Bediroğlu, Ş. (2024). A Comparison of Alternative GIS Data Model Methods for Landslide Susceptibility Mapping with XGBoost and SHAP. *The Black Sea Journal of Sciences*, 14(3), 1204-1224.

The Black Sea Journal of Sciences, 14(3), 1204-1224, 2024. DOI: 10.31466/kfbd.1446997

KFBD

Karadeniz Fen Bilimleri Dergisi
The Black Sea Journal of Sciences
ISSN (Online): 2564-7377 https://dergipark.org.tr/tr/pub/kfbd

Araştırma Makalesi / Research Article

A Comparison of Alternative GIS Data Model Methods for Landslide Susceptibility Mapping with XGBoost and SHAP

Şevket BEDİROĞLU[1*]

Abstract

Geographic Information Systems and machine learning algorithms suggest good alternatives for producing landslide susceptibility maps. In the process of producing these maps with machine learning, alternative data model options exist. Success rate of analyses may change according to the preferred data method. In this study, 6 different machine learning models were created by passing different data models with the XGBoost algorithm. Study area is located in the cities of Ordu and Giresun, Turkiye. 14 different factors and related geographic data layers were used. As a result of the study, the most successful model performance was achieved by taking the average values of all pixels of the combined landslide record polygons (Accuracy=0,88, Precision=0,86, F1 score=0,87). SHAP method was applied for better interpretation of machine learning results The susceptibility map produced with the ideal model, overlapped with 57.556 buildings in the region. The buildings were classified in 4 groups (low, moderate, high, and very high) and mapped, indicating their risk level.

Keywords: Landslide susceptibility mapping, machine learning, SHAP, GIS, geospatial data model.

XGBoost ve SHAP ile Heyelan Duyarlılık Haritalaması İçin Alternatif CBS Veri Modeli Yöntemlerinin Karşılaştırılması

Öz

Coğrafi Bilgi Sistemleri ve makine öğrenimi algoritmaları, heyelan duyarlılık haritalarının üretilmesi için iyi alternatifler önermektedir. Bu haritaların makine öğrenmesi ile üretilmesi sürecinde alternatif veri modeli seçenekleri mevcuttur. Tercih edilen veri yöntemine göre analizlerin başarı oranı değişebilir. Bu çalışmada XGBoost algoritması ile farklı veri modellerini geçerek 6 farklı makine öğrenmesi modeli oluşturulmuştur. Çalışma alanı Türkiye'nin Ordu ve Giresun illerinde bulunmaktadır. 14 farklı faktör ve ilgili coğrafi veri katmanları kullanıldı. Çalışma sonucunda en başarılı model performansı, birleştirilmiş heyelan kayıt poligonlarının tüm piksellerinin ortalama değerleri alınarak elde edilmiştir. Makine öğrenmesi sonuçlarının daha iyi yorumlanması için SHAP yöntemi uygulandı. İdeal model ile üretilen duyarlılık haritası, bölgedeki 57.556 bina ile örtüştü. Binalar 4 grupta (düşük, orta, yüksek ve çok yüksek) sınıflandırılarak risk düzeyleri belirtilerek haritalanmıştır.

Anahtar Kelimeler: Heyelan duyarlılık haritası, Makine öğrenmesi, SHAP, CBS, Coğrafi veri modeli

[1]Gaziantep University, City and Regional Planning Dep., Gaziantep, Türkiye, sbediroglu@gantep.edu.tr

[*]Sorumlu Yazar/Corresponding Author
Geliş/Received: 05.03.2024 **Kabul/Accepted:** 28.05.2024 **Yayın/Published:** 15.09.2024

1. Introduction

Landslide susceptibility (LS) is the spatial distribution of the probability of the occurrence of landslides as determined by various investigators (Youssef and Pourghasemi, 2021; Constantin et al., 2011). Landslide susceptibility mapping (LSM) is a functional method to avoid and reduce loses from landslide hazards (Hong, 2023). There are many sub-methods for creating LSM, including statistical methods, traditional machine learning (ML) methods, deep learning methods, etc. (Zhao et al., 2022). Due to the increase in flood events, there is a need to implement new models to enhance the prediction capability of flood and landslide hazards (Prasad et al., 2021; Hong et al. 2018). More hybrid and different models are applied in LSM. Each combination gives excellent prediction performance for LS mapping; besides this, further exploration and application of more set-based methods are needed (Chen and Li, 2020; Hong et al., 2020; Abedini et al., 2019; Pham et al., 2019; Wu et al., 2017).

ML and deep learning techniques have been proven to be powerful and promising tools in many geotechnical applications as well as in landslide identification (Wang et al., 2020; Li et al., 2019; Ching and Phoon, 2018; Lo and Leung, 2018; Papaioannou and Straub, 2017). Integration of terrain modelling and GIS analysis provides a toolset for rapid spatial prediction of landslide hazards (Gorsevski et al., 2006). ML methods are popular for detecting landslide areas. ML methods provide support in the process of determining the parameters of where landslide events will occur (Pham et al., 2020).

The quality of training datasets has a crucial influence on the accuracy of LSM (Hong et al., 2020). The performance of LS models is dependent on the number of training samples and their quality. This is more impactful when the training data is scarce (Sameen et al., 2020). As it is difficult to reach landslide records stored in the geographic data format in many parts of the world, these scarce data must be processed correctly. This process is relevant to ML studies for getting more accurate and effective results.

Most of the previous studies focused on the model development process, which consists of adjusting hyperparameters or obtaining hybrid models created by combining several methods (Fanos and Pradhan, 2019; Pradhan and Sameen, 2018). Other researchers have studied several approaches to improve the performance of LS maps (Sameen et al., 2020; Hong et al., 2018; Hussin et al., 2016; Mezaal et al., 2017; Yilmaz, 2010; Nefeslioglu et al., 2008). LS maps were created using machine learning in various scientific studies in Turkey. Orhan et al., 2020 aimed trying different machine learning models for creating LS maps. Akinci et al, 2020 measured Random Forest performance results for same purpose. Sahin 2022, studied on free and open-source semi-automatic feature engineering tools for creating LS maps. Kavzoglu and Teke 2022, measured performances of ensemble machine learning algorithms.

This paper is structured as follows. Ideal spatial data model was investigated for generating LS maps with Geographic Information Systems (GIS) and XGBoost algorithm. Two different crossover groups were created. In the first group, the presence data of the landslides is classified into groups such as recent, relict(old), or combined. In the second group, while transferring data from GIS dataset to ML, the data was separated into two classes. Central point of the landslide or average of all the pixels in polygon of the landslide record. By crossing three different comparison sample groups, 6 different analysis results were formed. In previous studies, there were fewer studies focusing on the behaviour of LS mapping performance on different data models than in this study. The results of these analyses were evaluated in a comparative manner.

2. Material and Methods

2.1. Study Area

Study area includes Ordu and Giresun cities located in the northeast region of Türkiye. Cities are located between latitudes of 40o13’ - 41o8’ and longitudes of 36o57’ - 39o14’ (Figure 1). Ordu city is 6001 km2 and Giresun city is 6934 km2. The characteristics of these two neighbouring cities such as topography, and land structure are similar.

