Landslides (2024) 21:2893–2911
DOI 10.1007/s10346-024-02336-3
Received: 9 January 2024
Accepted: 31 July 2024
Published online: 20 August 2024

corrected publication 2024

N. Dal Seno · D. Evangelista · E. Piccolomini · M. Berti

Comparative analysis of conventional and machine learning techniques for rainfall threshold evaluation under complex geological conditions

Abstract This research focuses on the essential task of defining rainfall thresholds in regions with complex geological features, specifically at a regional scale. It examines a variety of methodologies, from traditional empirical-statistical methods to cutting-edge machine learning (ML) techniques, for establishing these thresholds. The Emilia-Romagna region in Italy, known for its intricate geological structure and prevalence of weak rocks that often lead to large and deep-seated landslides, serves as the study area. The region's complex interplay between rainfall and landslide incidences poses a significant challenge in accurately determining rainfall thresholds. The effectiveness of ML methods is compared against conventional empirical-statistical approaches, evaluating factors such as prediction accuracy, model complexity, and the interpretability of results for use by regional landslide warning system operators. The findings indicate that machine learning techniques have an edge over traditional methods, yielding higher performance scores and fewer false positives. Nevertheless, these advancements are modest when considering the increased complexity of ML methods and the incorporation of additional rainfall parameters. This underlines the continued need for improvements in data quality and volume. The study stresses the importance of enhancing data collection and analysis techniques, especially in an era where advanced AI tools are increasingly available, to improve the accuracy of predicting rainfall thresholds for effective landslide warning systems.

Keywords Landslides · Northern Apennines · Machine learning · Rainfall threshold · Landslide prediction

Introduction

The definition of rainfall thresholds is critical for developing effective landslide warning systems, both regionally and locally (De Vita et al. 1998; Aleotti and Chowdhury 1999; Wieczorek and Glade 2007; Guzzetti et al. 2007, 2008). Rainfall is the primary trigger of landslides worldwide (Caine 1980) and identifying the critical rainfall conditions that trigger landslides is crucial for predicting their occurrence (Terlien 1998; Calvello and Piciullo 2016). Various methods exist for setting these thresholds, including empirical-statistical and physically based approaches (Segoni et al. 2018; Guzzetti et al. 2007). Empirical-statistical methods, which analyze historical landslide and rainfall data to establish a relationship, are widely used due to their simplicity and robustness (Berti et al. 2012; Aleotti 2004; Glade et al. 2000). They account for complex interactions affecting slope stability but are often specific to certain locations and may not hold up under changing climatic conditions (Crosta and Frattini 2022; Gariano and Guzzetti 2016).

Machine learning (ML) techniques have brought significant advancements to this field. These algorithms learn from existing data to identify patterns and predict future events, offering adaptability and versatility in analyzing various variables (Alpaydin 2010; Bishop 2006). Recent studies have demonstrated the effectiveness of machine learning in improving landslide warning systems by integrating various data types, including soil moisture and rainfall, to predict hydrologic responses and landslide occurrences (Orland et al. 2020; Liu et al. 2021; Mondini et al. 2023). These techniques represent a promising direction for enhancing the accuracy and applicability of rainfall thresholds in landslide prediction.

This study assesses the use of machine learning techniques for determining rainfall thresholds in Emilia-Romagna, Northern Italy, a region characterized by complex geology and weak rocks that lead to frequent, large landslides. The research focuses on evaluating the effectiveness of machine learning approaches in comparison with traditional empirical-statistical methods. Key aspects of this evaluation include prediction accuracy, model complexity, and the ease of interpreting and communicating findings to regional warning system staff, considering the intricate relationship between rainfall and landslides in the area.

Study area

The Emilia-Romagna region in Italy (Fig. 1) is highly prone to landslides due to several factors. The region's rocks are mostly weak and susceptible to weathering, and there is ongoing geological uplift, constantly reshaping the landscape. The lengthy rainy and snowy seasons further contribute to slope instability. Over 32,000 landslides have been identified in the area (Bertolini et al. 2005), with earthflows being the most prevalent type, especially in regions with clay rocks (Ligurian-Subligurian Domain, Fig. 1). These earthflows and slides typically move slowly, but can accelerate dramatically during heavy rain or snowmelt, sometimes reaching high speeds. Rockslides, another common type, are frequent in the flysch units (Tuscan-Romagna Domain, Fig. 1) and can form large, dangerous slides, especially when strata dip out of the slope. While less common than earthflows, rockslides pose a greater risk due to their higher speeds. Rockfalls occur in competent sandstone and limestone rocks (Epiligurian Domain, Fig. 1) but represent less than 1% of landslides in the region. Figure 2 provides illustrative examples of these types of landslides.

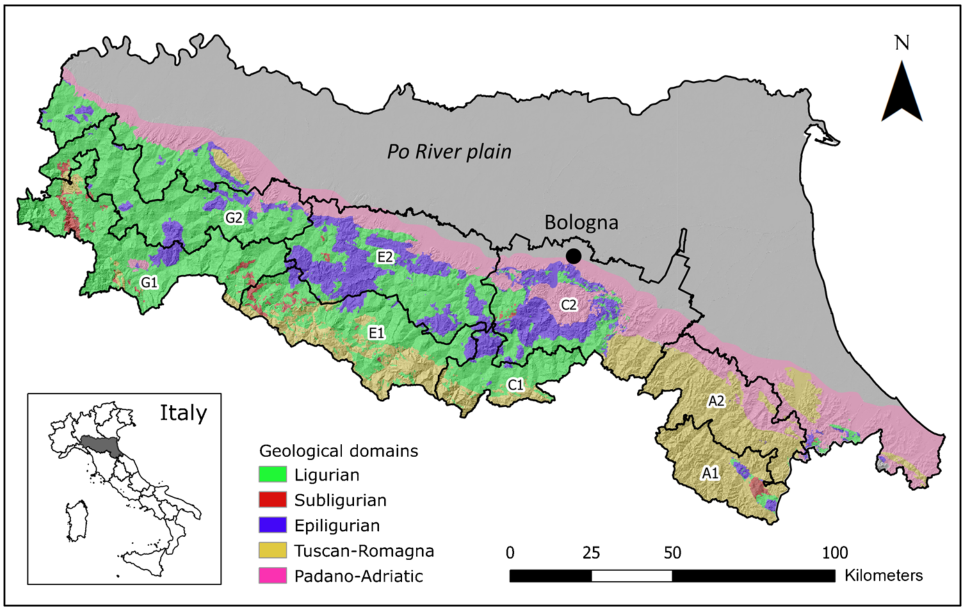


Fig. 1 Geographic overview of the Emilia-Romagna region (Italy) segmented into alert zones. Illustrative geological map of the region

Fig. 2 Examples of landslide types in the Emilia-Romagna Region. **a** Earthflows and **b** earth slides, commonly found in the clay rocks of the Ligurian Domain. **c** A typical rockslide observed in the flysch rocks of the Tuscan-Romagna Domain. **d** Rockfall characteristic of the sandstone rocks in the Epiligurian Domain

The Civil Protection Agency of Emilia-Romagna is responsible for managing the region's hydrogeological risk alert system. Since 1995, the agency has established and continuously improved the alert system, which comprises a prediction and monitoring phase. The prediction phase relies on meteorological forecasts for the subsequent 72 h and an evaluation of their corresponding effects. A team of specialized experts performs this crucial step, considering various rainfall threshold models and the current territory conditions. The technical group's assessment is reported in a daily Alert Bulletin that identifies three possible alert levels (ordinary, moderate, and elevate) based on the expected number and extent of landslides. Such indications enable the implementation of preparation plans at the municipality level. The forecasts cover eight alert areas (Fig. 1), which are territorial units characterized by homogenous meteo-climatic response and comparable hydrographic and orographic features. Alert areas are considerable in size (from 660 to 1900 km²) and exhibit significant geological variability within their boundaries, with different types of landslide phenomena. Nonetheless, the regional alert system aims to prepare municipalities for potential landslides on a large scale, rather than providing a forecast at the local level.

Methods

Data

The analysis in this study was based on data from two primary sources: a historical catalog of landslide events and precipitation records. A similar dataset was previously used by Berti et al. (2012) to determine probabilistic rainfall thresholds for Emilia-Romagna using a Bayesian approach. This current study differs in three key aspects: (i) it includes updated data on landslides from 2009 to 2022; (ii) it adopts a more representative approach for averaging rainfall over alert areas; (iii) it evaluates various traditional and machine learning algorithms for defining rainfall thresholds.

Landslide catalogue

The Emilia-Romagna Geological Survey has a register of past landslides that occurred in the regional territory (https://ambiente.regione.emilia-romagna.it/en/geologia/cartography/). The register is constantly updated and includes data from national archives, local press, and technical reports. Although the catalog is not entirely comprehensive, since minor events that occurred in remote areas may have gone unnoticed, it can be regarded as statistically complete from approximately 1951 (Rossi et al. 2010). The catalog stores the location, date of occurrence, characteristics, and consequences of each landslide with varying degrees of accuracy. The date occurrence is the most important parameter for our analysis since each landslide must be linked to a specific triggering precipitation event. To ensure accuracy, only landslides with a certain date of occurrence (daily accuracy) or with a maximum uncertainty of 1–7 days were considered.

Rainfall data

The Emilia-Romagna region features a rainfall network with over 200 tipping-bucket rain gauges spread uniformly across the area. Up to 2001, rainfall data was collected manually on a daily basis, but since then, automatic gauges have been used to record data every 30 min. In the civil protection system, rainfall measurements from an entire alert area (as shown in Fig. 1) are spatially interpolated and averaged to provide a single value for the whole area, due to the variability of precipitation over short distances. The averaging of daily rainfall data also matches with weather forecasts, which predict precipitation levels for each alert zone over 24, 48, and 72 h. The alert zones vary in size, from 660 km² (Zone A1) to 1900 km² (Zone C2), with each rain gauge monitoring an area of approximately 60 km².

Manual identification of triggering rainfall events

Assigning a specific rainfall event to each historical landslide proved to be a challenging task. The complexity of rainfall, which can consist of multiple episodes with varying intensities and durations, along with the impact of snowmelt not included in rainfall data, contributed significantly to this difficulty. The uncertainty surrounding the date of the landslide occurrence further complicated the procedure, making it impossible to utilize an automatic algorithm for identifying the triggering rainfall event. Therefore, a manual identification process was employed, requiring careful analysis of each single case. This process was complex, time-consuming, and involved the continuous re-analysis of more than 4000 historical records.

