

A Comparative Evaluation of Individual Machine Learning Models and Hybrid Ensemble Approaches for Short-Term Rainfall-Induced Landslide Prediction: Evidence from Rwanda

Emmanuel Hakizimana

ehakizimana55@gmail.com

University of Rwanda

Innocent Ngaruye

University of Rwanda

Research Article

Keywords: landslide prediction, machine learning, hybrid ensemble, stacking, rainfall-induced landslides, class imbalance, Rwanda

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11180425/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Comparative Evaluation of Individual Machine Learning Models and Hybrid Ensemble Approaches for Short-Term Rainfall-Induced Landslide Prediction: Evidence from Rwanda.

Emmanuel HAKIZIMANA ^{1*}, Innocent Ngaruye²

¹ *African Centre of Excellence in Data Science (ACE-DS), College of Business and Economics, University of Rwanda, P. O. Box 4285, Kigali, Rwanda*

² *Department of Mathematics, College of Science and technology, University of Rwanda, P. O. Box 3900, Kigali, Rwanda*

**Corresponding author: Emmanuel Hakizimana, ehakizimana55@gmail.com, Tel: +250785055504*

Abstract

Rainfall-induced landslides remain a persistent threat across Rwanda's mountainous districts, yet, testing whether the hybrid ensemble machine learning models offer a meaningful predictive advantage over individual models under real-world class imbalance had not been established. This study uses a dataset from the Ministry in Charge of Emergency Management (MINEMA) and Meteo Rwanda comprising 19,164 daily records (2018-2026) from Ngororero District, with 17 modelling predictors spanning rainfall, soil moisture, topography and land cover, and a landslide prevalence of only 1.69%. Data were split using random stratified sampling into training (64%), validation (16%) and test (20%) sets using a fixed random seed of 42.

A stacking ensemble combining five base learners (Logistic Regression, Support Vector Machine, Artificial Neural Network, Random Forest and Gradient Boosting) with a Logistic Regression meta-learner was benchmarked against each individual model. SMOTE addressed class imbalance in the training data only, and decision thresholds were tuned on the validation set by maximizing the F1-score. Random Forest achieved the strongest test performance (AUC = 0.841, PR-AUC = 0.144), with the Hybrid Ensemble performing comparably (AUC = 0.824, PR-AUC = 0.139); a DeLong test found no statistically significant difference between the ensemble and any individual model ($p > 0.05$ in every pairwise comparison). Operationally, even the best model detected only 24.6% of test-set landslide days at its optimal threshold, and predicted probabilities were severely miscalibrated (Brier scores approximately six times worse than the base-rate baseline).

Antecedent cumulative rainfall, particularly the 7-day total, was the strongest predictor, followed by seasonality and soil moisture, suggesting that stacking's benefits depend on base-learner quality and positive-case sample size rather than on flawed ensemble design. Random Forest is recommended for operational early-warning deployment, paired with a recall-favoring decision threshold and post-hoc probability recalibration. Broader, multi-district, prospective validation is recommended before nationwide deployment.

Keywords: landslide prediction; machine learning; hybrid ensemble; stacking; rainfall-induced landslides; class imbalance; Rwanda.

1. Introduction

Landslides rank among the deadliest natural hazards worldwide, responsible for an estimated 55,997 fatalities across 4,862 recorded events over little more than a decade, with rainfall as their most frequent trigger, particularly in tropical and mountainous regions where extreme precipitation is becoming increasingly erratic under climate change (Froude & Petley, 2018; Melillo et al., 2016). The Intergovernmental Panel on Climate Change projects an increase in extreme rainfall across East Africa, heightening landslide risk for already vulnerable communities and reinforcing the need for reliable early-warning systems (Calvin et al., 2023; Sättele et al., 2015). Physically, rainfall reduces the shear strength of hillslope soils, precipitating slope failure (Jiang et al., 2025).

Rwanda, nicknamed the ‘Land of a Thousand Hills’ for the mountainous Congo-Nile Divide that dominates its terrain, is particularly exposed to rainfall-triggered landslides because of its bimodal rainfall regime, comprising a long rainy season from March to May and a shorter one from mid-September to mid-December (Mind’je et al., 2020; Nsengiyumva et al., 2018). Between 2018 and 2023, landslides and associated floods killed hundreds of people, destroyed homes and farmland, and affected more than 150,000 people nationally (Idukunda et al., 2025),(Claude et al., 2020); the impact is disproportionately severe because roughly seven in ten Rwandans depend on hillside subsistence agriculture. The 2018 Gakenke tragedy, in which entire families were buried in a single night of intense rainfall, remains a stark illustration of this vulnerability. Vulnerability assessments indicate that more than 42% of Rwanda's territory lies within moderate-to-very-high landslide susceptibility zones (Byiringiro et al., 2024).

Conventional landslide-prediction approaches fall into two broad families. Physically based models, such as the Transient Rainfall Infiltration and Grid-Based Regional Slope Stability (TRIGRS) model, simulate the hydrological and geotechnical processes governing slope failure but require geotechnical parameters that are rarely available at regional scale in data-scarce contexts such as Sub-Saharan Africa (Ebrahim et al., 2025; Ye et al., 2025). Statistical and weighting-based approaches, including frequency-ratio modelling and the Analytic Hierarchy Process, assign static weights to conditioning factors and are consequently unable to represent the time-varying dynamics of rainfall. Machine learning methods, by contrast, have demonstrated a strong capacity to capture the non-linear relationships between environmental variables and landslide occurrence (Merghadi et al., 2020), yet the landslide-modelling literature remains fragmented across techniques, study areas and evaluation protocols (Sun et al., 2025), and the comparative performance of individual models against more elaborate hybrid ensembles remains understudied (Zeng et al., 2023), especially under the realistic class imbalance that characterizes short-term, operational prediction rather than the near-balanced samples typically used in susceptibility mapping. The temporal

distribution of landslide events recorded in Ngororero District between 2018 and 2026 is shown in Figure 1, providing the empirical motivation for a data-driven, short-term prediction approach.

Previous Rwanda studies have demonstrated the usefulness of individual machine learning models and antecedent rainfall, but to test whether the added complexity of stacking yields a statistically significant gain in predictive improvement when evaluated under the natural rarity of landslide events.

This study addresses that gap directly. Using a real-prevalence dataset (1.69% positive class) from Ngororero District, it evaluates whether a stacking-based hybrid ensemble of five machine learning models produces a statistically meaningful predictive improvement over the best individual model, thereby generating evidence on whether the added architectural complexity of ensemble stacking is justified under realistic operating conditions.

The study pursues four specific objectives: (1) to identify and examine the key factors influencing rainfall-induced landslides, including rainfall, slope, soil moisture and land cover; (2) to design and implement a stacking-based hybrid ensemble machine learning approach using five base classifiers and to test experimentally whether its predictive accuracy is significantly better than that of the individual classifiers; (3) to systematically compare the hybrid ensemble against the five individual models using evaluation measures suited to rare-event classification; and (4) to determine whether differences in predictive performance between the hybrid ensemble and individual models are statistically significant. The remainder of this paper presents the materials and methods (Section 2), the empirical results (Section 3), a discussion situating the findings within the wider literature (Section 4), and the conclusions and recommendations that follow (Section 5).

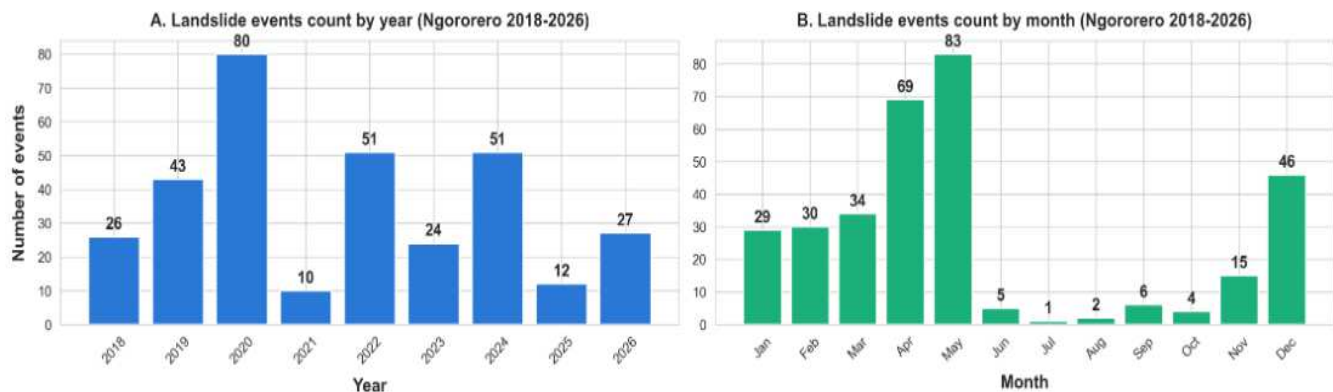


Figure 1: Annual and monthly distribution of recorded landslide events in Ngororero, Rwanda, 2018-2026.

As the left panel of Figure 1 shows, annual landslide counts varied markedly over the study period, from a low of 10 events in 2021 to a peak of 80 in 2020, with no clear monotonic trend, underscoring the year-to-year volatility that a static, weighting-based approach would struggle to represent. The right panel disaggregates these events by calendar month and reveals a pronounced seasonal concentration: occurrences peak sharply in April and May, coinciding with Rwanda's long rainy season, and rise again, though less steeply, in November and December during the short rainy season, while the June-to-October dry-season months account for a small minority of events. This

bimodal, rainfall-aligned pattern motivates the use of short-term antecedent rainfall variables as predictors and underlines why a model capable of capturing time-varying, non-linear dynamics is preferable to conventional static-weighting methods.

2. Materials and Methods

This study adopts a quantitative, comparative and experimental research design: quantitative because it relies on numerical rainfall, soil-moisture, topographic and land-cover measurements and statistical performance metrics; comparative because it systematically benchmarks five individual machine learning models against a stacking-based hybrid ensemble; and experimental because model hyperparameters and decision thresholds were deliberately varied and evaluated on held-out validation and test partitions under controlled, reproducible conditions (fixed random seed = 42).

2.1. Area of Study

Ngororero District, in Rwanda's Western Province, is a mountainous, predominantly agricultural district with elevations ranging from roughly 1,300 to 2,600 m (Irakoze & Gasirabo, 2025). Although the neighboring districts of Gakenke, Rutsiro and Nyabihu have also experienced landslide events, Ngororero has historically recorded the largest number of events nationally, making it a natural focal point for predictive modelling. (Figure 2) maps the district's slope profile, derived from a Digital Elevation Model (DEM); the steep escarpments and deep valleys evident in the north and center of the district, combined with high annual rainfall, make the area highly susceptible to slope failure. Because the study is confined to a single, uniformly mountainous district, the range of slope angles represented in the data is narrower than a multi-district sample would provide (approximately 2.8° – 32.73°), a constraint that is revisited when interpreting the slope results in Section 4.

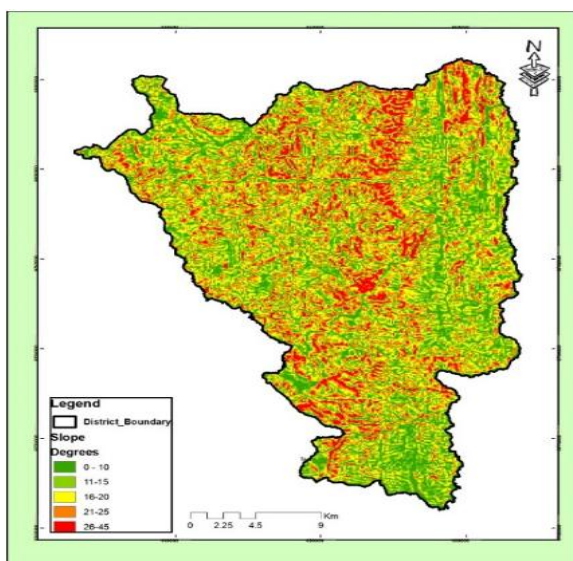


Figure 2: Slope map of Ngororero District, Rwanda, derived from a Digital Elevation Model.

