absence of multicollinearity).
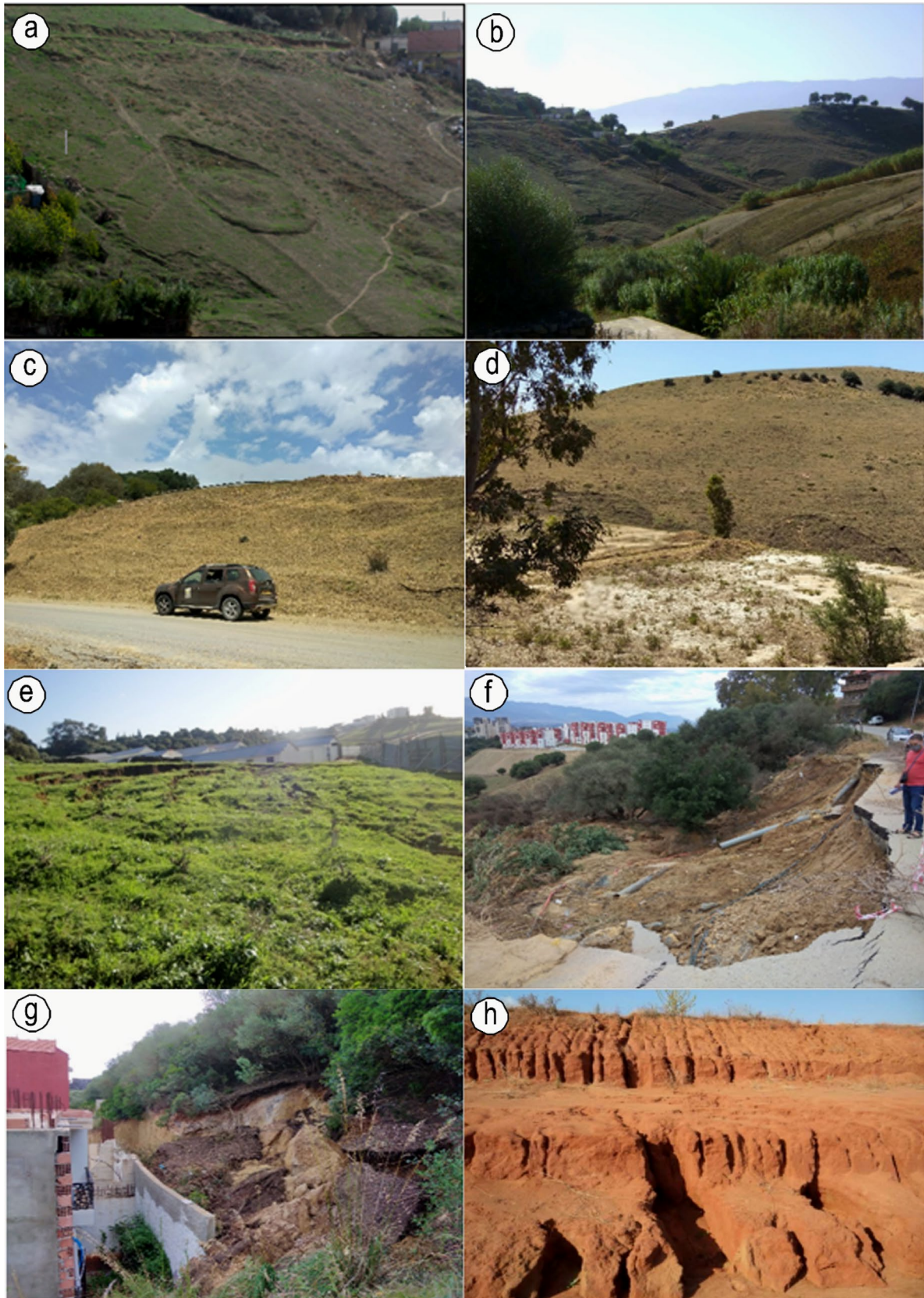
◄**Fig. 4** Types of observed landslides in the study area include: **a**, **b** landslides in the marly slopes at the south-west of Mahelma; **c**, **d** landslides in the marly slopes that caused damage to roads in the south of Rahmania; **e** Rupture of a sandy-clayey slope following a recent landslide at Sidi Bennour; **f** Intense ravine in a sandy-clayey slope, western Mahelma entrance

Landslide susceptibility modelling and mapping

The landslide susceptibility assessment in the study area was carried out using statistical and machine learning models such as FR, SI, WoE, and LR with the help of GIS techniques for the generation of LSMs. These models are usually used in geosciences, particularly in landslide susceptibility and hazard assessment, and are based on a statistical correlation between the landslide repartition and causal factors. The correlation is described by the equations of the mentioned models to determine the weighting factor values (Landslide Susceptibility Index, LSI) of each factor in order to produce the final LSMs. The resultant maps were categorized by dividing the weight value by two (LSI), mainly into five separate classes: very high, high, moderate, low, and very low susceptibility. Various methods have been developed in the literature for ranking weight values into susceptibility classes, including the equal interval method standard, the deviation method, and the natural break method.

Frequency ratio (FR) model

The FR model (Lee and Min 2001) analyses the spatial probability of landslide occurrence based on the relationship between the distribution of mass movements and their landslide conditioning factors. It expresses the relationship between the landslides in the class of landslide factors and the area in the class. The FR is defined as the ratio between the percentage of landslides in a given class and the percentage of the area in the same class:

$$\text{FR} = \left(\frac{\text{Landslide pixel class}}{\text{Area pixel class}} \right) = \left(\frac{\% \text{ of landslide area}}{\% \text{ of total area}} \right) \quad (2)$$

where FR is the frequency ratio, *L* spix is a landslide pixel in a factor class. *A* pix is the total pixel area of the class in the study area. *A* value of the FR ratio greater than 1 denotes a high correlation, whereas a value of the FR ratio less than 1 denotes a weaker correlation.

After calculating the FR for each factor using Microsoft Excel under GIS, the FR value for each class of each factor was attributed by the joint in the ArcGIS tool. Then, by using the Spatial Analysis Search Tool, the weighting landslide factors were rasterized. Afterwards, the

Landslide Susceptibility Index (LSI) is calculated by summation the frequency ratio of all factors as specified in Eq. (3):

$$\text{LSIFR} = \sum_{i=1}^{n} \text{FRij} \quad (3)$$

where LSI is the landslide susceptibility index, FR is the frequency ratio of each class i of factor j.

Following the calculation of the LSI, the index values were ranked into different landslide susceptibility classes in order to establish the final LSM using the standard deviation method in the ArcGIS tool.

Statistical index (SI) model

The statistical index method is a bivariate statistical analysis proposed by Van Westen (1997) based on a statistical relationship between the distribution of landslide areas and the predisposing factors. A weight value SI for each categorical factor is defined as the natural logarithm of the landslide density in the categorical class divided by the landslide density in the total area of the factor, as shown in Eq. (4) (Van Westen 1997):

$$\text{SIij} = In\left(\frac{Densclass}{Densmap} \right) = In\left(\frac{Lij}{L}\right) \quad (4)$$

where SIij is the weight of a class *i* of factor *j*; *In* is the natural logarithm used to consider the variation of the weights; *Lij* is the landslide density in the class *i* of factor *j* and *L* is the landslide density in the entire map of factor. When the SI is < 0.1, there is a low relationship between landslide and factors indicating a low probability of landslide occurrence. When the SI is > 0.1, this implies a high probability of landslide occurrence because the correlation between the factor and the landslide occurrences is high.

After the calculation and rasterization of the weighted SI for each class of each factor using Microsoft Excel and the GIS tool, the landslide susceptibility index (LSI) of the study area is calculated as in Eq. (5):

$$LSI_{SI} = \sum_{i=1}^{n} \text{SIij} \quad (5)$$

where LSI is the landslide susceptibility index, SI is the weighted SI of class *i* of factor *j*.

Where LSI and SI represent the landslide susceptibility index and the statistical index for each factor, respectively. A higher value of LSI, defines the higher probability of landslide occurrence. Following the calculation of the LSI, the index values were divided into different landslide susceptibility classes in order to establish the final LSM in the ArcGIS tool using the standard deviation method.
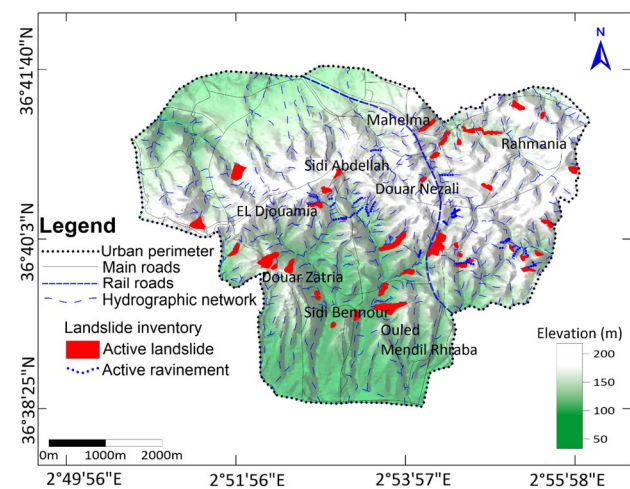
Fig. 5 The detailed landslide inventory map of the Sidi Abdellah zone

Weights of evidence (WoE) model

The WoE is a bivariate statistical method (Bonham-Carter et al. 1989) based on the log-linear form of the Bayesian probability model to estimate the posterior and prior probability (*P*) of landslide occurrence. The WoE model evaluates the spatial correlation between the landslide distribution (*L*) and the predisposing factors (*B*) within the area, based on the absence or presence of landslides (*L*) in the classes of a factor and in the form of positive (*W+*) and negative (*W−*) weights as follows (Bonham-Carter et al. 1989):

$$W^+ = In \frac{P(B/L)}{P(B/\overline{L})} \tag{6}$$

$$W^- = In \frac{P(\overline{B}/L)}{P(\overline{B}/\overline{L})} \tag{7}$$

where ln is the natural logarithm (logit) used in order to estimate the conditional probability of landslide occurrence. *P* is the probability of the ratio, *B* is the predictive factor, and *L* is the landslide. The overbar sign "‾" represents the absence of the class and/or landslide or predictive factor. The weights *W+* and *W−* weights represent the negative and positive relationships between the occurrence of landslides and the presence of landslide predisposing factors, respectively.