To address the subjectivity inherent in expert judgment, each rainfall event was classified based on its uncertainty level (refer to Fig. 3 for examples):

- Type 1) Low uncertainty. The triggering rainfall can be identified with a minimal degree of subjectivity as there is a distinct rainfall episode that occurred during or, at most, 1–3 days prior to the landslide, with a clearly defined beginning and end (Fig. 3a–c).
- Type 2) Moderate uncertainty. At the time of the landslide there is a heavy rainfall episode, but determining the exact beginning of the event is somewhat subjective due to the presence of multiple rainfall pulses with comparable duration and intensity (Fig. 3d–f).
- Type 3) High uncertainty. Determining the triggering rainfall with sufficient accuracy is not possible due to several reasons, such as (a) the landslide occurred at the beginning of a rainfall event (Fig. 3g); (b) the landslide occurred during a dry period, or after a light rainfall with minimal duration and intensity (Fig. 3h); (c) the landslide occurred after or during an extended period characterized by multiple rainfall events with similar duration and intensity (Fig. 3i). These records may indicate landslides with an incorrect date of occurrence, landslides triggered by snowmelt or rain-on-snow events, or situations where the relationship between rainfall and the landslide is particularly intricate because of the effect of antecedent rainfall.

For analytical purposes, Type 1 and Type 2 events are considered appropriate, while Type 3 events are excluded due to the high uncertainty about the duration and intensity of the triggering rainfall.

Regarding rockfalls, they were not explicitly excluded from our analysis; however, any events not clearly triggered by rainfall were categorized as Type 3 "High uncertainty" during the

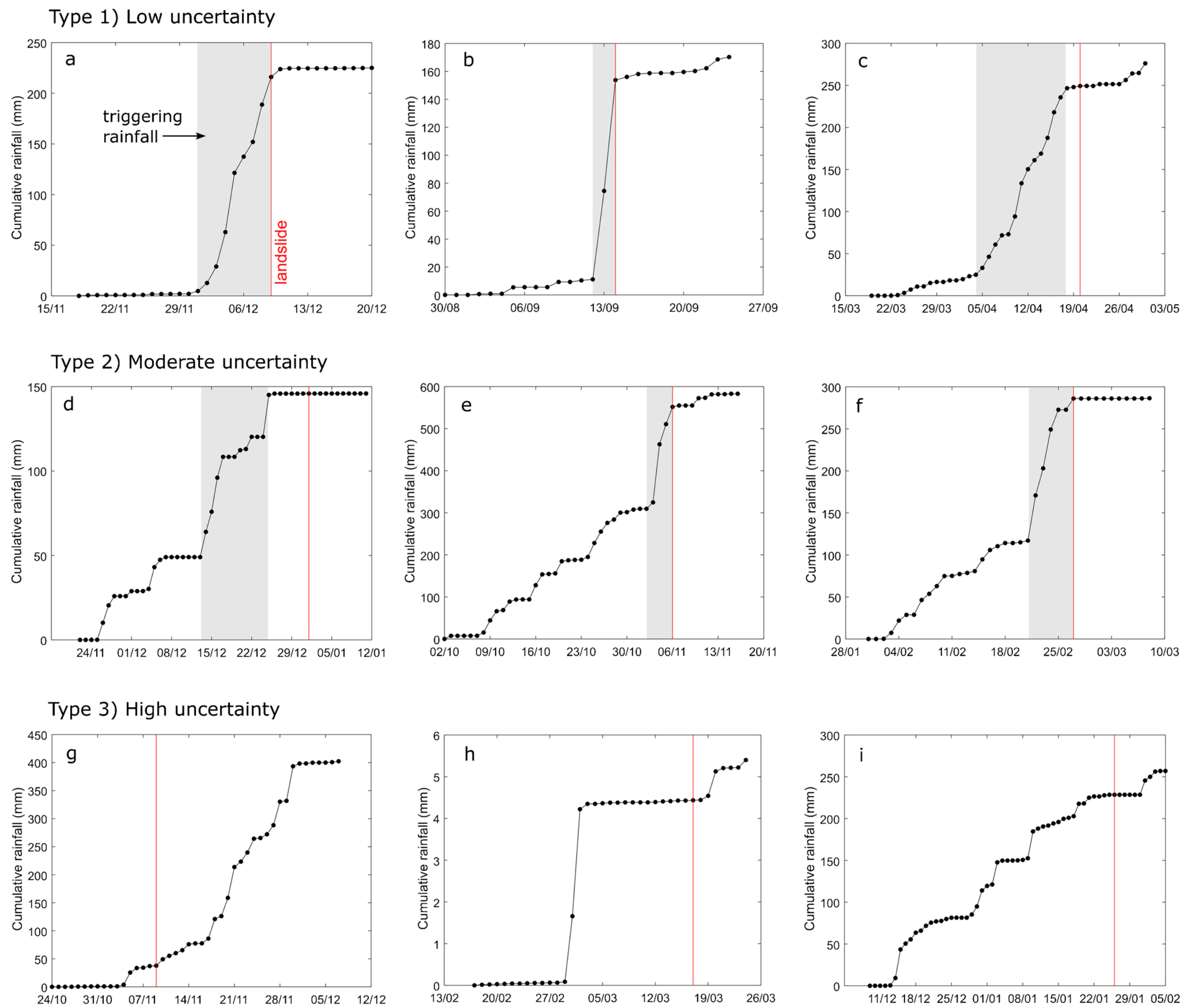


Fig. 3 Classification of rainfall events by uncertainty in expert assessment. **a**–**c** Type 1-low uncertainty: the triggering rainfall is clearly identified with minimal subjectivity, evident from a distinct rainfall episode occurring 1–3 days before the landslide. **d**–**f** Type 2-moderate uncertainty: heavy rainfall coincides with the landslide, but discerning the precise start is uncertain due to multiple similar-duration and intensity rainfall pulses. **g**–**i** Type 3-high uncertainty: pinpointing the triggering rainfall is challenging; the landslide could occur at the onset of a rainfall event (**g**), during a dry period (**h**), or after an extended period with several similar rainfall events (**i**)

manual identification process and were consequently excluded from further analysis.

Automated definition of rainfall events from precipitation data

To define rainfall thresholds via statistical techniques (discussed in the following section), it is essential to categorize both critical and non-critical rainfall events. Given the extensive span of the precipitation data, an automated process is necessary for isolating non-significant rainfall events within the dataset. A common approach in the literature is to set a threshold of duration D_{min} and precipitation amount P_{min} which serves to determine the start and end of a rainfall event (Corominas 2000; Saito et al. 2010). The values of D_{min} and P_{min} depend on the specific site conditions: short inter-event periods are typically used for shallow landslides and debris flows, that promptly react to rainfall, while longer periods are employed for deep-seated landslides in fine-grained materials (Wang and Sassa 2003). Regardless of the criteria used, it is essential to ensure consistency between the automated algorithm and the expert identification of critical rainfall. The algorithm, in fact, must be developed based on the same criteria and thresholds used by the experts to guarantee consistency between the datasets of critical and non-critical rainfall.

During the manual identification of the triggering rainfall, various combinations of D_{min} and P_{min} were tested to determine which one was most effective in replicating expert judgment. Ultimately, the best combination was determined to be $D_{min} = 3$ days and $P_{min} = 5$ mm (D3P5 algorithm hereafter). This means that a rainfall event is defined as a continuous period of precipitation that ends when the precipitation drops below 5 mm over 3 days. Berti et al. (2012) arrived at the same definition using a reduced version of the current dataset spanning the period 1939–2000. The D3P5 criterion was applied to the precipitation time series of the eight alert areas to extract all rainfall events in the period 1939–2022. Figure 4 shows an example of a rainfall time series (Alert Area A1, period February–September 1978) where rainfall events have been identified through the automatic recognition algorithm. The red polygon highlights a rainfall event that triggered landslides.

Dataset for analysis

The final dataset used for the analysis is organized into eight tables, each corresponding to a specific alert area. Every table row represents a distinct rainfall event, with various columns detailing key aspects of the event. These aspects include the duration of the event (in days), total precipitation (in millimeters), average daily rainfall intensity (mm/day), antecedent rainfall over preceding intervals of 7, 14, 30, and 60 days (in millimeters), and the day of the year (numbered 1–365, serving as an indicator of seasonal variation). Additionally, each row notes the number of landslides triggered in the alert area by the rainfall event, where zero indicates a non-critical event that did not result in any landslides.

While all events are geolocated and theoretically geological data could be included in the analysis, the traditional threshold methods used in this study are typically based only on two parameters—rainfall duration and intensity/cumulative. Including geological data would necessitate subdividing the dataset into geologically homogeneous areas, which would significantly reduce its statistical robustness. For machine learning methods, although feasible, we chose not to include geological factors to maintain a direct comparison with traditional methods, ensuring a more straightforward evaluation framework.

The dataset's primary focus is to investigate the link between rainfall characteristics and the occurrence of landslides on a larger scale, rather than delving into the broader environmental factors that may influence landslide risk. Hence, the analysis omits spatial details like geological makeup, slope gradients, and land use patterns, which are generally pivotal for understanding landslide genesis on a more localized, slope-specific scale.

Algorithms

In our study, we compared four traditional rainfall threshold determination methods (BART, Frequentist, Percentile, LDA) with seven advanced machine learning (ML) techniques (NN, LR, MLDA, MQDA, RF, XGBoost, TPOT), detailed in Table 1. This distinction highlights the shift from traditional methods, which rely on basic parameters like rainfall duration and intensity, to more complex, multiparametric ML techniques, as listed in Table 1. While traditional methods are straightforward, ML techniques, evolving from the statistical principles of traditional methods, offer greater analytical depth but are more complex to learn and apply.

Conventional methods

The traditional techniques considered in our analysis comprise four various Bayesian methods: BART, Frequentist, Percentile, and LDA.

BART

BART (Bayesian Analysis of Rainfall Thresholds) is a probabilistic method for defining rainfall thresholds (Berti et al. 2012, 2015). It calculates the probability of a landslide occurring in response to specific rainfall duration and intensity using a simplified application of Bayes' theorem as proposed by Berti et al. (2012):

$$P(A|B) \approx \frac{N(B|A)}{N_B} \tag{1}$$

where $N(B|A)$ is the number of rainfall events of magnitude B that resulted in landslides, and N_B is the number of rainfall events of

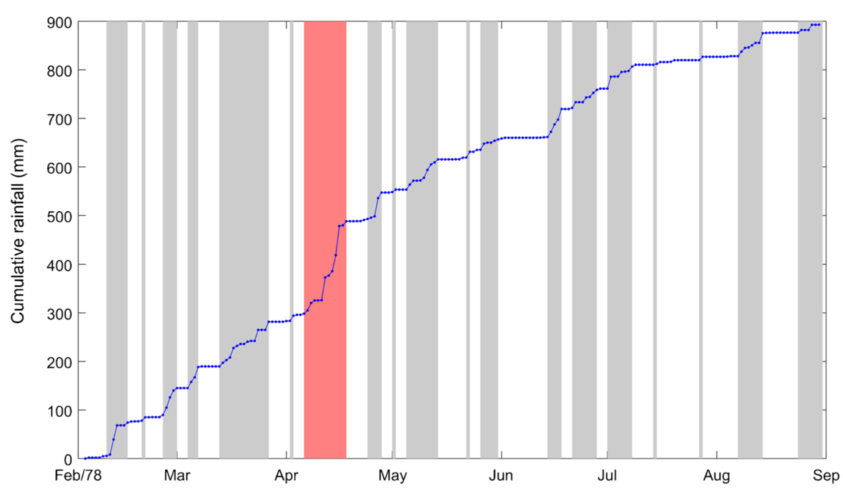


Fig. 4 Demonstration of automatic rainfall event identification in the rainfall time series for Alert Area A1 (February-September 1978). The rainfall events within this period were identified using an automated recognition algorithm. An event that resulted in landslides is emphasized with a red polygon