2.2. Data Acquisition and Landslide Inventory

2.2.1. Data Collection and Datasets

This study relies on secondary data provided by the Ministry in Charge of Emergency Management (MINEMA), the Rwanda Environment Management Authority (REMA) and Meteo Rwanda, used exclusively for research purposes with all sources duly acknowledged. Historical landslide-occurrence records were geo-referenced and encoded as a binary target (1 = landslide, 0 = no landslide). Daily meteorological observations were merged with these records to build a panel dataset of 19,164 daily observations spanning 2018-2026, described by 21 raw fields, comprising the binary landslide-occurrence target and 20 candidate predictor variables. After excluding two purely temporal identifiers (year and month) and one variable subsequently found to be a redundant linear average of the four soil-moisture depths (Section 2.2.4), a final set of 17 predictor variables was used to train all six models.

Data completeness was verified prior to modelling, the merged dataset contained no missing values and no duplicate records, and no outliers in physically plausible ranges of rainfall, soil moisture or slope were detected. (Figure 3) summarizes the resulting predictor framework, organizing the 17 modelling variables into four conceptual groups alongside the binary response variable.

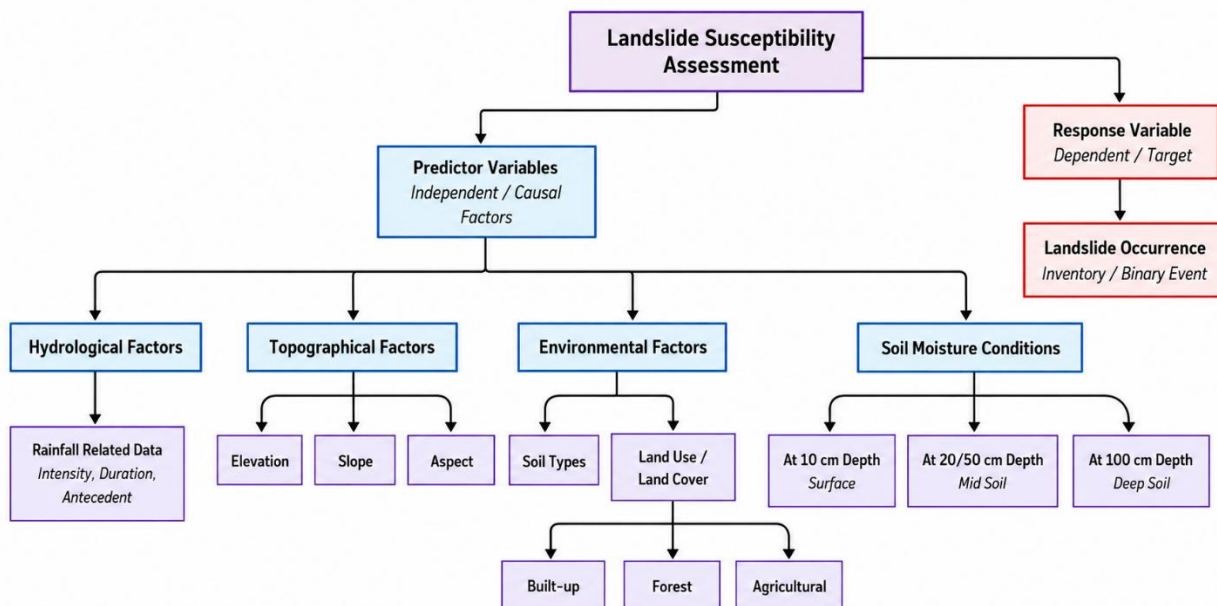


Figure 3: Framework showing the factors contributing to landslide occurrence

As Figure 3 illustrates, the predictor set is organized into four conceptual groups: hydrological factors, captured through rainfall intensity, duration and antecedent accumulation; topographical factors, comprising elevation, slope and aspect; environmental factors, spanning soil type and land use or land cover (built-up, forest and agricultural); and soil-moisture conditions, measured at surface (10 cm), mid-soil (20/50 cm) and deep-soil (100 cm) depths.

Landslide occurrence, encoded as a binary inventory event, forms the dependent response variable against which each of these predictor groups is evaluated. The following subsections describe each group in turn, beginning with the rainfall variables.

2.2.2. Rainfall Variables and Antecedent Rainfall

Daily precipitation data were obtained from Meteo Rwanda and used to derive both immediate and antecedent rainfall predictors: same-day rainfall, rainfall on the previous day and two days prior, rainfall rate (mm/hour), event duration, and cumulative antecedent rainfall totals over 3, 7, 14 and 30 days. These combined variables were included because prior Rwanda-based evidence indicates that antecedent moisture accumulation carries greater predictive power than single-day rainfall alone. To quantify this relationship empirically, point-biserial correlation coefficients were computed between the binary landslide indicator and each of the seven rainfall variables (Table 1)

Table 1: Point-biserial correlation between rainfall variables and landslide occurrence.

Rainfall Variable	Correlation with Landslide (r)
7-day cumulative rainfall	0.181
14-day cumulative rainfall	0.177
30-day cumulative rainfall	0.164
3-day cumulative rainfall	0.162
Previous-day rainfall	0.112
Daily (same-day) rainfall	0.098
Two-days-prior rainfall	0.093

All seven rainfall-related variables show a significant positive association with landslide occurrence, and the four multi-day cumulative totals are consistently more strongly correlated than the three single-day measures. The 7-day cumulative total is the strongest single correlate ($r = 0.181$), almost twice the correlation of same-day rainfall ($r = 0.098$), an early indication that antecedent moisture accumulation, rather than instantaneous rainfall intensity, is the dominant statistical correlate of landslide occurrence in this dataset

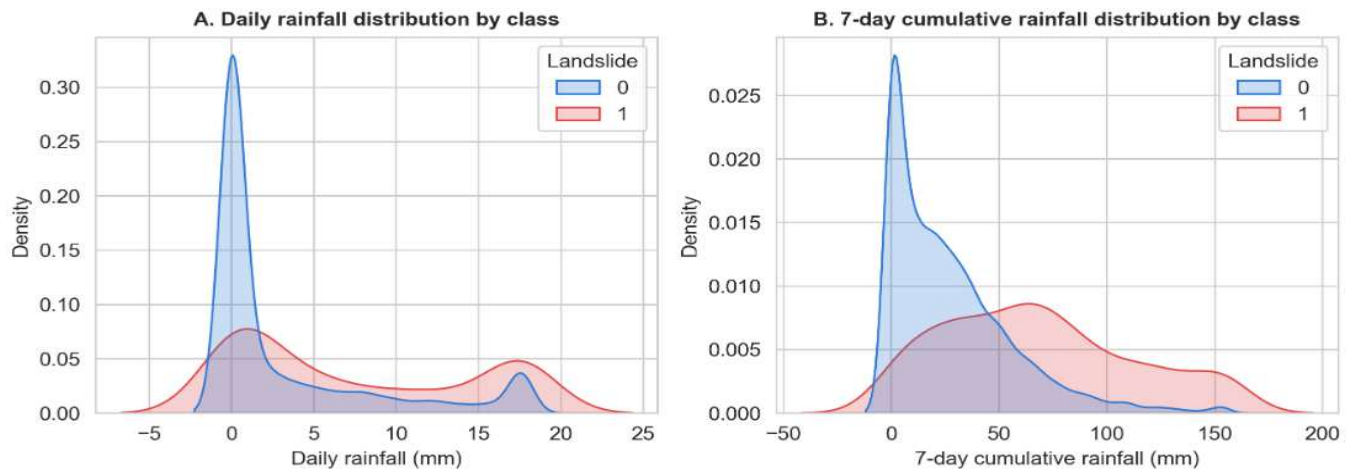


Figure 4: Distribution of daily and 7-day cumulative rainfall by landslide occurrence

Figure 4 shows that landslide days were associated with substantially higher rainfall, with mean daily rainfall of 7.34 mm compared with 3.19 mm on non-landslide days, while mean 7-day cumulative rainfall was 68.3 mm versus 27.3 mm. The right-shifted rainfall distributions for landslide days indicate that both intense and antecedent rainfall were associated with landslide occurrence

2.2.3. Slope and Other Topographic Factors

Topographic predictors comprised slope angle, elevation and aspect, extracted from the district DEM (Figure 2). Slope angle is widely reported in the literature as one of the most important topographic predisposing factors for landslides (Dusabimana et al., 2026), and was therefore retained as a modelling feature despite the restricted range of values imposed by the single-district study design (Figure 5).

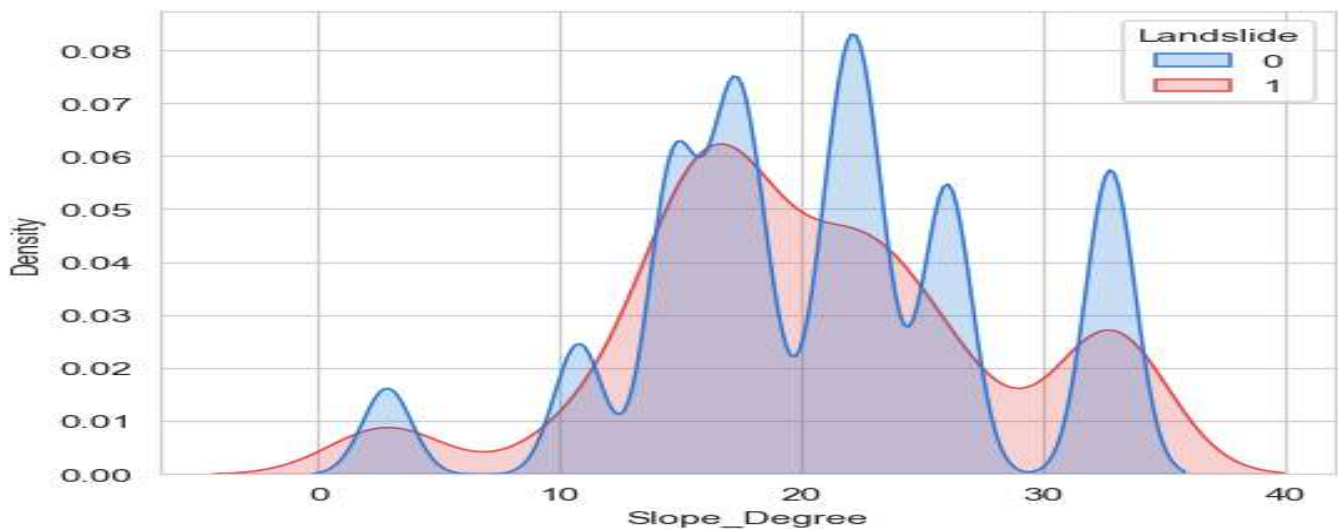


Figure 5: Density Distribution of Slope Angle for Landslide and Non-Landslide Events

2.2.4. Soil Moisture

Soil moisture was measured at four depths below the ground surface: 10 cm, 20 cm, 50 cm and 100 cm. The two shallower depths (10 cm, 20 cm) capture soil-water conditions that respond quickly to individual rainfall events, while the two deeper depths (50 cm, 100 cm) reflect longer-term moisture accumulation. A fifth soil-moisture variable, the cross-depth mean, was found through Variance Inflation Factor (VIF) analysis to be an exact linear average of the four depth-specific measures and was therefore excluded from the final 17-feature modelling set to avoid redundant multicollinearity.