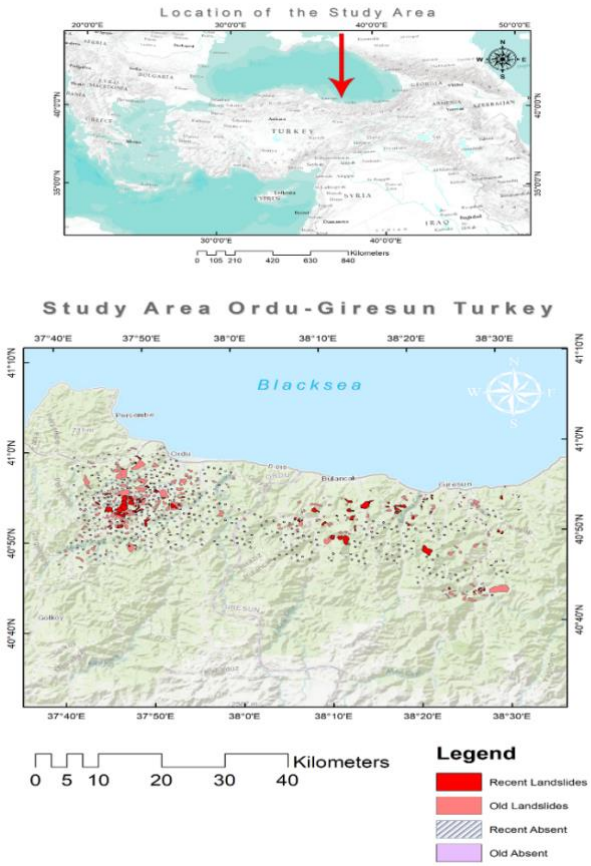

Figure 1. Study area

Ordu and Giresun are coastal cities and the mountains of this region are perpendicular to the sea. The region is rich in streams, and there are streams in all canyons. In the natural vegetation; there are spruce, pine (larice), alder, beech, carp, oak, and chestnut trees. Seedlings of hazelnuts and kiwis dominate agricultural land. According to the Köppen climate classification, the climate in the Black Sea region is classified as oceanic climate (Cfb) and subtropical rainy climate (Cfa) (Orman ve Su İşleri Bakanlığı, 2023). The area is rainy in all seasons and has a temperate climate with the thermal characteristics of the sea. The greatest amount of precipitation occurs during the fall months. The average annual precipitation is 1590 mm and the average number of rainy days is 163 days. The average annual temperature in the region is 14 oC. The region starts at sea level at an altitude of 0 and extends to an altitude of 2000. Land use and active population in the region are changeable due to the season. People live in coastal areas during the winter months and migrate to villages and highland settlements in high-altitude regions in the spring and summer months. Land use in the region is transforming rigidly due to human-made structures. Landslides occur frequently in the region. Figure 2 shows the settlement pattern in an active landslide zone and the studies carried out after the landslide in Ordu / Kabaduz. Due to floods and landslides that occurred in Giresun, 2020 or Ordu, 2023, a significant amount of life and property has been lost.

a- Post-landslide field studies from study area

b- A landslide zone from study area

Figure 2. Landslide risky areas and settlements in study area

2. 2. Methodology Schema

A methodology schema showing all steps of study is given in Figure 3. Python was preferred as programming language because it is compatible with rich ML libraries such as Scikit-Learn. Pandas library was also used. Jupyter-lab platform was used for creating and running ML. Scripts were written in Python language. Brief explanation of methods is given in the sections below. ArcMap was used at the GIS analysis and spatial data visualisation stages. AutoCAD was used for processing a small amount of CAD based data (almost all of the data was provided in GIS format).

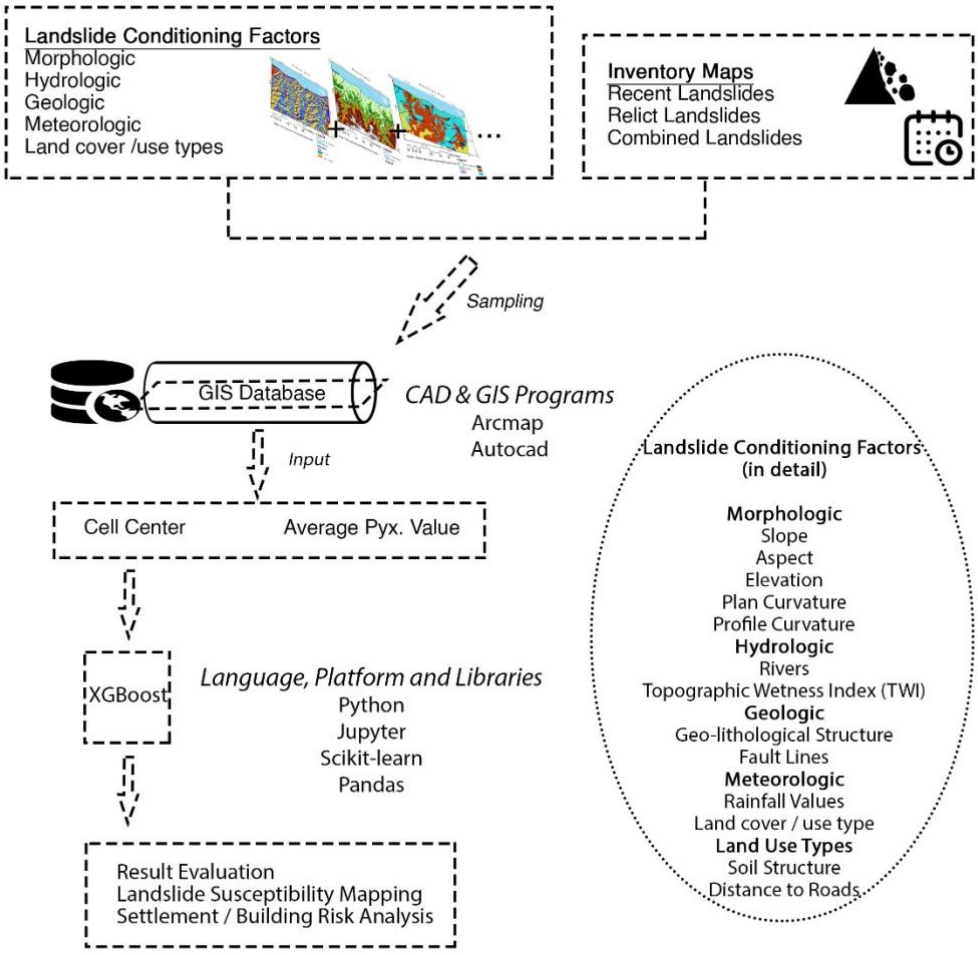

Figure 3. Methodology schema

2.3. Classification of Landslide Inventory Data

In era of GIS and integration of ML, one of the challenging problems that researchers need to solve is how the data model should be structured. In general, raw datasets in GIS cannot be directly ready for use in ML algorithms. Same problems arise in the process of producing LSM with ML. When converting GIS data into information, diversity of parameters is high, and these options will affect analysis results positively or negatively. The fact that datasets are produced according to different scales, sensitivity, accuracy, or data format (vector/raster) increases diversity. Spatiotemporal parameter of the preferred dataset is also an important criterion. Should we work with recent data, relict data or a combination of both? Also, preparation of positive / negative landslide data to be used by ML algorithms has an important effect, especially in solving classification problems. In the final stage, how to transfer information from the final datasets created in this study to the data of the factors affecting the landslide in raster format, and the choices to be made on pixel-based methods, can directly affect the results. It is aimed to state alternatives in this section and planned to compare alternative input and output choices in the following sections.

2.3.1. Presence Data

2.3.1.1. Recent Landslide Records

There are some classification types depending on whether the landslides are up-to-date or not. They vary by country with criteria such as records of landslide events, time of data collection, active-passive status, etc. In this study, records of landslide events in the last 10 years have been evaluated as "recent landslide records". 248 different "recent landslide records" in the pilot study area are organised in GIS data format. The data is collected from Türkiye's landslide public institutions producing records, MTA (Maden Tetkik Arama) and other different sources.

2.3.1.2. Relict Landslide Records

Landslide records that occurred more than 10 years ago and have not been repeated recently were evaluated as data for relict landslide records. In the area studied, 572 different landslides were collected according to this criterion.

2.3.1.3. Combined Landslide Records

Combined landslide record data is combination of recent and relict records in same data. Recent, relict landslide records were merged into same dataset (820 landslides) and this dataset. All these three datasets are shown in Figure 1.

2.3.2. Absence Data

Landslide absence data also play an important role in regional LSM based on statistical models, since they can suppress the statistical model's overestimation of the LS, thus enabling its ability to reasonably divide the area into high-susceptibility areas and low-susceptibility areas (Hong et al., 2019; Zhu et al., 2018). The reason behind creating and using absence data is to teach ML the difference between landslide and non-landslide areas. This way, the ML algorithm can define characteristics of independent variable parameters, such as LS. The perception of this difference helps ML make better decisions in the process of LSM. Absence data is important for teaching non-landslide areas to the computer and determining the parameters of the landslide event comparatively. In the study area, absence data for 248 recent landslides, 572 relict landslides, and 820 combined landslides were created randomly. It is considered important to ensure that the average sizes of

"presence landslide data" and "absence landslide data" are close to each other. Absence landslide record data takes 0 as an attribute before being used in ML.