The weight contrast WC represents the difference between the positive and negative weights for each class of each parameter analyzed:

$$WC = W^+ - W^- \tag{8}$$

The contrast magnitude WC indicates the overall spatial relationship between the landslides and predicted variables.

The WC value is generally between 0 and 2. When the WC value tends to zero, the presence of the parameter under consideration does not affect the occurrence of landslides in the area; whereas, when WC is close to two or more, the relationship is important.

In this work, after estimation and rasterization of the WC for each class of factors through the ArcGIS tool, the landslide susceptibility index (LSI) is calculated using the lookup tool in the spatial analysis as in Eq. (8):

$$LSI_{WC} = \sum_{i=1}^{n} WCij \tag{9}$$

where WC and LSI represent the contrast magnitude and the landslide susceptibility index for class *i* of factor *j* respectively.

After calculation of the LSI, the index values were hierarchized into different susceptibility classes to generate the final LSM using the standard deviation method in the Arc-GIS tool.

Binary logistic regression (LR) model

Logistic regression is one of the leading and most popular machine learning algorithms emerging in the field of statistics. The LR model, also known as the generalized linear model, is the most commonly used machine learning model in LSM due to its simple model design. Since 2000, there has been an increase in the number of research projects that apply LR in LSM, which coincides with the increased availability of high-resolution DEMs, developments in GIS platforms, and improved processing capacity.

LR is a specific sort of generalized linear model designed to produce a binary form of outcome. It is a parametric model that is often used to anticipate answers to classification problems using the concept of probability. The ability to identify an adequate fitting function to represent the non-linear relationship between the absence or presence of landslides and a collection of landslide conditioning factor data with essentially no "hyper-parameters" to tune in makes LR perfect for creating baseline models in predictive analysis. The ability to identify an adequate fitting function to represent the non-linear relationship between the absence or presence of landslides and a collection of landslide conditioning factor data with essentially no "hyper-parameters" to tune in makes LR perfect for creating baseline models in predictive analysis.

Moreover, Lani (2007) highlighted that in order to obtain precise LSM using an LR model, the following hypotheses must be satisfied: (i) The dependent variable must have a binary value; (ii) The number of duplicates in the input dataset should be maintained to a minimum; (iii) There should be little or no multicollinearity between the conditioning factors;
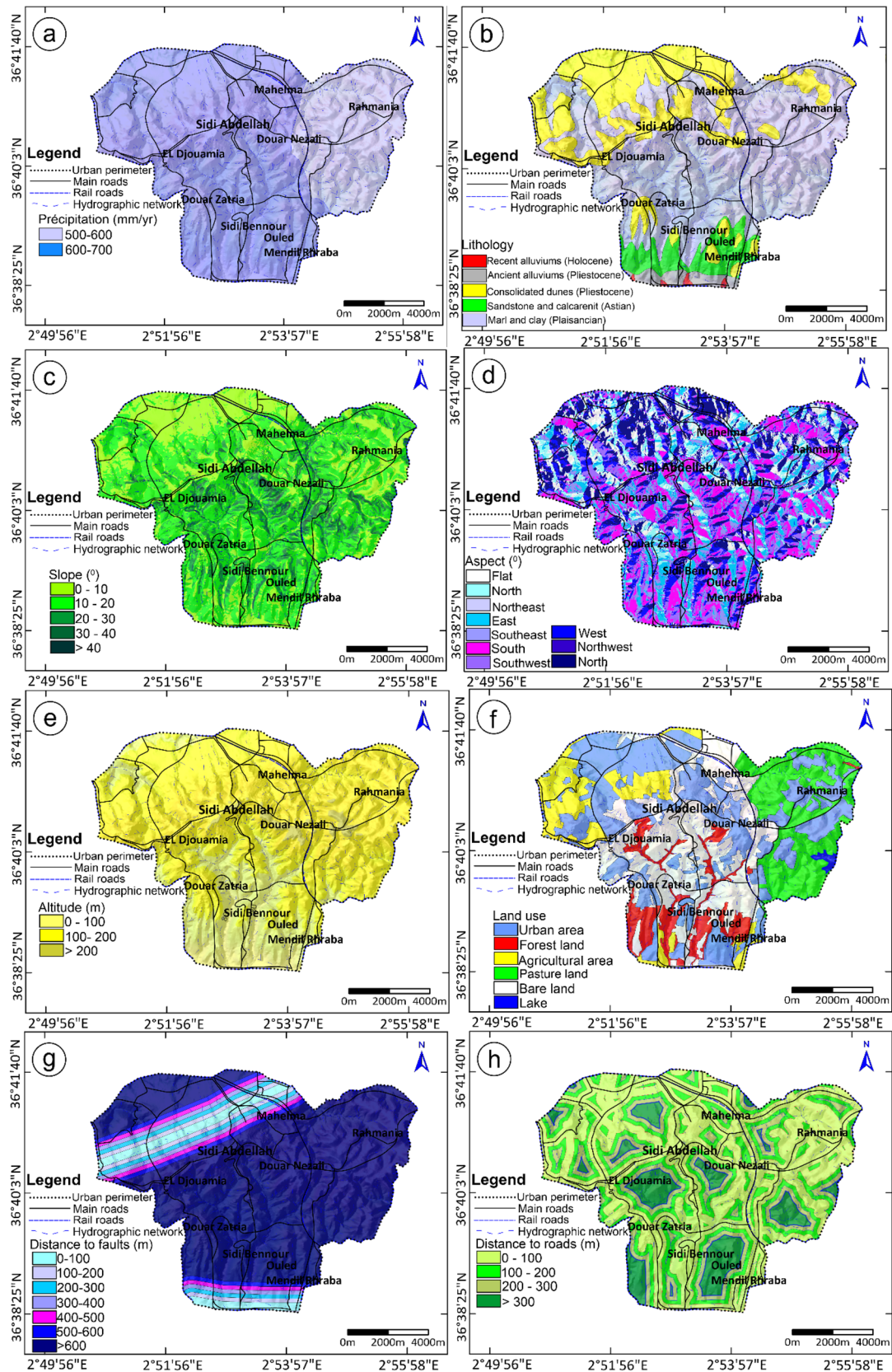
Fig. 6 Landslide predisposing factors in the Sidi Abdallah: **a** precipitation map; **b** lithological map; **c** slope angle map; **d** aspect map; **e** altitude map **f** Land use map, **g** Distance to rivers map; **h** Distance to faults; **i** Distance to roads map; **j** Stream density map; **k** Roads density map; and **l** Acceleration map
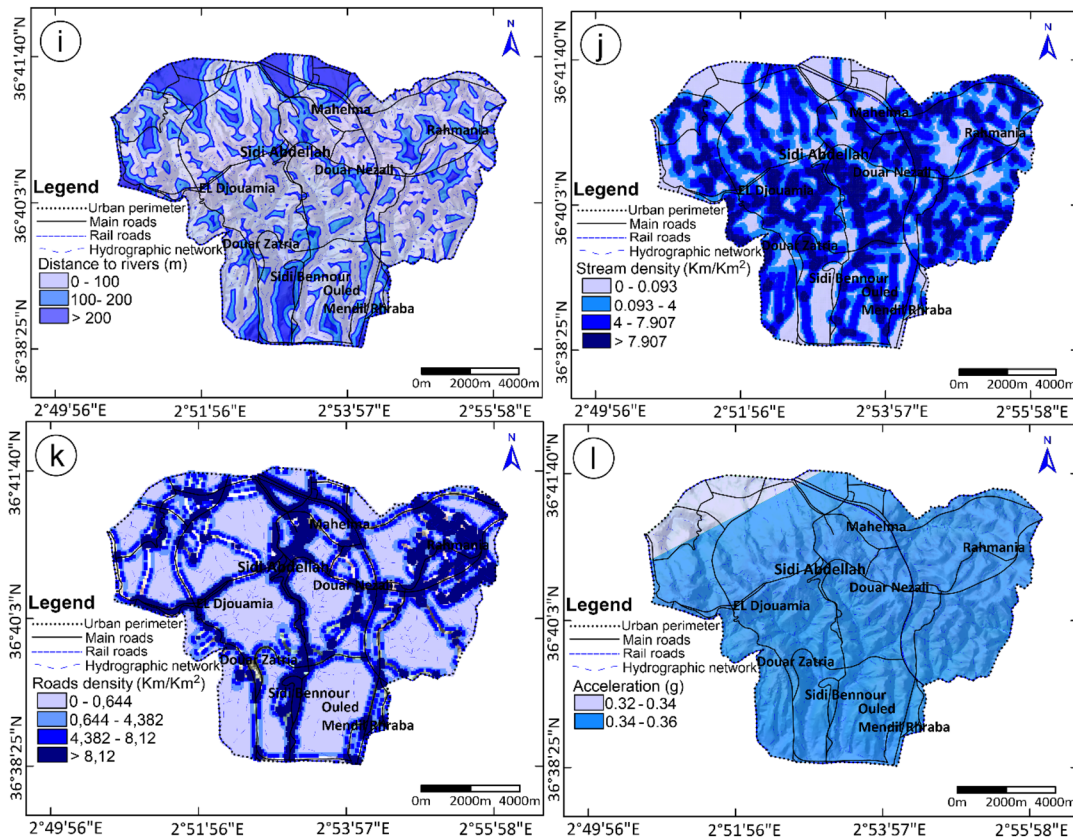
Fig. 6 (continued)

(iv) The conditioning factors and odds log should be in linear form; and (v) There should be a large sample size available.