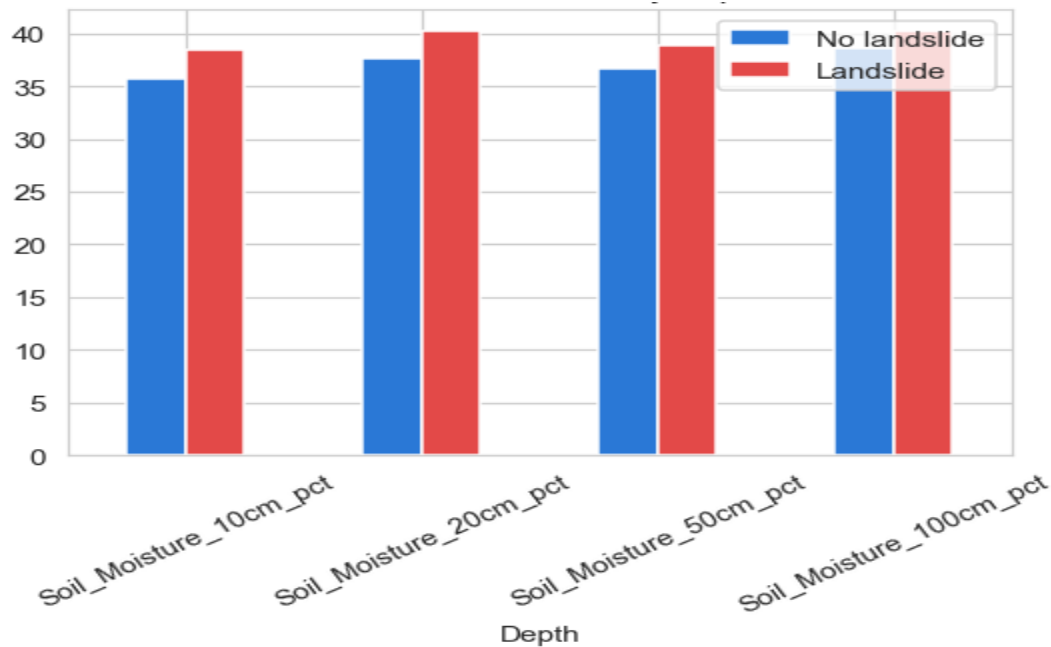


Figure 6: Soil Moisture Profile for Landslide and No Landslide Events

A post-removal multicollinearity check confirmed that this exclusion resolved the problem it was intended to fix: recomputing the Variance Inflation Factor (VIF) across all 17 retained predictors returned no infinite values, confirming that the exact linear dependency introduced by the derived average had been eliminated. Two of the four remaining soil-moisture depth columns, however, still exceeded the conventional severe-multicollinearity threshold of $VIF > 10$ (Soil Moisture at 50cm = 20.2; Soil Moisture at 10cm = 15.2). This is a milder, expected form of correlation: moisture at similar depths naturally tracks together after rainfall infiltration, but none of the four depth columns is mathematically determined by the others the way the derived average was. This residual depth-wise correlation is retained as a limitation of the final feature set rather than a basis for further feature removal, since the tree-based models used in this study are not materially sensitive to it.

2.2.5. Land Cover

Four land-cover classes shown in Figure 7 were represented in the dataset: Cropland Rainfed Herbaceous (code 11), Mosaic Cropland (code 30), Mosaic Tree and Shrubs (code 60), and Tree Broadleaved Deciduous (code 100). Land

cover and season were encoded as integer codes rather than one-hot encoded vectors to keep the feature space compact given the limited number of positive training examples ($n=207$). This simplification is more consequential for the linear and distance-based models (Logistic Regression, SVM) than for the tree-based models.

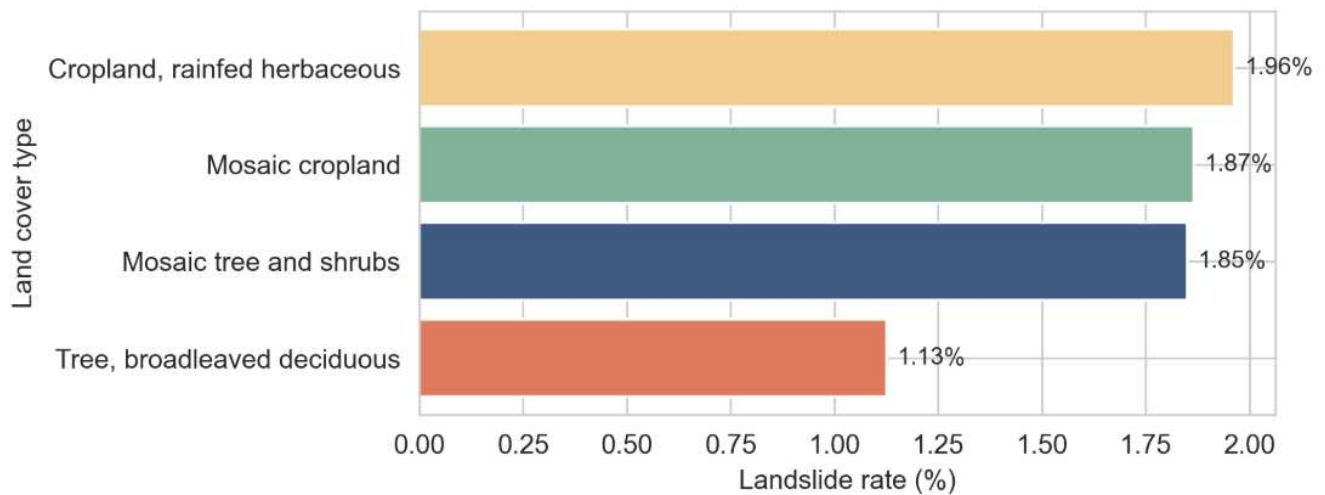


Figure 7: Percentage Distribution of Landslide Occurrence by Land Cover Category

2.2.6. Landslide Incidence (Target Variable) and Class Imbalance

The final modelling dataset comprised 19,164 daily observations collected between 2018 and 2026, with 17 predictor variables retained after preprocessing. Landslide occurrence was treated as a binary target and represented a rare event: only 324 observations (1.69%) corresponded to landslides, compared with 18,840 observations (98.31%) classified as non-landslide days. This pronounced class imbalance reflected the natural rarity of landslide events and motivated the use of stratified data splitting, moderate SMOTE oversampling applied exclusively to the training set, class weighting, and validation-tuned decision thresholds to improve minority-class detection while preserving the real-world prevalence in the validation and test sets (Fernandez et al., 2018).

The pie chart in Figure 8 illustrates the class distribution of the dataset used for machine learning modelling. It shows a highly imbalanced dataset, with landslide observations (positive class) accounting for only **1.69%** of the total observations, compared with **98.31%** for non-landslide observations (negative class).

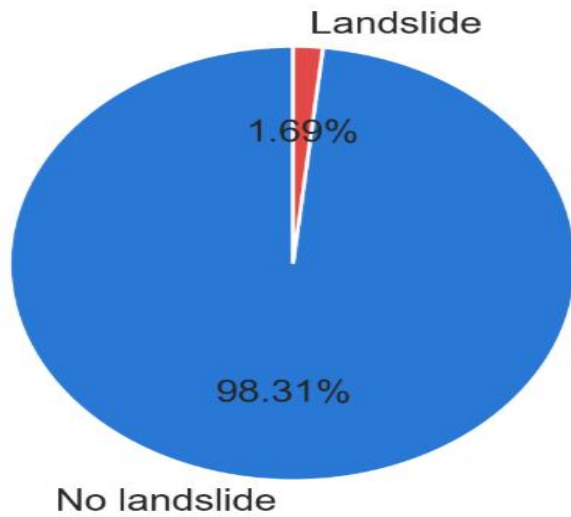


Figure 8: Distribution of landslide occurrence in the dataset (percentage count).

2.2.7. Splitting the Dataset

The dataset was partitioned into training, validation, and test subsets using random stratified sampling to preserve the true landslide prevalence in every partition, following a 64% / 16% / 20% split (Table 2), with a fixed random seed of 42 throughout to ensure reproducibility. Continuous predictors were standardized (zero mean, unit variance) using a scaler fitted on the training partition only and then applied to the validation and test partitions. To address class imbalance without inducing overfitting, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training partition only, at a conservative sampling ratio of 0.15 (raising the minority class to 15% of the majority class within training), rather than fully balancing the classes; this conservative ratio reflects established guidance that oversampling toward, rather than fully up to, parity reduces the risk of overfitting to synthetic samples, a risk that was particularly acute here given that only 207 genuine positive training examples were available to generate synthetic neighbors from. The validation and test partitions were never resampled and therefore retain the real-world landslide prevalence, so that reported test-set performance reflects expected operational behavior.

Table 2: Train / validation / test partition summary.

Partition	Number of Records	Share of Total	Landslide Prevalence
Training set	12,264	64%	1.69%
Validation set	3,067	16%	1.70%
Test set	3,833	20%	1.70%

A methodological caveat applies to this split. The 19,164 observations are not independent daily draws: they form a panel of 9 fixed monitoring sites (distinguished by unique elevation, slope, aspect combinations), each observed almost daily across the 2018–2026 period, and the 3/7/14/30-day cumulative rainfall predictors are trailing sums that share raw rainfall values with adjacent days at the same site. Under a random stratified split, a given site's day t can fall in training while day $t - 1$ or $t + 1$ from the same site falls in test, allowing the model to partially

learn from rainfall totals it will then be evaluated on a form of temporal leakage that can inflate the AUC and PR-AUC figures reported in Table 5 relative to genuine out-of-sample performance. As a robustness check, a site-blocked, rolling-origin split was constructed, in which the earliest $\approx 64\%$ of each site's chronological record was assigned to training, the next $\approx 16\%$ to validation, and only the latest $\approx 20\%$ to test, so that no site's observations are ever divided non-chronologically between partitions (Figure 9).