2.4. Landslide Conditioning Factors and Related GIS Dataset

13 different factors affecting the occurrence of landslide events were determined based on different literature studies. These factors are considered in five classes: morphological, hydrological, geological, meteorological, and land cover / use types. Morphological factors (Hong et al., 2020) are; slope aspect, elevation, plan curvature, and profile curvature. Hydrological factors (Yi et al., 2020; Fang et al., 2020; Wang et al., 2019; Shirzadi et al., 2018) are distance to the river and topographic wetness index (TWI). Geological factors (Yi et al., 2020; Nsengiyumva and Valentino, 2020; Pourghasemi et al., 2020) are geo-lithological structure and distance to fault. The meteorological factor is the maximum rainfall value in m2 per month. Land cover / use type factors (Nsengiyumva and Valentino, 2020; Wang et al., 2019) are land use types, distance to roads, and soil structure. Each landslide affecting factor corresponds to a GIS data layer in the database. Table 1 shows metadata information about the collected GIS dataset, such as source of data, accuracy of data, original and final format of data, and current literature studies using related data. Graphical views of collected datasets related to these factors are given in Figure 3. In addition to these landslide conditioning factors, landslide inventory data was collected and inserted into the GIS dataset.

2.5. Cell Center or Average Value within All Pixels Overlaying Landslide Record Polygons

In the process of analysing the factors affecting landslide occurrence with ML and producing LS maps, the choice of the correct data model is important. Landslide data is in polygon format and may cover small, medium or large areas. When transferring landslide data to ML as a dependent variable, it must be transferred as a single value. Because this value will be overlapped with pixel-based dependent variables (factors affecting landslide formation) in the ML environment. While obtaining this single value; there are alternatives such as taking the value of a point in the middle of the landslide polygon (geometric center), taking a random point or taking the average values of the pixels inside the polygon.

Table 1. Detailed information about GIS dataset

Data name	Source	Original Format / Accuracy	Last Format / Parameter	Current Literature Studies
Slope	General Command of Mapping (TR)	Vector / 5 mt	Raster / Degree	Lima et al., 2023; Hong, 2023; Xiao and Zhang, 2023; Saygin et al., 2023; Laura et al., 2023; Pham et al., 2020; Pourghasemi et al., 2020; Zhao et al., 2022;
Aspect	General Command of Mapping (TR)	Raster / 5 mt	Raster / Aspect Value	Chen and Li, 2020; Fang, 2020; Hong et al., 2020; Pham et al., 2020; Pourghasemi et al., 2020; Nsengiyumva and Valentino, 2020; Zhao et al., 2022; Lima et al., 2023; Hong, 2023; Laura et al., 2023
Elevation	General Command of Mapping (TR)	Vector / 5mt	Raster / Height Value	Hong, 2023; Xiao and Zhang, 2023; Laura et al., 2023; Arabameri et al., 2021; Chen and Chen, 2021; Ngo et al., 2021; Zhao et al., 2022; Lima et al., 2023
Plan Curvature	General Command of Mapping (TR)	Raster / 5 mt	Raster / Plann Curv. Value	Chen and Li, 2020; Fang, 2020; Hong et al., 2020; Pourghasemi et al., 2020; Nsengiyumva and Valentino, 2020; Arabameri et al., 2021; Chen and Chen, 2021; Ngo et al., 2021; Lima et al., 2023; Hong, 2023; Laura et al., 2023; Huang et al., 2023
Profile Curvature	General Command of Mapping (TR)	Raster / 5 mt	Raster / Profile Curv. Value	Chang et al.,2020; Chen and Li, 2020; Fang, 2020; Arabameri et al., 2021; Chen and Chen, 2021; Lima et al., 2023; Hong, 2023; Laura et al., 2023; Huang et al., 2023
River	Government Water Org.	Vector / 4 mt	Raster / Distance to River in meters	Hong et al., 2020; Nsengiyumva and Valentino, 2020; Pourghasemi et al., 2020; Sameen et al., 2020; Wang et al., 2020a; Yi et al., 2020;Arabameri et al., 2021; Chen and Chen, 2021; Zhao et al., 2022; Hong, 2023
TWI	Meteorology	Raster / 5 mt	Raster / Index Value	Wang et al., 2020a; Wang et al., 2020b; Arabameri et al., 2021; Chen and Chen, 2021; Hong, 2023
Geo-lithologic Structure	Mineral Exploration (MTA)	Vector / 10 mt	Raster / Attribute Type	Hong et al., 2020; Nsengiyumva and Valentino, 2020; Pham et al., 2020; Pourghasemi et al., 2020; Yi et al., 2020; Arabameri et al., 2021; Chen and Chen, 2021; Zhao et al., 2022; Lima et al., 2023; Hong, 2023; Laura et al., 2023
Fault Line	Mineral Exploration (MTA)	Vector / 10 mt	Raster / Distance to Fault Line in meters	Pham et al., 2020; Pourghasemi et al., 2020; Wang et al., 2020a; Yi et al., 2020; Arabameri et al., 2021; Ngo et al., 2021; Zhao et al., 2022; Lima et al., 2023; Hong, 2023; Xiao and Zhang, 2023
Land Use Type	Ministry of Agriculture and Food	Vector / 10 mt	Raster / Land Use Type Attribute	Sameen et al., 2020; Yi et al., 2020; Arabameri et al., 2021; Chen and Chen, 2021; Ngo et al., 2021; Zhao et al., 2022; Lima et al., 2023; Hong, 2023; Xiao and Zhang, 2023; Saygin et al., 2023; Laura et al., 2023
Soil	Ministry of Agriculture and Food	Vector / 10 mt	Raster / Soil Quality Attribute	Zhang et al., 2017; Chen and Li, 2020; Fang, 2020; Pourghasemi et al., 2020; Nsengiyumva and Valentino, 2020; Arabameri et al., 2021; Chen and Chen, 2021; Lima et al., 2023; Saygin et al., 2023; Laura et al., 2023
Road	Municipality	Vector / 1 mt	Raster / Distance to Roads	Yi et al., 2020; Arabameri et al., 2021; Chen and Chen, 2021; Ngo et al., 2021; Zhao et al., 2022; Lima et al., 2023; Hong, 2023; Xiao and Zhang, 2023
Landslide Records	Mineral Exploration (MTA)	Vector / 5 mt	Raster / Attribute Recent – Relict – Combined	Wang et al., 2020a; Arabameri et al., 2021; Chen and Chen, 2021; Ngo et al., 2021; Zhao et al., 2022; Xiao and Zhang, 2023
Rainfall	Meteorology	Excel / 3 mt	Raster / After Interpolation	Zhang et al., 2017; Chen et al., 2018; Shirzadi et al., 2018; Aghlmand et al., 2020; Chen and Li, 2020; Fang, 2020; Wang et al., 2020b; Arabameri et al., 2021; Ngo et al., 2021; Hong, 2023; Xiao and Zhang, 2023
Building Footprints	Municipality	Vector / 30 cm	Vector / Polygon	Singh et al., 2021; Fu et al., 2020; Ciampalini et al., 2014; Martha et al., 2013

2.6. Cell Center or Average Value within All Pixels Overlaying Landslide Record Polygons

In the process of analysing the factors affecting landslide occurrence with ML and producing LS maps, the choice of the correct data model is important. Landslide data is in polygon format and may cover small, medium or large areas. When transferring landslide data to ML as a dependent variable, it must be transferred as a single value. Because this value will be overlapped with pixel-based dependent variables (factors affecting landslide formation) in the ML environment. While obtaining this single value; there are alternatives such as taking the value of a point in the middle of the landslide polygon (geometric center), taking a random point or taking the average values of the pixels inside the polygon.