The LR allows for the evaluation of a multivariate regression correlation between a dependent (landslides) and an independent (landslide causative factors) variable (Lee and Pradhan 2007). In the LR model, the dependent variable is a binary variable that represents the absence (0) or presence (1) of a landslide; however, the independent variables can be continuous, discrete, dichotomous, or any combination of these. The LR model evaluates the probability (*P*) of landslide occurrence as a nonlinear dependency between the landslide occurrence (dependent) and the causative factor (independent) variables as expressed in Eq. (10) (Lee and Pradhan 2007) as follows:

$$P_r = \left(\frac{1}{1 + e^{-z}} \right), \tag{10}$$

where *Pr* is the probability of landslide occurrence, within the range (0 to 1) on an *S*-shaped curve; *z* represents the linear combination described by Eq. (10) which value ranges from—∞ to +∞:

$$Z = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_n X_n \tag{11}$$

where b_0 is the intercept of the LR model, b_2, b_3, and b_n, are the slope coefficients and, X_1, X_2, X_3, and X_n are the independent variables of the logistic regression model.

Model validation

The receiver operating characteristics (ROC) and statistical rules for spatially effective LSMs are the most available, valuable, and helpful approaches utilized in the literature to define the performance and quality of LSMs. The accuracy of the produced LSMs and the validity of the models were determined by comparing known landslide data with the LSMs (Chung and Fabbri, 2003; Yesilnacar and Topal 2005). The performance or accuracy as well as the validation process of the models and the produced LSMs were evaluated by comparing the known landslide location data with the obtained LSMs.

Validation of LSMs using ROC curve

The ROC prediction curve is one of the statistical methods that can be used to predict performance and accuracy as well as compare different models (Yesilnacar and Topal 2005).

Fig. 7 Density and percentage of landslides in each factor class: **a** Precipitation, **b** Lithology, **c** Land use, **d** Aspect, **e** Altitude, **f** Slope angle, **g** Distance to faults, **h** Distance to roads, **i** Distance to rivers, **j** Stream density, **k** Roads density, and **l** Acceleration (**g**)
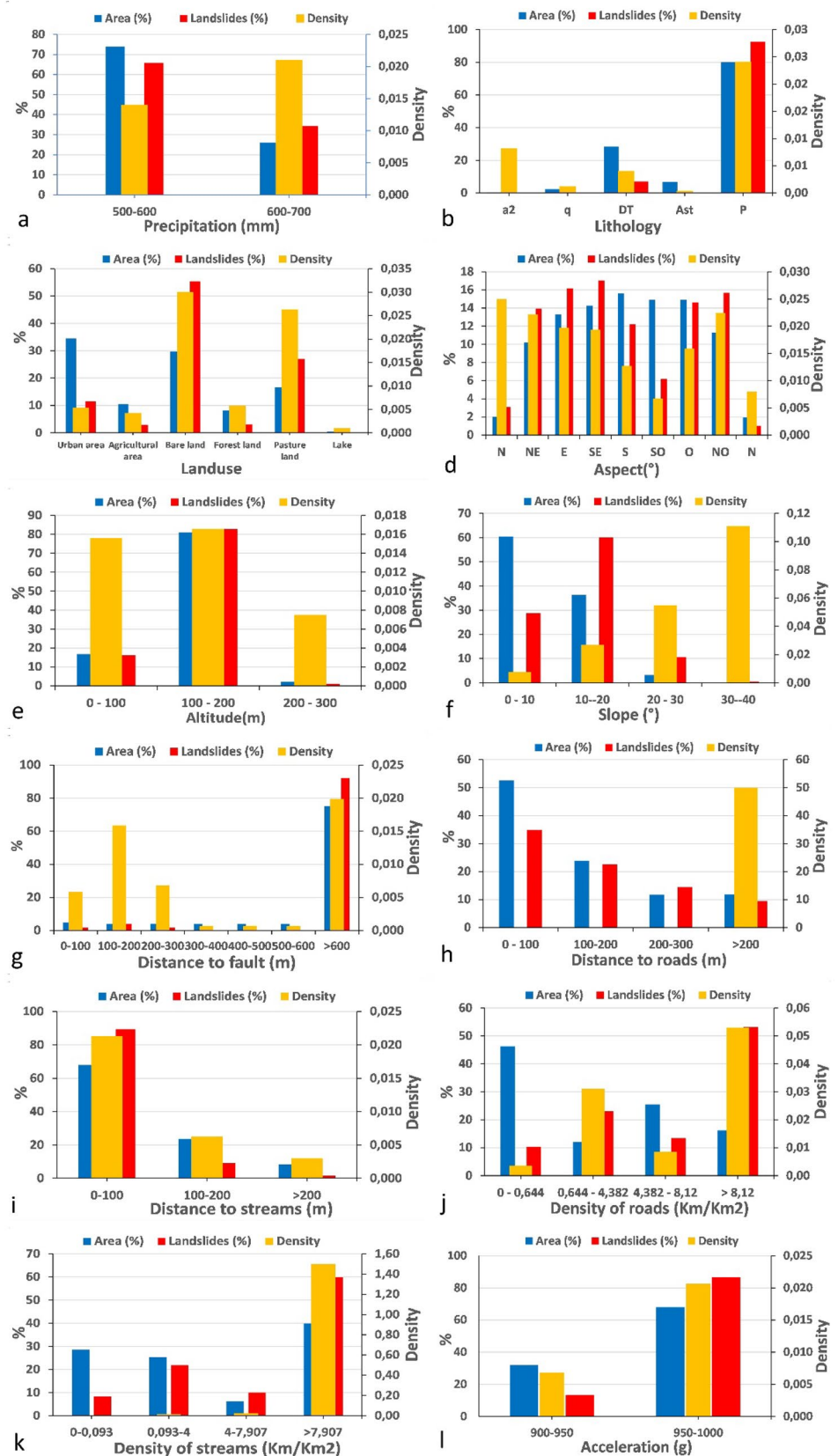
The ROC curve is a graph based on the "1 − specificity" as the x-axis and "sensitivity" as the y-axis. These statistical parameters of the ROC curve can be estimated using the following formulae (Sahana et al. 2020):

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{12}$$

$$\text{Speicificity} = \frac{TP}{TP + FN} \tag{13}$$

$$\text{AUC} = \frac{\sum TP + \sum TN}{P + N} \tag{14}$$

The value of area under the curve (AUC) is used to assess a forecast system's efficiency by defining the system's ability to accurately predict the non-occurrence or occurrence of a landslide (Chung and Fabbri 2003; Yesilnacar and Topal 2005). The AUC value of the ROC and associated performance model can be rated as follows (Yesilnacar and Topal 2005): 0.5–0.6 (poor performance), 0.6–0.7 (moderate performance), 0.7–0.8 (good performance), 0.8–0.9 (very good performance), and 0.9–1 (excellent performance).

Validation of LSMs using statistical rules

The accuracy of the LSMs can also be evaluated using two statistical rules for spatially effective LSMs (Pradhan and Lee 2010a) as follows: (i) percentage of landslides increased with the degree of susceptibility, where the smaller number of landslides were distributed in the low and very low susceptibility classes, and the higher number of landslides were distributed in the high susceptibility class of the LSMs, and (ii) the high susceptibility class should cover only small areas.

Results

Significant landslide conditioning factors assessment

Using training data, the correlation between the 12 predisposing factors is performed based on the SRCC. The correlation matrix (Table 2) depicts the results of the average SRCC values, which indicated the predictive capability of landslide predisposing factors. These results suggest that the distance to drainage and the precipitation (SRCC = 0.80), the distance to drainage and the elevation (SRCC = 0.70), and the distance to drainage and the slope (SRCC = 0.53), were highly correlated, with the absolute SRCC values greater than 0.5. This indicates that the paired landslide influence factors may have included redundant data. To investigate whether such highly correlated landslide influence factors would affect the performance of the landslide susceptibility assessment models, these paired influences were first removed separately and then simultaneously.

