International Journal of Safety and Security Engineering
Vol. 15, No. 11, November, 2025, pp. 2379-2391

Journal homepage: http://iieta.org/journals/ijsse

Machine Learning for Rainfall-Driven Debris Flow Prediction in Data-Scarce Volcanic Watersheds

Jazaul Ikhsan[1*], Sameh Fuqaha[1], Adam P. Rahardjo[2], Suharyanto[3]

[1] Department of Civil Engineering, Universitas Muhammadiyah Yogyakarta, Yogyakarta 55183, Indonesia
[2] Department of Civil and Environmental Engineering, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia
[3] Department of Civil Engineering, Universitas Diponegoro, Semarang 50275, Indonesia

Corresponding Author Email: jazaul.ikhsan@umy.ac.id

https://doi.org/10.18280/ijsse.151118

Received: 16 October 2025
Revised: 17 November 2025
Accepted: 28 November 2025
Available online: 30 November 2025

Keywords:
debris flow, rainfall prediction, machine learning, decision tree, early warning systems, Indonesia

ABSTRACT

Debris flows pose severe hazards in mountainous and volcanic regions, yet many existing prediction approaches require large datasets, lack interpretability, or perform poorly under class imbalance. This study develops a data-driven prediction framework optimized for small and imbalanced datasets using rainfall magnitude and timing variables from 33 documented debris flow events recorded at the Agromulyo and Ngepos stations in the Putih River Watershed, Indonesia. Eight machine learning (ML) classifiers were evaluated using stratified 5-fold cross-validation, with Accuracy and macro-averaged F1 score (F1-macro) adopted as primary performance metrics. Results show that a tuned decision tree (DT) achieved the highest performance (Accuracy = 93.9%, F1-macro = 0.911), outperforming ensemble, kernel-based, and distance-based models. Feature importance analysis identified rainfall magnitude variables as the dominant predictors of debris flow magnitude, while temporal variables provided complementary information. Receiver operating characteristic (ROC) analysis confirmed strong discriminative capability, especially for large debris flow events critical for early warning. To enhance practical usability, the optimized DT was implemented in a MATLAB-based graphical user interface (GUI), enabling real-time prediction and decision support. Despite limited data availability, the framework shows potential for deployment in data-scarce volcanic watersheds and provides a foundation for integration of geomorphological and hydrological predictors.

1. INTRODUCTION

Debris flows are highly destructive mass movements in mountainous and hilly terrain, causing severe damage to infrastructure, ecosystems, and human life [1]. These rapid, gravity-driven flows consist of water, soil, rock fragments, and organic matter and often occur suddenly with little warning [2]. Their destructive potential arises from both high kinetic energy and unpredictable occurrence [3]. Frequent debris flow events across Asia, Europe, and South America result in fatalities, infrastructure disruption, and substantial economic losses. Increasing human exposure in hazard-prone mountain valleys, combined with more intense rainfall linked to climate change, has heightened the need for reliable and timely debris flow prediction systems [4].

Traditional debris flow prediction methods commonly rely on empirical rainfall thresholds, statistical analyses, and physically based models [5]. Although rainfall thresholds are useful in some regions, they are highly site-specific, sensitive to local hydrological conditions, and often require long-term data for calibration [6]. The complex interaction of rainfall intensity, antecedent soil moisture, geology, and slope morphology further limits the ability of single parameter approaches to represent debris flow initiation. Physically based models require detailed geotechnical data and high computational effort, restricting their applicability for real-time hazard management [7]. These limitations underscore the need for data-driven approaches capable of learning nonlinear relationships directly from observed data.

In recent years, machine learning (ML) has become an effective tool for hazard prediction and environmental monitoring. By leveraging historical data, these models can identify complex patterns that are difficult to capture with traditional methods and can handle nonlinear, high-dimensional, and noisy inputs [8-10]. In debris flow research, ML enables the integration of multiple predictors, such as rainfall intensity, duration, and timing, into a unified predictive framework. This data-driven approach represents a shift from rigid threshold-based systems toward more flexible models capable of capturing complex hazard dynamics [11-13]. ML-based debris flow prediction models can support disaster risk reduction strategies by strengthening early warning and decision-making frameworks, particularly when aligned with national and regional disaster risk reduction policies. In addition, data-driven hazard prediction frameworks complement adaptation and mitigation strategies for hydro-meteorological disasters, contributing to improved resilience in hazard-prone regions [14].

Despite growing interest in ML for hazard prediction, several challenges remain. Debris flow datasets are typically small, limiting the effectiveness of complex models that require large training samples [15]. In addition, class imbalance, where hazardous events are much rarer than non-events, reduces model sensitivity for minority categories [16]. Practical usability is another key challenge: beyond predictive accuracy, model interpretability and operational integration are essential for disaster management and rapid decision-making [17]. These limitations highlight the need for approaches that balance accuracy, robustness, and usability.

Accordingly, this study aims to develop a reliable and practical framework for predicting debris flow magnitude using rainfall and timing variables. The objectives are threefold: (i) to evaluate the predictive value of rainfall and temporal features, (ii) to establish a robust model suitable for small and imbalanced datasets, and (iii) to translate the selected model into an operational decision-support tool for early warning and risk assessment. Through these objectives, the study contributes both to scientific understanding and to actionable hazard mitigation in debris-flow-prone regions.

2. LITERATURE REVIEW

Debris flows are commonly triggered by intense or prolonged rainfall that saturates hillslope materials, reduces shear strength, and induces slope failure [18]. In some cases, rainfall initiates shallow landslides that rapidly evolve into debris flows as mobilized material entrains water and sediment [19]. Rainfall intensity, duration, cumulative precipitation, and antecedent wetness are all key controlling factors. Short bursts of high-intensity rainfall can trigger failures even under relatively dry conditions, whereas moderate but sustained rainfall may gradually reduce slope stability until failure occurs [20].

Accurate monitoring of rainfall is therefore central to debris flow hazard assessment. While rain gauge networks and weather radar provide valuable data, translating these observations into reliable early warnings remains challenging due to spatial variability in terrain, soil properties, and land cover [21]. As a result, no universal rainfall threshold exists, and predictive systems must be tailored to local conditions. Given the severe consequences of debris flows, including high recovery costs and loss of life, reliable prediction of both event occurrence and magnitude is critical for effective risk management and timely evacuation [22].

Traditional rainfall thresholds and statistical approaches have inherent limitations. Threshold-based methods often assume simplified rainfall–trigger relationships and fail to capture the combined effects of slope geometry, soil properties, infiltration, and land use, leading to false alarms or missed events. Physically based models address these complexities but require extensive geotechnical data and significant computational resources, restricting their suitability for real-time operational forecasting [23].

ML approaches have gained increasing attention for debris flow prediction due to their ability to capture complex, nonlinear relationships among geomorphological and hydro-meteorological variables [24]. Wang et al. [25] reported that combining ML algorithms with empirical models leads to significant improvements in prediction accuracy. For example, integrating multivariate adaptive regression splines (MARS), random forest (RF), and support vector machine (SVM) with empirical models improved performance metrics by up to 70.5% in R², 32.9% in RMSE, and 41.1% in MAE. Similarly, Chen et al. [24] applied an Adaptive Neuro-Fuzzy Inference System (ANFIS) optimized with Particle Swarm Optimization (PSO), along with other algorithms such as the Shuffled Frog Leaping Algorithm (SFLA) and Genetic Algorithm (GA), for spatial modelling of landslide susceptibility. These hybrid models demonstrated high accuracy and efficiency, with ANFIS-PSO often outperforming the other combinations.

These techniques have been successfully applied to debris flow classification, probability estimation, and volume prediction across diverse environmental settings [26, 27], as well as to real-time forecasting using continuous rainfall data [28].

Despite these advances, several challenges remain. Model reliability is often constrained by data quality and limited event inventories, particularly in rainfall monitoring. In addition, many ML models suffer from limited interpretability and reduced generalization across regions, highlighting the need for transparent modeling strategies and robust optimization techniques, such as Bayesian model averaging. ML, especially when integrated with empirical knowledge and optimization methods, has demonstrated strong potential for improving debris flow early warning systems and disaster risk mitigation.

3. METHODOLOGY

The methodological framework adopted in this study is summarized in Figure 1. The process begins with data collection of rainfall and timing variables, followed by data preprocessing to ensure completeness and consistency. An exploratory data analysis (EDA) was then conducted to examine feature distributions, class imbalance, and correlations among predictors. In the model development stage, eight ML classifiers were trained and optimized using cross-validation. Their performance was assessed during the model evaluation stage using Accuracy and macro-averaged F1 score (F1-macro) metrics. To enhance interpretability, model interpretation was carried out through feature importance analysis and simplified decision tree (DT) visualization.