Table 1 Overview of the methods used to define rainfall thresholds. Parameters: I = rainfall intensity (mm/day), D = rainfall duration (days), dY = day of the year (1 to 365), C = total precipitation (mm), A7, A14, A30, A60 = antecedent rainfall in the previous 7, 14, 30, and 60 days (mm); Ac = alert area

Method	Parameters	Type of results	Reference
BART	I, D	Bayes probability	(Berti et al. 2012)
Frequentist	I, D	Exceedance probability	(Brunetti et al. 2010)
Percentile	I, D	Empirical distribution	(Guzzetti et al. 2007)
LDA	I, D	Binary classification	(Hastie et al. 2009)
NN	I, D, dY, C, A7, A14, A30, A60, Ac	Multiclass classification	(Huang and Xiang 2018)
LR	I, D, dY, C, A7, A14, A30, A60, Ac	Multiclass classification	(Frattini et al. 2009)
MLDA	I, D, dY, C, A7, A14, A30, A60, Ac	Multiclass classification	(Ramos-Cañón et al. 2016)
MQDA	I, D, dY, C, A7, A14, A30, A60, Ac	Multiclass classification	(Hastie et al. 2009)
RF	I, D, dY, C, A7, A14, A30, A60, Ac	Multiclass classification	(Hu et al. 2021)
XGBoost	I, D, dY, C, A7, A14, A30, A60, Ac	Multiclass classification	(Zhao et al. 2022)
TPOT	I, D, dY, C, A7, A14, A30, A60, Ac	Optimized pipeline	(Olson et al. 2016)

magnitude B. This approximation simplifies the computation, making it practical for implementation while maintaining the necessary accuracy for our analysis. Detailed derivations can be found in Berti et al. (2012).

Frequentist

The Frequentist method, developed by Brunetti et al. (2010) and based on Fisher's principles from 1936, analyzes the frequency of rainfall conditions historically leading to landslides, as shown in Fig. 5c–d. It involves plotting triggering rainfalls on a logarithmic scale (Fig. 5c), using the least squares method to find a fitting line (depicted in blue), and applying Kernel Density Estimation, as per Silverman (1986), to calculate the probability density function of regression residuals with a Gaussian model. The final step is setting a rainfall threshold for a specific exceedance probability. This optimal exceedance probability was identified on the training dataset by testing all intercepts corresponding to the Gaussian distribution of rainfall events that resulted in landslides while maintaining the slope of the best-fit line (blue line in Fig. 5c) which corresponds to 50% of the distribution. Based on the optimal linear residual (delta) score, the T20 line (red line in Fig. 5d), which represents 20% of the distribution, demonstrated superior performance on the training dataset and was subsequently employed during the testing phase.

Percentile

Guzzetti et al. (2007) employed a percentile method to determine the alert threshold for landslide occurrences in Central and Southern Europe. The percentile is a statistical technique used to determine the value below which a certain percentage of observations in a dataset fall (Moore and McCabe 1993). To perform the analysis, the rainfall events that resulted in landslides are plotted in the ID chart and the duration axis is divided into bins (Fig. 5e). The data falling within each bin are then sorted in ascending order based on rainfall intensity, and the point corresponding to a given percentile (e.g., 5%) is calculated. The rainfall threshold is finally obtained by linearly fitting the ID points of the considered percentile (Fig. 5f). For our study, after evaluating different percentiles, the 25th percentile was identified as the most effective threshold based on the optimal delta score.

LDA

Linear Discriminant Analysis (LDA), a dimensionality reduction and feature selection technique, aims to find a linear combination of features that effectively separates different classes (Fisher 1936; Brown 1998; Venables and Ripley 2002). Originally proposed by Fisher, LDA uses sample mean vectors and class variances to identify the optimal direction for class separation. It achieves this by maximizing between-class variance (S_B) and minimizing within-class variance (S_W), a process visualized in Fig. 5g–h. The separation is optimized by solving a generalized eigenvalue problem, with the formula:

$$J(W) = \frac{W^T S_B W}{W^T S_W W} \tag{2}$$

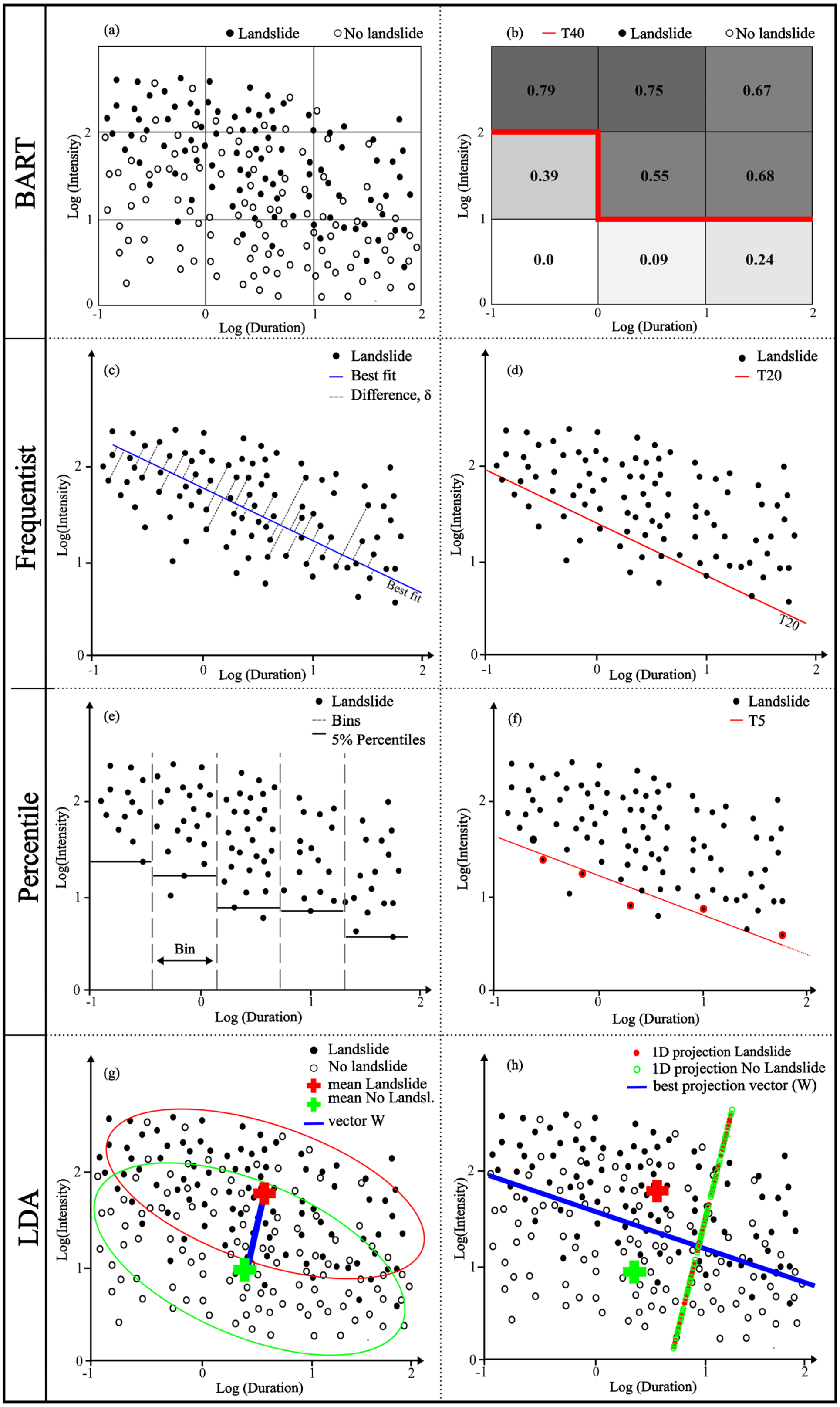


Fig. 5 Conceptual examples of conventional methods for rainfall threshold evaluation. BART: **a** scatter plot of rainfall events that did and did not result in landslides; **b** conditional probability plot of landslides for nine different combinations of intensity and duration, with the red line indicating the threshold with probability > 0.40. Frequentist: **c** rainfall events that resulted in landslides plotted with the best-fit line (blue); **d** rainfall threshold defined by shifting the regression line to the 20% exceedance probability (red line). Percentile: **e** percentile plot of rainfall events showing the 5th percentile for each bin; **f** rainfall threshold (T5) obtained from the best fit of the 5th percentile points (outlined in red). LDA: **g** mean vectors of critical and non-critical rainfall (red and green crosses) connected by the vector W (blue line); **h** optimized projection vector W (blue line) for best class separation

where W represents the projection vector. In our study, the best projection vector on W differentiates critical from non-critical rainfall, serving as the rainfall threshold in the ID plot.

Machine learning techniques

The four conventional methods were evaluated against seven algorithms from machine learning, using different numbers of input variables (Table 1).

Neural Networks

Neural Networks (NN), machine learning methods inspired by the human brain, are adept at solving complex problems (Goodfellow et al. 2016; Schmidhuber 2014). They consist of layers of interconnected neurons, learning through weight adjustments for each input (Lecun et al. 2015). In the context of landslide prediction from rainfall data, NNs assign weights to factors such as duration and intensity, using them in a prediction process that involves normalization, weight combination, and activation functions like sigmoid, to output a probability indicating landslide likelihood (Fig. 6a).

For our analysis, we implemented three NN architectures using the Keras Python library. The "shallow neural network" had four layers with 32, 16, 8, and 1 neuron; the "medium neural network" comprised five layers with 256, 64, 32, 8, and 1 neuron; and the "large neural network" also had five layers but with 512, 128, 64, 8, and 1 neuron. All used `ReLU` activation for hidden layers and `sigmoid` for the output layer. They were trained with the Adam optimizer (learning rate 1e − 4) and binary cross-entropy loss, considering class weights of 1 and 20 to balance data distribution between landslide and non-landslide classes.