Landslide susceptibility mapping

LSM generated by the FR model

Using the training data, in the study area, the frequency ratios of each class of predictive factor were calculated using Eq. (2), and the results are indicated in Table 3. Then the FR of each factor class was summed to obtain the landslide susceptibility index (LSI) using Eq. (3) in the GIS environment, which varied from 2.444 to 19.080. Due to the normal distribution of LSI values, the standard deviation method is appropriate and is used for dividing the weight values of LSI into classes. In this case, the mean value of LSI is used to determine the classes and permits us to divide the result into

Table 2 Correlation matrix of the independent landslide conditioning factors

|	PRE	SLO	ELE	DD	ASP	LAN	DS	DR	LIT	DF	ACC	DOR
Precipitation (PRE)	1											
Slope (SLO)	0.04	1										
Elevation (ELE)	−0.02	0.04	1									
Distance to drainage (DD)	0.80	0.53	0.70	1								
Aspect (ASP)	−0.01	0.02	0.00	0.02	1							
Landuse (LAN)	0.18	0.13	0.05	0.18	−0.03	1						
Density of streams (DS)	0.11	−0.16	−0.1	−0.80	−0.07	−0.20	1					
Distance to road (DR)	−0.04	0.08	−0.01	0.15	−0.01	0.03	−0.18	1				
Lithology (LIT)	0.34	0.23	0.02	−0.32	−0.01	0.30	0.32	0.08	1			
Distance to fault (DF)	0.31	0.16	−0.04	0.14	−0.02	0.22	0.14	−0.04	0.37	1		
Acceleration (ACC)	0.41	0.21	−0.14	0.21	−0.12	0.33	0.21	0.02	0.37	0.48	1	
Density of roads (DOR)	−0.07	−0.10	−0.02	−0.13	0.01	−0.02	−0.13	−0.50	−0.05	0.04	−0.05	1

Table 3 Statistical relationship between each factor and the landslides, with detail values of FR, SI, WoE, and LR calculated for each class of the selected landslide causal factors

Factors	Classes	% of totale area (a)	% of landslide area (b)	Density	Frequency ratio Fr (b/a)	Statistical index (Si)	Weights (W+)	Weights (W−)	Weight contrast WC	Logistic regression coefficients
Lithology	Recent alluviums (Holocene)	0.341	0.172	0.008	0.505	−0.683	−0.691	0.002	−0.692	1.04272
	Ancient alluviums (Pliestocene)	2.292	0.172	0.001	0.075	−2.589	−2.604	0.022	−2.626	
	Consolidated dunes (Pliestocene)	28.433	7.057	0.004	0.248	−1.394	−1.406	0.266	−1.672	
	Sandstone and calcarenit (Astian)	6.695	0.172	0.000	0.026	−3.661	−3.677	0.069	−3.746	
	Marl and clay (Plaisancian)	80.000	92.427	0.024	1.485	0.395	0.403	−1.620	2.023	
Slope (°)	0–10	60.481	28.744	0.008	0.475	−0.744	−0.753	0.603	−1.355	0.19523
	10–20	36.292	60.069	0.027	1.655	0.504	0.515	−0.473	0.988	
	20–30	3.152	10.671	0.055	3.386	1.220	1.260	−0.082	1.342	
	30–40	0.075	0.516	0.111	6.850	1.924	10.470	−0.005	10.465	
Aspect	North (0–22.5)	2.010	3.098	0.025	1.541	0.433	0.442	−0.011	0.453	1.28285
	Northeast (22.5–67.5)	10.199	13.941	0.022	1.367	0.313	0.319	−0.043	0.362	
	East (67.5–112.5)	13.287	16.179	0.020	1.218	0.197	0.201	−0.035	0.235	
	Southeast (112.5–157.5)	14.267	17.040	0.019	1.194	0.178	0.181	−0.033	0.214	
	South (157.5–202.5)	15.604	12.220	0.013	0.783	−0.244	−0.248	0.040	−0.288	
	Southwest (202.5–247.5)	14.900	6.196	0.007	0.416	−0.877	−0.989	0.118	−1.107	
	West (247.5–292.5)	14.900	14.630	0.016	0.982	−0.018	−0.019	0.003	−0.022	
	Northwest (292.5–337.5)	11.304	15.663	0.022	1.386	0.326	0.332	−0.051	0.384	
	North (337.5—360)	1.943	1.033	0.008	0.531	−0.632	−0.640	0.009	−0.649	
Land use	Urban area	34.477	11.532	0.005	0.334	−1.095	−1.106	0.306	−1.412	0.89775
	Agricultural area	10.511	2.926	0.004	0.278	−1.279	−1.291	0.083	−1.373	
	Bare land	29.809	55.422	0.030	1.859	0.620	−3.471	−0.460	−3.011	
	Forest land	8.166	3.098	0.006	0.379	−0.969	−5.085	0.055	−5.139	
	Pasture land	16.640	27.022	0.026	1.623	0.485	0.489	−0.133	0.621	
	Lake	0.396	0.000	0.001	0	0.000	0.000	0.002	−0.002	
Precipitation (mm)	500–600	73.887	65.749	0.014	0.891	−0.117	−0.119	0.276	−0.395	−0.65013
	600–700	26.113	34.251	0.021	1.309	0.271	0.276	−0.119	0.395	
Distance to streams (m)	0–100	67.974	89.329	0.021	1.314	0.273	0.278	−1.110	1.388	0.65039
	100–200	23.678	9.122	0.006	0.385	−0.954	−0.964	0.178	−1.142	
	>200	8.348	1.549	0.003	0.186	−1.684	−1.698	0.073	−1.770	
Distance to roads (m)	0—100	34.820	34.820	0.014	0.663	−0.134	−0.136	0.132	−0.269	−0.020136
	100–200	22.553	22.553	0.021	0.943	0.236	0.240	−0.089	0.329	
	200–300	14.427	14.427	0.018	1.235	0.087	0.088	−0.012	0.100	
	>200	9.437	9.437	50.000	0.797	−0.072	−0.074	0.009	−0.083	

Table 3 (continued)

Factors	Classes	% of totale area (a)	% of landslide area (b)	Density	Frequency ratio Fr (b/a)	Statistical index (Si)	Weights (W⁺)	Weights (W⁻)	Weight contrast WC	Logistic regression coefficients
Distance to faults (m)	0–100	4.796	1.721	0.006	0.359	−1.025	−1.035	0.032	−1.068	1.41045
	100–200	4.045	3.959	0.016	0.979	−0.022	−0.022	0.001	−0.023	
	200–300	4.096	1.721	0.007	0.420	−0.867	−0.876	0.025	−0.901	
	300–400	4.004	0.172	0.001	0.043	−3.147	−3.162	0.040	−3.202	
	400–500	3.970	0.172	0.001	0.043	−3.138	−3.154	0.039	−3.194	
	500–600	4.006	0.172	0.001	0.043	−3.147	−3.163	0.040	−3.203	
	>600	75.082	92.083	0.020	1.226	0.204	0.208	−1.158	1.365	
Acceleration (g)	0.32–0.34	31.928	13.425	0.007	0.420	−0.866	−0.876	0.245	−1.121	− 0.65013
	0.34–0.36	68.072	86.575	0.021	1.272	0.240	0.245	−0.876	1.121	
Altitude	0–100	16.804	16.179	0.016	0.963	−0.038	−0.039	0.008	−0.046	0.32911
	100–200	80.962	82.788	0.017	1.023	0.022	0.023	−0.102	0.125	
	200–300	2.234	1.033	0.008	0.462	−0.771	−0.780	0.012	−0.793	
Roads density (Km²/Km)	0–0.644	46.273	10.327	0.004	0.223	−1.500	−1.513	0.523	−2.036	−0.041099
	0.644–4.382	12.025	23.064	0.031	1.918	0.651	0.667	−0.136	0.803	
	4.382–8.12	25.498	13.425	0.009	0.527	−0.641	−0.649	0.153	−0.802	
Stream density (Km²/Km)	0–0.093	28.525	8.262	0.005	0.411	−1.576	−1.251	0.254	−1.505	− 0.61289
	0.093–4	25.253	21.859	0.014	0.326	−1.521	−0.147	0.045	−0.192	
	4–7.907	6.265	9.983	0.026	0.242	0.466	0.476	−0.041	0.517	
	>7.907	39.958	59.897	1.499	0.187	0.405	0.413	−0.409	0.822	
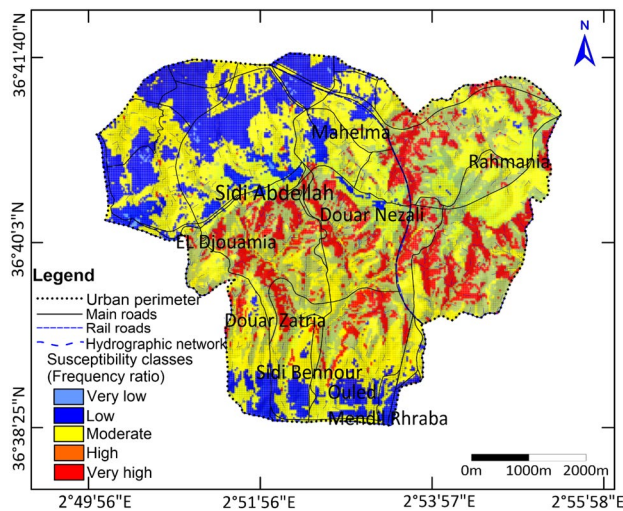
Fig. 8 Landslide susceptibility map obtained using FR model

five classes by subtracting or adding one standard deviation at a time. The five susceptibility classes with the LSI values are as follows: very low susceptibility (2.444 to 5.345), low susceptibility (5.345 to 7.790), moderate susceptibility (7.790 to 11), high susceptibility (11 to 12.8), and very high susceptibility (12.8 to 19.08). Figure 8 displays the obtained landslide susceptibility map.