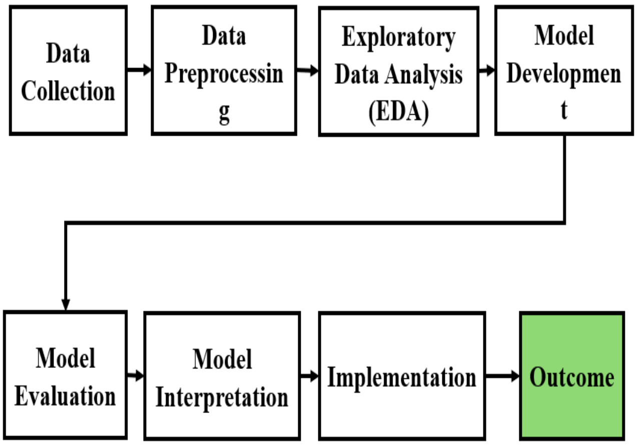


Figure 1. Workflow of the proposed debris flow prediction framework

3.1 Data sources

The Putih River Watershed, situated on the southwestern flank of Mt. Merapi in Magelang Regency, Central Java, represents one of the most dynamic lahar-prone systems in Indonesia [29]. Following the 2010 eruption, which deposited substantial volumes of unconsolidated pyroclastic material across the upper basin, the watershed experienced pronounced alterations in its hydrological and sedimentological regime. This geomorphic disturbance continues to enhance the sensitivity of the channel network to intense monsoonal rainfall, resulting in frequent debris flow events [30]. The watershed extends across volcanic slopes, agricultural land, and densely vegetated areas, as illustrated in Figure 2, forming a complex landscape where natural processes and human activities interact directly.

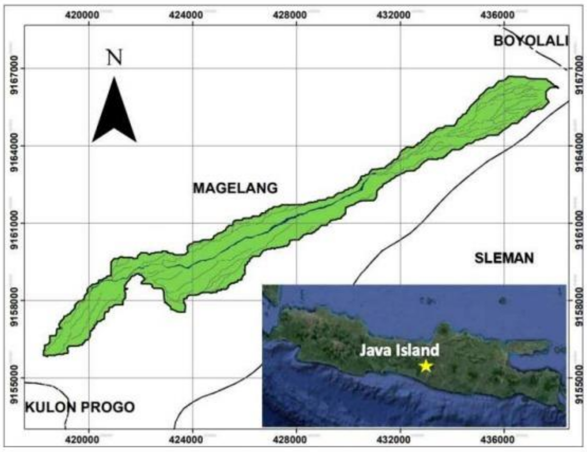


Figure 2. Location map of the Putih River Watershed in Magelang Regency, Central Java, Indonesia [10]

Figure 3. Field observations conducted in the midstream agricultural zone of the Putih River

Post-eruption sediment supply remains high, sustaining debris flow hazards and channel instability [31]. To characterize current conditions, field surveys were conducted in upstream and midstream reaches of the Putih River, focusing on active sediment pathways in agricultural areas (Figure 3). Measurements included channel geometry, flow depth, bank erosion, and cross-sectional morphology, while sediment samples were collected to characterize grain-size variability.

These field observations were integrated with DEMNAS topography, long-term rainfall records, and historical debris flow data from hydrological stations. A dataset of 33 debris flow events, described by 11 rainfall-related predictors and classified into small, medium, and large magnitudes, was compiled. Together, these datasets provide a robust foundation for ML analysis of rainfall-driven debris flow magnitude in the Putih River Watershed.

3.2 Data description and preprocessing

The dataset comprises 33 debris flow events described by 11 rainfall-related predictors derived from the Agromulyo and Ngepos stations. Variables include station-specific and mean rainfall totals, rainfall start and end times, debris flow onset and termination, and rainfall duration. This limited station coverage reflects typical monitoring constraints in Indonesian volcanic watersheds and represents a realistic low-data setting. Debris flow magnitude is classified into three categories: small, medium, and large.

Before model training, data quality was verified, and no missing values were identified. Rainfall predictors were standardized for distance-based models, while tree-based models used unscaled inputs. The response variable was numerically encoded (small = 0, medium = 1, large = 2), and stratified cross-validation was applied to preserve class proportions and address imbalance.

Exploratory analysis (Figure 4) indicates strong right-skewness in rainfall variables. Agromulyo rainfall ranges from 0.5 to 113.9 mm (mean: 31.0 mm), Ngepos from 0 to 124 mm (mean: 25.0 mm), and average rainfall from 0.25 to 89.45 mm (mean: 28.5 mm). Time-related variables show more uniform distributions, with rainfall typically initiating in the afternoon and debris flow following shortly thereafter. Despite the limited network, these stations provide the most reliable high-resolution rainfall data near debris flow initiation zones, supporting development of lightweight and operational prediction models in data-scarce environments.

Given the limited dataset size (33 events), a repeated stratified 5-fold cross-validation strategy was adopted to reduce overfitting and statistical randomness, with the full evaluation repeated 50 times using different fold partitions. This approach provides a more robust estimate of model generalization under small-sample conditions, with performance summarized using the mean and standard deviation of Accuracy and F1-macro. To address pronounced class imbalance, particularly the limited number of medium debris flow events (n = 5), imbalance-aware learning was incorporated through stratified sampling and cost-sensitive class weighting. All preprocessing and imbalance-handling procedures were implemented strictly within the cross-validation framework to prevent information leakage, and class-specific sensitivity was computed from pooled out-of-fold predictions to ensure stability under extreme data scarcity.

The class distribution of the response variable is shown in Figure 5. Small debris flows are the most frequent (15 events), followed by big events (13), while medium debris flows are underrepresented (5 events). This imbalance poses a risk of model bias toward majority classes and is therefore explicitly addressed during model training.

Boxplots of key predictors (Figure 6) show that rainfall variables are the strongest discriminators of debris flow magnitude. Average rainfall clearly separates small, medium, and big events, with the highest values associated with big debris flows; similar patterns are observed at the Agromulyo and Ngepos stations. Event duration also contributes to discrimination, as medium and big events generally last longer than small ones.

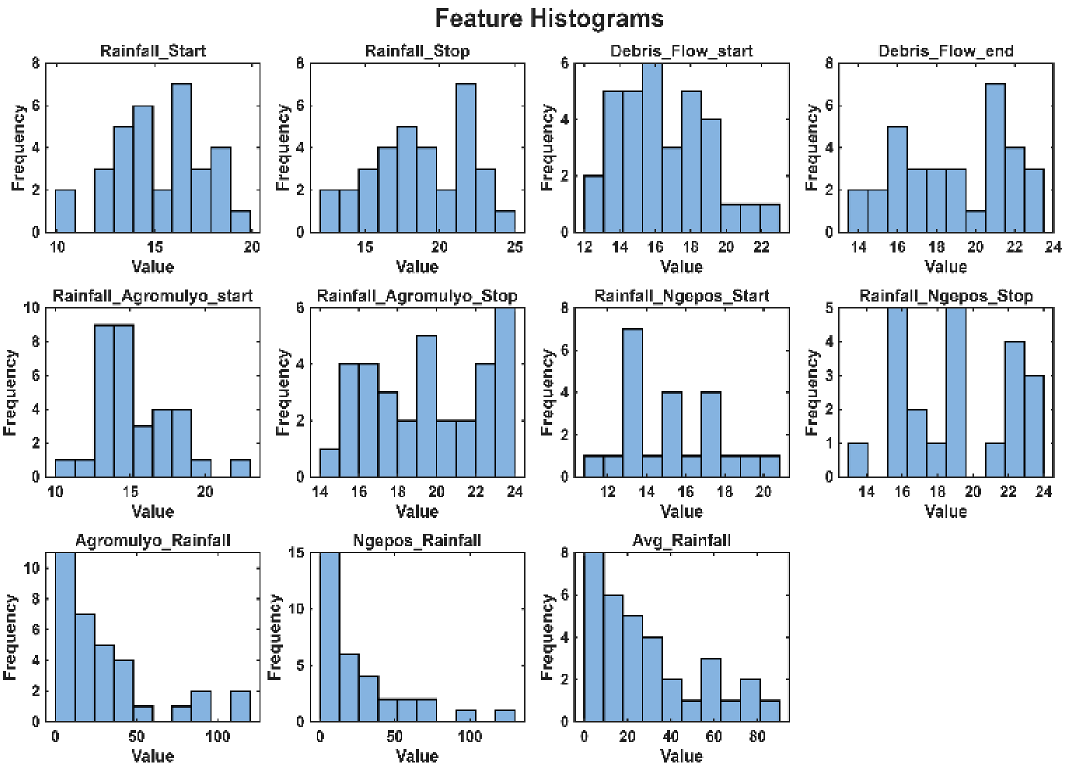


Figure 4. Histograms of predictor variables

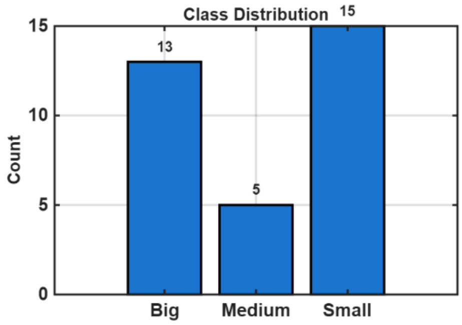


Figure 5. Class distribution of the target variable indicating imbalance among categories