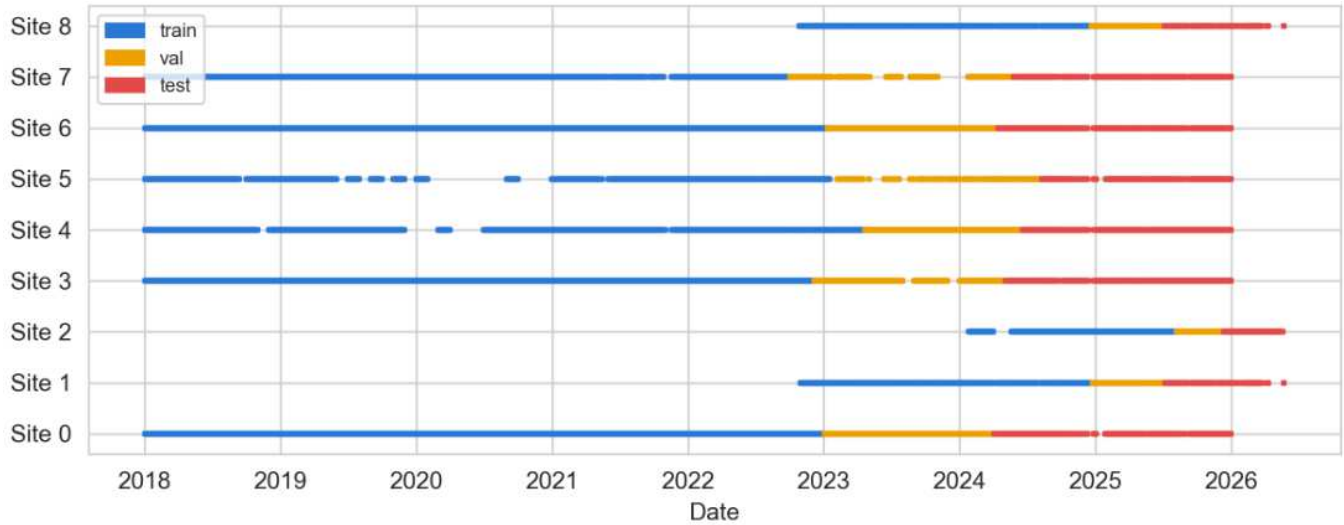


Figure 9: Chronological train, validation and test split across monitoring sites

2.3. Model Descriptions

Model development and evaluation were implemented in Python, using Pandas and NumPy for data manipulation, Matplotlib and Seaborn for visualization, scikit-learn for model training and evaluation, imbalanced-learn for SMOTE oversampling, and SHAP for model interpretability. Six models were compared: five individual base learners and a stacking-based hybrid ensemble that combines them. Other algorithms occasionally reported in the landslide-prediction literature (XGBoost, LightGBM, LSTM and CNN architectures) were deliberately excluded: Gradient Boosting was retained as a representative boosting-family algorithm, making a further XGBoost/LightGBM comparison largely redundant, while LSTM and CNN architectures require substantially more positive training examples than the 207 available here to train reliably without overfitting.

2.3.1. Hyperparameter Tuning

All six models were tuned manually rather than through exhaustive automated search (e.g., GridSearchCV), guided by validation-set AUC, PR-AUC and F1-score, primarily to control overfitting given the extremely small number of genuine positive training examples (207 of 12,264 training records). As a robustness check, a grid-search sensitivity test was additionally run for Random Forest, the study's best-performing model, to assess whether the manually selected settings were reasonable rather than arbitrary. Table 3 summarizes the key hyperparameters adopted for each model.

Concretely, a StratifiedKFold(3) grid search over `n_estimators` (200, 300), `max_depth` (4, 5, 7) and `min_samples_leaf` (15, 25, 40), scored by average precision on the SMOTE-resampled training folds, selected a deeper, less regularized configuration than the manually chosen one (`max_depth` = 7, `min_samples_leaf` = 15, `n_estimators` = 200; best cross-validated average precision = 0.741) rather than reproducing it exactly. Because this score is computed on resampled folds rather than the real, un-resampled training distribution or the held-out test set, it is not directly comparable to the test-set metrics reported in Table 5, and a less regularized tree of this kind carries a higher risk of overfitting to the 207 genuine positive training examples. The shallower, manually selected configuration (`max_depth` = 5, `min_samples_leaf` = 25, `n_estimators` = 300) was therefore retained for all reported results, with the grid search serving as a sanity check confirming that the manually chosen settings sit within a reasonable region of the hyperparameter space rather than as a basis for replacing them.

The five-fold cross-validation used for stacking (Table 3) generates out-of-fold predictions for the meta-learner as follows: for base learner k and training-fold partition $\{F_1, \dots, F_5\}$, the out-of-fold prediction for an observation i in fold F_j is

$$\hat{y}_i^k = f_k^{(-j)}(x_i)$$

where $f_k^{(-j)}$ denotes learner k fitted on all folds except F_j . The meta-learner (logistic regression) is trained on the stacked out-of-fold prediction matrix

$$Z = [y^1, \dots, y^2]$$

against the true labels y , ensuring it never sees a base learner's prediction on data that learner was trained on.

Table 3: Key hyperparameters adopted for each model

Model	Key Hyperparameters
Logistic Regression	$C = 0.3$; class weight = 'balanced'
SVM (RBF)	$C = 0.3$; gamma = 0.02; class weight = 'balanced'
Random Forest	Number of estimators = 300; max depth = 5; min samples/leaf = 25; class weight = 'balanced subsample'
Gradient Boosting	Number of estimators = 100; max depth = 2; learning rate = 0.03; subsample = 0.6; min samples/leaf = 30
ANN (MLP)	hidden layer size = 16; alpha = 0.3; early stopping = True
Hybrid Stacking Ensemble	5-fold cross-validated stacking; meta-learner: Logistic Regression ($C = 0.5$)

At inference time, each base learner is refit on the full training set before generating the base predictions passed to the meta-learner. One caveat: because SMOTE was applied once, prior to this internal cross-validation, a synthetic minority observation can be split from the real neighbors it was interpolated from across internal folds, a milder secondary leakage channel distinct from the one discussed in Section 2.2.7. Wrapping SMOTE inside each

internal fold (e.g., via an imbalanced-learn Pipeline) would remove this channel; given its likely small effect relative to the split-level issue in Section 2.2.7, this refinement was not implemented here.

2.3.2. Evaluation Metrics

Model performance was assessed using accuracy, precision, recall (sensitivity), specificity and F1-score, computed from the standard confusion-matrix counts (true/false positives and negatives). Because landslides are rare, the Precision-Recall Area Under the Curve (PR-AUC) was reported alongside the Receiver Operating Characteristic Area Under the Curve (ROC-AUC), since PR-AUC is not inflated by the very large number of true negatives characteristic of severe class imbalance (Jiang et al., 2025). Probability calibration was assessed with the Brier score, the mean squared difference between predicted probabilities and observed outcomes, with values closer to zero indicating better-calibrated probabilities. Finally, to determine whether differences in test-set AUC between the Hybrid Ensemble and each individual model were statistically significant, a DeLong test was applied to each pairwise comparison, with significance assessed at $\alpha = 0.05$.

2.3.3. Temporal Relationship Between Predictors and the Target Variable

Daily Rainfall (mm), the same-day rainfall total, is included as a predictor for the same day's landslide label, and the 3-day cumulative rainfall total likewise includes the current day's rainfall. The models therefore perform same-day detection of landslide occurrence rather than forecasting with a positive lead time: at the moment a prediction would be needed operationally, same-day rainfall may not yet be fully observed unless sourced from a real-time nowcast product. This distinction matters for how the recommended deployment in Section 5 should be read, the current models are best characterized as a same-day monitoring and decision-support tool rather than an early-warning system with an explicit lead time, unless same-day rainfall is dropped from the feature set and the models retrained and re-evaluated on a genuine 1-day-ahead prediction task.

3. Results

Results are organized around the study's four specific objectives: identification of the key factors influencing rainfall-induced landslides; construction of the hybrid ensemble; comparison of its predictive performance against the five individual base models; and assessment of the statistical significance of any differences.

3.1. Key Factors Influencing Landslide Occurrence

Landslide events in the dataset show a pronounced seasonal (bimodal) pattern, peaking sharply in April and May (69 and 83 events, respectively) with a secondary, smaller peak in December (46 events), and very few events between June and October, closely tracking Rwanda's long- and short-rain seasons. As reported in Table 1 (Section 2.2.2), multi-day cumulative rainfall is consistently more strongly correlated with landslide occurrence than same-day rainfall, with the 7-day cumulative total the strongest single correlate ($r = 0.181$). Soil moisture showed a

positive but comparatively weak correlation with landslide occurrence at every measured depth, marginally stronger at the near-surface layers (10 cm, 20 cm) than at depth. Slope angle showed an essentially null overall correlation with landslide occurrence ($r = -0.006$), with landslide and non-landslide observations sharing an almost identical slope range (2.8° - 32.73°) and near-identical means (20.23° versus 20.57°); as discussed in Section 4, this reflects the restricted range of slope values sampled within a single, uniformly mountainous district rather than a genuine absence of topographic influence. Land cover showed a modest gradient, with the two cropland classes recording the highest landslide rates (1.96% and 1.87%) and the mixed tree-and-shrub class the lowest (1.13%).

These bivariate patterns are corroborated by model-based feature-importance analysis. Figure 10 ranks the ten most influential predictors across the four interpretable base models, using absolute standardized coefficients for Logistic Regression, permutation importance for SVM, and impurity-based importance for Random Forest and Gradient Boosting. Complementary SHAP (SHapley Additive exPlanations) analysis of the two tree-based models (Table 4) confirms that daily rainfall and season code carry the largest mean absolute SHAP contributions in both Random Forest and Gradient Boosting, with the multi-day cumulative rainfall totals also retaining a positive, risk-increasing relationship, while topographic variables (slope, aspect, elevation) contribute negligible marginal signal once rainfall and seasonal information are available, consistent with the restricted-range explanation for slope's near-zero simple correlation.

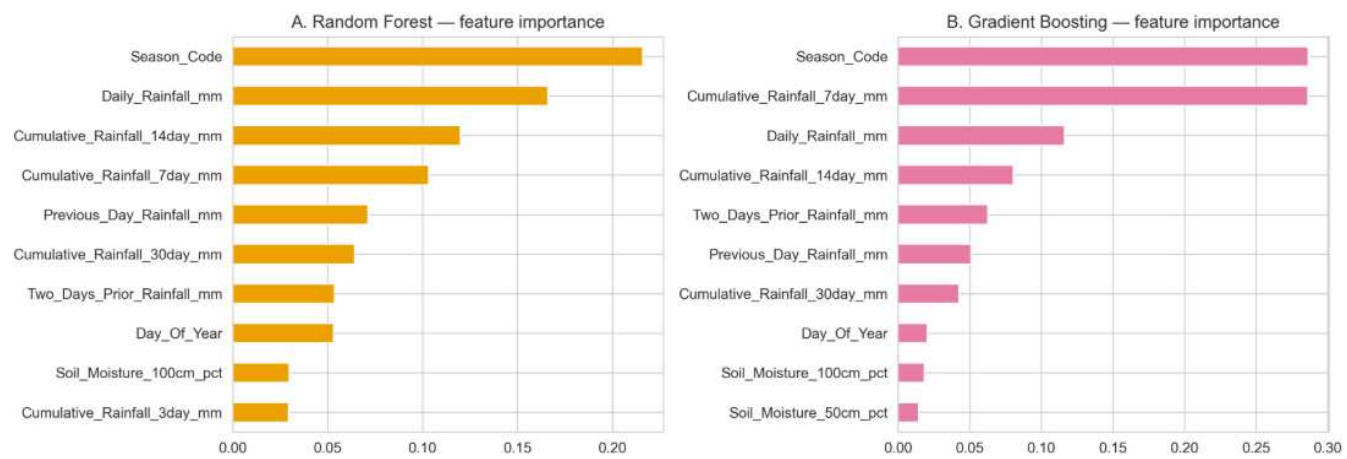


Figure 10: Top ten predictor variables ranked by model-based importance.

The mean absolute SHAP values were used to rank the predictors according to their contribution to the test-set predictions of the Random Forest and Gradient Boosting models. Table 4 presents the eight most influential predictors, with daily rainfall and season code showing the highest mean absolute SHAP values in both models.

Table 4: Top predictors ranked by mean absolute SHAP value (test set).