Table 3 shows that the slope angle is highly related to the forces involved. The 30–40° slope class has the highest value of FR (6,850). The lithological characteristics of the urban area represent a significant factor in landslide occurrence. The Plaisancian marl and clay units were found to be more susceptible and exhibited higher frequency ratio values (1,485). The north direction is the most exposed to landslides, and the landslide occurrence increased in degraded vegetation areas (FR value of 1,859 in bare land). Other influencing factors with a high probability of sliding occurrence are the distance to rivers (distances between 0 and 100 m) and the road proximity (high values for distances between 200 and 300 m), as well as the proximity to the fault. The higher FR values were also distributed in higher precipitation zones. This shows that landslide susceptibility increases with the quantity of precipitation. For road density, FR values indicate high values for moderate densities. However, for river density, the high values of FR are observed at low densities. Table 3 shows that higher FR values were distributed in areas of higher acceleration.

LSM generated by the SI model

The landslide density and the statistical index (SI) values of each parameter class calculated by using formula (4) of the SI model are presented in Table 3. Afterwards, all the results of the weighted values of all the layers were combined to

calculate the landslide susceptibility index (LSI) according to Eq. (5).

The results obtained from the SI models are very comparable to those obtained from the FR models. The most important active factors have the strongest correlation with landslides. A slope angle of $> 30°$ is the most susceptible area to landslides. Concerning lithology, the more susceptible classes are the Plaisancian marl and clay formations. In terms of landuse, bare soil yields a high value of SI and is more susceptible to landslides. For the aspect parameter, the N, NE, and NW directions are the most susceptible to sliding. Landslides are becoming more common near rivers, faults, and roads. The earthquake and rainfall constitute the main triggering factors. Landslides were more susceptible in classes with high acceleration and precipitation than in others.

The LSI was ranked into the following five distinct susceptibility classes using the standard deviation method, where the mean value of the LSI was considered as the reference point, and then the—and + standard deviations of the distribution were used as the limits of the classes, which is valid for normal distributions (Fig. 9): very low (-13.017 to -9.869), low (-9.869 to -6.096), moderate (-6.096 to -1), high (-1 to 1.6), and very high susceptibility (1.6 to 4.98).

LSM generated by the WoE model

Thematic maps of the major landslide-causing factors were superimposed, integrated with the landslide inventory map, and analyzed. We computed the weight contrast WC and the weights of evidence (WoE) probability values using Eqs. (6), (7), and (8) (Table 3). The chi-square values for each factor were calculated at the 95% significance
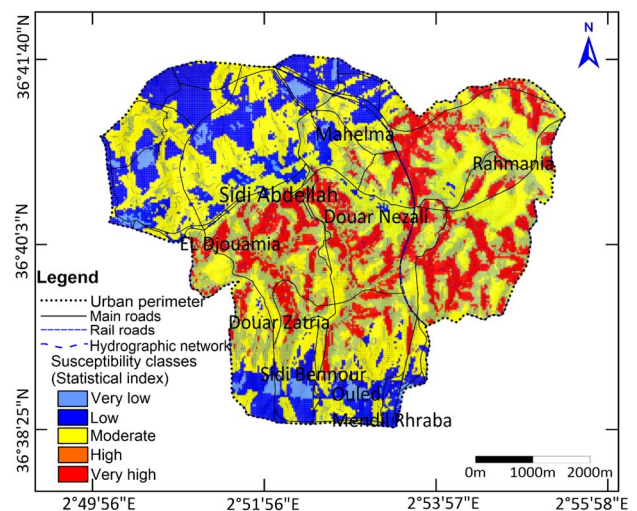
Fig. 9 Landslide susceptibility map obtained using SI model

level and one degree of freedom in order to evaluate the conditional independence between all pairs of binary patterns. Table 3 shows the implication of each factor in the occurrence of a landslide. The contrast C is positive for favorable factors for the landslide occurrence and negative for unfavorable factors for the landslide occurrence.

The results of the WoE susceptibility models show close similarity to the FR and SI models. They show that the most susceptible classes correspond to slope angles greater than 30°, a north slope aspect, Plaisancian geological formations, high precipitation, and high acceleration, which proves the good correlation with landslide occurrence. Finally, we assigned weights to each thematic layer's classes in order to generate weighted thematic maps. The thematic maps were then overlapped and numerically summed according to Eq. (15) in order to generate a landslide susceptibility index (LSI):

$$
\begin{aligned}
\text{LSI}_{WC} = \; & \text{WC Slope} + \text{WC Aspect} + \text{WC Altitude} \\
& + \text{WC Fault} + \text{WC River} + \text{WC Precip} \\
& + \text{WC Landuse} + \text{WC Road} + \text{WC Road Density} \\
& + \text{WC River Density} + \text{WC Acceleration}.
\end{aligned}
$$
(15)

Where LSI is the landslide susceptibility index and WC is the weight contrast of each landslide factor class.

Consequently, the landslide susceptibility map of Sidi Abdellah city (Fig. 10) is prepared from the respective LSI values showing the following five categories of landslide susceptibility: very low (−17.043 to −12.902), low (−12.902 to −7.482), moderate (−7.482 to 1), high (1 to 5.5) and very high (5.5 to 19.02).
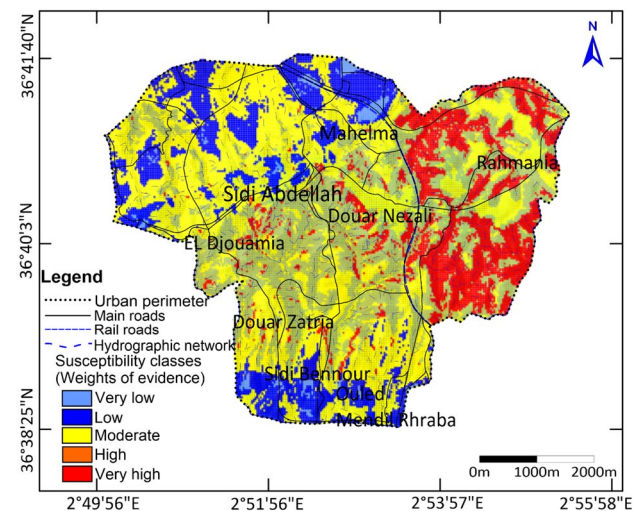
Fig. 10 Landslide susceptibility map obtained using WoE model

LSM generated by the LR model

Twelve conditioning factors and the presence or absence of landslides were converted into grid format and then into Excel data format files for use in the statistical package Real Statistics. The logistic regression model was run based on the percentage area of landslides in each factor to obtain the logistic regression coefficients. The coefficient of logistic regression for each controlling factor is presented in Table 3. The Hosmer and Lameshow tests exhibited the accepted and reliable fit of the equation because the significance of the chi-square is greater than 0.05 (17.420). A higher R-square value of Cox (0.75), Snell R^2 (1), and Nagelkerke R^2 (1) showed a better model. The ROC (Relative Operating Characteristic) value of 0.803 indicates a good correlation between the independent and dependent variables.

Finally, the binary logistic regression model and their respective coefficients are given in the following Eq. (16):

$$
\begin{aligned}
z - \; & 9,69727 - 0,22610 * \text{Precipitation} - 0,65013 \\
& * \text{Acceleration} + 1,04272 * \text{Lithology} \\
& + 0,32911 * \text{Altitude} + 0,19523 * \text{Slope} \\
& + 1,28285 * \text{Aspect} + 1,41045 * \text{Distance to Fault} \\
& + 0,89775 * \text{Land use} + 0,65039 * \text{Distance to streams} \\
& - 0,020136 * \text{Distance to Roads} \\
& - 0,041099 * \text{Density of Roads} \\
& - 0,61289 * \text{Density of streams}
\end{aligned}
$$
(16)

According to Eq. (16), precipitation, acceleration, distance to roads, and density of streams indicate a negative relationship with the landslide occurrence in the study region. However, lithology, altitude, slope, land use, distance to streams, aspect, and distance to fault are positively related to the occurrence of a landslide. In addition, the 'distance to fault' parameter is the most effective in landslide occurrence.