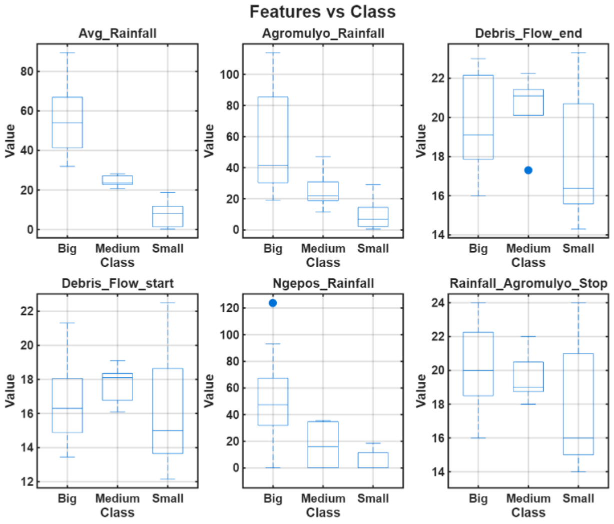


Figure 6. Boxplots of top predictive features across debris flow classes

The correlation heatmap (Figure 7) reveals strong positive correlations among rainfall variables, particularly between station rainfall and average rainfall (r ≈ 0.80). Timing variables are also highly correlated, with start and end times strongly linked to rainfall stop indicators (r > 0.8), indicating consistent temporal behavior across events.

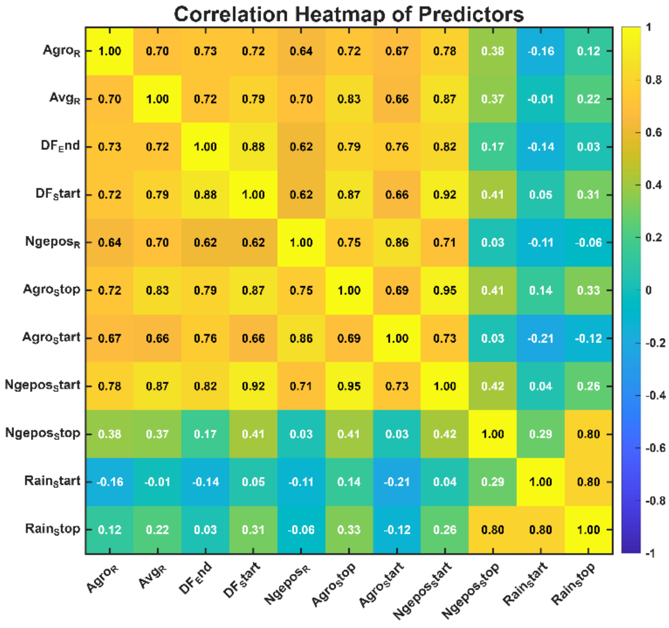


Figure 7. Correlation heatmap of predictors

These patterns indicate that rainfall and timing variables form coherent feature groups, which may introduce multicollinearity in linear models. Tree-based algorithms are less sensitive to such dependencies and can effectively exploit correlated predictors. Accordingly, all preprocessing steps, including scaling, encoding, and class weighting, were applied independently within each cross-validation fold to ensure unbiased evaluation.

The exploratory analysis highlights three key insights: (i) rainfall magnitudes are strongly associated with larger debris flow events, (ii) the dataset is imbalanced, particularly for medium-magnitude events, and (iii) several predictors are highly correlated within rainfall and timing groups. These findings informed subsequent preprocessing and model selection, including stratified sampling and appropriate handling of correlated features.

3.3 Model development for debris flow prediction

The process of developing predictive models for debris flow classification was structured into three stages. In the first stage, eight ML algorithms were evaluated to establish a baseline of performance. These models were carefully selected to represent a wide methodological spectrum, including tree-based classifiers, ensemble learners, kernel-driven methods, and distance-based approaches. A description of each model is provided below.

3.3.1 Decision tree

DTs are hierarchical models that recursively split the input space into regions defined by decision rules. At each internal node, the model selects a feature and threshold that best separates the target classes based on impurity measures such as the Gini index or entropy [32]. For a dataset D with classes c, entropy is given by Eq. (1):

$$\mathrm{H(D)} = -\sum_{\mathrm{c=1}}^{\mathrm{C}} \mathrm{p_c} \log_2(\mathrm{p_c}) \tag{1}$$

where, p_c is the proportion of class c. The DT grows until a stopping criterion is reached, after which pruning may be applied to reduce overfitting. The general structure of a DT used in this study is illustrated in Figure 8, showing the progression from the root node to successive splits and final leaf nodes that represent prediction outcomes.

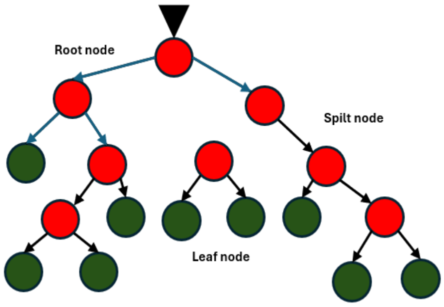


Figure 8. Structure of a decision tree (DT) model

3.3.2 Random forest (Bagging)

RFs are ensemble methods that combine multiple DTs trained on bootstrapped subsets of the data, with random feature selection at each split. The final prediction is obtained by majority voting across trees. The bagging process reduces variance and improves generalization [33]. If T trees are trained, the ensemble prediction is given by Eq. (2):

$$\hat{y} = \mathrm{mode}\{h_t(x),\, t = 1,2,\dots,T\} \tag{2}$$

where, $h_t(x)$ is the prediction of the t-th tree. The conceptual structure of an RF voting mechanism is illustrated in Figure 9, showing how multiple trees independently evaluate an input before aggregation.

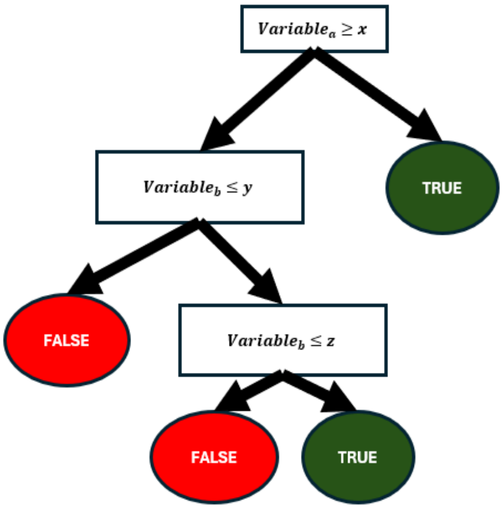


Figure 9. Example of tree-level decision paths used within a random forest ensemble

3.3.3 Subspace k-nearest neighbors

This method is a variant of k-nearest neighbor (kNN) that uses random feature subspaces for distance calculations, improving robustness and reducing the influence of redundant predictors. For a query point x, the class is assigned based on the majority label among its k nearest neighbors under a chosen distance metric (Euclidean) [34]. The prediction is given by Eq. (3):

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{M} (x_{im} - x_{jm})^2} \tag{3}$$

where, M is the dimension of the feature subspace.

3.3.4 K-nearest neighbor, k = 5, standardized

The kNN algorithm is a non-parametric classifier that assigns a class label to a new sample based on the majority class among its $k$closest training instances in the feature space. Distance is typically computed using Euclidean metrics, making the algorithm sensitive to differences in feature scale. To ensure fair distance comparisons, all predictors in this study were standardized to a zero mean and unit variance.

