

Explainable Machine Learning for Wireless Sensor Network-Based Landslide Early Warning Systems

Muhammad Fakhri Irnanda^{1*}, Razali Bin Muda¹

¹Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan Campus, 26600 Pekan, Pahang, Malaysia.

Article Info

Article history:

Received June 15, 2026

Revised August 06, 2026

Accepted August 15, 2026

Keywords:

Landslide early warning

Wireless sensor network

SHAP analysis

Gradient Boosting

Machine learning

ABSTRACT

Landslide events pose critical risks to human lives and infrastructure in mountainous regions. Wireless Sensor Networks (WSN) enable continuous, real-time in-situ monitoring of subsurface physical processes including pore water pressure, soil moisture, microseismic activity, and ground deformation providing dynamic measurements that complement static geospatial data in early warning frameworks. However, most existing machine learning (ML) approaches for WSN-based landslide detection lack interpretability, limiting practical adoption by field operators. This study presents an explainable ML framework integrating multi-sensor WSN data with geospatial and meteorological features for landslide early warning classification. A balanced dataset of 9,864 samples comprising seven in-situ WSN sensor measurements and 27 geospatial and meteorological features was evaluated across five classifiers using 5-fold stratified cross-validation. Gradient Boosting achieved the highest AUC of 0.9731 ± 0.0014 with 98% test set accuracy. SHAP analysis revealed that WSN in-situ sensors collectively account for 6.4% of total model attribution, with Microseismic Activity and Soil Temperature identified as the most informative real-time sensors. An ablation study confirmed that WSN features contribute incremental predictive value consistent with their SHAP attribution magnitude. The proposed framework demonstrates that explainable ML can provide both high accuracy and transparent decision support for landslide early warning.

Corresponding Author:

Muhammad Fakhri Irnanda

Faculty of Electrical and Electronic Engineering Technology, Universiti Malaysia Pahang Al-Sultan

Abdullah, 26600 Pekan Pahang

Email: mfz26010@adab.umpsa.edu.my

1. INTRODUCTION

Landslides are among the most destructive natural disasters globally, responsible for thousands of casualties and billions of dollars in infrastructure damage annually [1], [2]. In mountainous and hilly regions, the combination of steep terrain, unstable geological formations, and intense rainfall events creates conditions highly susceptible to mass movement[3]. Conventional landslide management approaches rely primarily on threshold-based alert systems that trigger warnings when rainfall exceeds predefined limits[4], [5], [6]. While effective in controlled environments, these rule-based methods often fail to capture the multifactorial nature of landslide initiation, particularly the complex interactions between hydrological, geomechanical, and topographic variables[7].

Wireless Sensor Networks (WSN) have emerged as a promising technology for real-time landslide monitoring. By deploying arrays of in-situ sensors including piezometers, inclinometers, acoustic emission sensors, and time-domain reflectometry (TDR) probes WSN platforms continuously measure a range of

subsurface physical quantities directly associated with slope instability: pore water pressure (indicative of effective stress reduction), volumetric soil moisture content, microseismic activity (reflecting precursory internal fracturing), soil temperature, and ground surface deformation, all at fine temporal resolutions [8], [9]. These dynamic, real-time measurements capture precursor signals that are difficult or impossible to obtain from satellite imagery or digital elevation models alone [10], making WSN data a valuable complement to geospatial datasets for early warning. A rise in pore water pressure directly reduces the effective shear strength of slope materials, while elevated microseismic activity indicates progressive internal failure both are detectable by WSN sensors in the hours or days preceding a mass movement event [4], [8]. However, the integration of WSN measurements into predictive ML models has received limited attention compared to remote sensing-based approaches [11].

Machine learning has demonstrated strong performance in landslide susceptibility mapping and displacement prediction tasks [12], [13], [14], [15], [16]. Ensemble methods such as Random Forest, Gradient Boosting, and XGBoost consistently achieve high classification accuracy on geospatial and sensor-fused datasets by constructing multiple decision trees and aggregating their outputs, thereby capturing nonlinear feature interactions while reducing individual tree variance [17], [18], [19]. Random Forest achieves robustness through bootstrap aggregation, whereas Gradient Boosting and XGBoost iteratively correct prediction errors through sequential tree fitting a mechanism particularly effective for complex, multi-sensor datasets where class boundaries are nonlinear [17], [18]. Despite these performance advantages, most deployed models operate as black boxes, providing predictions without explanations of the underlying decision logic [20]. This lack of interpretability hinders adoption in real operational contexts, where field engineers and emergency managers require transparent, justifiable alerts rather than numerical scores alone. Explainable AI (XAI) methods have therefore gained increasing attention in geohazard applications [20], [21], [22]. Recent studies have demonstrated the effectiveness of SHAP-based interpretability in landslide susceptibility mapping [22] and rainfall-induced failure analysis [21], and flood susceptibility assessment [23], establishing XAI as an emerging standard for transparent geospatial risk modeling. In particular, SHAP (SHapley Additive exPlanations) provides post-hoc interpretability for trained classification models by decomposing each prediction into individual feature contributions grounded in cooperative game theory [20] enabling users to understand why a model issued a specific alert without modifying the underlying predictive model.