Logistic regression

Logistic regression (LR) (Cox 1958; Hastie et al. 2009), is a generalized linear model using the sigmoid function to predict event probabilities. Like neural networks, LR assigns weights to input

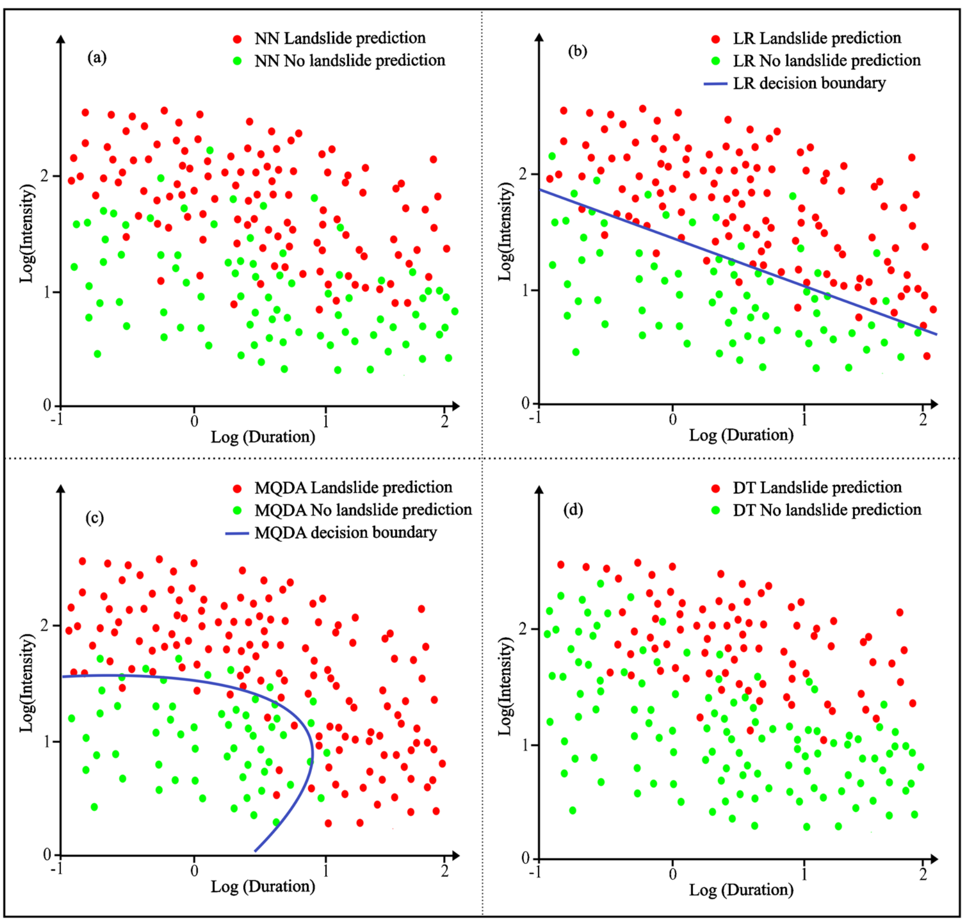


Fig. 6 Differences in outputs from machine learning algorithms for rainfall threshold evaluation. **a** Neural network: this approach uses multiclass analysis for the individual classification of each rainfall event, lacking a distinct rainfall threshold. **b** Logistic regression: the decision boundary can be depicted as a line in the 2D plot. Note that a critical rainfall event may appear below this line, influenced by other parameters not represented in this 2D visualization. **c** Multiparametric quadratic discriminant analysis (MQDA): the decision boundary can be shown as a curve in the 2D plot. Similar to logistic regression, the impact of additional parameters is not visible in this plot. **d** RF, XGBoost, and TPOT: these methods employ multiclass Decision Tree (DT) approaches, which do not have clearly defined boundaries in their outputs

features during training and calculates the probability of a binary outcome by applying the sigmoid function to the sum of weighted inputs. LR aims to optimally set weights to minimize prediction errors, following the formula:

$$p(Y = 1|x) = \frac{1}{(1 + e^{(-w^T x)})} \tag{3}$$

where $p(y=1|x)$ is the probability of the outcome being 1 given input x, with w as the weight vector.

In our study, LR was implemented for binary classification using Scikit-learn's `LogisticRegression` module with "`l2`" Ridge regularization. We chose "`lbfgs`" as a solver to handle the optimization. To address class imbalance, we set `class_weight` to {0:1, 1:18.7}, reflecting the prevalence of class 0 over class 1. The LR model's conceptual approach (Fig. 6b) was effectively employed in our analysis.

MLDA

Multiparametric linear discriminant analysis (MLDA) extends the concept of LDA (Fisher 1936; Venables and Ripley 2002). Unlike LDA, which considers only two predictor variables, MLDA accommodates multiple predictor variables, making it suitable for complex tasks like predicting rainfall-induced landslides.

In our study, we implemented MLDA using Scikit-learn's `LinearDiscriminantAnalysis` module with the "`lsqr`" solver, `priors` set to [0.4, 0.6], and a `shrinkage` parameter of 0.4 for regularization.

MQDA

Multiparametric quadratic discriminant analysis (MQDA) extends MLDA for curved decision boundaries (Hastie et al. 2009), making it valuable for modeling nonlinear rainfall thresholds (Fig. 6c). In our study, we used Scikit-learn's `QuadraticDiscriminantAnalysis` module with customized settings, including `priors` [0.3, 0.7] for class probabilities and `reg_param` 0.7 for regularization control.

Random forest

Random Forest is a machine learning algorithm that improves accuracy by combining predictions from multiple decision trees, each trained on different data subsets (bootstrapping) (Breiman 2001). Decision trees, used for classification and regression, are flowchart-like structures (Breiman et al. 1984).

In our study, we used Scikit-learn's `RandomForestClassifier` module with hyperparameter settings for optimization. For instance, `n_estimators` were set to 10 to limit the number of trees, the criterion used entropy-based information gain, and hyperparameters like `max_depth`, `min_samples_split, and min_samples_leaf` were adjusted for tree control and preventing overfitting. To address the class imbalance, `class_weight` was set to {0:1, 1:18.7} to assign higher weights to the minority class (Pedregosa et al. 2011).

RF classifies data into "landslide" or "no-landslide" without providing a specific rainfall threshold for separating critical and non-critical events in the ID plane (Fig. 6d).

XGBoost

Extreme Gradient Boosting (XGBoost) is a powerful machine-learning algorithm that improves accuracy by combining predictions from multiple decision trees through gradient boosting (Chen and Guestrin 2016). Like Random Forest, XGBoost builds multiple decision trees, but it uses gradient boosting to enhance performance. This technique corrects errors from previous trees, leading to higher accuracy and better model performance (Friedman 2001). XGBoost excels with large datasets and numerous input variables, making it a widely used algorithm for predictive modeling.

In our analysis, we employed the XGBoost algorithm using the `XGBClassifier` module from the `xgboost` Python package. We optimized the model by adjusting hyperparameters such as booster (`gbtree`), `max_depth` (6), and `scale_pos_weight` (18.7) to enhance its ability to capture complex data relationships.

TPOT

TPOT (Tree-Based Pipeline Optimization Tool) is not a model but an automated machine-learning library designed to optimize decision tree-based models (TPOTClassifier) (Olson et al. 2016). TPOT employs genetic programming for autonomous model optimization (Olson and Moore 2019). It evolves through generations of randomly generated models, converging on an optimized decision tree-based pipeline.

In our study, TPOT was configured with specific parameters: 150 `generations`, a `population_size` of 150, a custom F2 score, and `early_stop` after 15 generations. This involved evaluating 22,500 pipeline configurations. It's crucial to note that TPOT operates as an automated optimization tool, providing limited user control beyond initial parameter settings, distinguishing it from traditional machine learning algorithms.

The optimal pipeline identified by TPOT is shown below:

```
model = make_pipeline(
 StackingEstimator(estimator=RandomForestClassifier(bootstrap=True, criterion="entropy", max_features=0.1, min_samples_leaf=15, min_samples_split=3, n_estimators=100)),
 StackingEstimator(estimator=MLPClassifier(alpha=0.1, learning_rate_init=0.1)),
 SelectFwe(score_func=f_classif, alpha=0.032),
 StackingEstimator(estimator=DecisionTreeClassifier(criterion="entropy", max_depth=1, min_samples_leaf=13, min_samples_split=17)),
 GaussianNB())
```

Considerations on Bayesian network

Bayesian network machine learning (BN-ML) techniques are powerful tools for modeling complex dependencies and providing insights into variable importance (Neapolitan 2003). However, in this study, BN-ML was not included due to several reasons. Firstly, the complexity of determining the network structure and the computational cost of performing inference with large datasets make these techniques less practical for our analysis. Secondly, BN-ML techniques are less frequently represented in related literature, which limits their comparability with the more commonly used methods. Lastly, to maintain the focus and manageability of our study, we opted to use methods that are more straightforward to implement and have established performance metrics for our specific application.

Experiments and performance evaluation

In our analysis, the dataset was split into a training set and a test set with an 80/20 ratio (Alpaydin 2010). Since the dataset is highly imbalanced with a ratio of 18:1 for the target variable (landslide events), we used stratified sampling to ensure that the training and validation sets have the same proportion of critical and non-critical events. Stratified sampling randomly selects samples from each group in a way that maintains the overall class balance (Kohavi 1995).

Algorithms were then evaluated based on the four values of the confusion matrix (Bishop 2006).

Among these, the F_β -score, a generalized form of the *F*-score, was specifically chosen for its ability to weigh recall and precision according to our research priorities. The general formula for the F_β-score is given by:

$$F_\beta = \left(1+\beta^2\right) \cdot \frac{\text{Precision} \bullet \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}} \quad (4)$$

where.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})}$$

For our specific application in landslide prediction, we prioritized recall over precision due to the critical nature of not missing actual landslide events. Thus, we chose a beta of 2, leading to the F2-score, which weighs recall higher than precision:

$$F2 = \left(1+2^2\right) \cdot \frac{\text{Precision} \bullet \text{Recall}}{(2^2 \cdot \text{Precision}) + \text{Recall}} \quad (5)$$

Among the many available performance metrics, we chose the F2-score because it assigns more weight to false negatives (FN), which represent missed landslide events. Having a low number of missed alarms is of particular importance in landslide prediction. A high F2 score reflects a model's preference to issue false alarms (FP in this case) over missing alarms by significantly minimizing false negatives (FN). This scoring approach is particularly advantageous in situations where the costs of missing a true positive are much higher than those of handling false positives. The range of the F2 score is between 0 and 1, with a higher score indicating better performance.

The ROC AUC (area under the receiver operating characteristic curve) (Fawcett 2006) was used to have a graphical representation of the algorithm performance based on the comparison between true positive rate on the *y*-axis, $\text{TPR} = \text{TP}/(\text{TP}+\text{FN})$, and false positive rate on the *x*-axis, $\text{FPR} = \text{FP}/(\text{FP}+\text{TN})$. By plotting the ROC curve, we can visually assess the model performance. A model with superior performance will exhibit an ROC curve that is closer to the top-left corner of the graph, indicating higher TPR and lower FPR.

Both metrics offer valuable insights into model performance, even in the presence of class imbalance.

The `roc_curve` and `roc_auc_score` functions from the scikit-learn Python library have been used to analyze the algorithms' performance by means of the ROC curve. Traditional calibration and testing procedures were applied for conventional methods. For machine learning techniques, optimal performance was achieved through independent heuristic tuning of the algorithms' hyperparameters.