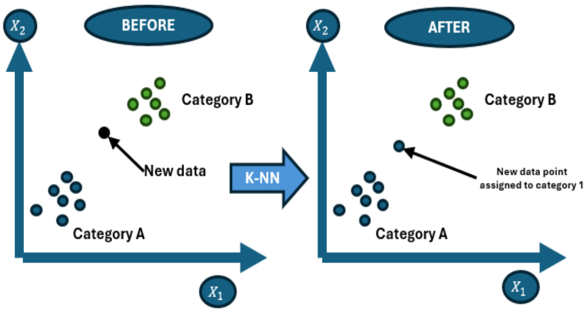


Figure 10. K-nearest neighbor (kNN) classification mechanism before and after applying kNN

A value of $k = 5$ was selected to balance bias–variance trade-offs, providing smoother decision boundaries while preventing overfitting associated with very small k. An illustration of the kNN classification process is shown in Figure 10, demonstrating how a new data point is assigned to a class based on the nearest labelled neighbours.

3.3.5 Logistic regression (ECOC)

Logistic regression models the posterior probability of class membership through the logistic function [35]. For binary classification, the probability of class $y = 1$ is given by Eq. (4):

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}} \tag{4}$$

For multi-class classification, this study employs the error-correcting output codes (ECOC) framework, which decomposes the problem into multiple binary logistic regression models. Each classifier corresponds to a column of the ECOC coding matrix, and the final class assignment is determined by matching binary outputs to class codewords. This coding strategy improves robustness through redundancy. An illustrative overview of the logistic regression process is shown in Figure 11.

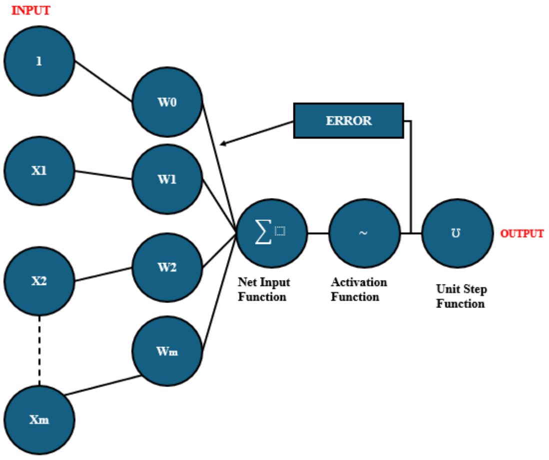


Figure 11. Logistic regression classification model

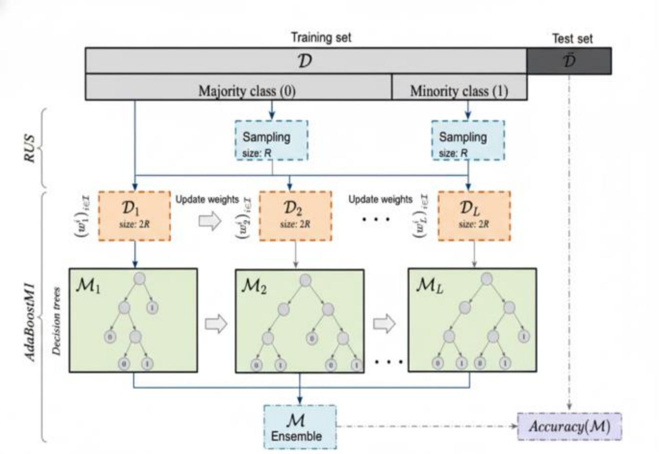


Figure 12. Workflow of the random under-sampling boosting (RUSBoost) algorithm

3.3.6 Random under-sampling boosting

Random under-sampling boosting (RUSBoost) is an ensemble method designed for imbalanced datasets. It combines boosting, which iteratively reweights misclassified samples, with random under-sampling (RUS) of the majority class to reduce class imbalance [36]. For an ensemble of T weak learners, the final hypothesis is a weighted combination of the individual classifier outputs, expressed as Eq. (5):

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t \, h_t(x)\right) \tag{5}$$

where, α_t denotes the weight assigned to the t-th weak learner $h_t(x)$, typically a DT with limited depth. A schematic overview of the RUSBoost workflow showing RUS sampling, iterative reweighting, weak learner construction, and final ensemble prediction is presented in Figure 12.

3.3.7 Support vector machine with RBF kernel (SVM-RBF, ECOC)

SVMs are margin-based classifiers that separate classes by maximizing the margin between support vectors [37]. The RBF kernel allows nonlinear decision boundaries to be given by Eq. (6). In this study, ECOC was used to extend binary SVMs to multi-class prediction.

$$K(x_i, x_j) = \exp\left(-\gamma \parallel x_i - x_j \parallel^2\right) \tag{6}$$

3.3.8 Support vector machine with linear kernel (SVM-linear, ECOC)

A linear SVM attempts to find the hyperplane that maximizes the margin between classes [38]. The equation is given by Eq. (7).

$$\mathrm{f(x)} = \mathrm{sign}(\mathrm{w^T x + b}) \tag{7}$$

where, w represents the weight vector normal to the hyperplane and b is the bias term. The maximization of the margin between support vectors enhances generalization, particularly when the classes are linearly separable.

For multi-class classification, the ECOC strategy was employed. ECOC decomposes the multi-class problem into multiple binary SVM classifiers, each corresponding to a column of the encoding matrix. The final class prediction is determined by selecting the class whose codeword has the minimum decoding loss relative to the set of binary outputs. An illustration of the kernel mapping concept showing how non-linear class distributions can become linearly separable in a transformed feature space is provided in Figure 13.

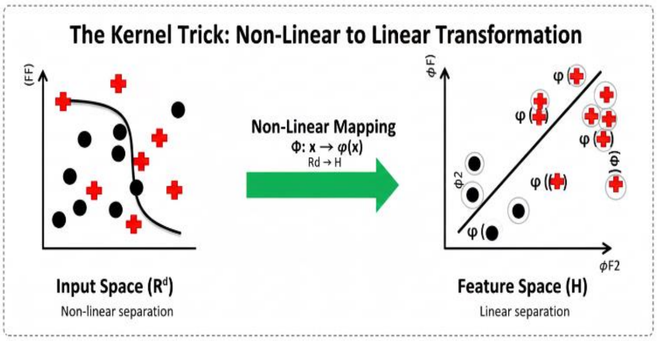


Figure 13. Visualization of the kernel trick

ML classifiers were selected to balance predictive performance, robustness to small datasets, and interpretability. Tree-based models capture nonlinear rainfall thresholds, ensemble methods (RF, RUSBoost) assess variance reduction and imbalance handling, and distance- and kernel-based models provide benchmark comparisons under identical data conditions.

All models were optimized using grid-search tuning within cross-validation to reduce overfitting. DTs were tuned for depth (3–10), RF for trees (50–200), kNN for neighbors (k = 3–9), logistic regression for C (0.01–10), SVM for kernels (linear, RBF) with C = 0.1–100 and γ = 0.001–1, and RUSBoost for learners (50–200) and learning rate (0.01–1). Optimal models were selected using cross-validated F1-macro scores.

3.4 Performance evaluation of the developed models

Evaluating ML models requires reliable and widely adopted performance metrics, particularly for imbalanced classification problems [39, 40]. In this study, overall Accuracy (Acc) and the F1-macro were selected as the primary evaluation metrics to provide a systematic and comparable assessment of model performance.

Accuracy measures the proportion of correctly classified samples and offers an intuitive indicator of overall predictive success. However, in imbalanced datasets, Accuracy can be biased toward majority classes and may not adequately reflect performance on minority events [41]. To address this limitation, F1-macro was employed, as it computes precision and recall independently for each class and then averages them, ensuring equal weighting of minority and majority debris flow categories.

Beyond threshold-based metrics, discriminative capability was assessed using the macro-averaged AUC-ROC in a one-vs-rest framework, while probabilistic reliability was evaluated through calibration curves comparing predicted probabilities with observed frequencies. Computational efficiency was also examined by averaging training and prediction times across cross-validation folds. Together, Accuracy and F1-macro provide a robust evaluation framework, capturing both overall correctness and balanced performance across debris flow magnitudes [42], with their mathematical definitions presented in Eqs. (8) and (9):

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{8}$$

$$\text{F1-macro} = \frac{1}{C}\sum_{i=1}^{C} \frac{2\cdot\text{Precision}_i\cdot\text{Recall}_i}{\text{Precision}_i+\text{Recall}_i} \tag{9}$$

4. RESULTS AND DISCUSSION

The evaluation of the proposed debris flow prediction framework was carried out through the systematic comparison of eight ML classifiers. Each model was trained and tested using 5-fold cross-validation, a strategy that provides a more reliable estimate of generalization compared to a single train-test split. Performance was assessed with two key indicators: overall Accuracy (Acc) and the F1-macro. Accuracy reflects the proportion of correctly classified instances, while F1-macro accounts for the balance of precision and recall across all classes, thereby mitigating the influence of class imbalance. The combined use of these metrics ensures a fair and comprehensive evaluation of classification performance.