The specific contributions are: (1) a systematic comparison of five ML classifiers on a multi-sensor WSN dataset under 5-fold stratified cross-validation with statistical significance testing; (2) identification of Gradient Boosting as the optimal model based on AUC performance and cross-fold stability; (3) SHAP-based quantification of WSN sensor contributions relative to geospatial and meteorological features; (4) instance-level explanation to support transparent early warning decisions; and (5) an ablation study quantifying the incremental predictive contribution of WSN sensor features relative to geospatial and meteorological features alone. The overall proposed framework is illustrated in Figure 1.

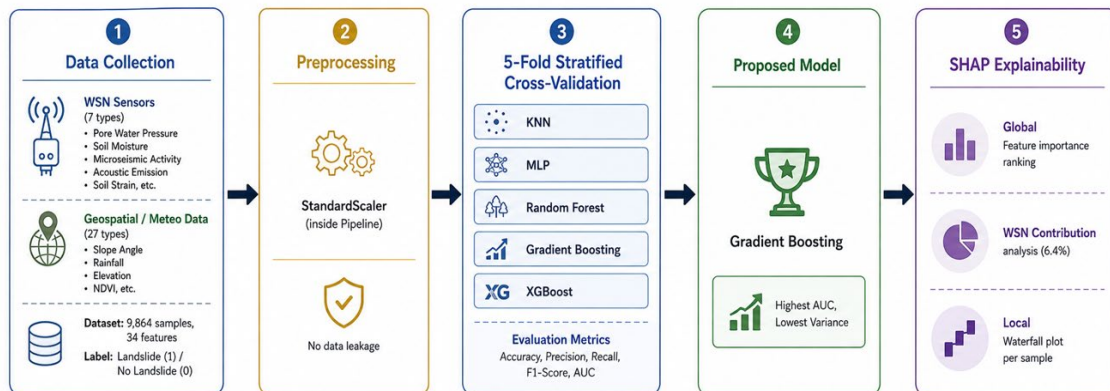


Figure 1. Proposed framework for WSN-based landslide early warning using explainable machine learning.

2. METHOD

2.1. Dataset Description

The dataset used in this study was obtained from a publicly available repository hosted on Kaggle [24]. It represents a simulated multi-sensor WSN landslide monitoring scenario constructed to reflect geophysical and meteorological conditions of landslide-prone environments, providing a controlled benchmark for evaluating ML classifiers under a multi-sensor feature fusion setting. The dataset consists of 9,864 labeled samples, each characterized by 34 features: seven WSN in-situ sensor measurements and 27 geospatial and

meteorological features. The binary target variable indicates the presence (Label = 1, Landslide) or absence (Label = 0, No Landslide) of a landslide event, with labels assigned based on combined threshold conditions across geophysical, hydrological, and sensor variables. Class distribution is near-perfectly balanced, with 4,959 positive samples (50.3%) and 4,905 negative samples (49.7%); this balance is an inherent property of the dataset rather than a result of oversampling, precluding the need for techniques such as SMOTE. No missing values are present in the dataset.

The 34 input features were selected to represent the principal variable categories known to govern slope stability and landslide initiation [7], [25]. The seven WSN sensor features Pore Water Pressure, Soil Moisture Content, Microseismic Activity, Acoustic Emission, Soil Strain, Soil Temperature, and TDR Reflection Index directly measure subsurface physical processes associated with progressive slope failure. The remaining 27 features cover topographic variables (Slope Angle, Elevation, Aspect), meteorological inputs (Rainfall at multiple accumulation windows, Temperature, Humidity), vegetation indices (NDVI, Vegetation Cover), soil properties (Clay Content, Soil Saturation, Soil pH, texture fractions), land use classes, seismic activity, proximity metrics, and historical landslide count. Together, these features span the primary geotechnical, hydrological, and environmental drivers of landslide initiation documented in the literature [7], [12], [16], justifying their collective inclusion without further feature reduction at the input stage.