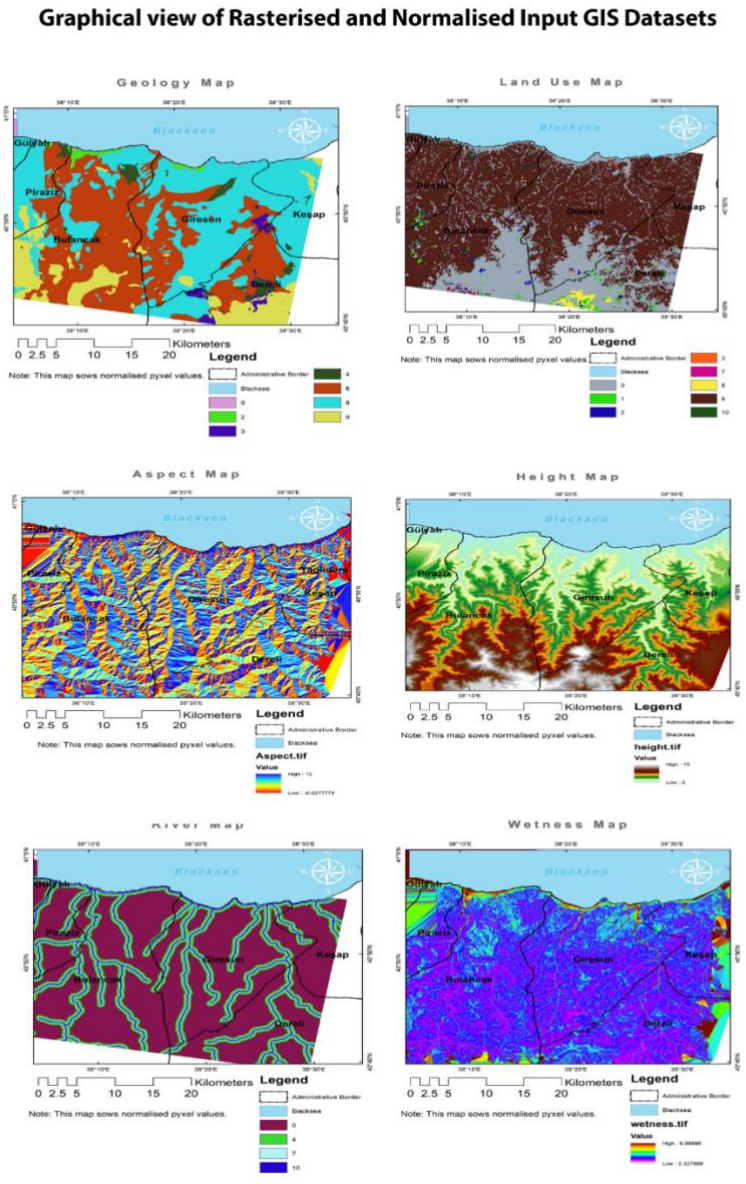

Figure 4. Graphical view of input GIS dataset

In this study, the geometric center of the landslide was taken as an alternative to represent the landslide areas. Alternatively, the mean value of all pixels in overlapping the relevant level with the landslide area was taken (Figure 5). The data obtained from calculations made with these two different methods and all the input layers have been prepared for ML. Presence and absence data of Recent, Relict and a were arranged and made available in 3 different ML algorithms to be tested comparatively.

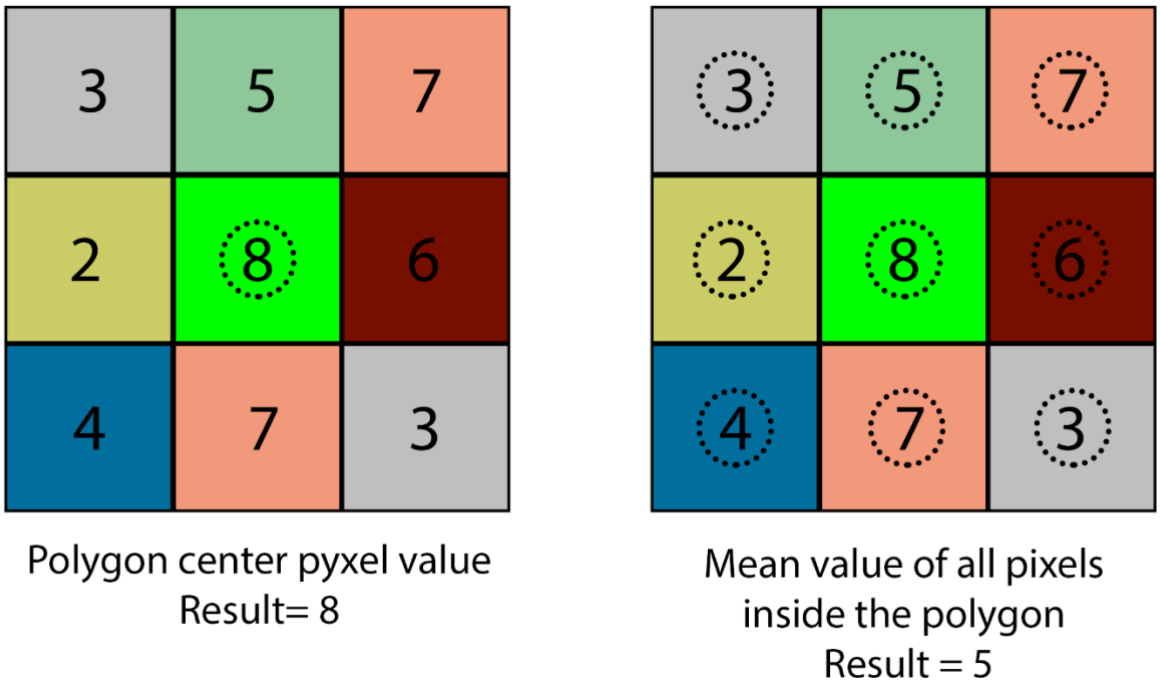

Figure 5. Differences between cell center and average value methods

2.7. Gradient Boosting and XGBoost

Gradient Boosting Decision Tree (GBDT) is a sub-group of decision forests that includes models like XGBoost, CatBoost, and LightGBM (Sagi and Rokach, 2021). XGBoost, or eXtreme Gradient Boosting, stands as a seminal advancement in machine learning, renowned for its exceptional predictive performance and versatility across diverse domains. Among the machine learning methods used in practice, gradient tree boosting is one technique that shines in many applications. Tree boosting has been shown to give state-of-the-art results on many standard classification benchmarks (Li., 2010; Chen and Guestrin, 2016). XGBoost was mainly designed for speed and performance using gradient-boosted decision trees. It represents a way for machine boosting, or in other words applying boosting to machines, initially done by Tianqi Chen (Dhaliwal

et al., 2018). Gradient tree boosting is also known as gradient boosting machine (GBM) or gradient boosted regression tree (GBRT) problems.

2.8. Building GIS Based Machine Learning Model

ML model is shown at methodology section. ML learning model of this study has 6 main alternatives. These alternatives occur while crossing A) landslide inventory data type (3 alternatives) B) cell center or average pixel value data, the method used to transfer information inside the input landslide polygon to ML (2 alternatives). So, crossing these alternatives (3*2) generates 6 different alternatives.

Normalising the values of landslide conditioning factors using the min–max method helps getting better results (Sameen et al., 2020). In this method, the largest and smallest values in a group of data are handled. All other data is normalized to these values. Calculation formula of normalisation min-max method is given below.

$$x' = \frac{x - xmin}{xmax - xmin} \quad (1)$$

During ML analyses normalized independent variables (X values) were used. Training / test split method was applied for the validation process (70% train, 30% test). In addition, the k-fold cross validation method was applied as k = 10 and compared with the training / test split method. Since there is no significant difference in terms of performance, the training / test split method, which is a practical method, was preferred in all following analyses. Primitive models were defined directly with independent variables. After performing the error tests and performance analyses, the model tuning phase was started. At this stage, the most appropriate parameters were found with the GridSearchCV method by assigning values at different intervals to the parameter "C", and the analyses were performed with that parameter.

3. Results and Discussion

3.1. Model Performance Comparison of Alternatives

This paper aims producing LSM with high prediction accuracy. Assessment of prediction accuracy and performance of the models there are alternative methods (Pourghasemi et al., 2020; Rahmati et al., 2017). Cross-validation results can be produced quantitatively and graphically by means of Accuracy, Precision, F1 Score or confusion matrix. According to the analysis methodology applied in the study, performance analyses were carried out under 2 different classes. A) landslide inventory data type B) cell center or average pixel value data. For this purpose, XGBoost tuned model

result success scores were compared. Table 2 shows model performances and success scores such as Accuracy, Precision and F1 score.

When the model performance is evaluated on the basis of the landslide date, the findings are as follows. In all combinations of methods and landslide analysis models, the analyses performed with "Combined Landslide Data" gave the highest performance. The best performance in terms of Accuracy, Precision and F1Score criteria was in the analyses performed with “Combined Landslide Data”. Analyses made with "Recent Landslide Data" ranked second in terms of performance. The lowest performance is the analysis with "Relict Landslide Records". The interpretation of this result assumes that two different reasons may cause this situation. First, changes in the topographical structure and, on the other hand, the fact that the relict landslide zones became more stable compared to the past decreased the function of the "Recent Landslide Records" data. This prediction is currently in the assumption phase and could become a widely accepted rule if supported and validated in several future studies

The result of two different methods applied in the process of transferring landslide data from landslide polygon to the ML environment is the following. Accuracy and precision increases and decreases when averaging all pixels of the landslide data polygon. In this method, the accuracy was 0.86. When the values of the pixel at the center of gravity or the geometric midpoint of the polygon are used, the accuracy decreases to 0.81 (Table 2).