The prediction of landslide occurrence in the study area was evaluated using the above logistic regression coefficients and according to Eq. (10). The probability ranges from $5.324*10^{-12}$ to 1. Based on the cumulative percentage of the observed slide occurrences against the probability index values, the subsequent LSM was obtained. The probability map has been divided into five susceptibility classes using the standard deviation method (Fig. 11) as follows: very low (1.474E−05–2.696E−05), low (2.696E−05–6.35E−05), moderate (6.35E−05–6.59E−05), high (6.59E−05–6.642E−05) and very high (6.642E−05–6.698E−05). Based on the standard deviation method, the LSMs were classified into
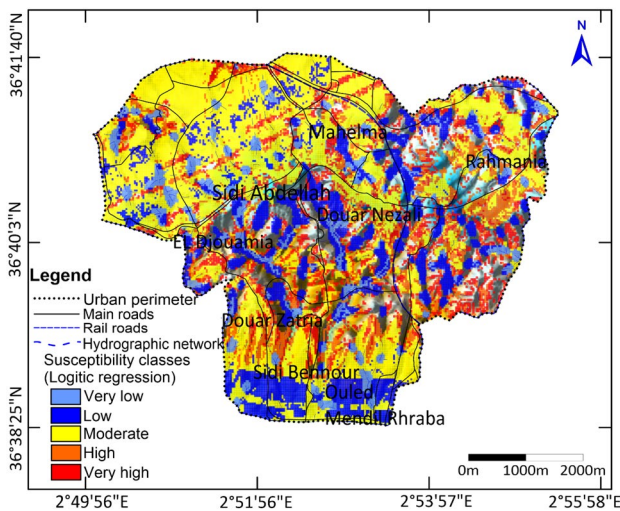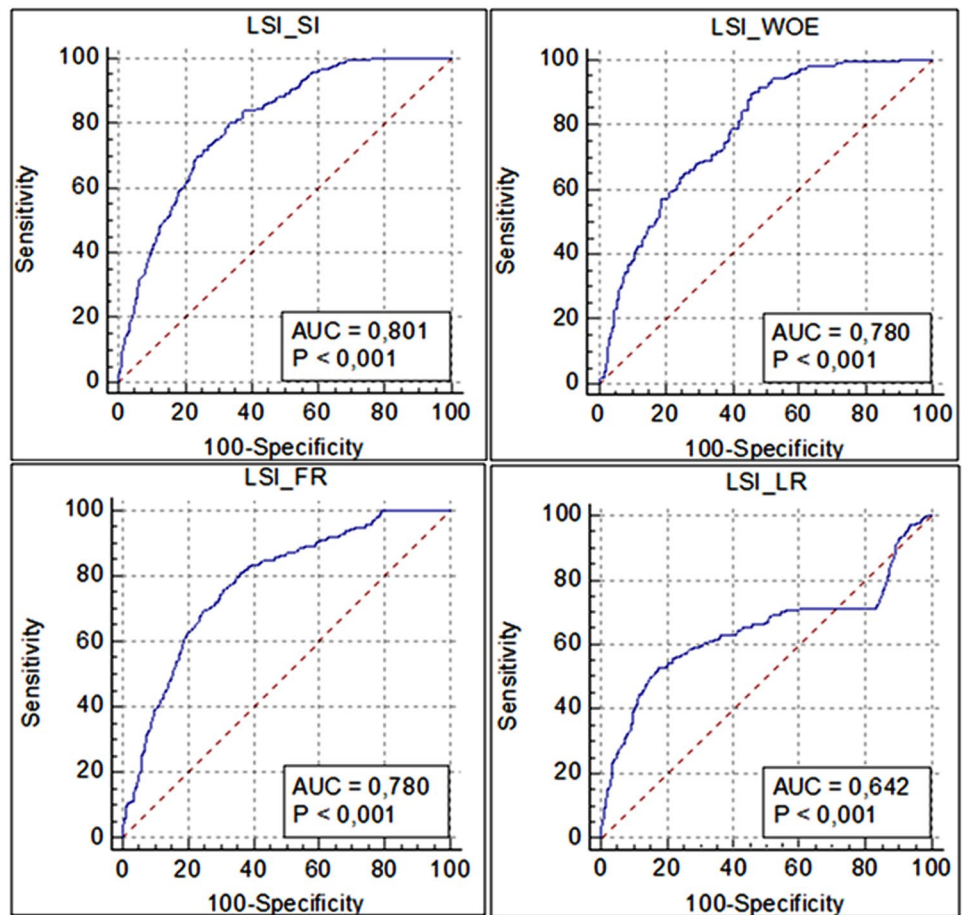
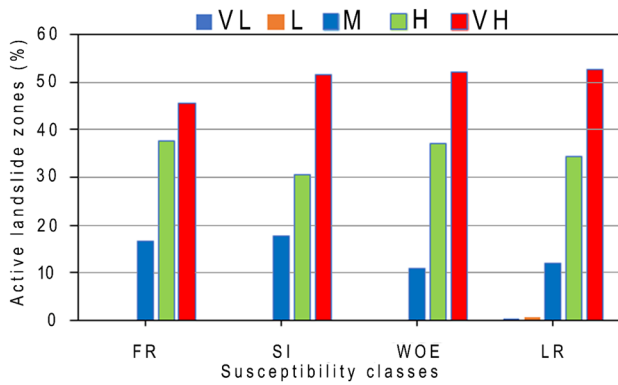**Fig. 11** Landslide susceptibility map obtained using LR model

five classes because the data values obtained in the LSI indicated a normal distribution.

## Validation and comparison of the LSMs

To validate the performance of the four used models, the landslide area was randomly divided into the following two categories: 30% for model validation and 70% for training, taking into account spatial allocation and using the random division technique. Comparing the landslide training pixels (30%) with the four LSMs yielded ROC curves in this study, and the area under the curves was calculated using Eqs. (12), (13), and (14). The AUC of the ROC curves is presented in Fig. 12 for the four applied models. The validation results showed that the SI model has a higher accuracy of prediction (AUC = 80.1%) than WoE (AUC = 78.2%), FR (AUC = 8%) and LR (AUC = 64.2%). According to these results, the used models presented almost good accuracy in predicting the landslide susceptibility, except for the LR model, which shows a moderate accuracy of prediction. The results also offered a theoretical framework for the use of statistical methods (e.g., SI) in landslide prevention, mitigation, and urban planning so as to provide an adequate response to the increasing demand for effective and low-cost tools in landslide susceptibility assessments.

Fig. 12 ROC curves and AUC for accuracy prediction of the used models (FR, SI, WoE and LR)

Fig. 13 Percentage of active landslide zones in each susceptibility classes
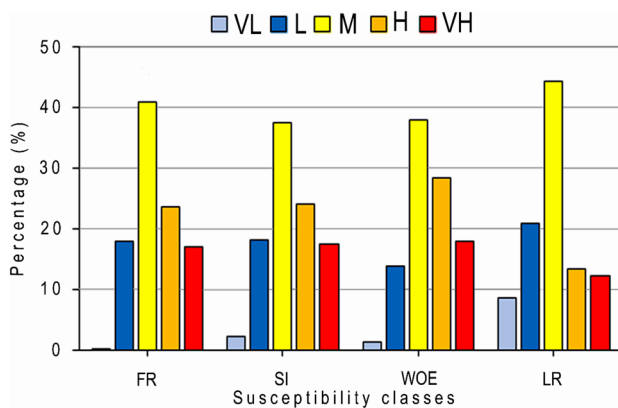
Fig. 14 The comparative distribution of different susceptibility classes in different landslide susceptibility maps

For spatially effective LSMs, the obtained LSMs were also tested and confirmed using two rules (Pradhan and Lee 2010a). Figure 13 demonstrates that all LSMs' high and very high susceptibility classes contain 82 to 88% of the active landslide zones, while the moderate zones contain 10 to 16% of the active landslide zones. Less than 1% of the active landslide zones in all LSMs are in the low and very low susceptibility zones. The statistics in Fig. 13 show that the percentages of landslides increase from classes with very low to very high susceptibility, with the high susceptibility class comprising only small areas. The acquired results showed clearly that the four commonly used statistical and machine learning models are very suitable for landslide susceptibility mapping.

A comparison between the obtained susceptibility maps using FR, SI, WoE, and LR models was performed. As shown in Fig. 14, the areas of the landslide susceptibility zones, ranging from very low to very high, were determined for each classifier. The LSM produced with the LR method contains 9% and 21% of the total area, which are designated

to be of very low and low susceptibility, respectively. The area that falls into the categories of moderate, high, and extremely high susceptibility is divided into 44%, 13%, and 12%, respectively. The LSM obtained by using the FR model, which included 0.3% of the total area, is classified as having very low landslide susceptibility. Areas with low to very high vulnerability make up about 18%, 41%, 23%, and 17% of the entire area, respectively. The LSM produced by applying the SI shows that a significant majority of the susceptibility zones are in the high and very high levels, with 24 and 17%, respectively, while the percentages of moderate, very low, and low susceptibility areas are 37%, 18%, and 2%, respectively. The LSM created with the WoE model, which involved 1% of the total urban area, is identified as having very low landslide susceptibility. The low and moderate susceptibility classes take up 14% and 38% of the total area, respectively. The high and very high zone values are close to 28% and 18%, respectively (Fig. 14).

Discussion

Landslides are the most recurrent and progressive geological hazards in many districts of Sidi Abdelah city, where the geomorphological, climatic, seismotectonic, and anthropogenic factors are most favorable. They pose severe threats to the planning and development of urban areas, which necessitates conducting local landslide susceptibility mapping for effective risk management and long-term development planning in landslide-prone areas. However, providing a reliable spatial prediction of landslides is regarded as one of the most challenging aspects of landslide hazard and risk assessment. Despite the variety of methods used, their accuracy is still debated. In this study, we tackled this issue by assessing and comparing the performance of the four statistical and machine-learning models, including the SI, WoE, FR, and LR, for the spatial prediction of landslide susceptibility in the urban areas of Sidi-Abdelah.