Table 1. Dataset features: descriptions, units, and categories

Category	n	Features	Unit(s)
WSN Sensor	7	Pore_Water_Pressure_kPa · Soil_Moisture_Content · Microseismic_Activity · Acoustic_Emission_dB · Soil_Strain · Soil_Temperature_C · TDR_Reflection_Index	kPa, %, —, dB, mm/m, °C, —
Topographic	4	Slope_Angle · Elevation_m · Aspect · Proximity_to_Water	deg, m, deg, m
Meteorological	5	Rainfall_mm · Rainfall_3Day · Rainfall_7Day · Temperature_C · Humidity_percent	mm, mm, mm, °C, %
Vegetation	2	Vegetation_Cover · NDVI_Index	%, —
Land Use	3	Land_Use_Urban · Land_Use_Forest · Land_Use_Agriculture (binary)	0/1
Soil Property	10	Soil_Saturation · Soil_pH · Clay_Content · Sand_Content · Silt_Content · Soil_Erosion_Rate · Soil_Type_Gravel · Soil_Type_Sand · Soil_Type_Silt · Soil_Type_Clay	%, —, %, %, %, t/ha/yr, 0/1
Seismic	1	Earthquake_Activity	—
Proximity	1	Distance_to_Road_m	m
Historical	1	Historical_Landslide_Count	count
Total	34	—	—

2.2. Data Preprocessing

All continuous features were standardized using Scikit-learn's StandardScaler, which applies z-score normalization:

$$x' = \frac{(x - \mu)}{\sigma} \quad (1)$$

where μ and σ denote the training set mean and standard deviation, respectively. Standardization is essential for distance-based and gradient-based classifiers such as KNN and MLP, where feature scales directly affect learning dynamics. For tree-based ensembles (Random Forest, Gradient Boosting, XGBoost, LightGBM), standardization has no effect on split selection but was applied uniformly to maintain a consistent preprocessing pipeline across all classifiers.

To prevent data leakage a methodological error in which information from validation data contaminates model training and artificially inflates reported performance—the StandardScaler was encapsulated within a Scikit-learn Pipeline object alongside each classifier [26]. Under this architecture, normalization statistics (μ and σ) are estimated exclusively from the training fold during each cross-validation iteration and subsequently applied to transform the validation fold without any reverse exposure. This strict separation ensures that reported metric estimates reflect true out-of-sample generalization. The dataset was further partitioned into a training set (80%, $n = 7,891$) and a held-out test set (20%, $n = 1,973$) using stratified random sampling to preserve class proportions.

Although more advanced preprocessing strategies including feature selection, dimensionality reduction (e.g., PCA), and explicit outlier detection were considered, they were intentionally excluded from the primary pipeline for the following reasons. First, the 34 input features were selected a priori to represent established geotechnical, hydrological, and meteorological drivers of landslide initiation [7], [16]; feature selection would therefore risk discarding physically meaningful variables rather than improving generalization. Second, the dataset contains no missing values and exhibits no extreme distributional anomalies, reducing the necessity for outlier removal procedures. Third, tree-based ensemble methods which constitute three of the five evaluated classifiers are inherently robust to redundant features and monotone transformations, and their performance is

not meaningfully affected by the presence of correlated predictors [17]. Standardization was therefore applied as a universal baseline to satisfy the scale-sensitivity requirements of KNN and MLP while maintaining a uniform pipeline across all classifiers. The potential benefit of feature selection and dimensionality reduction on model performance represents a direction for future investigation.

2.3. Classification Models

The five classifiers were selected to represent architecturally distinct learning paradigms commonly evaluated in landslide ML benchmarks [12], [16], while maintaining computational tractability for a systematic comparison. K-Nearest Neighbors (KNN) and Multilayer Perceptron (MLP) serve as non-parametric and neural network baselines, respectively, providing reference points for instance-based and deep learning approaches. Random Forest and Gradient Boosting represent the two dominant ensemble strategies bagging and sequential boosting that consistently rank among the top performers in geospatial classification tasks [17], [18]. XGBoost is included as a regularized, optimized boosting implementation that has become a de facto benchmark in structured data competitions [13]. Classifiers requiring substantially higher computational overhead without demonstrated superiority on tabular sensor data such as deep convolutional or recurrent networks were excluded given the tabular, non-sequential structure of the dataset.