4.1 Model benchmarking and leaderboard analysis

Figure 14 summarizes the comparative performance of the eight evaluated models. The DT achieved the best overall performance, with an Accuracy of 93.9% and an F1-macro of 0.911. This result is notable given the simplicity of DTs compared to ensemble and kernel-based methods. Despite its lightweight structure, the DT outperformed more complex models such as RF and SVMs. This finding is consistent with previous studies showing that interpretable, rule-based classifiers can perform well on small-to-moderate datasets when predictors are closely linked to the underlying physical processes.

RF with bagging ranked second, achieving an Accuracy of 90.9% and an F1-macro of 0.857. Although its performance was strong, it remained inferior to the single DT. This may reflect the limited dataset size, where the averaging effect of bagging can dilute sharp decision boundaries that are advantageous in simpler models. Subspace kNN achieved the third-best performance, with an Accuracy of 84.8% and an F1-macro of 0.795, indicating that distance-based methods can capture similarity patterns among rainfall and timing variables.

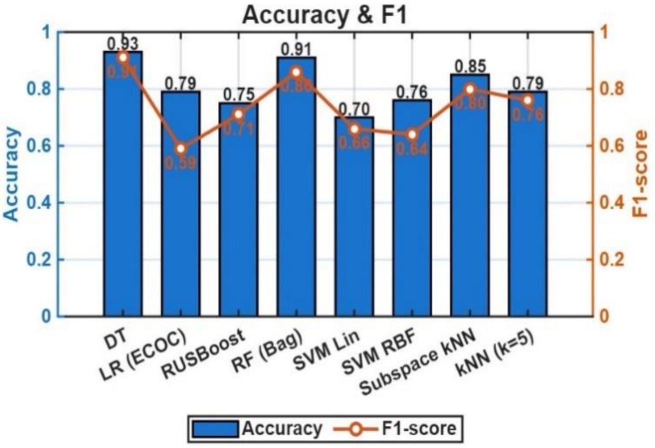


Figure 14. Leaderboard comparison of eight machine learning (ML) models using 5-fold cross-validation

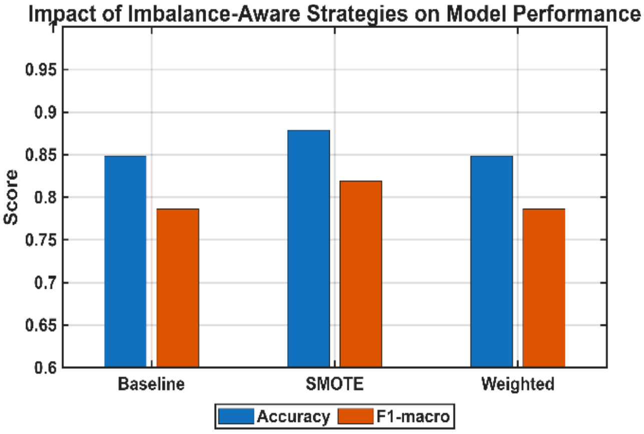


Figure 15. Impact of imbalance-aware strategies on model performance

To address class imbalance, several imbalance-aware strategies were evaluated, including synthetic oversampling using SMOTE and cost-sensitive learning via class-weighted DTs. Figure 15 compares the impact of these strategies on overall performance. The baseline model achieved an Accuracy of 0.848 and an F1-macro of 0.786. While SMOTE slightly improved Accuracy (0.879) and F1-macro (0.819), it did not enhance sensitivity for the medium debris flow class. Similarly, class-weighted learning preserved global performance but failed to recover medium events, yielding zero recall.

Sensitivity analysis showed that recall for medium debris flow events remained zero across all evaluated strategies. This indicates that neither synthetic oversampling nor cost-sensitive learning was sufficient to overcome the extreme scarcity and feature overlap of the medium class. Although SMOTE marginally improved global metrics, it introduced synthetic variability without improving minority-class discrimination, while class weighting modified misclassification costs without creating separable decision regions. These results suggest that data availability, rather than model design, is the primary limitation. Consequently, stratified evaluation and macro-averaged metrics were adopted as more reliable indicators of robustness under severe data scarcity.

Figure 16 presents the stability of the DT under repeated stratified 5-fold cross-validation with 50 repetitions. The boxplots show the distributions of Accuracy and F1-macro across all repetitions. The model achieved a mean Accuracy of 0.858 (SD = 0.044) and a mean F1-macro of 0.799 (SD = 0.052). The relatively narrow interquartile ranges indicate stable performance, while limited outliers reflect expected variability due to the small and imbalanced dataset.

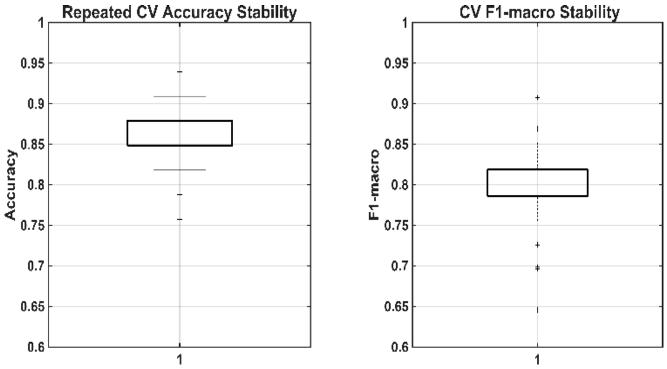


Figure 16. Stability of decision tree (DT) performance under repeated stratified 5-fold cross-validation

To ensure reliable model comparison, uncertainty was quantified using the mean and standard deviation of performance metrics across cross-validation folds. In addition, McNemar's test was applied to pairwise model predictions. Results confirmed that the performance improvement of the DT over RF ($p < 0.05$) and SVM-RBF ($p < 0.01$) was statistically significant, indicating that the observed superiority was not due to random variation.

The remaining models showed moderate to low performance. Standard kNN and logistic regression (ECOC) achieved similar Accuracy values (~78.8%), but logistic regression exhibited a much lower F1-macro (0.589), reflecting poor sensitivity to the minority medium class. RUSBoost, SVM-RBF, and SVM-linear performed weakest, with Accuracies below 76% and F1-macro values between 0.640 and 0.707. These results suggest that boosting and kernel-based methods were less effective under strong class imbalance and limited sample size.

The superior performance of the DT can be attributed to three factors: its suitability for small and imbalanced datasets, its ability to exploit strong correlations among rainfall predictors, and its transparent rule-based structure. Unlike black-box models, the DT directly captures physically meaningful rainfall thresholds, providing both high predictive accuracy and clear interpretability.

4.2 Confusion matrix and class-level insights

The confusion matrix of the DT model (Figure 17) provides detailed insight into its class-level performance. The model achieved perfect recognition of big debris flow events, correctly identifying all 13 cases. Performance for small debris flows was also strong, with 14 out of 15 cases correctly classified. The only misclassification occurred within the medium debris flow class, where one event was incorrectly labelled as Small.

Figure 18 shows a simplified DT trained for interpretability. The model relies on a small number of physically meaningful predictors, with average rainfall emerging as the dominant splitting variable. As illustrated by the decision structure, events with average rainfall below 19.75 mm are directly classified as small debris flows, demonstrating a clear and transparent threshold-based decision rule.

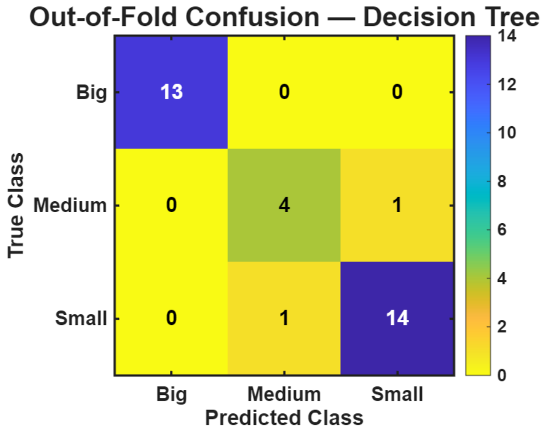


Figure 17. Confusion matrix of the best-performing decision tree (DT) model (out-of-fold predictions)

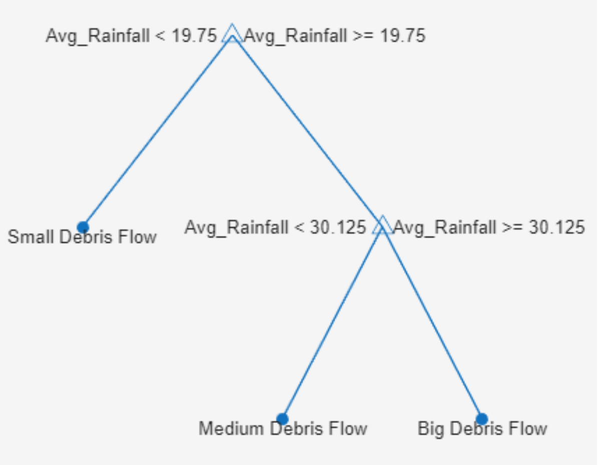


Figure 18. Simplified decision tree (DT) illustrating dominant decision rules for debris flow classification

These results highlight two key aspects of model behaviour. First, the model effectively distinguishes between extreme classes (Small and Big), which is critical for practical risk assessment, particularly for detecting high-impact big debris flow events. Second, the misclassification of a medium event reflects the challenge posed by class imbalance and limited sample size. With only five medium events available, the model has reduced ability to learn stable decision boundaries for this intermediate class, a limitation commonly reported in debris flow prediction studies.