Rank	Feature	Random Forest Mean SHAP	Gradient Boosting Mean SHAP
1	Daily rainfall (mm)	0.0889	0.3616

2	Season code	0.0810	0.3449
3	14-day cumulative rainfall (mm)	0.0471	0.1007
4	Previous-day rainfall (mm)	0.0443	0.1650
5	7-day cumulative rainfall (mm)	0.0351	0.2193
6	2-days-prior rainfall (mm)	0.0294	0.1413
7	30-day cumulative rainfall (mm)	0.0273	0.0278
8	Day of year	0.0264	0.0118

3.2. Predictive Performance: Individual Models versus the Hybrid Ensemble

Table 5 reports the test-set performance of the five individual models and the Hybrid Ensemble, evaluated at each model's own validation-tuned decision threshold. Because 98.31% of test-set observations are non-landslide days, accuracy alone is a poor performance indicator here (even a trivial always-negative classifier would score 98.31%); the discussion therefore centers on AUC and PR-AUC. Random Forest achieved the highest test-set AUC (0.841) and PR-AUC (0.144), with Gradient Boosting (AUC = 0.827) and the Hybrid Ensemble (AUC = 0.824) close behind. No single model dominated every metric: Gradient Boosting achieved the highest accuracy and precision, Random Forest the highest AUC and PR-AUC, and SVM the highest recall (0.308).

Table 5: Test-set performance of the five individual models and the Hybrid Ensemble.

Model	Accuracy	Precision	Recall	Specificity	F1	AUC	PR-AUC
Random Forest	0.966	0.163	0.246	0.978	0.196	0.841	0.144
Gradient Boosting	0.978	0.231	0.138	0.992	0.173	0.827	0.127
Hybrid Ensemble	0.965	0.153	0.231	0.978	0.184	0.824	0.139
ANN	0.949	0.114	0.292	0.961	0.164	0.817	0.095
Logistic Regression	0.964	0.080	0.108	0.979	0.092	0.813	0.081
SVM	0.931	0.083	0.308	0.941	0.131	0.800	0.078

The ROC curves in Figure 11 show that Random Forest achieved the highest discrimination, with an AUC of 0.841, followed by Gradient Boosting (AUC = 0.827) and the Hybrid Ensemble (AUC = 0.824). All three models maintained ROC curves well above the diagonal reference line, indicating better-than-chance classification performance, although the differences in AUC were relatively small: the three curves are visually near-indistinguishable across most of the false-positive-rate range, separating only modestly at low false-positive rates, where Random Forest sits marginally above Gradient Boosting and the Hybrid Ensemble.

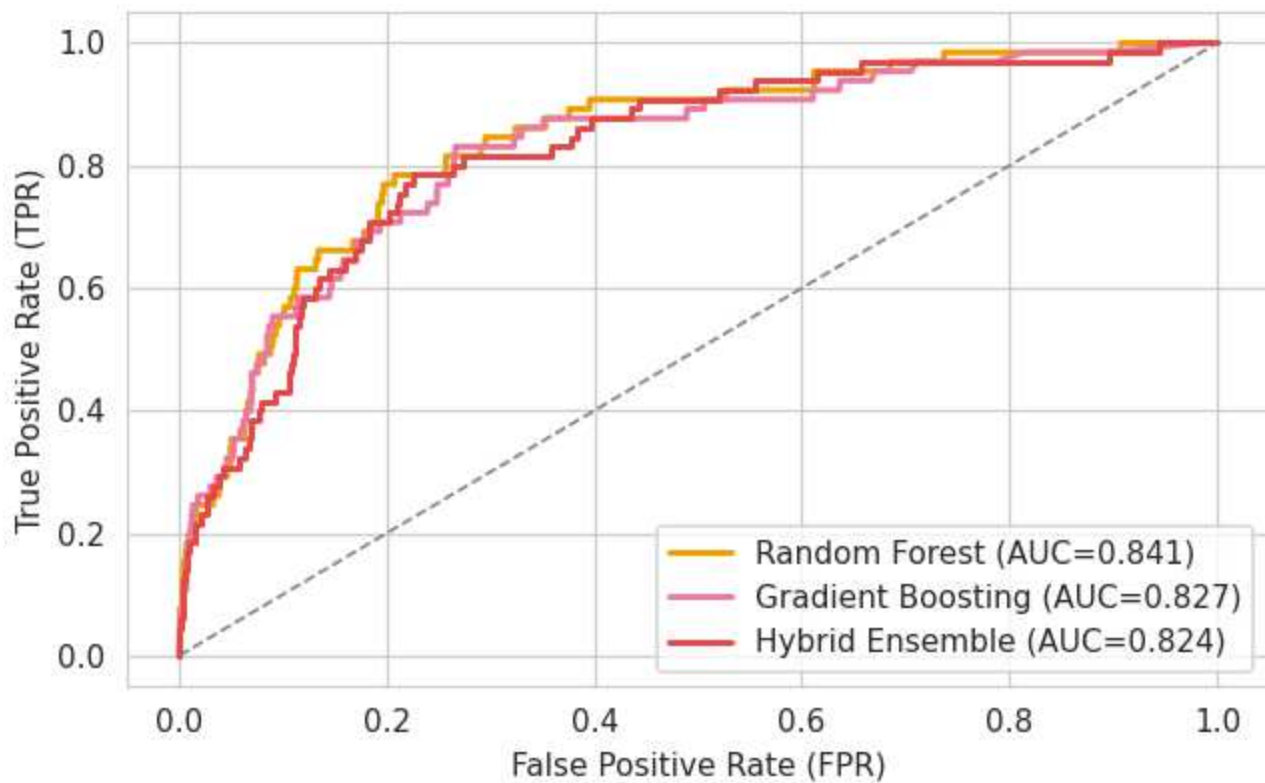


Figure 11: Test-set ROC curves for Random Forest, Gradient Boosting and the Hybrid Ensemble.

Because the class prevalence is only 1.69%, the corresponding precision-recall curves in Figure 12 provide a harsher, arguably more operationally realistic view: PR-AUC values (0.078-0.144) are an order of magnitude smaller than the ROC-AUC values, and Random Forest's curve encloses the largest area, meaning it achieves the highest precision at any given recall level of the six models compared.

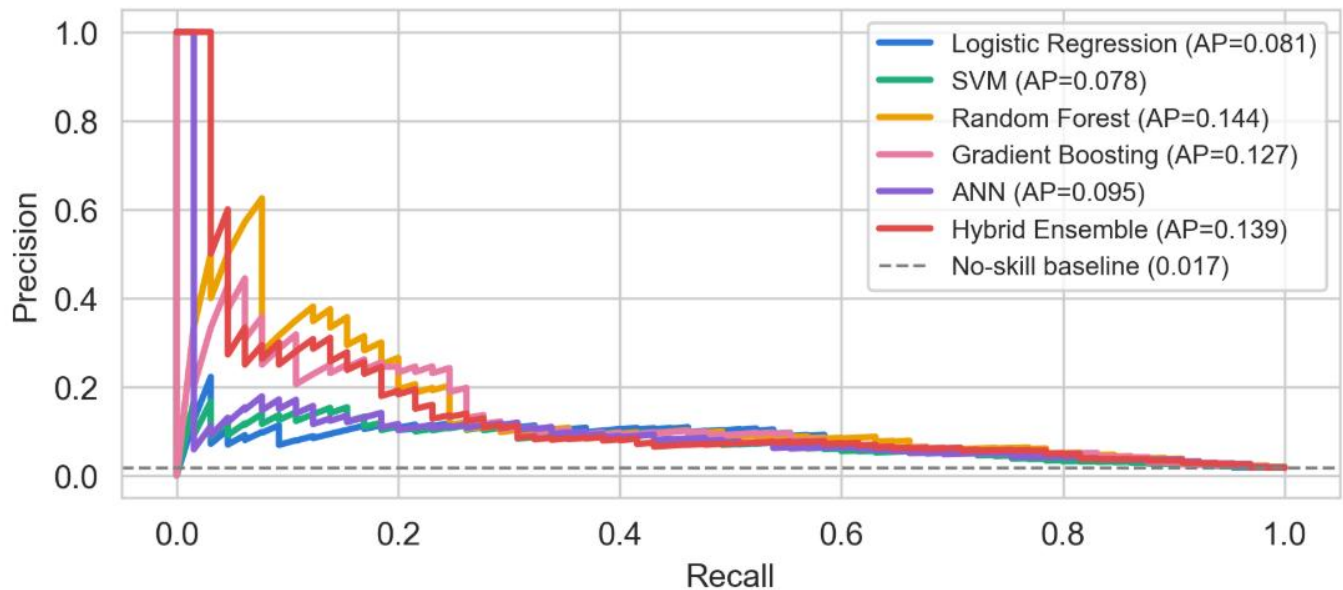


Figure 12: Precision-recall curves for all six models on the test set.

3.3. Confusion Matrices at Validation-Tuned Thresholds

Because AUC and PR-AUC are threshold-free, Figure 13 reports the confusion matrices for Random Forest and the Hybrid Ensemble, the two leading models, at their respective validation-tuned decision thresholds (0.810 and 0.913). Out of 3,833 test-set days, including 65 landslide days, Random Forest correctly flagged 16 landslide days (recall = 24.6%) and missed 49, while producing 82 false alarms among 3,768 safe days; the Hybrid Ensemble correctly flagged 15 landslide days (recall = 23.1%), missed 50, and produced 83 false alarms, a difference of essentially one detected event between the two models. Both optimal thresholds are considerably higher than the conventional default of 0.50, a direct consequence of maximizing F1-score under severe class imbalance, where even a small number of false positives imposes a large relative cost on precision.

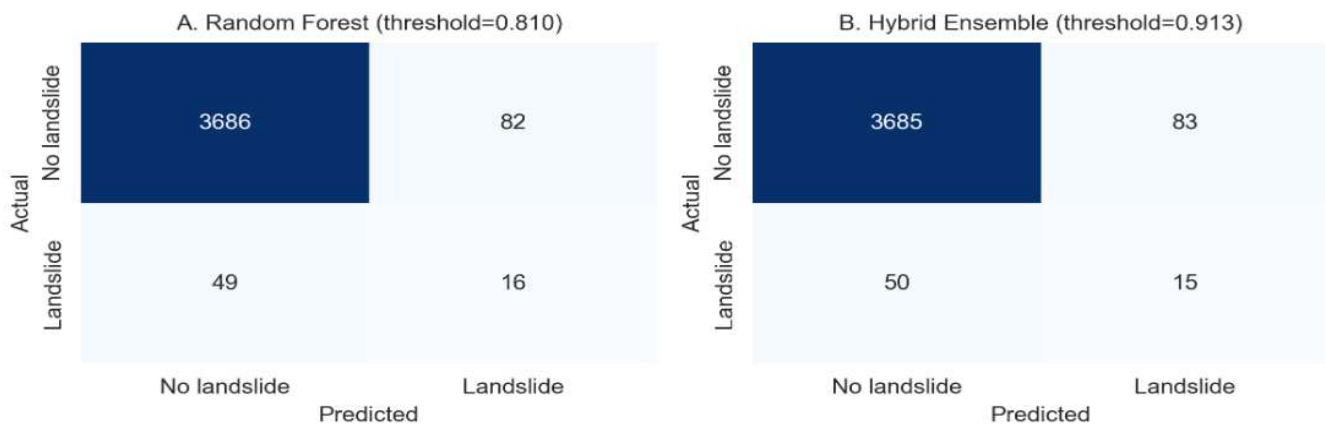


Figure 13: Test-set confusion matrices for Random Forest and the Hybrid Ensemble.

3.4. Probability Calibration of the Leading Models

For an early-warning application, a good ranking of relative risk is not sufficient on its own: the predicted probabilities should also be reliable in an absolute sense, so that a forecast of, say, 70% approximates an observed landslide frequency near 70%. Figure 14 reports reliability diagrams and Brier scores for Random Forest and the Hybrid Ensemble on the test set. Both models are substantially overconfident across virtually the entire probability range, not only above the 0.5 mark: Random Forest's predicted-probability bin centered near 0.52 corresponds to an observed landslide frequency of only about 3%, and its bin centered near 0.20 corresponds to an observed frequency of well under 1%. The Hybrid Ensemble shows the same pattern, with a bin centered near 0.50 corresponding to an observed frequency under 5%.