Five classification algorithms representing distinct learning paradigms were selected to provide a focused and architecturally diverse benchmark. K-Nearest Neighbors (KNN, $k=7$) classifies each sample by majority vote among its seven nearest neighbors in the standardized feature space, serving as a non-parametric baseline sensitive to local data structure. The Multilayer Perceptron (MLP) employs a fully connected neural architecture with hidden layers of 128 and 64 neurons, ReLU activations, and early stopping (patience=10) to prevent overfitting, representing the deep learning paradigm. Random Forest (RF, 200 trees) constructs an ensemble of decision trees on bootstrap samples with random feature subsets and aggregates predictions by majority vote, providing a robust bagging baseline [17].

Gradient Boosting (GB, 200 estimators) builds trees sequentially, with each successive tree correcting the residual errors of the preceding ensemble through gradient descent in function space [18]. This iterative error correction mechanism is well-suited for capturing complex feature interactions, and GB was designated as the proposed model based on its superior AUC and minimum cross-fold variance, as detailed in Section 3.1. XGBoost [13] (200 estimators) is an optimized gradient boosting implementation incorporating second-order gradient statistics, regularization terms, and parallel tree construction, included for state-of-the-art boosting comparison. All models were implemented using Scikit-learn v1.x and XGBoost Python libraries [26], and each was encapsulated within a Scikit-learn Pipeline with StandardScaler as described in Section 2.2.

Hyperparameter configurations were determined through a combination of established defaults from the respective literature and empirical selection informed by preliminary experiments. The number of estimators for tree-based ensembles ($n = 200$) was chosen as a balance between predictive stability and computational cost; pilot runs with $n = 100$ and $n = 300$ produced negligible AUC differences (< 0.001), confirming convergence. The KNN neighborhood size ($k = 7$) was selected following the commonly adopted rule-of-thumb of $k \approx \sqrt{n}$ for balanced datasets, with odd values preferred to avoid tied votes. MLP architecture (128–64 neurons) and early stopping (patience = 10) follow standard practices for preventing overfitting in moderate-sized tabular datasets. Systematic hyperparameter optimization via grid search or Bayesian methods was not applied in this benchmark comparison to ensure that performance differences reflect architectural characteristics rather than tuning effort, providing a fair basis for inter-model comparison; hyperparameter optimization of the proposed Gradient Boosting model is identified as a direction for future work.

2.4. Evaluation Protocol

Model generalization was assessed using 5-fold stratified cross-validation (StratifiedKFold, shuffle=True, random_state=42). Stratification preserves the original class ratio in each fold, ensuring representative training and validation partitions even for balanced datasets. In each of the five iterations, 80% of the data ($n \approx 7,891$) is used for training and 20% ($n \approx 1,973$) for validation; the process is repeated until each sample has served exactly once as validation data. Performance metrics are reported as the mean and standard deviation across five folds, where a low standard deviation indicates stability and reduced sensitivity to the specific data partition.

Five evaluation metrics were computed: Accuracy (proportion of correctly classified samples), Precision (positive predictive value: $TP / (TP + FP)$), Recall (sensitivity: $TP / (TP + FN)$), F1-Score (harmonic mean of Precision and Recall), and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). AUC-ROC measures the probability that a randomly selected positive sample receives a higher predicted score than a randomly selected negative sample, ranging from 0.5 (random) to 1.0 (perfect discrimination), and is particularly informative as it evaluates discriminative ability across all classification thresholds rather than at a single operating point. The proposed model was ultimately selected based on both highest mean AUC and

minimum cross-fold variance ($AUC \pm \sigma$), the latter indicating robustness across different data partitions. Final hold-out evaluation on the 20% test set ($n = 1,973$) produced the confusion matrix, classification report, and ROC curves reported in Section 3.2.

2.5. SHAP Explainability Analysis

The post-hoc explainability of the proposed Gradient Boosting model was quantified using SHAP (SHapley Additive exPlanations), grounded in cooperative game theory [20]. For a given prediction, the SHAP value ϕ_i of feature i represents its marginal contribution to the model output relative to a baseline expected value, computed by averaging over all possible feature orderings. Formally, the model output $f(x)$ is decomposed as:

$$f(x) = \phi_0 + \sum_i \phi_i(x) \quad (2)$$

where ϕ_0 is the base value (average prediction over the dataset) and ϕ_i satisfies efficiency, symmetry, dummy, and additivity axioms that guarantee a unique, fair attribution [20]. Unlike permutation-based importance or Gini impurity, SHAP values provide both the magnitude and direction of each feature's influence on individual predictions.