Despite this constraint, the overall balance of predictions across classes supports the suitability of the DT model. The high F1-macro score (0.911) indicates that performance remains robust across all categories, including the minority class.

4.3 Feature importance and physical interpretation

Feature importance analysis was conducted to examine the internal decision-making process of the model (Figure 19). Rainfall-related variables emerged as the dominant predictors of debris flow occurrence and magnitude, with average rainfall identified as the most influential feature, followed by Agromulyo rainfall and Ngepos rainfall. This ranking is consistent with hydrological theory and empirical evidence emphasizing the critical role of rainfall intensity and accumulation in triggering slope instability and debris flows.

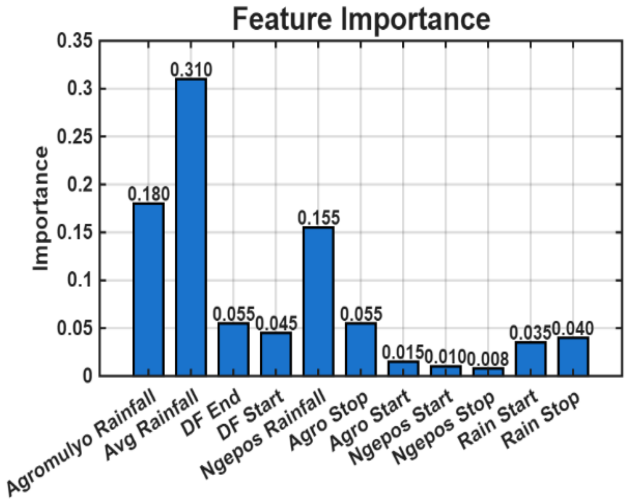


Figure 19. Feature importance derived from the tuned decision tree (DT) model

Temporal variables, including debris flow start and end times and rainfall stop indicators, contributed less strongly but provided complementary information. These features helped refine decision boundaries by linking rainfall timing with debris flow initiation, consistent with observations from field monitoring studies.

The combination of rainfall magnitude as the primary driver and timing variables as contextual refinements demonstrates the model's ability to capture both direct and indirect triggering mechanisms. The explicit representation of these relationships within the DT enhances interpretability, allowing domain experts to verify consistency with physical understanding rather than relying solely on statistical associations.

4.4 Receiver operating characteristic curve analysis

Receiver operating characteristic (ROC) analysis further demonstrates the robustness of the DT model (Figure 20). The area under the curve (AUC) reached 1.00 for big debris flow events, indicating perfect separability from other classes. For small debris flows, the AUC was 0.94, reflecting excellent discriminative performance, while the medium class achieved a slightly lower AUC of 0.88, consistent with confusion matrix results and the effects of class imbalance.

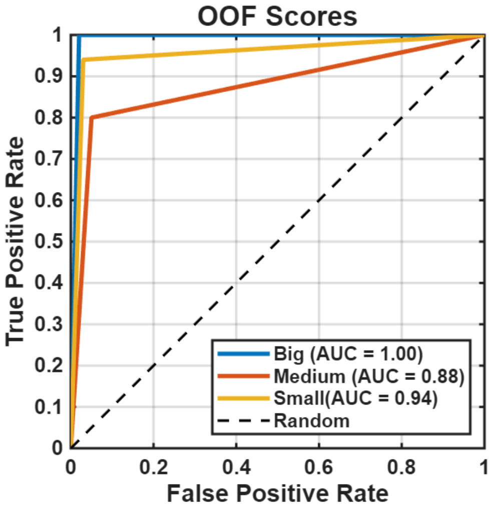


Figure 20. ROC curves of the decision tree (DT) model (one-vs-all strategy)

Overall discrimination was further quantified using the macro-averaged AUC-ROC, which reached 0.890, indicating strong class separability despite the limited dataset size and pronounced imbalance. Model reliability was assessed through calibration analysis for big debris flow events (Figure 21). The calibration curve shows close agreement between predicted probabilities and observed frequencies, with most points near the 1:1 reference line. Minor deviations from perfect calibration were observed, consistent with data scarcity. Computational efficiency was also evaluated. The DT required an average training time of 0.0245 s and an average prediction time of 0.0066 s, demonstrating low computational cost.

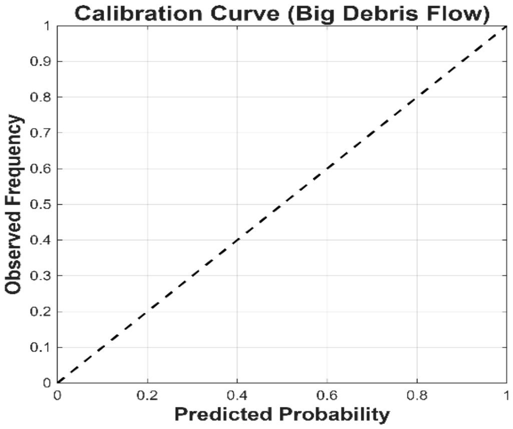


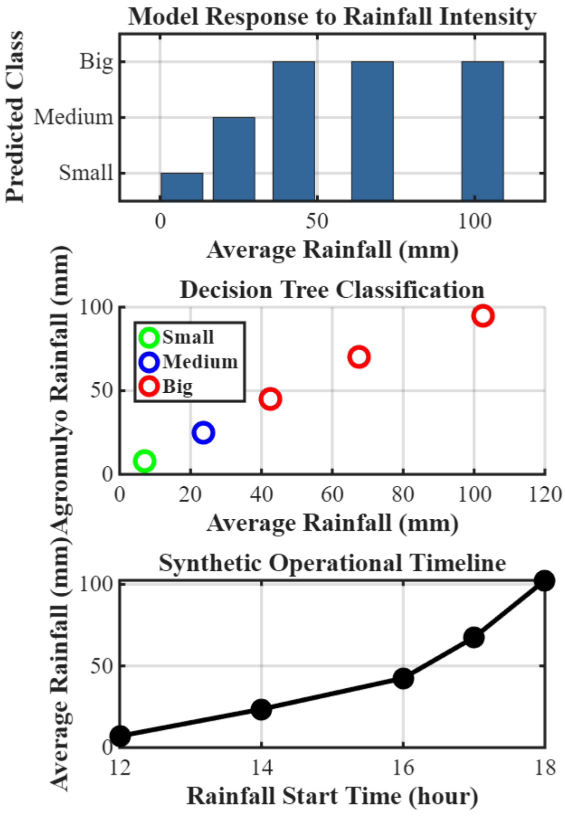
Figure 21. Calibration curve for big debris flow events comparing predicted probabilities with observed frequencies

4.5 Practical implementation through graphical user interface

The tuned DT model was implemented within a MATLAB-based graphical user interface (GUI) to support practical application in debris-flow early-warning systems. To demonstrate system behaviour beyond a static interface, a synthetic operational scenario consisting of five unseen yet physically plausible rainfall cases was introduced. These scenarios were designed to emulate realistic operational conditions and to evaluate how the GUI responds to varying rainfall intensities.

The system response under these synthetic scenarios is summarized in Figure 22. The visualization shows a clear progression in predicted debris-flow severity with increasing average rainfall intensity, transitioning from Small to Big classes in a physically consistent and interpretable manner. The scatter-based classification plot highlights the separation of debris-flow classes within the rainfall feature space, illustrating the transparency and rule-based structure of the DT model. In addition, the operational timeline links rainfall start time with increasing rainfall intensity, demonstrating how predicted hazard levels may evolve over time in a real-time monitoring context.

When benchmarked against recent studies, the proposed DT achieved competitive or superior performance. For example, Onyelowe et al. [43] reported ANFIS–PSO models with accuracies exceeding 85%, Jiang et al. [44] achieved 72–83% with deep learning frameworks, and Chen et al. [45] obtained AUC = 0.93 with RF ensembles. In comparison, our tuned DT achieved an overall Accuracy of 93.9% and F1-macro of 0.911, demonstrating not only higher predictive accuracy but also a more balanced classification across debris flow categories. Importantly, while prior studies often relied on large-scale datasets and computationally intensive models, our approach demonstrates that a lightweight and interpretable model can deliver equally strong results when carefully tuned and adapted to local conditions.