The Hybrid Ensemble is, in fact, somewhat better calibrated than Random Forest overall (Brier score = 0.107 versus 0.119), even though it trails Random Forest on AUC and PR-AUC. To put these Brier scores in perspective, a trivial model that always predicts the landslide prevalence (1.69%) would achieve a baseline Brier score of 0.0166. Both models achieve Brier scores roughly six to seven times higher than this baseline (0.107 and 0.119), confirming that their predicted probabilities are severely poorly calibrated and heavily overconfident across the probability spectrum. In practice, this means neither model's raw predicted probability should be read as a literal event

likelihood anywhere on the scale; only the relative ranking of observations by risk is reliable, motivating the post-hoc recalibration step recommended in Section 5.

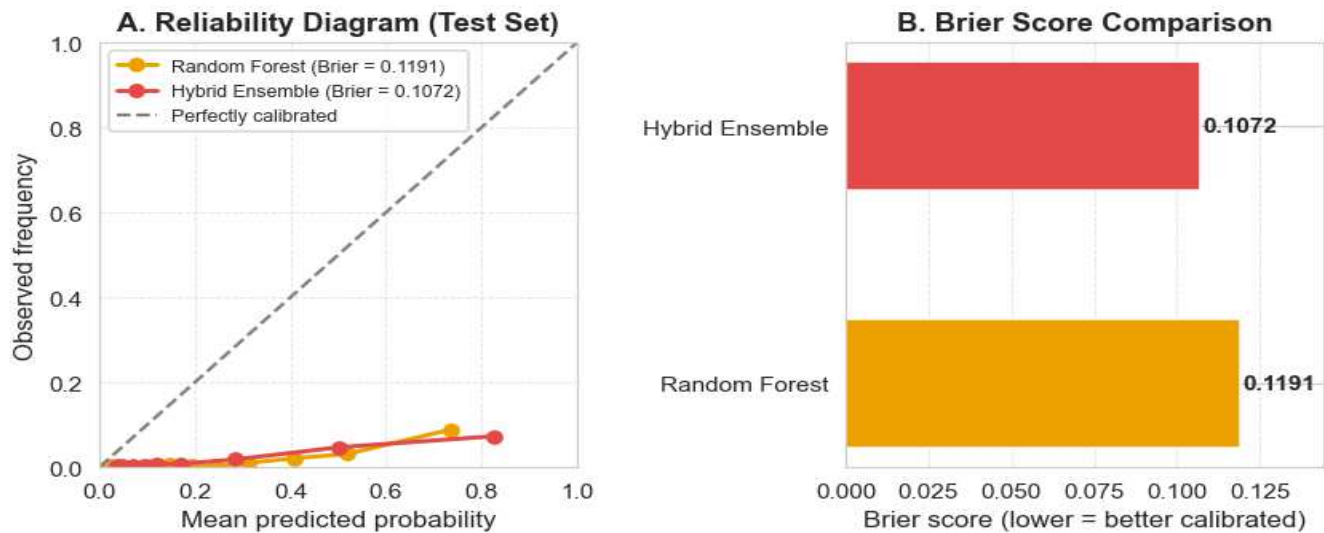


Figure 14: Reliability diagrams and Brier-score comparison for Random Forest and the Hybrid Ensemble.

3.5. Decision-Threshold Sensitivity

Figure 15 plots precision and recall across every possible decision threshold for Random Forest and the Hybrid Ensemble, with each model's validation-tuned, F1-maximizing threshold marked as a dotted vertical line. Both curves show the expected trade-off: recall is near 1.0 at a threshold of 0, while precision is near 0.0, and the two lines converge around 0.5-0.6. Both models' optimal thresholds (0.810 for Random Forest and 0.913 for the Hybrid Ensemble) fall within a narrow, steep zone above approximately 0.75, in which precision rises rapidly from below 0.2 to 0.5-0.7 over a short range while recall drops equally quickly toward zero. Operating in this steep region is risky in practice: small shifts in the fitted threshold, or retraining the model on new data, could substantially change precision and recall. This sensitivity is a further reason to treat the point-estimate metrics in Table 5 as approximate indicators rather than fixed operating characteristics, and reinforces the case, developed further in Section 4, for an operational threshold that favors recall over the statistically optimal F1-maximizing point.

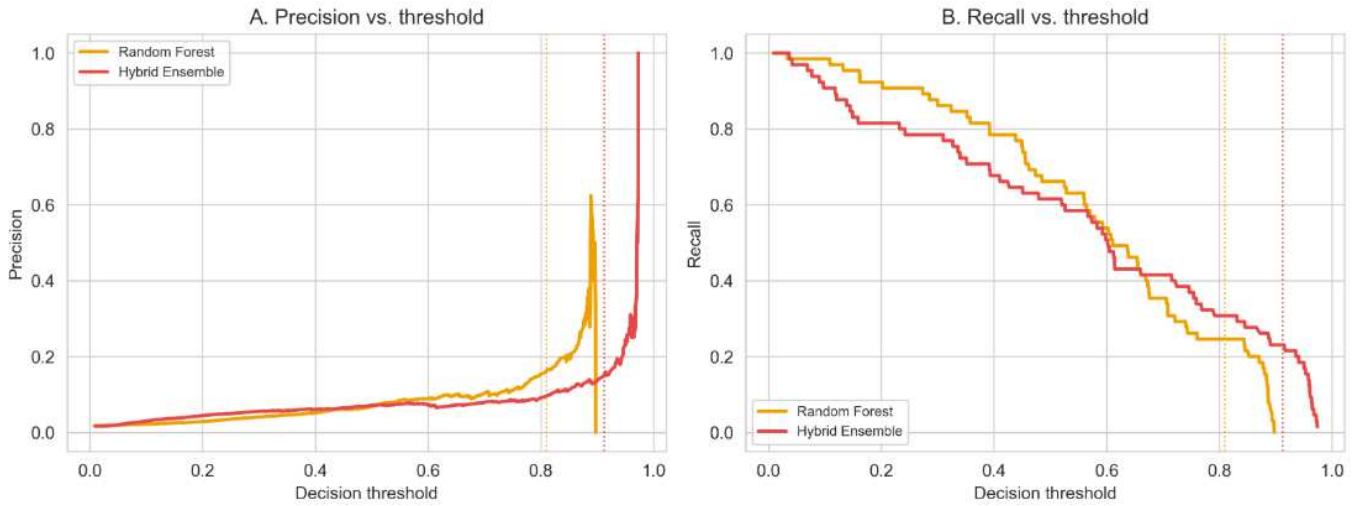


Figure 15: Precision and recall as a function of decision threshold for Random Forest and the Hybrid Ensemble.

3.6. Statistical Significance of the Hybrid Ensemble's Improvement

Table 6 reports the DeLong test comparing the Hybrid Ensemble's test-set AUC against each of the five individual models. None of the five pairwise comparisons reach statistical significance at $\alpha = 0.05$. This is a central empirical finding of the study: although the Hybrid Ensemble numerically outperforms three of the five individual models and is itself numerically outperformed by Random Forest and Gradient Boosting, none of these differences are statistically distinguishable from zero at this sample size (only 65 positive events in the test set). The comparison against Random Forest returns the smallest p-value (0.075), the closest any comparison comes to conventional significance, and it runs in Random Forest's favor rather than the Hybrid Ensemble's.

First, five pairwise DeLong comparisons were run against a single $\alpha = 0.05$ threshold without adjustment; at this threshold the family-wise probability of at least one false-positive finding across five tests is approximately 23% rather than 5%, so a multiple-comparison correction (Holm's step-down procedure: reject hypothesis (i), if:

$$P_i < \frac{\alpha}{m - i + 1}$$

for p-values ordered ascending, $m = 5$) has been applied and is reported alongside the raw p-values in Table 6 below. Second, DeLong's test assumes independent test-set observations; because the test set is drawn from only 9 repeated sites over multiple years, this independence assumption is unlikely to hold exactly, so the reported p-values should be treated as indicative rather than exact.

A site-level cluster bootstrap of the AUC difference (2,000 resamples of whole sites with replacement, rather than individual rows) was therefore run as a companion to the DeLong test, giving a confidence interval robust to this dependence; the results are reported in Figure 16 alongside the DeLong confidence intervals.

Table 6: DeLong test: Hybrid Ensemble vs. each individual model (test-set AUC).

Comparison	AUC (Indiv idual)	AUC (Hybr id)	p- value	Significa nt ($\alpha = 0.05$)	95% CI (DeLong)	Holm- adjust ed p	Significant after Holm ($\alpha = 0.05$)
Hybrid vs Logistic Regression	0.813	0.824	0.656	No	−0.037, 0.059	1.000	No
Hybrid vs SVM	0.800	0.824	0.258	No	−0.018, 0.066	1.000	No
Hybrid vs Random Forest	0.841	0.824	0.075	No	−0.035, 0.002	0.375	No
Hybrid vs Gradient Boosting	0.827	0.824	0.841	No	−0.027, 0.022	1.000	No
Hybrid vs ANN	0.817	0.824	0.739	No	−0.033, 0.047	1.000	No

Figure 16 compares the AUC differences between the Hybrid Ensemble and each individual model using DeLong and site-cluster bootstrap 95% confidence intervals. Although the Hybrid Ensemble shows higher AUC than ANN, SVM, and Logistic Regression, and lower AUC than Random Forest and Gradient Boosting, the confidence intervals generally overlap zero, indicating no clear statistically significant AUC differences

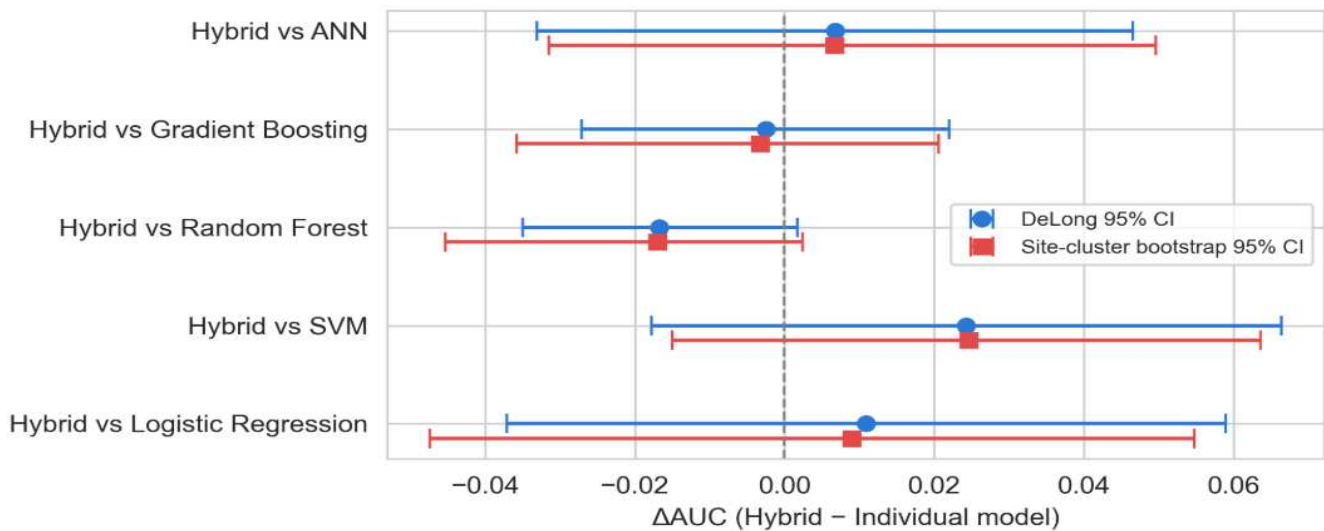


Figure 16: AUC differences between the Hybrid Ensemble and individual models with 95% confidence intervals.