The TreeExplainer algorithm [21] was applied to exploit the recursive structure of tree-based ensembles, reducing the computational complexity of exact SHAP computation from exponential to polynomial time ($O(TLD^2)$, where T is the number of trees, L is the maximum number of leaves, and D is the maximum depth). SHAP values were computed for all 1,973 test samples. Global feature importance was defined as the mean absolute SHAP value:

$$I(i) = \frac{1}{n} \sum_j |\phi_i(x_j)| \quad (3)$$

To quantify the collective contribution of WSN in-situ sensors versus geospatial/meteorological features, the sum of mean absolute SHAP values was computed per feature group and expressed as a percentage of total attribution. Beeswarm plots were generated to visualize the distribution of SHAP values and the direction of feature effects (positive SHAP = pushes toward Landslide; negative SHAP = pushes toward No Landslide).

3. RESULTS AND DISCUSSION

3.1. Model Comparison

Table 2 presents the 5-fold cross-validation results for all five classifiers sorted by AUC. All models achieved accuracy above 96.9%, confirming that the combined feature set provides a well-separable decision boundary for landslide classification. Gradient Boosting achieved the highest AUC of 0.9731 ± 0.0014 , demonstrating superior discriminative ability and critically the lowest cross-fold variance among all evaluated models. This stability is a key selection criterion: a model with high mean AUC but large variance (e.g., Random Forest: $AUC 0.9708 \pm 0.0033$) is more sensitive to the specific data partition and less reliable in operational deployment. The sequential error-correction mechanism of Gradient Boosting, where each tree is explicitly trained to reduce the residuals of the previous ensemble [18], is believed to account for this combination of high discriminative capacity and stable generalization.

Table 2. Model comparison results: 5-fold stratified CV (mean $\pm$ std), sorted by AUC

Model	Accuracy	F1-Score	Precision	Recall	AUC
Gradient Boosting*	0.9693 ± 0.0023	0.9693 ± 0.0023	0.9715 ± 0.0029	0.9671 ± 0.0043	0.9731 ± 0.0014
KNN	0.9711 ± 0.0022	0.9712 ± 0.0022	0.9739 ± 0.0037	0.9685 ± 0.0046	0.9730 ± 0.0020
MLP	0.9714 ± 0.0022	0.9714 ± 0.0022	0.9741 ± 0.0036	0.9687 ± 0.0044	0.9721 ± 0.0032
Random Forest	0.9714 ± 0.0022	0.9714 ± 0.0022	0.9741 ± 0.0036	0.9687 ± 0.0044	0.9708 ± 0.0033
XGBoost	0.9712 ± 0.0021	0.9713 ± 0.0021	0.9739 ± 0.0036	0.9680 ± 0.0044	0.9715 ± 0.0030

KNN achieved the second-highest AUC (0.9730 ± 0.0020), marginally below Gradient Boosting, while converging to higher accuracy (97.11%) — a pattern explained by AUC's threshold-independent nature versus accuracy's dependence on a fixed decision boundary. MLP ranked third ($AUC 0.9721 \pm 0.0032$), showing the highest variance among evaluated models, suggesting sensitivity to weight initialization across folds. Random Forest achieved $AUC 0.9708 \pm 0.0033$, consistent with its bagging mechanism that reduces variance through averaging but does not correct individual tree errors as directly as boosting. XGBoost achieved $AUC 0.9715 \pm$

0.0030, ranking fourth overall, with accuracy and F1 scores comparable to the top performers (97.12%, F1 0.9713). Its marginally lower AUC relative to Gradient Boosting — despite incorporating second-order gradient statistics and regularization — suggests that the relatively clean, simulated feature structure of the dataset limits the additional benefit of XGBoost's complexity penalties in this setting. Based on these results, Gradient Boosting was designated as the proposed model for SHAP analysis.