Table 2. Model performance evaluation, accuracy, precision and F1 score

	Accuracy	Precision	F1 Score
Recent L. Pixel Average	0,74	0,72	0,80
Relict L. Pixel Average	0,69	0,68	0,68
Combined L. Pixel Average	0,88	0,86	0,87
Recent L. Pixel Center	0,71	0,70	0,76
Relict L. Pixel Center	0,68	0,67	0,66
Combined L. Pixel Center	0,84	0,81	0,83

3.2. Landslide Conditioning Factor Effect Analysis and SHAP Method

An important issue in multi-factor analyses made with GIS is to detect the individual effects of the input layers on the result. It would be meaningless to use factors that affect the result below acceptable values. In addition, these factors in ML can lead to overfitting problems. In the classification of factors according to the domain, sensitivity analysis, Pearson correlation coefficient (Wang et al., 2020; Chang et al., 2020), spatial heterogeneity (Hong et al., 2019), factor importance or partial response curves (Chen et al., 2018). There are different methods such as (Pourghasemi et

al., 2020). Input factors should be evaluated one by one before evaluating the success performance of the model in ML analysis.

SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. This approach helps us for better understanding and interpretation of machine learning results. The local explanation using the SHAP values via each individual SHAP value which explains why the ML model gives its decision and the contributions of the predictors/features (Le et al., 2022). The results showed that SHAP analysis can effectively improve machine learning transparency (Ou et al., 2020; Zhang et al., 2023). As it is given at previous page best score was achieved by applying "Combined Landslide Data" and "averaging all pixels of the landslide data polygon". Results of data model and analysis via XgBoost was evaluated with SHAP method. SHAP method is applied after inserting Python-SHAP analysis codes to Jupyter platform.

Density scatter plot of SHAP values is created for each feature to identify how much impact each feature has on the model output for individuals in the validation dataset (Figure 6). Figure 6 shows SHAP value magnitudes across all samples and we can understand that slope and wetness factors are most dominant factors and curvature, rainfall follow them. Less affecting factors are proximity to river and roads. Figure 7 also shows dependence scatter plot graphics of each input factor used for XGBoost based LSM analysis.

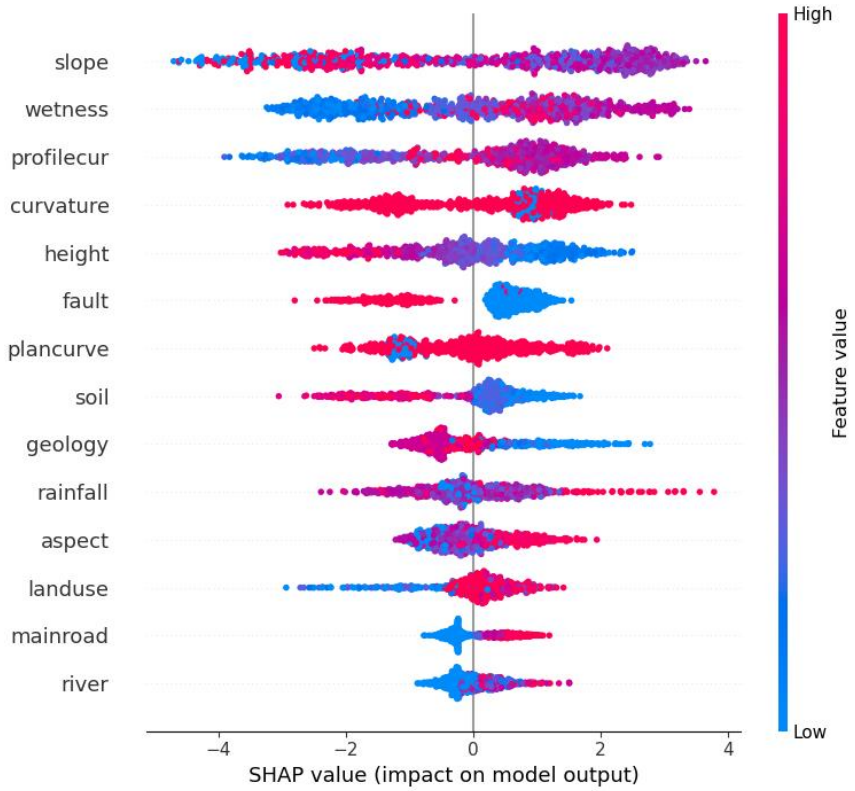

Figure 6. SHAP values (impact on model output)

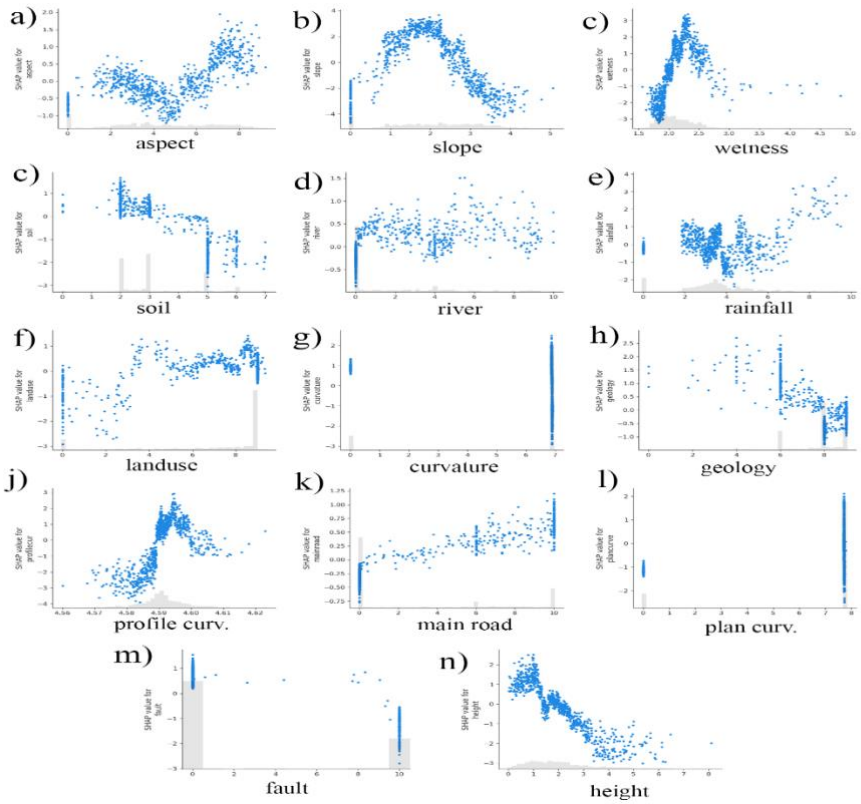

Figure 7. SHAP dependence scatter plots

3.3. Generating LS Maps and Landslide Risk Assessment of Buildings

More accurate susceptibility map can reduce the cost and damage of environmental disasters such as landslides (Chen and Li, 2020). LS map was produced according to the XGBoost results with best data model alternative. While producing this map, ArcMap 10.6 / Spatial Analyst / Map Algebra / Raster Calculator tool was used. Map Algebra is a simple and powerful algebra with which you can execute all Spatial Analyst tools, operators, and functions to perform geographic analysis. Map Algebra supports basic mathematical calculation with overlaying raster based GIS layers such as; multiplying, adding or dividing pixel values. In addition to basic math operations map algebra allows conditional, trigonometric and logarithmic math calculations. The LS map produced by multiplying coefficients with the normalized input layers (Figure 8).