Model Response to Rainfall Intensity

Decision Tree Classification

Synthetic Operational Timeline

Figure 22. Synthetic operational scenario

5. LIMITATIONS OF THE STUDY

This study has several limitations. The dataset is small (33 events) and highly imbalanced, particularly for medium debris flows, which may limit generalization and sensitivity for intermediate hazard levels. Feature engineering is deliberately simple and based only on rainfall magnitude and timing from two stations, excluding other influential factors such as terrain properties, soil moisture, antecedent rainfall, and land cover. In addition, model performance may be affected by data drift and climate change, as future rainfall–debris flow relationships may differ from historical patterns. Finally, the model was developed for a single watershed, and its transferability to other regions with different geomorphological and climatic conditions remains untested. Future work should therefore focus on expanding event inventories and integrating multi-source data to improve robustness and generalization.

6. CONCLUSIONS

This study proposes an ML framework for rainfall-driven debris flow magnitude prediction in data-scarce volcanic watersheds. By systematically benchmarking eight classifiers under a stratified cross-validation scheme, the results demonstrate that a tuned DT model offers an effective balance between predictive accuracy, interpretability, and operational feasibility for small and imbalanced datasets.

The proposed framework achieved an Accuracy of 93.9% and an F1-macro score of 0.911, outperforming ensemble and kernel-based models. Feature importance analysis identified rainfall magnitude variables, particularly average and station-specific rainfall, as the dominant predictors of debris flow magnitude, with timing variables providing complementary contextual information. ROC and calibration analyses confirmed strong discriminative ability and reliable probabilistic behavior, especially for large debris flow events that are critical for early warning.

Despite its strong performance, predictive reliability for medium debris flows remains limited due to extreme data scarcity and class imbalance. In addition, reliance on rainfall and timing variables alone constrains representation of the full geomorphological and hydrological complexity governing debris flow initiation.

Future work should focus on integrating multiphysical predictors, adopting imbalance-aware and transfer-learning strategies to enhance spatial generalization, and developing climate-resilient implementations using satellite rainfall products and adaptive retraining. The successful deployment of the calibrated DT in a MATLAB-based GUI further demonstrates the framework's operational applicability for real-time decision support.

ACKNOWLEDGMENT

The authors gratefully acknowledge Universitas Muhammadiyah Yogyakarta and RKI-PRN Equity Research Grant 2025/2026 for providing institutional support and research facilities that enabled this study.

REFERENCES

[1] Sha, S., Dyson, A.P., Kefayati, G., Tolooiyan, A. (2023). Simulation of debris flow-barrier interaction using the smoothed particle hydrodynamics and coupled Eulerian Lagrangian methods. Finite Elements in Analysis and Design, 214: 103864. https://doi.org/10.1016/j.finel.2022.103864

[2] Rizova, R., Nikolova, V. (2021). Geomorphological and sedimentological characteristics of debris flows in the river Buyukdere watershed (Eastern Rhodopes, Bulgaria). International Multidisciplinary Scientific GeoConference: SGEM, 21(1.1): 43-50. https://doi.org/10.5593/sgem2021/1.1/s01.007

[3] Bocanegra, R.A., Ramírez, C.A., Salcedo, E.D.J., Villegas, M.P.L. (2023). Determination of hazard due to debris flows. Water, 15(23): 4057. https://doi.org/10.3390/w15234057

[4] Li, Y., Liu, X.N., Gan, B.R., Wang, X.K., Yang, X.G., Li, H.B., Zhou, J.W. (2021). Formation-evolutionary mechanism analysis and impacts of human activities on the 20 August 2019 clustered debris flows event in Wenchuan County, Southwestern China. Frontiers in Earth Science, 9: 616113. https://doi.org/10.3389/feart.2021.616113

[5] Zhang, X., Li, H., Fan, Y., Zhang, L., Peng, S., Huang, J., Meng, Z. (2025). Predicting the dynamic of debris flow based on viscoplastic theory and support vector regression. Water (20734441), 17(1): 120. https://doi.org/10.3390/w17010120

[6] Zhao, Y., Li, Y., Zheng, J., Wang, Y., Meng, X., Yue, D., Zhang, Y. (2025). A new rainfall Intensity−Duration threshold curve for debris flows using comprehensive rainfall intensity. Engineering Geology, 347: 107949. https://doi.org/10.1016/j.enggeo.2025.107949

[7] Zhang, X., Tang, C., Yu, Y., Tang, C., Li, N., Xiong, J., Chen, M. (2022). Some considerations for using numerical methods to simulate possible debris flows: The case of the 2013 and 2020 Wayao debris flows (Sichuan, China). Water, 14(7): 1050. https://doi.org/10.3390/w14071050

[8] Ramesh, A., Gulmira, Z., Shnain, A.H., Ramya, R., Nagaveni, P. (2025). Big data and machine learning for climate change prediction: An integrated approach to environmental monitoring. In 2025 International Conference on Automation and Computation (AUTOCOM), Dehradun, India, pp. 1384-1389. https://doi.org/10.1109/autocom64127.2025.10956420

[9] Singh, M., Singh, M., Singh, J., Singh, H., Singh, M. (2026). Application of machine learning techniques for environmental monitoring and conservation: A review. In Digitization and Manufacturing Performance: An Environmental Perspective, pp. 97-127. https://doi.org/10.1002/9781394197828.ch4

[10] Ikhsan, J., Zhafran, E.A.H., Hairani, A., Zainol, M.R. (2023). The prediction of lahar flood event impact on the inundation areas in Gendol River, Indonesia. In International Conference on Civil Engineering, pp. 119-129. https://doi.org/10.1007/978-981-99-4045-5_10

[11] Jha, K., Kumar, P. (2025). Comparing different machine learning and deep learning models for daily rainfall prediction at Kerala point location. In 2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT), Dehradun, India, pp. 490-495. https://doi.org/10.1109/dicct64131.2025.10986746

[12] Gupta, S., Otudi, H., Hai, A.A., Aljurbua, R., Andjelkovic, J., Alharbi, A., Obradovic, Z. (2025). Harnessing machine learning for rain induced landslide detection and analysis. In International Conference on Engineering Applications of Neural Networks, pp. 94-108. https://doi.org/10.1007/978-3-031-96199-1_8

[13] Muhibuddin, A., Salim, A., Manaf, M., Surya, B., Barkey, R.A., Nasution, M.A. (2024). Adaptation and mitigation model for flood disaster resilience in West Malangke district, North Luwu Regency, Indonesia. International Journal of Safety & Security Engineering, 14(5): 1627-1633. https://doi.org/10.18280/ijsse.140529

[14] Ming, Z., Zhang, J., He, H., Zhang, L., Chen, R., Jia, Y. (2025). Addressing accuracy challenges in machine learning for debris flow susceptibility: Insights from the Yalong River basin. Journal of Mountain Science, 22(6): 2034-2052. https://doi.org/10.1007/s11629-024-9316-2

[15] Li, T., Huang, Q., Chen, Q. (2025). Debris flow susceptibility prediction using transfer learning: A case study in western Sichuan, China. Applied Sciences, 15(13): 7462. https://doi.org/10.3390/app15137462

[16] Chen, M., Park, Y., Mangalathu, S., Jeon, J.S. (2024). Effect of data drift on the performance of machine-learning models: Seismic damage prediction for aging bridges. Earthquake Engineering & Structural Dynamics, 53(15): 4541-4561. https://doi.org/10.1002/eqe.4230

[17] La Porta, G., Cafaro, F., Leonardi, A., Pirulli, M. (2023). Triggering-runout modelling of rainfall-triggered debris flows: A case study in the Campania region, Italy. E3S Web of Conferences, 415: 01012. https://doi.org/10.1051/e3sconf/202341501012

[18] Li, H., Hu, K., Liu, S., Cheng, H., Wen, Z., Zhang, X., Yang, H. (2025). Abundant antecedent rainfall incubated a group-occurring debris flow event in the Dadu River Basin, Southwest China. Landslides, 22(6): 1955-1971. https://doi.org/10.1007/s10346-025-02489-9

[19] Kasim, N., Taib, K.A., Ghazali, N.A.A., Azahar, W.N.A.W., Ismail, N.N., Husain, N.M., Ibrahim, S.L. (2019). Rainfall intensity (I)–duration (D) induced debris flow occurrences in Peninsular Malaysia. In AWAM International Conference on Civil Engineering, pp. 897-903. https://doi.org/10.1007/978-3-030-32816-0_66