4. Discussion

4.1. Seasonality and the Dominance of Antecedent Rainfall

The concentration of landslide events in April-May, and to a lesser extent December, mirrors Rwanda's bimodal rainfall regime and is consistent with previously documented seasonal clustering of fatal landslides in the country (Dusabimana et al., 2026; Uwihirwe et al., 2020), including the cluster of more than 700 simultaneous shallow

landslides that struck the Karongi region on the night of 6-7 May 2018 at the peak of the long rains (Byiringiro et al., 2024). A comparable seasonal concentration of catastrophic rainfall-triggered landslides has been reported outside Rwanda in climatically similar East African highland settings, including the Gofa Zone disaster in southern Ethiopia, which killed at least 257 people on 22 July 2024 during that region's principal wet season (Shano et al., 2024), lending external validity to the seasonal pattern identified here.

The finding that multi-day cumulative rainfall outperforms same-day rainfall as a correlate of landslide occurrence is consistent with the hydro-meteorological threshold literature, in which antecedent rainfall accumulation, by raising soil moisture and pore-water pressure toward saturation, is treated as the dominant control on slope-failure timing rather than instantaneous rainfall intensity alone (Bogaard & Greco, 2018; Monsieus et al., 2019). Prior Rwanda-specific work similarly found that a 10-day antecedent precipitation total, combined with event-day rainfall, offered the strongest discriminatory power for distinguishing landslide from non-landslide days nationally (Guzzetti et al., 2022), a result this study's independent dataset and district closely corroborate. This antecedent-rainfall and seasonal signal is not an artefact of simple bivariate correlation alone. The SHAP analysis (Table 4) shows daily rainfall and season code as the dominant contributors in both tree-based models once all 17 predictors interact freely, providing convergent evidence, across a bivariate statistic, an impurity-based ranking, and a game-theoretic attribution method, for the same substantive conclusion. This finding is consistent with previous work showing that antecedent and forecast rainfall inputs, combined with Random Forest, provided the strongest-performing configuration among several machine-learning algorithms (Fang et al., 2021).

A distinction is worth making explicit here between the SHAP-based feature attributions reported above (Table 4, Figure 10) and causal inference. SHAP values quantify each feature's contribution to a fitted model's output under the Shapley-value axioms of local accuracy, consistency, and missingness; they describe what the model has learned to associate with landslide occurrence, not the causal effect of that feature within the physical slope-failure process. Two limitations follow. First, given the correlation among the four soil-moisture depth predictors (Section 2.2.4), SHAP attribution can be distributed across correlated features even when one carries the dominant physical influence, so relative SHAP rankings among correlated predictors should not be read as relative causal importance. Second, because the models perform same-day detection rather than lead-time forecasting (Section 2.3.3), same-day rainfall's high SHAP contribution is consistent with it being either a trigger of landslide occurrence or a correlate of the same storm system that produces both, and the present design cannot distinguish between these; establishing a causal triggering relationship would require a study design with explicit temporal precedence between predictor and outcome, which is outside the scope of the present predictive-modeling study.

4.2. Slope Angle: A Restricted-Range Effect, Not an Absence of Effect

The near-zero correlation between slope angle and landslide occurrence ($r = -0.006$) should not be interpreted as evidence that slope is globally unimportant for landslide initiation. Rather, this result is a direct consequence of the

study's spatiotemporal design. Slope_Degree is derived from a static Digital Elevation Model and is therefore time-invariant for each of the nine monitoring sites, meaning it takes only nine unique values across the entire 19,164-day panel. In a daily prediction task, a static site characteristic cannot, by itself, explain the day-to-day timing of slope failure; its predictive value is necessarily diluted by temporally dynamic variables such as antecedent rainfall and soil moisture, which vary substantially from day to day. Furthermore, because Ngororero District was deliberately selected as a uniformly mountainous, high-risk area, the sampled slopes are restricted to a narrow range (2.8° – 32.73°), with landslide and non-landslide days sharing near-identical means (20.23° versus 20.57°). This restriction-of-range effect, combined with the predictor's temporal invariance, explains both the null bivariate correlation and the near-zero SHAP contribution once rainfall and seasonal information are accounted for. The divergence from Kuradusenge et al. (Kuradusenge et al., 2020), who identified slope as the most important predictor in the same district using an earlier subset, is therefore attributable to differences in sampling design: their 1:1 landslide/non-landslide sampling ratio amplified topographic differences, whereas the present study preserves the real-world 1.69% landslide prevalence, in which most days across all sites are non-landslide days, rendering slope a weak discriminator of event timing. In operational contexts where the goal is daily prediction rather than static susceptibility mapping, time-varying hydrometeorological triggers, particularly antecedent rainfall will inevitably dominate over static topographic factors, regardless of their theoretical importance in slope-stability mechanics.

4.3. Land Cover

The higher landslide rates observed in the two cropland classes relative to the mixed tree-and-shrub class are broadly consistent with the wider landslide-susceptibility literature (Dou et al., 2021), including studies of landslide susceptibility and conditioning factors in Rwanda (Mind'je et al., 2020; Panelli et al., 2024), which identifies land cover and vegetation structure, rather than the mere presence of trees, as a significant predisposing factor; notably, the deciduous-forest class in this dataset did not show the same comparatively protective pattern as the mixed tree-and-shrub class, despite its woody cover, suggesting that vegetation type and management practice, not canopy alone, shape the land-cover-landslide relationship in this district.

4.4. Why the Hybrid Ensemble Did Not Significantly Outperform the Individual Models

Contrary to the expectation that stacking would better capture non-linearities than individual models, the Hybrid Ensemble did not achieve a statistically significant AUC improvement over any single base model, and was itself numerically outperformed by Random Forest and Gradient Boosting. This pattern is consistent with the general recognition in the ensemble-learning literature that stacking performs poorly when base learners vary substantially in competence, or when too little positive-class data is available to train the meta-learner reliably (Zeng et al., 2023), with only 207 true positive training examples, the meta-learner had a limited basis for learning an optimal combination rule across five learners of uneven strength (SVM and ANN trailed Random Forest and Gradient Boosting by a wide margin). The Hybrid Ensemble also exhibited the largest train-test AUC gap of the six models

(0.118), though re-scoring on the real, un-resampled training distribution narrows this gap to approximately 0.06, indicating that much of the apparent overfitting is a SMOTE-resampling artefact rather than genuine overfitting alone, with a modest residual gap still attributable to the very limited positive-class sample.

Ranking models by a single metric can also be misleading in this setting: Gradient Boosting led on accuracy and precision, Random Forest on AUC and PR-AUC, and SVM on recall, at roughly triple Random Forest's false-positive rate. Because a missed landslide (false negative) is considerably costlier than a false alarm (Taherdangkoo et al., 2026) in a disaster early-warning context, recall deserves particular weight when selecting an operational model (Guzzetti et al., 2022); on this combined basis, Random Forest offers the best overall balance among the six models evaluated, ranking first on AUC and PR-AUC and third on recall, closely followed by the Hybrid Ensemble.

The AUC values obtained here (0.80-0.84) are lower than the very high accuracy figures often reported in the wider landslide-susceptibility literature, but this reflects a difference in task difficulty rather than weaker modelling: most benchmark studies use static susceptibility-mapping designs that sample non-landslide locations near a 1:1 ratio, a substantially easier task than the 1.69% real-world prevalence preserved throughout this study for operational realism. Because both model training and this comparison are based exclusively on Ngororero District, the reported performance and feature-importance rankings should be read as representative of that district's specific rainfall regime, soil and topography rather than as universally applicable across Rwanda; districts with different geology, land cover or rainfall seasonality, such as the Eastern Province's flatter, drier terrain, would likely require model retraining or local recalibration before deployment.

4.5. Probability Calibration and Operational Reliability

A further practical caveat concerns probability calibration: both leading models were found to be substantially overconfident across virtually the entire probability range once calibration was assessed directly (Brier scores of 0.119 for Random Forest and 0.107 for the Hybrid Ensemble; Figure 14). By contrast, a recent calibrated stacked-ensemble landslide susceptibility model achieved near-ideal calibration (Brier = 0.054) only after applying fold-wise isotonic post-hoc calibration (Taherdangkoo et al., 2026), meaning their raw predicted probabilities should be treated as a risk-ranking score rather than a literal event likelihood, reinforcing the case for post-hoc recalibration before any operational use. The decision-threshold sensitivity analysis in Figure 15 compounds this caveat: both models' F1-maximizing thresholds sit in a steep region of the precision-recall trade-off, where small shifts in the fitted threshold or retraining on new data could substantially change operational precision and recall.

5. Conclusions

This study developed and evaluated a hybrid ensemble machine learning framework for short-term, rainfall-induced landslide prediction in Ngororero District, Rwanda, using 19,164 daily observations spanning 2018-2026. Landslide occurrence was confirmed to be strongly seasonal and better predicted by cumulative antecedent rainfall, especially the 7-day total, than by same-day rainfall, corroborating prior Rwanda-focused research and directly satisfying the study's first objective. A five-learner stacking ensemble was successfully designed, implemented and benchmarked against its constituent base models, satisfying the second objective. However, contrary to initial expectations, the Hybrid Ensemble did not significantly outperform the best individual model: Random Forest achieved the highest test-set AUC (0.841) and PR-AUC (0.144), and a DeLong test found no statistically significant difference between the Hybrid Ensemble and any of the five base models ($p > 0.05$ throughout), thereby answering the study's third objective and central research question.

This outcome should not be read as a failure of the research design but as a substantive finding: stacking's benefits appear contingent on base-learner diversity and quality and, critically, on the availability of a sufficiently large number of positive events for the meta-learner to combine base-model predictions reliably. This reinforces a broader recognition in the rare-event-classification literature that simpler, well-understood individual models can match, or exceed, more architecturally complex ensembles when positive-class data are scarce, a consideration relevant to model selection in other rare natural-hazard prediction domains beyond landslides.

On this basis, Random Forest, rather than the more complex Hybrid Ensemble, is recommended for operational deployment by MINEMA, since the added complexity of stacking did not yield a statistically significant predictive gain. Because a missed landslide event carries a far higher cost than a false alarm, a recall-favoring decision threshold, rather than the statistically optimal but overly conservative F1-maximizing threshold identified in this study, is recommended for operational use, alongside a post-hoc probability recalibration step (for example, Platt scaling or isotonic regression) before predicted probabilities are presented to decision-makers as literal likelihoods. Meteo Rwanda is encouraged to continue strengthening antecedent-rainfall monitoring given it demonstrated predictive importance, and MINEMA and Meteo Rwanda are both encouraged to explore satellite-based remote sensing to build a more complete landslide inventory, given the sparse, ad hoc nature of current ground-based reporting. This recommendation, and the reported test-set metrics generally, should be read in light of two caveats: the same-day (rather than lead-time) prediction framing established in Section 2.3.3 and the temporal-validation caveat in Section 2.2.7. A site-blocked or rolling-origin re-evaluation (Figure 9), alongside the recall-favoring threshold and post-hoc recalibration already recommended, is a prerequisite rather than an optional refinement before any operational recommendation to MINEMA is finalized.