Results

Rainfall-landslide correlations

The landslide archive of the Emilia-Romagna region documents 6699 instances of landslides that occurred between 1939 and 2022 (Table 2, column 1). Out of these landslides, 4019 have a known triggering date with a daily precision or uncertainty of fewer than 7 days (column 4). The distribution of these landslides is not uniform across the eight alert areas, with around half of the occurrences happening in areas E1–E2. This unevenness is largely attributed to the unfavorable geological conditions in these regions, but it could also be due to more accurate detection of landslides made by local authorities.

The triggering rainfall for each of the 4019 documented landslides was manually identified (see "Data" section). As mentioned earlier, this process was challenging, time-consuming, and involved a considerable degree of subjectivity due to the complexity of rainfall time series. Columns 5–6-7 in Table 2 summarize the results of this expert evaluation, indicating the number of triggering rainfalls identified with low uncertainty (type 1), moderate uncertainty (type 2), and those that could not be identified with the necessary objectivity for various reasons (type 3). The events suitable for the analysis are 2718 (type 1 + type 2) and account respectively for 39% (1532 landslides) and 30% (1186 landslides) of the 4019 historical records (column 9). On the other hand, those unsuitable for the analysis (type 3) account for 31% (1186 landslides). The fact that roughly one-third of the historical data is unsuitable for analysis represents a significant limitation that will be discussed later.

Critical rainfall only represents a small fraction of the overall rainfall events. During the 83 years of analysis, an average of 2000 rainfall events with varying intensity and duration were identified in each alert area using the D3P5 algorithm (see "Data" section). Only a subset of these, ranging from 50 to 200, triggered landslides (Table 3). The percentage of critical rainfall events varied widely across the eight alert areas, ranging from 1.9% in area A1 to a maximum of 8.4% in areas E1 and E2, indicating that the dataset is heavily unbalanced. Further examination of the data reveals that while the number of rainfall events is roughly similar in all eight areas, there are significant differences in both precipitation amounts and the number of landslides (Fig. 7). For instance, area E1, which experiences the highest annual rainfall (1596 mm/year), also has the highest number of landslides (761), whereas A2, which has much lower precipitation levels (664 mm/year), has relatively fewer landslides (158). However, area E2, which receives a similar

Table 2 Overview of the Emilia-Romagna landslide dataset archive from 1939 to 2022

Alert area	Landslides triggered	Precision of the date of occurrence			Triggering rainfall				
		1 day	1 week	Less than 1 week	Type 1	Type 2	Type 3	No data	Type 1+2
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*
A1	144	49	37	86	34	26	24	2	60
A2	494	125	136	261	95	63	72	31	15
C1	491	119	257	376	147	129	100	0	276
C2	734	141	372	513	216	114	183	0	330
E1	1761	426	611	1037	465	296	276	0	761
E2	1720	382	588	970	318	307	328	17	625
G1	523	111	211	322	132	113	77	0	245
G2	832	137	317	454	125	138	187	4	263
Total	6699	1490	2529	4019	1532	1186	1247	54	2718

The table summarizes the precision of the date of landslide occurrence and the type of triggering rainfall across the eight alert areas. *Type 1* = low uncertainty; *Type 2* = moderate uncertainty; *Type 3* = high uncertainty

Table 3 Overview of rainfall events detected by the automatic algorithm D3P5 during the 83-year analysis period

Area	Mean annual rainfall (mm)	Overall rainfall events (N)	Critical rainfall events (N)	Number of triggered landslides	Percentage of critical rainfall (%)
A1	954	2034	39	60	1.9
A2	664	1980	49	158	2.5
C1	1490	2537	115	276	4.5
C2	868	2474	106	330	4.3
E1	1596	2532	212	761	8.4
E2	651	1851	156	625	8.4
G1	1250	1954	95	245	4.9
G2	743	1817	96	263	5.3
Total	1027	17,179	868	2718	5.1

The table illustrates the average number of identified rainfall events, their intensity, and duration in each alert area. Notably, only a subset of these events, resulted in triggered landslides

annual rainfall to A2 (651 mm/year), has the second-highest number of historical landslides (625). These disparities underscore the complex relationship between rainfall and landslides in the Emilia-Romagna region. This complexity is further underscored by the intensity-duration plot (Fig. 8), which displays no distinct boundary between critical and non-critical rainfall events, making it challenging for any algorithm to find a distinct separation boundary.

To further understand the impact of various predictive features on landslide occurrence, we conducted a feature importance analysis using the machine learning models in this study. As illustrated in Fig. 9, the analysis highlights that the total rainfall during the event is by far the most significant predictive feature. This was determined by calculating the feature importance based on the F2 score across 100 iterations for each method. The other parameters, while contributing positively, have relatively lower informative values, with their importance scores ranging between 0.0 and 0.1. This suggests that while total rainfall is a crucial factor in landslide prediction, other variables still play a role, albeit a smaller one.

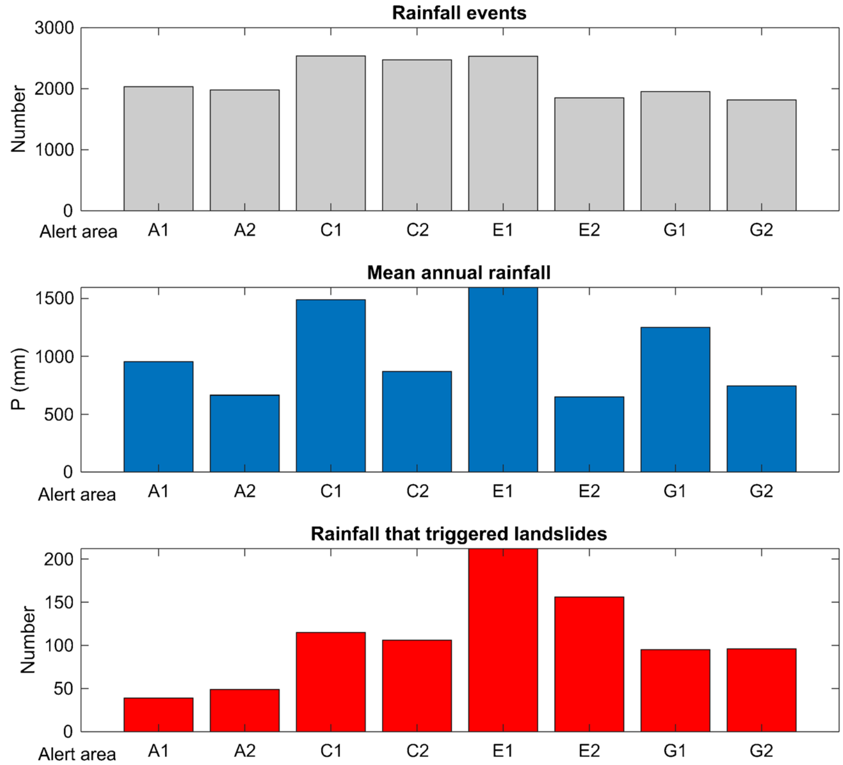


Fig. 7 Variations in the total number of rainfall events (upper), mean annual precipitation (middle), and number of triggered landslides (lower) across the eight alert areas of the Emilia-Romagna region in the period 1939–2022

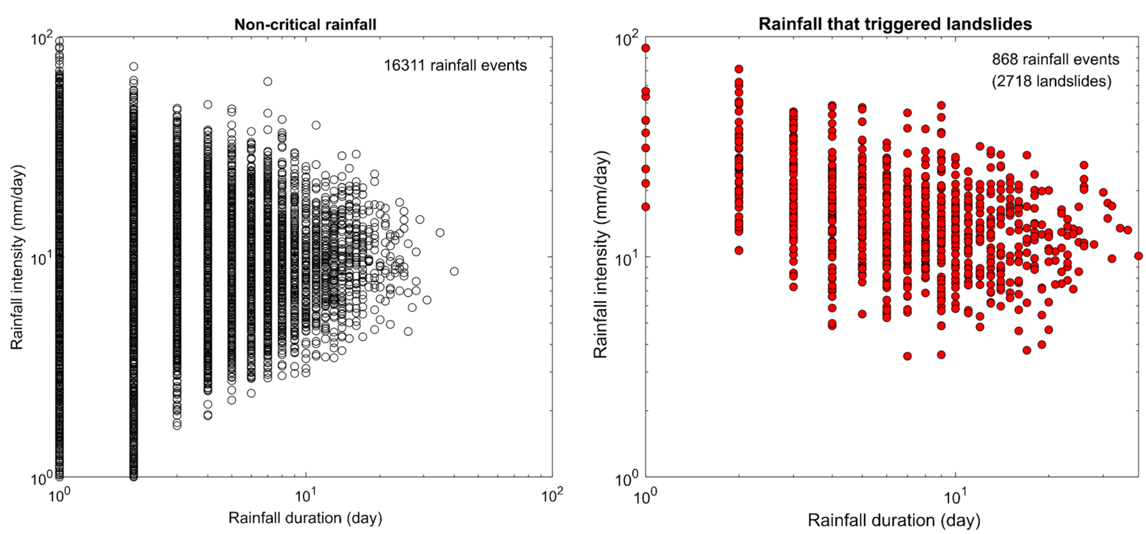


Fig. 8 Log duration-log intensity charts showing the distribution of the total rainfall events that occurred in the Emilia-Romagna region in the period 1939–2022 (left) and the critical rainfall that resulted in landslides (right)

Comparative evaluation of rainfall thresholds algorithms

In this section, we detail the results from the comparative analysis of algorithms discussed in the "Algorithms" section, focusing on their accuracy in forecasting landslides as presented in Table 4 and Fig. 10, which include confusion matrices, F2 scores, precision scores, and ROC AUC scores, detailed in the "Experiments and performance evaluation" section.

Based on the results presented in Table 4, Figs. 10 and 11, we make the following condensed observations:

Conventional methods (BART, Frequentist, Percentile, LDA).

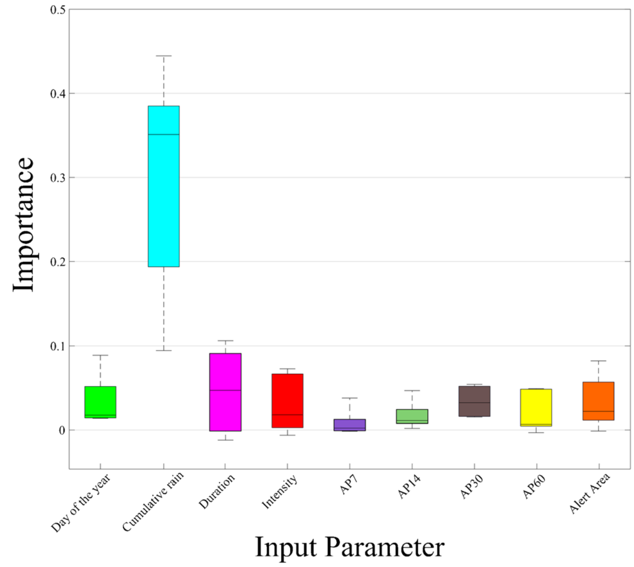


Fig. 9 Feature importance analysis for machine learning models: This boxplot illustrates the feature importance for all the machine learning methods used in this study, calculated based on the F2 score across 100 iterations for each method. The parameters AP7, AP14, AP30, and AP60 refer to the antecedent precipitation over 7, 14, 30, and 60 days, respectively

- BART demonstrates the highest accuracy among conventional methods with the highest F2 score (0.477) and ROC AUC score (0.806). It predicts a triggering rainfall event accurately once in every 5 positive predictions reflecting a low precision (0.182).
- Frequentist and Percentile methods display lower efficiency mainly due to their higher rate of false positives.
- LDA, while performing similarly to BART in terms of triggering event predictions, shows slightly better precision (0.188) with fewer false positives.