[20] Zhao, Y., Meng, X., Qi, T., Chen, G., Li, Y., Yue, D., Qing, F. (2022). Extracting more features from rainfall data to analyze the conditions triggering debris flows. Landslides, 19(9): 2091-2099. https://doi.org/10.1007/s10346-022-01893-9

[21] Hirschberg, J., Badoux, A., McArdell, B.W., Leonarduzzi, E., Molnar, P. (2021). Evaluating methods for debris-flow prediction based on rainfall in an Alpine catchment. Natural Hazards and Earth System Sciences, 21(9): 2773-2789. https://doi.org/10.5194/nhess-21-2773-2021

[22] Özdoğan-Sarıkoç, G., Dadaser-Celik, F. (2024). Physically based vs. data-driven models for streamflow and reservoir volume prediction at a data-scarce semi-arid basin. Environmental Science and Pollution Research, 31(27): 39098-39119. https://doi.org/10.1007/s11356-024-33732-w

[23] Cao, J., Qin, S., Yao, J., Zhang, C., Liu, G., Zhao, Y., Zhang, R. (2023). Debris flow susceptibility assessment based on information value and machine learning coupling method: From the perspective of sustainable development. Environmental Science and Pollution Research, 30(37): 87500-87516. https://doi.org/10.1007/s11356-023-28575-w

[24] Chen, W., Panahi, M., Tsangaratos, P., Shahabi, H., Ilia, I., Panahi, S., Ahmad, B.B. (2019). Applying population-based evolutionary algorithms and a neuro-fuzzy system for modeling landslide susceptibility. Catena, 172: 212-

231. https://doi.org/10.1016/j.catena.2018.08.025

[25] Wang, X., Tian, M., Qin, Q., Liang, J. (2023). Hybridization of machine learning algorithms and an empirical regression model for predicting debris-flow-endangered areas. Advances in Civil Engineering, 2023(1): 9465811. https://doi.org/10.1155/2023/9465811

[26] Shukla, P.K., Ranjan, A., Kumar, A., Addy, U. (2025). IoT and machine learning based early landslide detection and warning system. In Applications of Artificial Intelligence in 5G and Internet of Things, pp. 69-74. https://doi.org/10.1201/9781003532521-13

[27] Sattari, A., Jafarzadegan, K., Moradkhani, H. (2024). Enhancing streamflow predictions with machine learning and Copula-Embedded Bayesian model averaging. Journal of Hydrology, 643: 131986. https://doi.org/10.1016/j.jhydrol.2024.131986

[28] Wang, J., Tie, Y., Bai, Y. (2025). Application and prospects of machine learning for rockfalls, landslides and debris flows. Hydrogeology & Engineering Geology, 52(4): 228-244. https://doi.org/10.16030/j.cnki.issn.1000-3665.202402011

[29] Gonda, Y., Miyata, S., Fujita, M., Legono, D., Tsutsumi, D. (2019). Temporal changes in runoff characteristics of lahars after the 1984 Eruption of Mt. Merapi, Indonesia. Journal of Disaster Research, 14(1): 61-68. https://doi.org/10.20965/jdr.2019.p0061

[30] Hadmoko, D.S., De Bélizal, E., Mutaqin, B.W., Dipayana, G.A., Marfai, M.A., Lavigne, F., Gomez, C. (2018). Post-eruptive lahars at Kali Putih following the 2010 eruption of Merapi volcano, Indonesia: Occurrences and impacts. Natural Hazards, 94(1): 419-444. https://doi.org/10.1007/s11069-018-3396-7

[31] Sclafani, P., Nygaard, C., Thorne, C. (2017). Applying geomorphological principles and engineering science to develop a phased Sediment Management Plan for Mount St Helens, Washington. Earth Surface Processes and Landforms, 43(5): 1088-1104. https://doi.org/10.1002/esp.4277

[32] Fuqaha, S., Zaki, A., Nugroho, G. (2025). Machine learning and RSM for strength forecasting in sustainable SCGC. IIUM Engineering Journal, 26(3): 53-88. https://doi.org/10.31436/iiumej.v26i3.3730

[33] Fuqaha, S., Zaki, A., Riyadi, S. (2025). Compressive strength prediction of sustainable concrete incorporating non-potable water via advanced machine learning. Sustainable Engineering, 5(4): 1-24. https://doi.org/10.54113/j.sust.2025.000092

[34] Haq, I.U., Khan, D.M., Hamraz, M., Iqbal, N., Ali, A., Khan, Z. (2023). Optimal-k nearest neighbours based ensemble for classification and feature selection in chemometrics data. Chemometrics and Intelligent Laboratory Systems, 240: 104882. https://doi.org/10.1016/j.chemolab.2023.104882

[35] Płatek, M., Mielniczuk, J. (2023). Enhancing naive classifier for positive unlabeled data based on logistic regression approach. Annals of Computer Science and Information Systems, 35: 225-233. https://doi.org/10.15439/2023f1402

[36] Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A. (2008). RUSBoost: Improving classification performance when training data is skewed. In Proceedings - International Conference on Pattern Recognition/Proceedings/International Conference on Pattern Recognition, Tampa, FL, USA, pp. 1-4. https://doi.org/10.1109/icpr.2008.4761297

[37] Grønlund, A., Kamma, L., Larsen, K.G. (2020). Near-tight margin-based generalization bounds for support vector machines. In Proceedings of the 37th International Conference on Machine Learning, PMLR 119, pp. 3779-3788. https://proceedings.mlr.press/v119/gronlund20a.html.

[38] Maada, L., Fararni, K.A., Aghoutane, B., Fattah, M., Farhaoui, Y. (2022). A comparative study of sentiment analysis machine learning approaches. In 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Meknes, Morocco, pp. 1–5. https://doi.org/10.1109/iraset52964.2022.9738346

[39] Alqaraleh, M., Alzboon, M.S., Al-Batah, M.S., Wahed, M.A., Abuashour, A., Alsmadi, F.H. (2024). Harnessing machine learning for quantifying vesicoureteral reflux: A promising approach for objective assessment. International Journal of Online and Biomedical Engineering (iJOE), 20(11): 123-145. https://doi.org/10.3991/ijoe.v20i11.49673

[40] Tian, Y., Zeng, Z., Wen, M., Liu, Y., Kuo, T., Cheung, S. (2020). EvalDNN: A toolbox for evaluating deep neural network models. In 2020 IEEE/ACM 42nd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), pp. 45–48. https://doi.org/10.1145/3377812.3382133

[41] Călin, S. (2025). Handling imbalanced data: The SMOTE technique. In 2025 17th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Targoviste, Romania, pp. 1-5. https://doi.org/10.1109/ecai65401.2025.11095450

[42] Mir, T.A., Banerjee, D., Aggarwal, P., Pokhariya, H.S. (2024). Proactive management of shrinkage defects using deep learning analytics. In 2024 Asia Pacific Conference on Innovation in Technology (APCIT), Mysore, India, pp. 1-6. https://doi.org/10.1109/apcit62007.2024.10673656

[43] Onyelowe, K.C., Moghal, A.A.B., Ahmad, F., Rehman, A.U., Hanandeh, S. (2024). Numerical model of debris flow susceptibility using slope stability failure machine learning prediction with metaheuristic techniques trained with different algorithms. Scientific Reports, 14(1): 19562. https://doi.org/10.1038/s41598-024-70634-w

[44] Jiang, H., Zou, Q., Zhu, Y.Q., Li, Y., et al. (2024). Deep learning prediction of rainfall-driven debris flows considering the similar critical thresholds within comparable background conditions. Environmental Modelling & Software, 179: 106130. https://doi.org/10.1016/j.envsoft.2024.106130

[45] Chen, Y., Li, N., Xing, F., Xiang, H., Chen, Z. (2025). Study on debris flow vulnerability of ensemble learning model based on spy technology A case study of upper Minjiang river basin. Scientific Reports, 15(1): 22480. https://doi.org/10.1038/s41598-025-03479-6

NOMENCLATURE

Acc	overall Accuracy
AUC	area under the curve
Avg_Rain	average rainfall

DT	decision tree
ECOC	error-correcting output codes
F1-macro	macro-averaged F1 score
GUI	graphical user interface
k	number of nearest neighbors
kNN	k-nearest neighbor
RF	random forest
RUSBoost	random under-sampling boosting
SVM	support vector machine

Greek symbols

γ	kernel width parameter in RBF-SVM
μ	mean value
σ	standard deviation

Subscripts

A	Agromulyo rainfall station
N	Ngepos rainfall station
Avg	average value
i	i-th sample
test	testing dataset