The crucial point in landslide susceptibility modeling is the evaluation of the correlation between historical landslide events and causal factors based on the main concept that landslide occurrences in the past and present determine their occurrence in the future (Guzzetti et al. 1999). A landslide inventory map and many causative factors should be collected in the first step. The statistical and machine learning models used are then tested and validated using the two training and testing datasets. The resulting models are then used to calculate the probability of a landslide's occurrence.

The FR and SI models are simple and easy to apply, and the obtained results are comprehensible. The results indicate the FR and SI models are very comparable and satisfactory. It was found that, the highest values are observed in the classes of marl and clay, at (30–40°) of slope, on bare

land, in (600–700 mm) of precipitation at (0,34–0,36 g) of acceleration, at (0–100 m) of distance to streams, and at (100–200 m) of altitude (Table 3).

The WoE approach has numerous advantages over other statistical methods because it is data-driven and primarily employs the Bayesian probability model. It is a quantitative (data-driven) technique used to combine datasets. The combination of elements that are associated with landslides suggests that the causal factors are conditionally independent of the landslides.

The LR approach is an alternative and appropriate theory that was used for the prediction of landslide susceptibility in the study area. The goal is to forecast the likelihood of a dichotomous event based on a set of variables that can be discrete, continuous, or both in combination. The regression coefficients and model statistics are frequently used in the model to calculate the accuracy and qualified importance of the causal factors. The obtained results from the LR model indicate that distance to faults, aspect, lithology, distance to roads, land use, and slope are the most important factors. The landslide susceptibility analysis performed for a relatively large area (urban area) necessitates a large number of landslide pixels. Therefore, the accuracy of the LR model can be improved if additional landslide data is included in the analysis.

In general, the FR, SI, WoE, and LR models produced significantly better results than the other models in terms of overall classification accuracy. However, the FR model has more regularity in terms of negative and positive predictive values (Table 3). The choice of conditioning factors, which is an important aspect influencing the quality of landslide susceptibility models, is carried out based on the analysis of the landslide types and the characteristics of the study area. As a result, lithology, land use, and stream density are considered the most important factors that influence the landslide occurrences in this study. This seems reasonable because most of the landslides in the study area occurred on bare land in altered Plaisancian marls and near streams.

Our results from the used models show good agreement with other case studies throughout the world. Yalcin (2008) showed that the FR model gave a more accurate representation of landslide susceptibility than the LR. Yilmaz (2009) specified that the prediction accuracy of LSMs produced by the FR and LR was 82.60% and 84.20%, respectively. According to Pradhan and Lee (2010b), the accuracy observed for the FR, LR regression, and ANN models was 86.41, 89.59, and 83.55 percent, respectively. Pradhan and Youssef (2010) showed that the accuracy of the FR model (89.25%) was better in predicting landslides than that of the LR model (85.73%). Tien Bui et al. (2011) indicated almost equal predicting accuracy for the Si (94,6%) and the LR (95,0%) models. Park et al. (2013) showed generally similar overall accuracies of 65.27% and 65.51% for the FR and

LR models, respectively. According to Mohammady et al. (2012), the FR and WoE models have an accuracy of 80.1% and 74.6%, respectively. For mapping landslide susceptibility in the Sultan Mountains in NE Turkey (Ozdemir and Altural 2012), the accurate predictions of the FR, LR, and WoE models are 97.6%, 95.2%, and 93.7%, respectively. The findings of mapping landslide susceptibility at Zonouz Plain (Iran) obtained by Nourani et al. (2014) showed that the prediction accuracy of LSMs, produced by the FR and LR, was 87.57% and 89.42%, respectively. According to Bourenane et al. (2016), the FR approach is more accurate (86.59%) than the WoE (82.38%), Wf (77.58%), and LR (70.45%) methods. Regmi et al. (2014) stated that the FR and WoE models have a success rate of 76.8% and 75.6%, respectively. The results of validated landslide susceptibility prediction in the Setif Region (NE Algeria) specified that the FR model provides a more accurate prediction (86%) than the LR (84%) and WoE (79%) models, according to Karim et al. (2019).

The obtained results may serve as a first tool for landslide risk prevention and land use planning in the city of Sidi Abdellah, ensuring long-term development in this area. The results validated previous observations, demonstrating the effectiveness of statistical approaches. The following mitigating strategies may be suggested based on the derived LSMS to reduce the consequences of current landslides and to control urban growth: (i) restricting development in landslide-prone areas, based on the LSM_S, (ii) controlling, by means of codes and urban rules, human activity (i.e., excavation, construction, vegetation clearance, cutting slopes, landscaping, etc.) in the landslide-prone areas; (iii) protecting existing developments and the population through physical mitigation measures (drainage, counterfort, and protective barriers); (iv) developing and implementing monitoring and warning systems; (v) prohibit all new construction in the high landslide susceptibility area, and (vi) protecting the existing developments by physical mitigation measures (such as drainage, down counterfort berms) that serve as buttresses, and protective barriers, river bed widening and cleaning to allow easy drainage of flood waters during heavy rains.

Conclusion

Landslides constitute a serious constraint to the development and urban planning of the new city of Sidi Abdellah (Northern Algeria). Consequently, landslide susceptibility mapping is seen as a vital undertaking that can help authorities reduce landslide disaster losses by serving as a guideline for durable landuse planning, such as the restriction of urban extension in hazardous zones. In this framework, many scientific methods have been developed and applied for landslide

susceptibility and hazard mapping. The bivariate and multivariate statistical approaches are considered most suited and more objective for large-scale landslide susceptibility and hazard mapping, which integrates GIS techniques and remote sensing for spatial data acquisition, processing, management, and analysis in order to assess and predict landslide susceptible areas.

In the present work, we applied and compared the worldwide developed statistical (represented by FR, SI, and WoE) and machine learning models methods (represented by LR) to investigate the LSMs performance of frequently used data-based models and and generate more satisfactory landslide susceptibility mapping. The first step consisted of identifying landslide locations based on satellite image analysis supported by field survey investigations. The unstable urban perimeter covers a total area of approximately 522.9 hectares, which represents about 1.62% of the urban area. Then, twelve landslide causative factors, including slope, altitude, aspect, land use, lithology, precipitation, seism, distance to a drainage network, distance to a fault, distance to a road network, the density of roads, and the density of streams, have been derived from high-resolution satellite images, geologic maps, DEM, and precipitation data. The LSMs were produced using each of the four methods and classified into the following five classes: very low, low, medium, high, and very high. Subsequently, the results have been validated using the receiver operating characteristic technique (ROC) by comparing the obtained susceptibility maps with known landslide event sites. According to the obtained AUC values of the ROC curve, the SI model has a higher prediction performance (80.10%) than the WoE (78.2%), FR (78.2%), and LR (64.20%) models. This signifies that the statistical models provide a high accuracy for landslide susceptibility prediction in the study region than the machine learning model, which has a moderate accuracy. Moreover, for both statistical and machine learning models the statistical rules show that landslide density increased from the low to very high susceptibility zone, with the highest percentage of landslides being observed in the high susceptibility zone. These clearly demonstrated that the four commonly used statistical and machine learning models provided satisfactory results and depicted a high accuracy level for landslide susceptibility mapping.

The established LSM$_S$ is considered a necessary tool for sustainable urban planning and development in landslide prone areas by identifying the expected landslide occurrence zones. They serve as a useful tool for planners, engineers, and decision-makers in slope management and future expansion planning in the urban area. As our results are specified on a large-scale map, the precise extent of the landslide areas and the details of high susceptibility areas are defined. This will be helpful for further site-specific studies.

The development of urbanization in landslide-prone areas can be avoided if a landslide susceptibility map is available.

Data availability The corresponding author will provide data to back up the study's findings upon a reasonable request.

Declarations

Conflict of interests The author have no competing interests to declare that are relevant to the content of this article.

Consent to participate All authors have given their consent for their research to be published in this journal.

References

ANRH, Agence Nationale des Ressources Hydrauliques (2005) Precipitation map performed from precipitation database from local meteorological stations located in the vicinity of the study area by using Kriging interpolation

Aymé A (1954) Service de la Carte Géologique de l'Algérie, Étude des terrains néogènes de la cluse du Mazafran:(Sahel d'Alger)

Bonham-Carter GF, Agterberg FP, Wright DF (1989) Weights of evidence modelling: a new approach to mapping mineral potential. In: Agterberg FP, Bonham-Carter GF (eds) Statistical applications in earth sciences. Geological Survey of Canada, Ottawa, pp 171–183

Bourenane H, Guettouche MS, Bouhadad Y, Braham M (2016) Landslide hazard mapping in the Constantine city, Northeast Algeria using frequency ratio, weights factor, logistic regression, weights of evidence, and analytical hierarchy process methods. Arab J Geosci 9:154. https://doi.org/10.1007/s12517-015-2222-8

CGS (2009) Centre National de Recherche Appliquée en Génie Parasismique. Seismic Microzonation study of the city of Sidi Abdellah. Unpublished Internal Report.