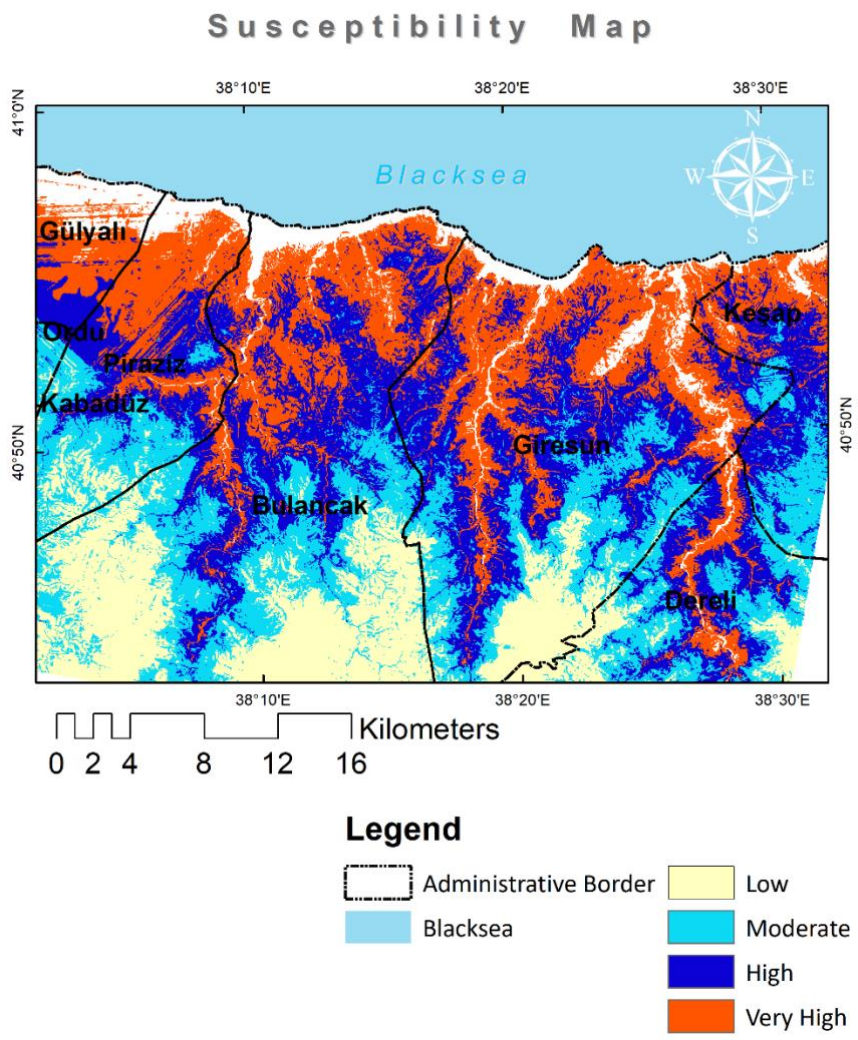

Figure 8. LS map of study area

There are 57.601 buildings in the study area. The analysis of susceptibility to landslides was overlapped with building footprints prepared in GIS environment. The average value of the landslide risk for each building has been calculated. Based on the risk analysis, it was determined that the buildings in the region are: 59% (34.006) low,30% (17.342) moderate, 10% (5872) high-risk and 1% (291) in very high in terms of landslide risk. Each building is labelled according to its level of risk (Figure 9). The production of risk maps and the pre-determination of potential threats with technology will significantly contribute to the decision-making processes of decision-makers, planners, and managers. Therefore, life and property lose will be reduced and welfare in developing countries will be more sustainable.

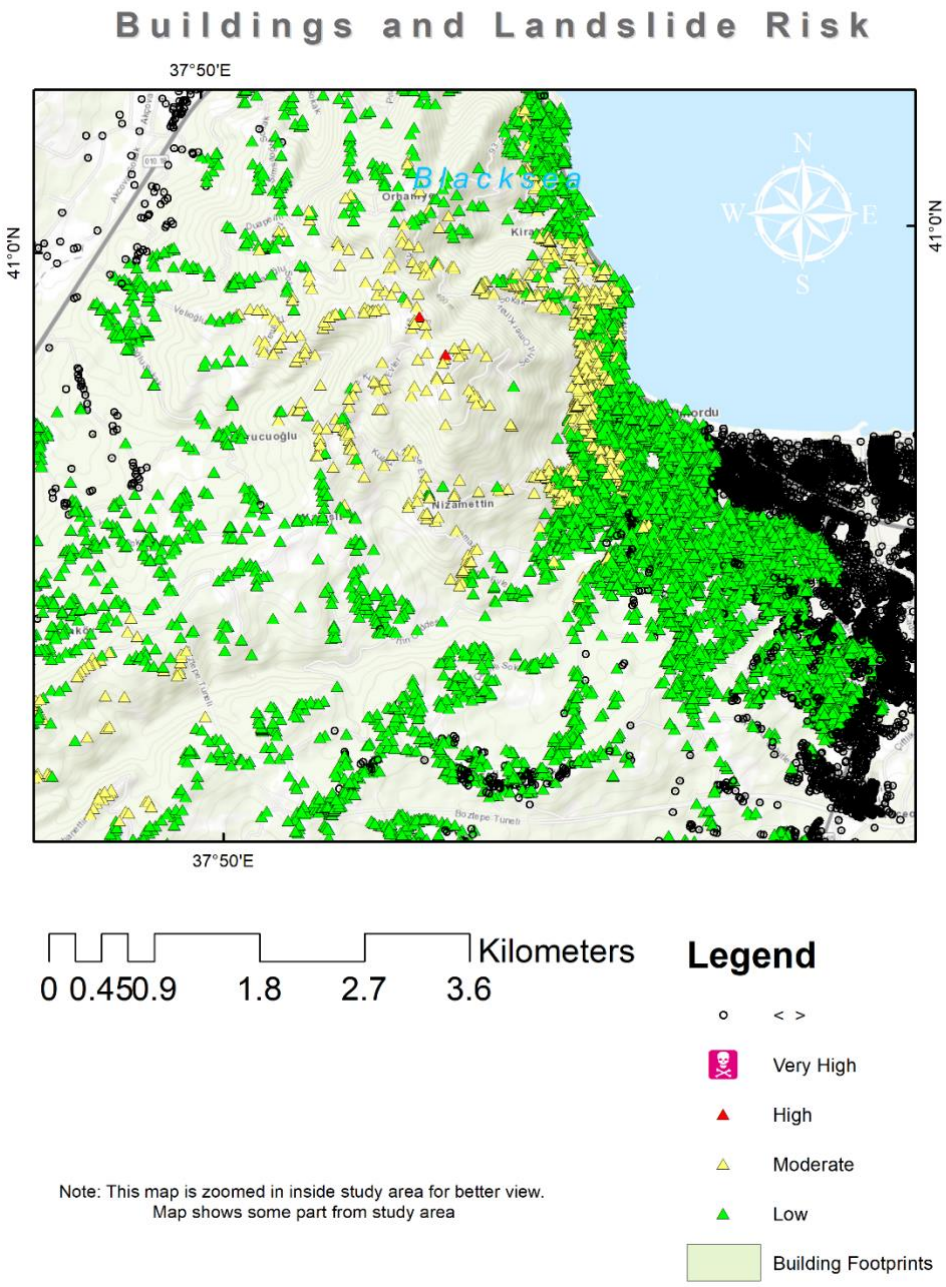

Figure 9. Building risk assessment

4. Conclusions and Recommendations

In this study, LS analysis were made by using XGBoost algorithm. The study was compared by applying different model combinations to be formed based on the characteristic structure of the landslide data and the method of transferring it to the ML environment. Model success comparisons were made using Accuracy, Precision and F1 score. As a result of the study, the most successful model performance was achieved by taking the average values of all pixels of the landslide data polygon in the XGBoost algorithm, combined landslide data. In addition, the study has shown in a

practical way that ML is an effective tool in GIS analysis with multi-criteria structure such as LS map generation and in solving similar GIS problems. Study shows that GIS supported machine learning methods may give efficient results in the process of producing landslide susceptibility maps. The methods / parameters used and the parameters obtained in study are in a structure that can be used directly in regions with landslide risk anywhere in the world. SHAP method is an efficient tool for evaluation and visualization of input ML factors. Each factor may be evaluated in separate or opposing approaches with the help of SHAP. The methods used in the study can be used with some model revisions in analysing not only landslide events but also other types of disaster. In addition, as a result of the hybrid use of machine learning models with multi-criteria decision support systems, it will be beneficial in solving multi-layered problems such as appropriate site or facility location selection for settlement or another purposes. In the future, testing and comparing ML algorithms with different data model combinations will increase our model success performance and bring us closer to ideal solutions and right decisions.

Authors' Contributions

All the paper is designed, created and revised by one author. Author contributed equally to the study.

Statement of Conflicts of Interest

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The author declares that this study complies with Research and Publication Ethics.