Machine learning methods (MLDA, NN, LR, MQDA, RF, XGBoost, TPOT).

- MLDA and MQDA slightly outperform the conventional LDA with higher F2 scores (0.504 for MLDA and 0.483 for MQDA) and ROC AUC scores, indicating the limited effectiveness of the other parameters.
- Neural Networks (NN) and Random Forest (RF) exhibit similar F2 and ROC AUC scores but with a significant difference in the number of false negatives handled.
- Logistic regression (LR) exhibits a good F2 performance (0.522) with the lowest number of false positives, reflecting its precision in event prediction.
- XGBoost, the top-performing method, not only has the highest F2 score (0.531) and ROC AUC score (0.848) but also maintains good precision (0.223), accurately predicting a triggering rainfall event once in every four instances.
- TPOT achieved good results with the highest F2 (0.531) and precision score (0.266) among the methods tested. However, for our purposes, its reliability is compromised due to the high number of false negatives (51).

Table 4 Performance evaluation of conventional methods (Bayes, Frequentist, Percentile, LDA) and machine learning techniques (NN, LR, MLDA, MQDA, RF, XGBoost, TPOT)

Method	Scores			Confusion matrix	
	F2	Precision	ROC AUC	TN	FP
				FN	TP
BART	0.477	0.182	0.806	2631	631
				34	140
Frequentist	0.398	0.131	0.762	2328	934
				33	141
Percentile	0.384	0.133	0.738	2433	829
				47	127
LDA	0.465	0.188	0.783	2710	552
				46	128
NN	0.500	0.197	0.815	2677	582
				34	143
LR	0.522	0.221	0.822	2775	487
				36	138
MLDA	0.504	0.198	0.822	2681	581
				31	143
MQDA	0.483	0.183	0.811	2629	633
				32	142
RF	0.500	0.193	0.823	2654	608
				29	145
XGBoost	0.531	0.223	0.848	2772	490
				33	141
TPOT	0.531	0.266	0.801	2923	339
				51	123

In summary, machine learning techniques generally outperform conventional methods as shown by their F2 scores, yet the high number of True Negatives in our unbalanced dataset skew ROC AUC scores, as illustrated in Fig. 12.

The observed low precision across models results directly from prioritizing F2 during training, which allows more false positives to reduce false negatives, reflecting our strategic focus on minimizing missed alarms.

Figure 13 provides a comparative analysis of seven machine learning techniques using 2D plots, which depict the probability that a rainfall event is classified as critical by each model within the two-dimensional space defined by rainfall intensity and duration during the training process. Given the multidimensional nature of these ML models, the plots serve only as approximations of how these variables influence classifications. This simplification allows

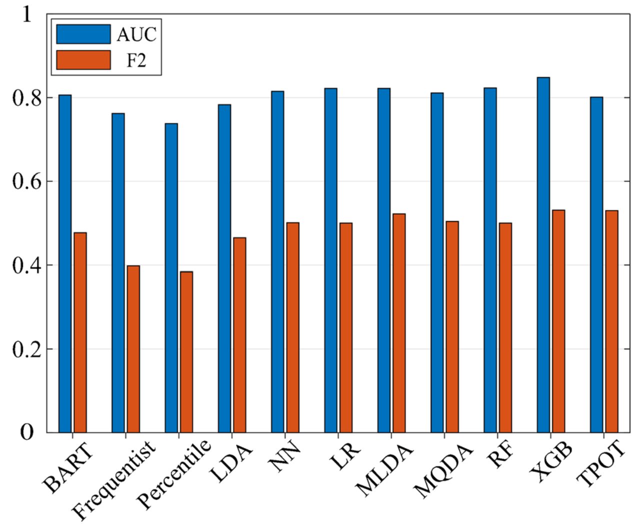


Fig. 10 Distribution of ROC AUC and F2 scores for different predictive models. The *x*-axis represents the methods, and the *y*-axis displays the corresponding AUC, in blue bars, and F2, in red bars, scores

us to visualize complex interactions in a more comprehensible form, illustrating how ML algorithms align with established threshold trends of conventional methods, albeit in a simplified manner that may not fully capture the models' capabilities. It's notable how the best-performing models, in terms of scores, exhibit distinct behaviors. TPOT, as shown in Fig. 13g, assigns a binary output and, according to its confusion matrix, adopts a higher tolerance threshold compared to conventional methods. This results in better performance but also introduces higher risk. Conversely, XGBoost (Fig. 13f) seems to maintain a more conservative profile. Both models classify a significant number of events as non-landslide which are well above conventional tolerance levels. This demonstrates how additional parameters influence decisions, leading to a decision-making process that resembles the BART threshold for short durations and high intensities, yet aligns with lower-slope thresholds for long-duration events.

Discussion

Performance and usability of machine learning methods

The results from our study demonstrate that machine learning methods, particularly XGBoost, have outperformed conventional methods such as BART, which is currently employed as a decision-support tool within the Civil Protection Agency of the Emilia-Romagna Region (Italy) (Berti et al. 2015). For instance, when comparing the performance of XGBoost with BART, a reduction in the number of false positives—from 631 in BART to 490 in XGBoost—is observed as detailed in Table 4. Moreover, the variables used in

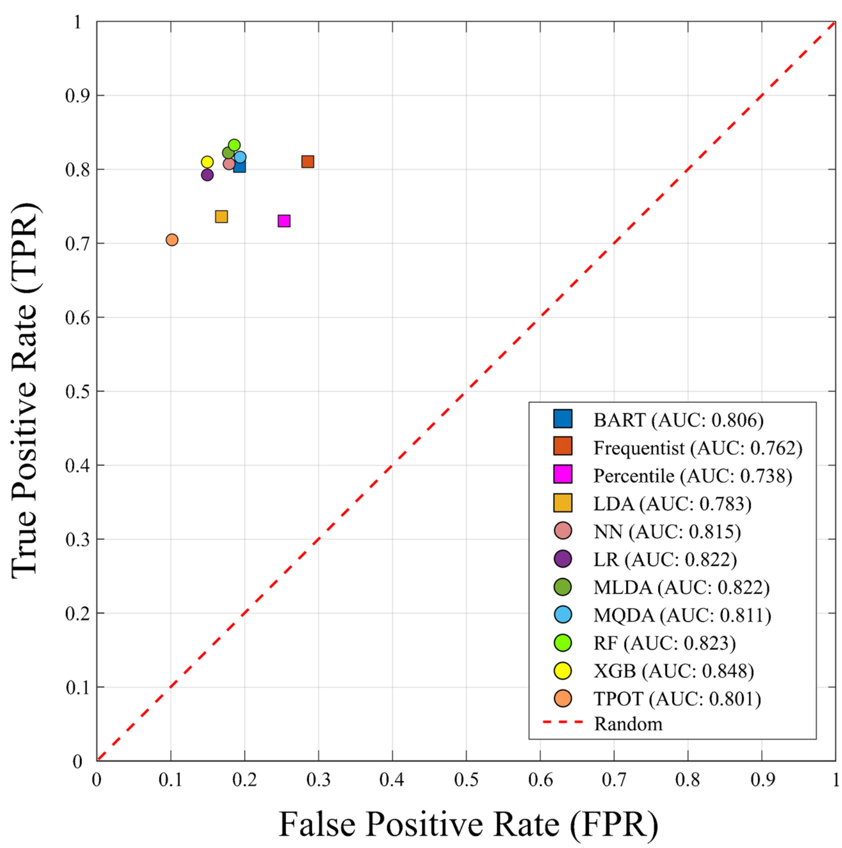


Fig. 11 Comparison of ROC AUC optimal operating points for each method

this study for the ML methods are readily available in real time, making the integration feasible and potentially beneficial.

In scenarios where warnings do not lead to costly actions, such as evacuations or road closures, as is the case with the Civil Protection Agency of the Emilia-Romagna Region, as outlined in the Regional Council Decree No. 1761 (dated November 30, 2020), a certain level of false positives may be considered acceptable. In some cases, indeed, tolerating a higher rate of False Positives might even be preferable, as it could alleviate the burden of responsibility on risk managers in the event of unforeseen events. Therefore, while our improvement in terms of reducing false positives may not be decisive, it remains an enhancement.

Beyond mathematical scores, other aspects need consideration, particularly in terms of method communication and practicality. Conventional methods, with their threshold approach in a 2D plane, offer ease of use in practical settings through a simple, visual boundary line. In contrast, while machine learning techniques are potentially more powerful, they vary significantly in their levels of interpretability. While some ML algorithms, like neural networks and deep learning, often function as "black boxes" with limited visibility into their decision-making processes, many others, such as decision trees, linear models, and Bayesian networks, are inherently more interpretable and provide greater insight into their workings. This variability in transparency can be a challenge as users may need to trust the model without direct control over the prediction mechanism. However, the level of this challenge depends on the specific ML technique employed. Although complexity and difficulty in usage might be acceptable with superior results, in our case, the performance improvement of machine learning methods over traditional techniques does not seem sufficiently clear-cut to universally justify these challenges.

Efficacy of multiparametric methods

Multiparametric models, including those based on machine learning, can capture the complexity of real-world problems, often surpassing simpler models in effectiveness. These methods, by integrating a broader range of variables, have the potential to deliver more precise and detailed predictions. They excel at identifying intricate patterns and relationships, as well as managing nonlinear interactions among variables—abilities that are particularly valuable in the complex domain of predicting rainfall-induced landslides. However, the observation that multiparametric methods do not significantly outperform simple two-parameter models in our study is somewhat unexpected.

In our analysis, the seven machine learning methods were applied using seven additional parameters beyond the conventional metrics of rainfall duration and intensity, including factors like seasonality and antecedent rainfall. Despite incorporating these extra parameters, the improvement in predictive performance was modest. To further explore this, we also ran the machine learning algorithms using only rainfall intensity and duration, paralleling the approach of traditional models. The results, as detailed in Table 5, reveal that machine learning algorithms, when constrained to this simpler framework, scored lower in both F2 and ROC AUC metrics compared to their multiparametric versions.

This outcome underscores that the true strength of machine learning lies in its capacity to process and leverage complex, multi-variable datasets. Consequently, the limited effectiveness of ML in our study implies that the extra parameters included were not particularly impactful. This underscores a crucial, albeit intuitive, principle: the quality and relevance of input parameters are more crucial than the model's complexity in enhancing the predictive accuracy of rainfall thresholds. In our specific instance, although the additional parameters used in the multiparametric methods were chosen with care, they did not significantly enhance the distinction between critical and non-critical rainfall events. On the other hand, slope stability is influenced by numerous factors beyond rainfall, such as stress history, soil moisture, subsurface flows, human activities, changes in vegetation, and others, making it a challenge to pinpoint the most relevant factors when working at a regional scale.