Chung CJF, Fabbri AG (2003) Validation of spatial prediction models for landslide hazard mapping. Nat Hazards 30:451–472. https://doi.org/10.1023/B:NHAZ.0000007172.62651.2b

Djerbal L, Khoudi I, Alimrina N, Melbouci B, Bahar R (2017) Assessment and mapping of earthquake-induced landslides in Tigzirt City, Algeria. Nat Hazards 87:1859–1879. https://doi.org/10.1007/s11069-017-2831-5

Goetz JN, Brenning A, Petschko H (2015) Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Comput Geosci 81:1–11. https://doi.org/10.1016/j.cageo.2015.04.007

Guirous L, Dubois L, Melbouci B (2014) Contribution à l'étude du mouvement de terrain de la ville de Tigzirt (Algérie). Bull Eng Geol Environ. https://doi.org/10.1007/s10064-014-0624-6

Guzzetti F, Carrara A, Cardinali M, Reichenbach P (1999) Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. Geomorphology 31:181–216. https://doi.org/10.1016/S0169-555X(99)00078-1

Hadji R, Boumazbeur A, Limani Y, Baghem M, Chouabi A, Demdoum A (2013) Geologic, topographic and climatic controls in landslide hazard assessment using GIS modeling: a case study of Souk Ahras region, NE Algeria. Quatern Int 302:224–237. https://doi.org/10.1016/j.quaint.2012.11.027

Hallal N, Dubois L, Bougdal R, Djouder F (2017) Instabilités gravitaires dans la région de Béjaïa (Algérie): Inventaire et appréciation de l'importance relative des différents paramètres conduisant au déclenchement, au maintien ou à l'activation des instabilités. Bull Eng Geol Environ. https://doi.org/10.1007/s10064-017-1050-3

Harbi A, Maouche S, Ayadi A, Benouar D, Panzag F, Benhallou H (2004) Seismicity and tectonic structures in the site of algiers and its surroundings: a step towards microzonation. Pure Appl Geophys 161(2004):949–967. https://doi.org/10.1007/s00024-003-2502-1

Huang F, Cao Z, Guo J, Jian SH, Li S, Guo Z (2020) Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping. CATENA 191:104580. https://doi.org/10.1016/j.catena.2020.104580

Huang J, Wu X, Ling S, Li X, Wu Y, Peng L, He Z (2022) A bibliometric and content analysis of research trends on GIS-based landslide susceptibility from 2001 to 2020. Environ Sci Pollut Res 29:86954–86993. https://doi.org/10.1007/s11356-022-23732-z

Karim Z, Hadji R, Hamed Y (2019) GIS based approaches for the landslide susceptibility prediction in Setif Region (NE Algeria). Geotech Geol Eng 37(1):359–374. https://doi.org/10.1007/s10706-018-0615-7

Koica (2003) Master Plan project of the city of Sidi Abdellah established by KOREA International Cooperation Agency (KOICA). Unpublished Internal Report

Lani J (2007) Generalized linear models. Organ. Res. Methods 7:127-150

Laribi A, Walstra J, Ougrine M, Seridi A, Dechemi N (2014) Use of digital photogrammetry for the study of unstable slopes in urban areas: Case study of the El Biar landslide. Algiers Eng Geol 187(2015):73–83. https://doi.org/10.1016/j.enggeo.2014.12.018

Lee S, Min K (2001) Statistical analysis of landslide susceptibility at Yongin, Korea. Eng Geol 40:1095–1113. https://doi.org/10.1007/s002540100310

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

Meghraoui M (1988) Géologie des zones sismiques du Nord de l'Algérie : Paleosismologie, tectonique active et synthèse sismotectonique. Thèse de Doctorat, spécialité Sciences Naturelles, mention Géologie Structurale, Université de Paris Sud, Centre Orsay N°.3495 Paris le 23 juin 1988.

Merghadia A, Yunusb AP, Douc JD, Whiteleye J, Thai Phamg B, Tien Buihn B, Avtari R, Boumezbeur A (2020) Machine learning methods for landslide susceptibility studies : A comparative overview of algorithm performance. Earth-Sci Rev. https://doi.org/10.1016/j.earscirev.2020.103225

Mohammady M, Pourghasemi HR, Pradhan B (2012) Landslide susceptibility mapping at Golestan Province Iran: a comparison between frequency ratio, Dempster-Shafer, and weights-of-evidence models. J Asian Earth Sci 61:221–236. https://doi.org/10.1016/j.jseaes.2012.10.005

Moulouel H, Bouchelouh A, Bensalem R, Tebbouche MY, Ait Benamar D, Oubaiche E, Gharbi S, Machane D, Benamghar A (2020) The Mahelma fault: a secondary structure of the Sahel anticline? Arab J Geosci 13:715. https://doi.org/10.1007/s12517-020-05694-z

Nourani V, Pradhan B, Ghaffari H, Sharifi SS (2014) Landslide susceptibility mapping at Zonouz Plain, Iran using genetic programming and comparison with frequency ratio, logistic regression, and artificial neural network models. Nat Hazards 71:523–547. https://doi.org/10.1007/s11069-013-0932-3

Ozdemir A, Altural T (2012) A comparative study of frequency ratio, weights of evidence and logistic regression methods for landslide susceptibility mapping: Sultan Mountains SW Turkey. J Asian Earth Sci 2:2. https://doi.org/10.1016/j.jseaes.2012.12.014

Park S, Choi C, Kim B (2013) Landslide susceptibility mapping using frequency ratio, analytic hierarchy process, logistic regression, and artificial neural network methods at the Inje area. Korea Environ Earth Sci 68:1443–1464. https://doi.org/10.1007/s12665-012-1842-5

Pourghasemi HR, Moradi HR, Fatemi Aghda SM (2013) Landslide susceptibility mapping by binary logistic regression, analytical hierarchy process, and statistical index models and assessment of their performances. Nat Hazards 69:749–779. https://doi.org/10.1007/s11069-013-0728-5

Pradhan B, Lee S (2010a) Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. Environ Model Softw 25:747–759. https://doi.org/10.1016/j.envsoft.2009.10.016

Pradhan B, Lee S (2010b) Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia. Landslides 7:13–30. https://doi.org/10.1007/s10346-009-0183-2

Pradhan B, Youssef A (2010) Manifestation of remote sensing data and GIS on landslide hazard analysis using spatial-based statistical models. Arab J Geosci 3:319–326. https://doi.org/10.1007/s12517-009-0089-2

Regmi AD, Yoshida K, Pradhan B, Pourghasemi HR, Khumamoto T, Akgun A (2014) Application of frequency ratio, statistical index and weights-of-evidence models, and their comparison in landslide susceptibility mapping in Central Nepal Himalaya. Arab J Geosci 7(2):725–742. https://doi.org/10.1007/s12517-012-0807-z

Rodgers JL, Nicewander WA (1988) Thirteen ways to look at the correlation coefficient. Am Stat 42(1):59–66

Sahana M, Pham BT, Shukla M, Costache R, Thu D, Chakrabortty R, Satyam N, Nguyen HD, Phong TV, Le HV, Pal SC, Areendran G, Imdad K, Prakash I (2020) Rainfall induced landslide susceptibility mapping using novel hybrid soft computing methods based on multi-layer perceptron neural network classifier. Geocarto Int 37:2747–2771. https://doi.org/10.1080/10106049.2020.1837262

Thiery Y, Malet JP, Sterlacchini S, Puissant A, Maquaire O (2007) Landslide susceptibility assessment by bivariate methods at large scales: application to a complex mountainous environment. Geomorphology 92(1–2):38–59. https://doi.org/10.1016/j.geomorph.2007.02.020

Tien Bui D, Lofman O, Revhaug I, Dick O (2011) Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. Nat Hazards 59:1413–1444. https://doi.org/10.1007/s11069-011-9844-2

Varnes DJ (1978) Slope movement, types and processes. In: Schuster RL, Krizek RJ (Eds) Landslides, analyses and control. National Academy of Science, Report 176, Washington, DC, pp 11–35.

Van Westen CJ (1997) Statistical landslide hazard analysis. ILWIS 2.1 for Windows application guide. ITC publication, Enschede, The Netherlands, pp 73–84.

Xiao T, Yin K, Yao T, Yin K (2019) Spatial prediction of landslide susceptibility using GIS-based statistical and machine learning models in Wanzhou County, Three Gorges Reservoir, China. Acta Geochim 38:654–669. https://doi.org/10.1007/s11631-019-00341-1

Yalcin A (2008) GIS-based landslide susceptibility mapping using analytical hierarchy process and bivariate statistics in Ardesen (Turkey): comparisons of results and confirmations. CATENA 72(1):1–12. https://doi.org/10.1016/j.catena.2007.01.003

Yesilnacar E, Topal T (2005) Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). Eng Geol 79(3–4):251–266. https://doi.org/10.1016/j.enggeo.2005.02.002

Yılmaz I (2009) Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: a case study from Kat landslides (Tokat-Turkey). Comput Geosci 35(6):1125–1138

Zhou G, Esaki T, Mitani Y, Xie M, Mori J (2003) Spatial probabilistic modeling of slope failure using an integrated GIS Monte Carlo simulation approach. Eng Geol 68:373–386

Zhou S, Zhang Y, Tan X, Abbas SM (2021) A comparative study of the bivariate, multivariate and machine-learning-based statistical models for landslide susceptibility mapping in a seismic-prone region in China. Arab J Geosci 14:440. https://doi.org/10.1007/s12517-021-06630-5