References

Abedini M, Ghasemian B, Shirzadi A, Shahabi H, Chapi K, Pham BT, Bin Ahmad B, and Tien Bui D. 2019. A novel hybrid approach of Bayesian Logistic Regression and its ensembles for landslide susceptibility assessment. Geocarto International. 34(13):1427-1457.

Aghdam IN., Varzandeh MHM., and Pradhan B. (2016). Landslide susceptibility mapping using an ensemble statistical index (Wi) and adaptive neuro-fuzzy inference system (ANFIS) model at Alborz Mountains (Iran). Environmental Earth Sciences. 75(7):553.

Aghlmand, M., Onur M. İ. and Talaei R. (2020). Heyelan Duyarlılık Haritalarının Üretilmesinde Analitik Hiyerarşi Yönteminin ve Coğrafi Bilgi Sistemlerinin Kullanımı. Avrupa Bilim ve Teknoloji Dergisi Özel Sayı, S. 224-230, Nisan 2020

Akinci H., Kilicoglu C., and Dogan S. (2020). Random Forest-Based Landslide Susceptibility Mapping in Coastal Regions of Artvin, Turkey. ISPRS International Journal of Geo-Information. 2020; 9(9):553

Althuwaynee OF., Pradhan B., Park H-J., and Lee JH. (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.

Althuwaynee OF., Pradhan B., and Lee S. (2016). A novel integrated model for assessing landslide susceptibility mapping using CHAID and AHP pair-wise comparison. International Journal of Remote Sensing. 37(5):1190-1209.

Arabameri, A., Chandra Pal, S., Rezaie, F., Chakrabortty, R., Saha, A., Blaschke, T., di Napoli, M., Ghorbanzadeh, O., and Thi Ngo, P. T. (2022). Decision tree based ensemble machine learning approaches for landslide susceptibility mapping. Geocarto International, 37(16), 4594–4627. https://doi.org/10.1080/10106049.2021.1892210

Atkinson PM., and Massari R. (1998). Generalised Linear Modelling of Susceptibility to Landsliding in the Central Apennines, ITALY. Computers & Geosciences. 24(4):373-385.

Beguería S. (2006). Validation and Evaluation of Predictive Models in Hazard Assessment and Risk Management. Natural Hazards. 37(3):315-329.

Breiman L. (2001). Random Forests. Machine Learning. Kluwer Academic Publishers. 45(1):5-32.

Chang Z., Du Z., Zhang F., Huang .F, Chen J., Li W., and Guo Z. (2020). Landslide Susceptibility Prediction Based on Remote Sensing Images and GIS: Comparisons of Supervised and Unsupervised Machine Learning Models. Remote Sensing. 12(3).

Ciampalini, A., Bardi, F., Bianchini, S., Frodella, W., del Ventisette, C., Moretti, S., and Casagli, N. (2014). Analysis of building deformation in landslide area using multisensor PSInSARTM technique. International Journal of Applied Earth Observation and Geoinformation, 33, 166–180.

Chen W., Peng J., Hong H., Shahabi H., Pradhan B., Liu J., Zhu AX., Pei X., and Duan Z. (2018). Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. Science of The Total Environment. 626:1121-1135.

Chen W. and Li Y. (2020). GIS-based evaluation of landslide susceptibility using hybrid computational intelligence models. CATENA. 195:104777.

Chen T. and Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794.

Ching J. and Phoon K-K. (2019). Constructing Site-Specific Multivariate Probability Distribution Model Using Bayesian Machine Learning. Journal of Engineering Mechanics. 145(1):04018126.

Constantin M., Bednarik M., Jurchescu MC., and Vlaicu M. (2011). Landslide susceptibility assessment using the bivariate statistical analysis and the index of entropy in the Sibiciu Basin (Romania). Environmental Earth Sciences. 63(2):397-406.

Dai FC., Lee CF., and Zhang XH. (2001). GIS-based geo-environmental evaluation for urban land-use planning: a case study. Engineering Geology. 61(4):257-271.

Dehnavi A., Aghdam IN., Pradhan B., and Morshed Varzandeh MH. (2015). A new hybrid model using step-wise weight assessment ratio analysis (SWARA) technique and adaptive neuro-fuzzy inference system (ANFIS) for regional landslide hazard assessment in Iran. CATENA. 135:122-148.

De Sy V., Schoorl JM., Keesstra SD., Jones KE., and Claessens L. (2013). Landslide model performance in a high resolution small-scale landscape. Geomorphology. 190:73-81.

Fang, Z., Wang, Y., Peng, L., and Hong, H. (2020). Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping. Computers & Geosciences, 139, 104470.

Fanos, A. M. and Pradhan, B. (2019). A novel rockfall hazard assessment using laser scanning data and 3D modelling in GIS. CATENA, 172, 435–450.

Feizizadeh B., Shadman Roodposhti M., Jankowski P., and Blaschke T. (2014). A GIS-based extended fuzzy multi-criteria evaluation for landslide susceptibility mapping. Computers & Geosciences. 73:208-221.

Froude M. and Petley D. (2018). Global fatal landslide occurrence 2004 to 2016. Natural Hazards and Earth System Sciences Discussions.1-44.

Fu, S., Chen, L., Woldai, T., Yin, K., Gui, L., Li, D., Du, J., Zhou, C., Xu, Y., and Lian, Z. (2020). Landslide hazard probability and risk assessment at the community level: a case of western Hubei, China. Nat. Hazards Earth Syst. Sci., 20(2), 581–601.

Girshick, R. (2015). Fast R-CNN. Proceedings of the IEEE international conference on computer vision 1440–1448.

Greedy F. J. function approximation: a gradient boosting machine. Annals of Statistics, 29(5):1189{1232, 2001.

Gorsevski, P.V., Gessler, P.E., Foltz, R.B. and Elliot, W.J. (2006), Spatial Prediction of Landslide Hazard Using Logistic Regression and ROC Analysis. Transactions in GIS, 10: 395-415.

Guzzetti F., Reichenbach P., Ardizzone F., Cardinali M., and Galli M. (2006). Estimating the quality of landslide susceptibility models. Geomorphology. 81(1):166-184.

Hong H., Miao Y., Liu J., and Zhu AX. (2019). Exploring the effects of the design and quantity of absence data on the performance of random forest-based landslide susceptibility mapping. CATENA. 176:45-64.

Hong H., Naghibi SA., Pourghasemi HR., and Pradhan B. (2016). GIS-based landslide spatial modeling in Ganzhou City, China. Arabian Journal of Geosciences. 9(2):112.

Hong H., Pradhan B., Sameen MI., Kalantar B., Zhu A., and Chen W. (2018). Improving the accuracy of landslide susceptibility model using a novel region-partitioning approach. Landslides. 15(4):753-772.

Hong H., Liu J., and Zhu AX. (2020). Modeling landslide susceptibility using LogitBoost alternating decision trees and forest by penalizing attributes with the bagging ensemble. Science of The Total Environment. 718:137231.

Hong, H. (2023). Assessing landslide susceptibility based on hybrid Best-first decision tree with ensemble learning model. Ecological Indicators, 147, 109968.

Huang, W., Ding, M., Li, Z.; Zhuang, J., Yang, J., Li, X., Meng, L., Zhang, H., and Dong, Y. An Efficient User-Friendly Integration Tool for Landslide Susceptibility Mapping Based on Support Vector Machines: SVM-LSM Toolbox. Remote Sens. 2022, 14, 3408.

Hussin HY., Zumpano V., Reichenbach P., Sterlacchini S., Micu M., van Westen C., and Bălteanu D. (2016). Different landslide sampling strategies in a grid-based bi-variate statistical susceptibility model. Geomorphology. 253:508-523.

Kavzoglu, T.,and Teke, A. (2022). Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost). Arab J Sci Eng 47, 7367–7385

Lary DJ., Alavi AH., Gandomi AH., and Walker AL. (2016). Machine learning in geosciences and remote sensing. Geoscience Frontiers. 7(1):3-10.

Lecun Y., Bottou L., Bengio Y., and Haffner P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE. 86(11):2278-2324.

Li X., Zhang L., Xiao T., Zhang S., and Chen C. (2019). Learning failure modes of soil slopes using monitoring data. Probabilistic Engineering Mechanics. 56:50-57.

Li P. (2010). Robust Logitboost and adaptive base class (ABC)Logitboost. In Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artifcial Intelligence(UAI'10), pages 302{311, 2010.

Liu Y., Fan B., Wang L., Bai J., Xiang S., and Pan C. (2018). Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. ISPRS Journal of Photogrammetry and Remote Sensing. 145:78-95.