Machine learning methods, however, now offer unprecedented capabilities to analyze large arrays of interrelated factors efficiently. This capability encourages the collection and analysis of a broader range of factors more comprehensively than ever before, potentially leading to more insightful and accurate models in complex scenarios like rainfall-induced landslide prediction.

Enhancing the reliability of the rainfall threshold

To improve the predictive capability of the rainfall threshold, it is essential to go beyond the mere computational power of the algorithm. While more advanced methodologies, such as machine learning, offer powerful tools for predictions, their effectiveness can be compromised by several significant sources of uncertainty. These uncertainties contribute to the overlap of critical and non-critical rainfalls shown in Fig. 8, emphasizing the complexity of the prediction task. Key sources of data uncertainty include:

Initiation date ambiguity

The lack of precision in pinpointing the exact date of a landslide's onset leads to variability in the dataset, primarily because it complicates the identification of the triggering rainfall event. This ambiguity often stems from the challenge of recognizing the true beginning of the landslide, as it is typically recorded in chronicles only upon being noticed.

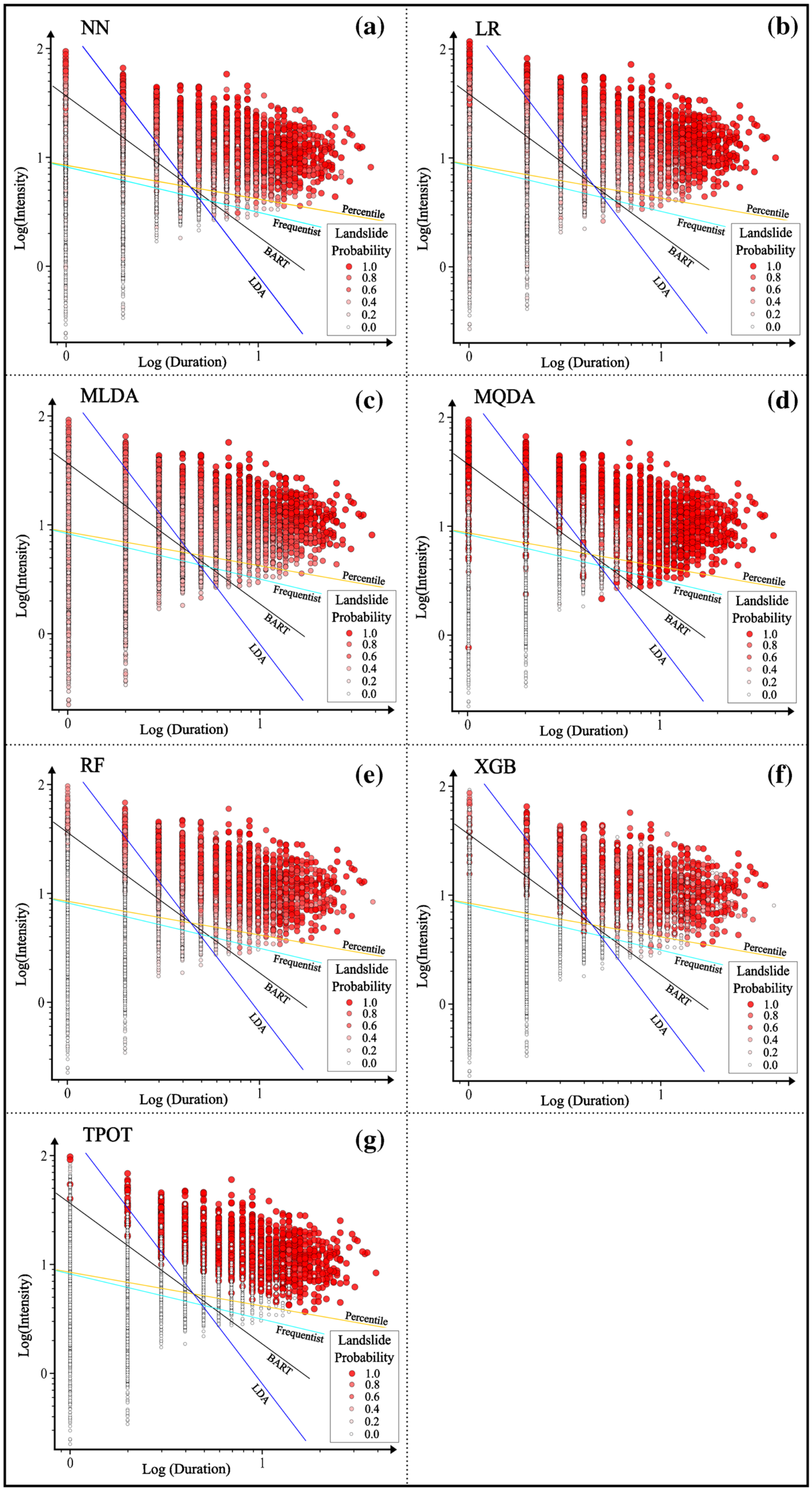
Variable response to rainfall

The way landslides react to similar rainfall events varies, influenced by their distinct geological, geomorphological, and hydrogeological attributes, along with their stress history and any human interventions. When defining rainfall thresholds at a regional scale, it is impractical to isolate and analyze each of these factors separately, leading to an inherent increase in uncertainty.

Definition of the rainfall event

The methodology employed in defining a rainfall event significantly influences its determined duration and mean intensity. Currently, the absence of established guidelines for selecting this criterion results in its definition being highly subjective. The choice of parameters, such as the specific start and end times of the rainfall, the threshold values for intensity, and the inclusion of antecedent moisture conditions, greatly affect the characterization of the event. This variability in definition can lead to diverse interpretations and assessments of the same rainfall data.

In summary, enhancing the predictive capacity of the rainfall threshold necessitates an investment in data quality. This involves a



NN
(a)
LR
(b)
MLDA
(c)
MQDA
(d)
RF
(e)
XGB
(f)
TPOT
(g)
Log(Intensity)
Log (Duration)
Percentile
Frequentist
BART
LDA
Landslide Probability
1.0
0.8
0.6
0.4
0.2
0.0

◀**Fig. 12** Comparison between conventional and Machine Learning methods in the rainfall Intensity-Duration plot. Rainfall events are denoted as points, with their color varying from white to red, indicating the probability of being classified as critical (on a scale from 0 to 1) by the different ML algorithms during their training process. Lines indicate the optimal rainfall thresholds obtained by the conventional methods. The equations for the optimal rainfall thresholds in the conventional methods are as follows: BART in black ($I = -1 \cdot \log(D) + 1.7347$), LDA in blue ($I = -1\ .8602 \cdot \log(D) + 2.3542$), Frequentist approach in light blue ($I = -0.3159 \cdot \log(D) + 1.2504$), and Percentile method in orange ($I = -0.2341 \cdot \log(D) + 1.243$)

more detailed characterization of the outcomes following a rainfall event. While historical data may be limited in this regard, future efforts could focus on improving data quality, allowing machine learning methods to fully deploy their potential.

Conclusions

This study leads to several key insights:

- Identifying rainfall thresholds in regions with intricate geological features, such as the Emilia-Romagna Region in Italy, is notably challenging. This complexity primarily arises from the difficulty in differentiating between rainfall events that are critical for landslide initiation and those that are not. These events often overlap significantly on the rainfall duration-intensity chart, making clear differentiation a complex task.
- In our analysis, we observed significant variations in the threshold equations obtained by different conventional methodologies, particularly when comparing early machine learning approaches (like BART and LDA) to traditional mathematical-statistical methods (such as Frequentist and Percentile approaches).
- Machine learning techniques, especially the eXtreme Gradient Boosting (XGBoost) algorithm, demonstrated superior performance over conventional two-parameter methods. Their enhanced effectiveness is reflected in both improved prediction scores and a significant reduction in False Alarms.
- The boost in predictive accuracy provided by machine learning methods was however relatively modest, considering that these methods utilized eight additional rainfall parameters.
- The marginal impact of these additional parameters implies that their informational value, particularly concerning antecedent rainfall over periods of 7 to 60 days, is somewhat limited when compared to more direct factors like rainfall intensity or cumulative duration.
- Progress in predictive modeling for landslide risks in the future should focus on improving the quality and specificity of data. Developing a dataset that is more detailed and informative has the potential to create models that are both more accurate and reliable, offering greater benefits than the mere application of advanced methods.

In summary, this study underscores the ongoing challenges faced in accurately predicting landslides caused by rainfall in complex geological conditions. It brings to light the continuous necessity for improvements in both the quality and volume of data, especially in the current context where advanced artificial intelligence

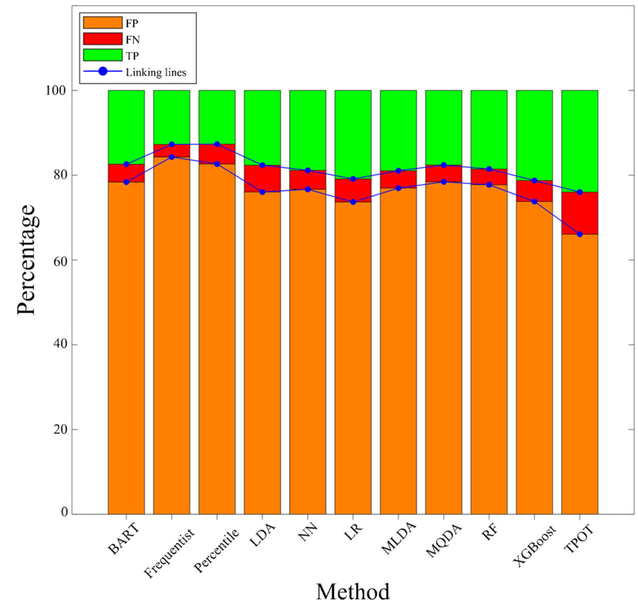


Fig. 13 Normalized confusion matrix components (excluding true negatives) comparison. This figure presents the normalized percentages of false positives (FP), false negatives (FN), and true positives (TP) for the methods investigated in this study. The connecting lines illustrate the trends and comparative performance of the methods across the different components

Table 5 Influence of the number of parameters on the prediction capability

Method	Score			
	2 parameters (I, D)		9 parameters (I, D, dY, C, A7) (A14, A30, A60, Ac)	
	F2	AUC	F2	AUC
BART	0.477	0.806	/	/
Frequentist	0.398	0.762	/	/
Percentile	0.384	0.738	/	/
LDA	0.465	0.738	/	/
NN	0.450	0.789	0.500	0.815
LR	0.468	0.790	0.522	0.822
MLDA	/	/	0.504	0.822
MQDA	0.437	0.790	0.483	0.811
RF	0.451	0.793	0.500	0.823
XGBoost	0.455	0.794	0.531	0.848
TPOT	0.451	0.749	0.531	0.801

The table compares the F2 and AUC scores obtained by the conventional methods (gray cells) with those obtained by the machine learning methods executed with different numbers of input parameters techniques have become widely accessible. This emphasis on data refinement is crucial because the effectiveness of these powerful AI methods is highly dependent on the quality and depth of the dataset they utilize. Therefore, as we advance in employing sophisticated AI tools for environmental analysis and prediction, equal emphasis must be placed on enhancing the data that forms the foundation of these predictive models.