This study was confined to a single district, which restricted the range of slope angles available for analysis and limits the generalizability of the reported performance and feature-importance rankings to Rwanda's other, topographically and climatically distinct districts. The rarity of landslide events (1.69% prevalence) further constrained achievable recall across all six models despite SMOTE oversampling. Future research should prioritize

multi-district or nationwide validation, which would also restore the slope-angle variability this single-district design could not capture; a longer historical record and a larger number of confirmed positive events, to give the stacking meta-learner a more reliable basis for combining base models; and systematic exploration of additional predictors, such as geological substrate, groundwater dynamics and human land-use change, alongside cost-sensitive threshold optimization that better reflects the asymmetric costs of missed detections in disaster early-warning practice.

Author Contributions

Conceptualization, E.H. and I.N.; methodology, E.H.; software, E.H.; validation, E.H. and I.N.; formal analysis, E.H.; investigation, E.H.; data curation, E.H.; writing original draft preparation, E.H.; writing review and editing, E.H. and I.N.; visualization, E.H.; supervision, I.N.; project administration, I.N. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding. The study was conducted as part of the authors' academic work at the African Centre of Excellence in Data Science (ACE-DS), University of Rwanda.

Institutional Review Board Statement

Not applicable. This study did not involve humans or animals; it relied exclusively on secondary, organizational meteorological and disaster-management records.

Informed Consent Statement

Not applicable.

Data Availability Statement

The landslide-inventory and meteorological data used in this study were provided by the Ministry in Charge of Emergency Management (MINEMA), and Meteo Rwanda under a research-use agreement and are not publicly available due to third-party ownership restrictions. Data may be made available by the corresponding author upon reasonable request and with permission from MINEMA and Meteo Rwanda.

Acknowledgments

The authors thank the Ministry in Charge of Emergency Management (MINEMA) and Meteo Rwanda for providing the data used in this study, and the African Centre of Excellence in Data Science (ACE-DS), University of Rwanda, for the academic environment and resources that supported this research.

Conflicts of Interest

The authors declare no conflict of interest.

Clinical Trial Number:

Not Applicable

REFERENCES:

- Bogaard, T., & Greco, R. (2018). Invited perspectives: Hydrological perspectives on precipitation intensity-duration thresholds for landslide initiation: proposing hydro-meteorological thresholds. *Natural Hazards and Earth System Sciences*, 18(1), 31–39. <https://doi.org/10.5194/nhess-18-31-2018>
- Byiringiro, F.-V., Jolivet, M., Dauteuil, O., Arvor, D., & Hitimana Niyotwambaza, C. (2024). Exceptional Cluster of Simultaneous Shallow Landslides in Rwanda: Context, Triggering Factors, and Potential Warnings. *GeoHazards*, 5(4), 1018–1039. <https://doi.org/10.3390/geohazards5040049>
- Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P. W., Trisos, C., Romero, J., Aldunce, P., Barrett, K., Blanco, G., Cheung, W. W. L., Connors, S., Denton, F., Diongue-Niang, A., Dodman, D., Garschagen, M., Geden, O., Hayward, B., Jones, C., ... Ha, M. (with Lee, H.). (2023). *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.* (P. Arias, M. Bustamante, I. Elgizouli, G. Flato, M. Howden, C. Méndez-Vallejo, J. J. Pereira, R. Pichs-Madruga, S. K. Rose, Y. Saheb, R. Sánchez Rodríguez, D. Ürgen-Vorsatz, C. Xiao, N. Yassaa, J. Romero, J. Kim, E. F. Haïtes, Y. Jung, R. Stavins, ... C. Péan, Eds.; First). Intergovernmental Panel on Climate Change (IPCC). <https://doi.org/10.59327/IPCC/AR6-9789291691647>
- Claude, M. J., Martin, N. V., Abias, M., Francoise, M., Johnson, U., Tonny, K., & Martine, U. (2020). Mapping Landslide Susceptibility and Analyzing Its Impact on Community Livelihoods in Gakenke District, Northern Rwanda. *Journal of Geoscience and Environment Protection*, 08(05), 41–55. <https://doi.org/10.4236/gep.2020.85003>

- Dou, J., Yunus, A. P., Merghadi, A., Wang, X., & Yamagishi, H. (2021). A Comparative Study of Deep Learning and Conventional Neural Network for Evaluating Landslide Susceptibility Using Landslide Initiation Zones. In F. Guzzetti, S. Mihalić Arbanas, P. Reichenbach, K. Sassa, P. T. Bobrowsky, & K. Takara (Eds.), *Understanding and Reducing Landslide Disaster Risk* (pp. 215–223). Springer International Publishing. https://doi.org/10.1007/978-3-030-60227-7_23
- Dusabimana, J. D., Dewitte, O., Uwihirwe, J., Bogaard, T., Bugenimana, E. D., Musemakweri, J., Vanmaercke, M., Van Weverberg, K., & Reinoso Rondinel, R. (2026, March 13). *Susceptibility-informed hydro-meteorological thresholds for rainfall-triggered landslides in Rwanda*. <https://doi.org/10.5194/egusphere-egu26-6632>
- Ebrahim, K., Gomaa, S. M. M. H., Zayed, T., & Alfalah, G. (2025). Rainfall-Induced Landslide Prediction Models, Part I: Empirical–Statistical and Physically Based Causative Thresholds. *Water*, 17(22), 3273. <https://doi.org/10.3390/w17223273>
- Fang, Z., Wang, Y., Peng, L., & Hong, H. (2021). A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping. *International Journal of Geographical Information Science*, 35(2), 321–347. <https://doi.org/10.1080/13658816.2020.1808897>
- Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61, 863–905. <https://doi.org/10.1613/jair.1.11192>
- Froude, M. J., & Petley, D. N. (2018). Global fatal landslide occurrence from 2004 to 2016. *Natural Hazards and Earth System Sciences*, 18(8), 2161–2181. <https://doi.org/10.5194/nhess-18-2161-2018>
- Guzzetti, F., Gariano, S. L., Peruccacci, S., Brunetti, M. T., & Melillo, M. (2022). Rainfall and landslide initiation. In *Rainfall* (pp. 427–450). Elsevier. <https://doi.org/10.1016/B978-0-12-822544-8.00012-3>
- Idukunda, C., Henry, S., Twarabamenye, E., De Longueville, F., & Michellier, C. (2025). *Evaluating Disaster Risk Management System: A Case Study of Rwanda's Response to the 2nd–3rd May 2023 disaster event*. Risk Assessment, Mitigation and Adaptation Strategies, Socioeconomic and Management Aspects. <https://doi.org/10.5194/egusphere-2025-2112>
- Irakoze, S., & Gasirabo, A. (2025). Mapping Landslide Susceptibility Areas and Influencing Factors in Ngororero District, Rwanda. *East African Journal of Environment and Natural Resources*, 8(3), 29–39. <https://doi.org/10.37284/eajenr.8.3.3651>

- Jiang, X., Yang, Z., Mei, H., Zheng, M., Yuan, J., & Wang, L. (2025). An Ensemble Learning Approach for Landslide Susceptibility Assessment Considering Spatial Heterogeneity Partitioning and Feature Selection. *Remote Sensing*, 17(16), 2875. <https://doi.org/10.3390/rs17162875>
- Kuradusenge, M., Kumaran, S., & Zennaro, M. (2020). Rainfall-Induced Landslide Prediction Using Machine Learning Models: The Case of Ngororero District, Rwanda. *International Journal of Environmental Research and Public Health*, 17(11), 4147. <https://doi.org/10.3390/ijerph17114147>
- Melillo, M., Brunetti, M. T., Peruccacci, S., Gariano, S. L., & Guzzetti, F. (2016). Rainfall thresholds for the possible landslide occurrence in Sicily (Southern Italy) based on the automatic reconstruction of rainfall events. *Landslides*, 13(1), 165–172. <https://doi.org/10.1007/s10346-015-0630-1>
- Merghadi, A., Yunus, A. P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D. T., Avtar, R., & Abderrahmane, B. (2020). Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Science Reviews*, 207, 103225. <https://doi.org/10.1016/j.earscirev.2020.103225>
- Mind'je, R., Li, L., Nsengiyumva, J. B., Mupenzi, C., Nyesheja, E. M., Kayumba, P. M., Gasirabo, A., & Hakorimana, E. (2020). Landslide susceptibility and influencing factors analysis in Rwanda. *Environment, Development and Sustainability*, 22(8), 7985–8012. <https://doi.org/10.1007/s10668-019-00557-4>
- Monsieurs, E., Dewitte, O., Depicker, A., & Demoulin, A. (2019). Towards a Transferable Antecedent Rainfall—Susceptibility Threshold Approach for Landsliding. *Water*, 11(11), 2202. <https://doi.org/10.3390/w11112202>
- Nsengiyumva, J., Luo, G., Nahayo, L., Huang, X., & Cai, P. (2018). Landslide Susceptibility Assessment Using Spatial Multi-Criteria Evaluation Model in Rwanda. *International Journal of Environmental Research and Public Health*, 15(2), 243. <https://doi.org/10.3390/ijerph15020243>
- Panelli, C., Marsala, V., Paglia, G., Galli, A., Colasante, P., & Miccadei, E. (2024). Landslide susceptibility of Rwanda (Central Africa). *Journal of Maps*, 20(1), 2428654. <https://doi.org/10.1080/17445647.2024.2428654>
- Sättele, M., Bründl, M., & Straub, D. (2015). Reliability and effectiveness of early warning systems for natural hazards: Concept and application to debris flow warning. *Reliability Engineering & System Safety*, 142, 192–202. <https://doi.org/10.1016/j.ress.2015.05.003>
- Shano, L., Gisila, B., Jerene, W., Ekaso, D., Baye, T., Degife, T., Chameno, G., Dosa, Z., & Jothimani, M. (2024). Fatal landslides in Kencho, Shacha & Gozdi villages, Gofa zone, Ethiopia: A detailed investigation (Geological, Geotechnical, geophysical & geospatial) of the July 22, 2024 catastrophe and its socioeconomic repercussions. *Quaternary Science Advances*, 16, 100241. <https://doi.org/10.1016/j.qsa.2024.100241>

- Sun, Q., Du, L., Zhang, W., Chen, J., Wang, J., Zhang, H., Qin, Y., Yin, H., & Zhao, Y. (2025). Investigation into the evolutionary mechanisms of catastrophic bedding rock landslides in deeply incised canyon regions: A case study of the Zhisu Mountain landslide. *Landslides*, 22(8), 2677–2692. <https://doi.org/10.1007/s10346-025-02521-y>
- Taherdangkoo, R., Franz, L.-S., Arab, A., Wichert, J., & Butscher, C. (2026). Reliable probabilistic landslide susceptibility mapping using a calibrated stacked ensemble with bootstrap-based uncertainty. *Bulletin of Engineering Geology and the Environment*, 85(5), 333. <https://doi.org/10.1007/s10064-026-05004-z>
- Uwihirwe, J., Hrachowitz, M., & Bogaard, T. A. (2020). Landslide precipitation thresholds in Rwanda. *Landslides*, 17(10), 2469–2481. <https://doi.org/10.1007/s10346-020-01457-9>
- Ye, C., Wu, H., Oguchi, T., Tang, Y., Pei, X., & Wu, Y. (2025). Physically Based and Data-Driven Models for Landslide Susceptibility Assessment: Principles, Applications, and Challenges. *Remote Sensing*, 17(13), 2280. <https://doi.org/10.3390/rs17132280>
- Zeng, T., Wu, L., Peduto, D., Glade, T., Hayakawa, Y. S., & Yin, K. (2023). Ensemble learning framework for landslide susceptibility mapping: Different basic classifier and ensemble strategy. *Geoscience Frontiers*, 14(6), 101645. <https://doi.org/10.1016/j.gsf.2023.101645>