Lo MK., and Leung YF. (2018). Bayesian updating of subsurface spatial variability for improved prediction of braced excavation response. Canadian Geotechnical Journal. 56(8):1169-1183.

Mathew J., Jha VK., and Rawat GS. (2009). Landslide susceptibility zonation mapping and its validation in part of Garhwal Lesser Himalaya, India, using binary logistic regression analysis and receiver operating characteristic curve method. Landslides. 6(1):17-26.

Martha, T. R., van Westen, C. J., Kerle, N., Jetten, V., and Vinod Kumar, K. (2013). Landslide hazard and risk assessment using semi-automatically created landslide inventories. Geomorphology, 184, 139–150.

Mezaal MR., Pradhan B., Sameen MI., Mohd Shafri HZ., and Yusoff ZM. (2017). Optimized Neural Architecture for Automatic Landslide Detection from High-Resolution Airborne Laser Scanning Data. Applied Sciences. 7(7).

Nefeslioglu HA., Gokceoglu C., and Sonmez H. (2008). An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps. Engineering Geology. 97(3):171-191.

Nguyen H-L., Le T-H., Pham C-T., Le T-T., Ho LS., Le VM., Pham BT., and Ly H-B. (2019). Development of Hybrid Artificial Intelligence Approaches and a Support Vector Machine Algorithm for Predicting the Marshall Parameters of Stone Matrix Asphalt. Applied Sciences. 9(15):3172.

Nsengiyumva, J. B., and Valentino, R. (2020). Predicting landslide susceptibility and risks using GIS-based machine learning simulations, case of upper Nyabarongo catchment. Geomatics, Natural Hazards and Risk, 11(1), 1250–1277. https://doi.org/10.1080/19475705.2020.1785555

Orhan, O., Bilgilioglu, S. S., Kaya, Z., Ozcan, A. K., and Bilgilioglu, H. (2022). Assessing and mapping landslide susceptibility using different machine learning methods. Geocarto International, 37(10), 2795–2820. https://doi.org/10.1080/10106049.2020.1837258

Ou C., Liu J., Qian Y., Chong W., and He X. (2020). Rupture risk assessment for cerebral aneurysm using interpretable machine learning on multidimensional data. Front. Neurol., 11.

Papaioannou I, Straub D. (2017). Learning soil parameters and updating geotechnical reliability estimates under spatial variability – theory and application to shallow foundations. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards. 11(1):116-128.

Park S., and Kim J. (2019). Landslide Susceptibility Mapping Based on Random Forest and Boosted Regression Tree Models, and a Comparison of Their Performance. Applied Sciences. 9(5).

Pham BT., Phong TV., Nguyen-Thoi T., Trinh PT., Tran QC., Ho LS., Singh SK., Duyen TT., Nguyen LT., and Le HQ. (2020). GIS-based ensemble soft computing models for landslide susceptibility mapping. Advances in Space Research. 66(6):1303-1320.

Pham BT., Prakash I., Singh SK., Shirzadi A., Shahabi H., Tran T-T., and Bui DT. (2019). Landslide susceptibility modelling using Reduced Error Pruning Trees and different ensemble techniques: Hybrid machine learning approaches. CATENA. 175:203-218.

Pradhan, B., and Sameen, M. I. (2018). Manifestation of SVM-Based Rectified Linear Unit (ReLU) Kernel Function in Landslide Modelling. In W. Suparta, M. Abdullah, & M. Ismail (Eds.), Space Science and Communication for Sustainability (pp. 185–195). Springer Singapore. https://doi.org/10.1007/978-981-10-6574-3_16

Pourghasemi HR., Kornejady A., Kerle N., and Shabani F. (2020). Investigating the effects of different landslide positioning techniques, landslide partitioning approaches, and presence-absence balances on landslide susceptibility mapping. CATENA. 187:104364.

Prasad P., Loveson VJ., Das B., and Kotha M. (2021). Novel ensemble machine learning models in flood susceptibility mapping. Geocarto International.1-23.

Rahmati O., Tahmasebipour N., Haghizadeh A., Pourghasemi HR., and Feizizadeh B. (2017). Evaluating the influence of geo-environmental factors on gully erosion in a semi-arid region of Iran: An integrated framework. Science of The Total Environment. 579:913-927.

Sagi Ö. and Rokach L. (2021). Approximating XGBoost with an interpretable decision tree, Information Sciences, Volume 572, Pages 522-542, ISSN 0020-0255.

Sahin, E.K. (2023). Implementation of free and open-source semi-automatic feature engineering tool in landslide susceptibility mapping using the machine-learning algorithms RF, SVM, and XGBoost. Stoch Environ Res Risk Assess 37, 1067–1092

Sameen MI., Pradhan B., Bui DT., and Alamri AM. (2020). Systematic sample subdividing strategy for training landslide susceptibility models. CATENA. 187:104358.

Singh, A., Pal, S., and Kanungo, D. P. (2021). An integrated approach for landslide susceptibility–vulnerability–risk assessment of building infrastructures in hilly regions of India. Environment, Development and Sustainability, 23(4), 5058–5095. https://doi.org/10.1007/s10668-020-00804-z

T.C. Orman ve Su Işleri Bakanlığı. (2016). Meteoroloji Genel Müdürlüğü, Köppen iklim siniflandirmasina göre Türkiye iklimi, Climatology Report.

Thi-Thu-Huong L., Kim H., Kang H. and Kim H. (2022). "Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method" Sensors 22, no. 3:

Url, 2. https://en.wikipedia.org/wiki/Logistic_regression, 2020

Van Westen CJ., Van Asch TWJ., and Soeters R. (2006). Landslide hazard and risk zonation—why is it still so difficult? Bulletin of Engineering Geology and the Environment. 65(2):167-184.

Wang H., Zhang L., Yin K., Luo H., and Li J. (2021). Landslide identification using machine learning. Geoscience Frontiers. 12(1):351-364.

Wang Y., Fang Z., and Hong H. (2019). Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China. Science of The Total Environment. 666:975-993.

Wang Y., Feng L., Li S., Ren F., and Du Q. (2020). A hybrid model considering spatial heterogeneity for landslide susceptibility mapping in Zhejiang Province, China. CATENA. 188:104425.

Wu Z., Wu Y., Yang Y., Chen F., Zhang N., Ke Y., and Li W. (2017). A comparative study on the landslide susceptibility mapping using logistic regression and statistical index models. Arabian Journal of Geosciences. 10(8):187.

Yamaguchi, S., and Kasai, M. (2022). A new index representative of seismic cracks to assess post-seismic landslide susceptibility. Transactions in GIS, 26, 1040– 1061.

Yi Y., Zhang Z., Zhang W., Jia H., and Zhang J. (2020). Landslide susceptibility mapping using multiscale sampling strategy and convolutional neural network: A case study in Jiuzhaigou region. CATENA. 195:104851.

Yilmaz I. (2010). The effect of the sampling strategies on the landslide susceptibility mapping by conditional probability and artificial neural networks. Environmental Earth Sciences. 60(3):505-519.

Youssef AM., and Pourghasemi HR. (2021). Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. Geoscience Frontiers. 12(2):639-655.

Zhang W., Goh ATC., Zhang Y., Chen Y., and Xiao Y. (2015). Assessment of soil liquefaction based on capacity energy concept and multivariate adaptive regression splines. Engineering Geology. 188:29-37.

Zhang W., and Goh ATC. (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geoscience Frontiers. 7(1):45-52.

Zhao, Z., He, Y., Yao, S., Yang, W., Wang, W., Zhang, L., and Sun, Q. (2022). A comparative study of different neural network models for landslide susceptibility mapping. Advances in Space Research, 70(2), 383–401.

Zhang, K., Wu, X., Niu, R., Yang, K. and Zhao, L. (2017). The assessment of landslide susceptibility mapping using random forest and decision tree methods in the Three Gorges Reservoir area, China. Environmental Earth Sciences. Vol: 11 – 76, pages 1-20

Zhang J., Ma X., Zhang J., Sun D., Zhou X., Mi C. and Wen H. (2023). Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model, Journal of Environmental Management, Volume 332, 2023, 117357, ISSN 0301-4797.

Zhu AX., Miao Y., Yang L., Bai S., Liu J., and Hong H. (2018). Comparison of the presence-only method and presence-absence method in landslide susceptibility mapping. CATENA. 171:222-233.