Acknowledgements

This study was carried out within the RETURN Extended Partnership and received funding from the European Union Next-GenerationEU (National Recovery and Resilience Plan–NRRP, Mission 4, Component 2, Investment 1.3–D.D. 1243 2/8/2022, PE0000005) and from Fondazione Cassa di Risparmio in Bologna (Fondazione CARISBO).

Authors' contributions

Conceptualization: Matteo Berti and Nicola Dal Seno; data collection and material preparation: Matteo Berti and Nicola Dal Seno; methodology: Matteo Berti, Nicola Dal Seno, Davide Evangelista, and Elena Piccolomini; formal analysis and investigation: Matteo Berti, Nicola Dal Seno; writing—original draft preparation: Nicola Dal Seno; writing—review and editing: Matteo Berti, Nicola Dal Seno, Davide Evangelista and Elena Piccolomini; funding acquisition: Matteo Berti; resources: Matteo Berti; supervision: Matteo Berti. All authors read and approved the final manuscript.

Funding

Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

Data Availability

The authors are pleased to make the dataset used in this study available to interested researchers upon request. Please feel free to reach out to the corresponding author for access to the data.

Code availability

The authors are pleased to make the software code utilized in this experiment available to interested researchers upon request. Please feel free to reach out to the lead author for access to the code.

Declarations

Competing interests The authors declare that they have no competing interests.

References

Aleotti P (2004) A warning system for rainfall-induced shallow failures. Eng Geol 73:247–265. https://doi.org/10.1016/j.enggeo.2004.01.007

Aleotti P, Chowdhury R (1999) Landslide hazard assessment: summary review and new perspectives. Bull Eng Geol Environ 58:21–44. https://doi.org/10.1007/s100640050066

Alpaydin E (2010) Introduction to machine learning, 2nd edn. MIT Press, Cambridge

Berti M, Martina MLV, Franceschini S, Pignone S, Simoni A, Pizziolo M (2012) Probabilistic rainfall thresholds for landslide occurrence using a Bayesian approach. J Geophys Res: Earth Surf 117(F4). https://doi.org/10.1029/2012JF002367

Berti M, Martina MLV, Franceschini S, Pignone S, Simoni A, Pizziolo M (2015) Implementation of a probabilistic model of landslide occurrence on a civil protection alert system at regional scale. In: Lollino G et al. (ed), Engineering geology for society and territory, vol 2. Springer, Cham. https://doi.org/10.1007/978-3-319-09057-3_110

Bertolini G, Guida M, Pizziolo M (2005) Landslides in Emilia-Romagna region (Italy): strategies for hazard assessment and risk management. Landslides 2:302–312. https://doi.org/10.1007/s10346-005-0020-1

Bishop CM (2006) Pattern recognition and machine learning, 1st edn. Springer

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and Regression Trees. Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470

Brown CE (1998) Multiple logistic regression. Appl Multivar Stat Geohydrol Relat Sci 147–151:147. https://doi.org/10.1007/978-3-642-80328-4_12

Brunetti MT, Peruccacci S, Rossi M, Luciani S, Valigi D, Guzzetti F (2010) Rainfall thresholds for the possible occurrence of landslides in Italy. Nat Hazards Earth Syst Sci 10:447–458. https://doi.org/10.5194/nhess-10-447-2010

Caine N (1980) The rainfall intensity-duration control of shallow landslides and debris flows. Geogr Ann Ser A 62(1–2):23–27. https://doi.org/10.2307/520449

Calvello M, Piciullo L (2016) Assessing the performance of regional landslide early warning models: the EDuMaP method. Nat Hazard 16(1):103–122. https://doi.org/10.5194/nhessd-3-6021-2015

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

Corominas J (2000) Landslides and climate. In: Bromhead E, Dixon N, Ibsen ML (eds) Proceedings of the 8th international symposium on landslides. A. A. Balkema, Cardiff, pp 1-33

Cox DR (1958) The regression analysis of binary sequences. J R Stat Soc Ser B (Methodological) 20(2):215–242. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

Crosta GB, Frattini P (2022) Rainfall thresholds for triggering soil slips and debris flow. In: Proceedings of the 2nd EGS Plinius Conference on Mediterranean Storms: Publication CNR GNDCI, vol 2547, pp 463–487

De Vita P, Reichenbach P, Bathurst JC, Borga M, Crozier GM, Glade T, Guzzetti F, Hansen A, Wasowski J (1998) Rainfall-triggered landslides: a reference list. Environ Geol 35(2):219–233. https://doi.org/10.1007/S002540050308

Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874. https://doi.org/10.1016/J.PATREC.2005.10.010

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188. https://doi.org/10.1111/J.1469-1809.1936.TB02137.X

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451

Gariano SL, Guzzetti F (2016) Landslides in a changing climate. Earth Sci Rev 162:227–252. https://doi.org/10.1016/J.EARSCIREV.2016.08.011

Glade T, Crozier M, Smith P (2000) Applying probability determination to refine landslide-triggering rainfall thresholds using an empirical "Antecedent Daily Rainfall Model." Pure Appl Geophys 157(6–8):1059–1079. https://doi.org/10.1007/S000240050017

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge. Available online: https://www.deeplearningbook.org/. Accessed on 26 June 2007

Guzzetti F, Peruccacci S, Rossi M, Stark CP (2007) Rainfall thresholds for the initiation of landslides in central and southern Europe. Meteorol Atmos Phys 98(3):239–267. https://doi.org/10.1007/S00703-007-0262-7

Guzzetti F, Peruccacci S, Rossi M, Stark CP (2008) The rainfall intensity-duration control of shallow landslides and debris flows: an update. Landslides 5(1):3–17. https://doi.org/10.1007/S10346-007-0112-1

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Stanford University, Stanford. https://doi.org/10.1007/978-0-387-84858-7

Hu X, Wu S, Zhang G, Zheng W, Liu C, He C, ... Zhang H (2021) Landslide displacement prediction using kinematics-based random forests method: a case study in Jinping Reservoir Area, China. Eng Geol 283:105975. https://doi.org/10.1016/j.enggeo.2020.105975

Huang L, Xiang Ly (2018) Method for meteorological early warning of precipitation-induced landslides based on deep neural network. Neural Process Lett 48:1243–1260. https://doi.org/10.1007/s11063-017-9778-0

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol 2. Morgan Kaufmann Publishers Inc., San Francisco, pp 1137–1143. https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf. Accessed 12 Dec 2023

Lecun Y, Bengio Y, Hinton G (2015) Deep learning. In Nature, vol 521, Issue 7553. Nature Publishing Group, pp 436–444. https://doi.org/10.1038/nature14539

Liu Z, Gilbert G, Cepeda JM, Lysdahl AOK, Piciullo L, Hefre H, Lacasse S (2021) Modelling of shallow landslides with machine learning algorithms. Geosci Front 12(1):385–393. https://doi.org/10.1016/J.GSF.2020.04.014

Mondini AC, Guzzetti F, Melillo M (2023) Deep learning forecast of rainfall-induced shallow landslides. Nat Commun 14:2466. https://doi.org/10.1038/s41467-023-38135-y

Moore DS, McCabe GP (1993) Introduction to the practice of statistics. WH Freeman/Times Books/Henry Holt & Co., New York

Neapolitan RE (2003) Learning Bayesian networks. Prentice Hall, Upper Saddle River

Olson RS, Moore JH (2019) TPOT: a tree-based pipeline optimization tool for automating machine learning. pp. 151–160, https://doi.org/10.1007/978-3-030-05318-5_8

Olson RS, Bartley N, Urbanowicz RJ, Moore JH (2016) Evaluation of a tree-based pipeline optimization tool for automating data science. GECCO 2016 - proceedings of the 2016 genetic and evolutionary computation conference, pp 485–492. https://doi.org/10.1145/2908812.2908918

Orland E, Roering JJ, Thomas MA, Mirus BB (2020) Deep learning as a tool to forecast hydrologic response for landslide-prone hillslopes. Geophys Res Lett 47:e2020GL088731. https://doi.org/10.1029/2020GL088731

Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, Vanderplas J, Cournapeau D, Pedregosa F, Varoquaux G, Gramfort A, Thirion B, Grisel O, Dubourg V, Passos A, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Ramos-Cañón AM, Prada-Sarmiento LF, Trujillo-Vela MG et al (2016) Linear discriminant analysis to describe the relationship between rainfall and landslides in Bogotá, Colombia. Landslides 13:671–681. https://doi.org/10.1007/s10346-015-0593-2

Rossi F, Witt A, Guzzetti F, Malamud BD, Peruccacci S (2010) Analysis of historical landslide time series in the Emilia-Romagna Region, Northern Italy. Earth Surface Process Landf (print) 35:1123–1137

Saito H, Nakayama D, Matsuyama H (2010) Relationship between the initiation of a shallow landslide and rainfall intensity—duration thresholds in Japan. Geomorphology 118:167–175. https://doi.org/10.1016/j.geomorph.2009.12.016

Schmidhuber J (2014) Deep learning in neural networks: an overview. Neural Netw 61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Segoni S, Piciullo L, Gariano SL (2018) A review of the recent literature on rainfall thresholds for landslide occurrence. Landslides 15(8):1483–1501. https://doi.org/10.1007/S10346-018-0966-4

Silverman BW (1986) Density estimation for statistics and data analysis. https://doi.org/10.1201/9781315140919

Venables WN, Ripley BD (2002) Modern applied statistics with S. https://doi.org/10.1007/978-0-387-21706-2

Wang G, Sassa K (2003) Pore-pressure generation and movement of rainfall-induced landslides: effects of grain size and fine-particle content. Eng Geol 69(1–2):109–125. https://doi.org/10.1016/S0013-7952(02)00268-5

Wieczorek GF, Glade T (2007) Climatic factors influencing occurrence of debris flows. Debris-Flow Hazards Relat Phenom 325–362. https://doi.org/10.1007/3-540-27129-5_14

Zhao Y, Meng X, Qi T, Li Y, Chen G, Yue D, Qing F (2022) AI-based rainfall prediction model for debris flows. Eng Geol 296:106456. https://doi.org/10.1016/j.enggeo.2021.106456

N. Dal Seno (✉) · M. Berti
Department of Biological, Geological, and Environmental Sciences (BiGeA), University of Bologna, Via Zamboni 67, Bologna, Italy

N. Dal Seno
Email: nicola.dalseno@unibo.it

D. Evangelista · E. Piccolomini
Department of Informatics: Science and Engineering (DISI), University of Bologna, Bologna, Italy