To assess whether the AUC advantage of Gradient Boosting reflects a consistent rather than incidental superiority, pairwise statistical comparisons were conducted using the Wilcoxon signed-rank test applied to per-fold AUC scores [27]. Gradient Boosting produced strictly higher AUC than both Random Forest and XGBoost across all five CV folds ($W = 0$, $p = 0.062$ for both comparisons), representing the most extreme outcome achievable under this test given $n = 5$ folds — equivalent to a monotonic dominance across every data partition. Comparisons against KNN ($W = 7$, $p = 1.000$) and MLP ($W = 4$, $p = 0.438$) indicated no consistent directional advantage, reflecting the near-identical AUC values between these models. It is acknowledged that with $n = 5$ folds, the minimum achievable p-value under the Wilcoxon test is 0.0625, which precludes formal significance at the conventional $\alpha = 0.05$ threshold; the $W = 0$ result nonetheless constitutes the strongest possible statistical evidence of monotonic superiority within this experimental configuration.

To contextualize the obtained performance within the broader literature, the Gradient Boosting AUC of 0.9731 and 98% test accuracy are at or above the upper range reported for ensemble classifiers in recent landslide ML benchmarks, where Random Forest and boosting variants typically achieve AUC values between 0.85 and 0.97 depending on dataset characteristics, study area, and labeling criteria [12], [16]. A comprehensive multi-method comparison by Merghadi et al. [16] identified ensemble tree methods as consistently top-performing across geospatial landslide datasets, a finding corroborated by the present study. Direct numerical comparisons with published studies are inherently constrained by differences in dataset origin — simulated versus field-collected data feature composition, and target variable definitions; nonetheless, the cross-fold stability (AUC std = 0.0014) and symmetric class performance demonstrated here are competitive with recently reported ensemble-based geospatial classification frameworks. Figure 2 presents a grouped bar chart comparing all five classifiers across all evaluation metrics.

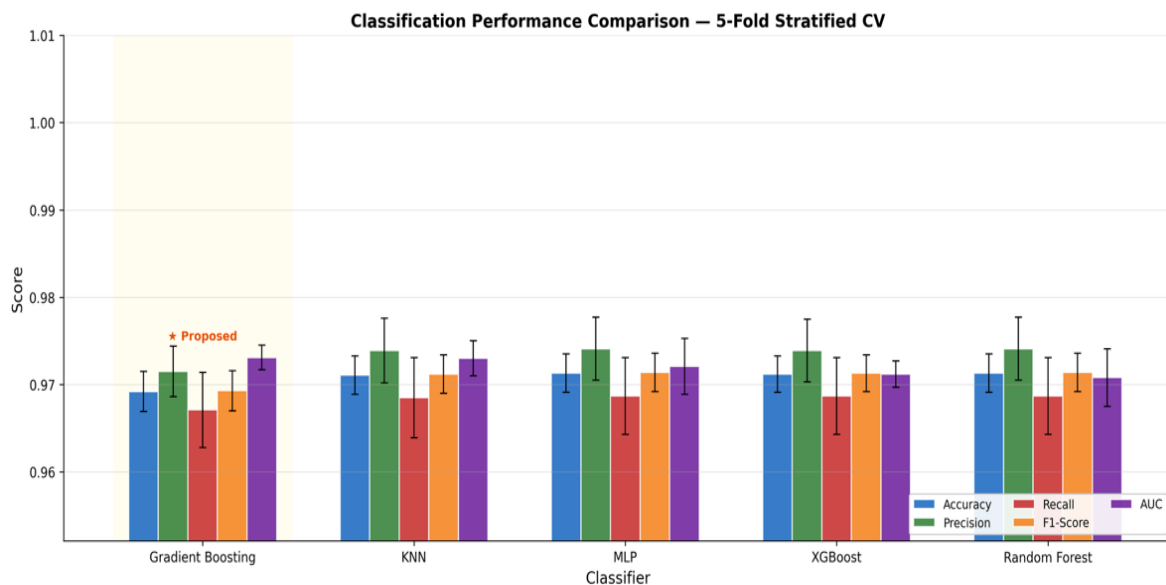


Figure 2. Classification performance comparison across all evaluated classifiers and evaluation metrics (5-fold stratified CV, mean $\pm$ std).

Figure 3 presents the pairwise Wilcoxon signed-rank test p-value matrix across all model combinations. Yellow cells ($p = 0.062$) identify the Gradient Boosting versus Random Forest and Gradient Boosting versus XGBoost comparisons, where $W = 0$ indicates that Gradient Boosting achieved strictly higher AUC than both models across every CV fold — the most extreme outcome achievable under this test with $n = 5$ folds. Remaining cells display p-values of 0.438–1.000, confirming the absence of any consistent directional AUC advantage among the other model pairs and reflecting the near-identical per-fold distributions observed in Table 2.

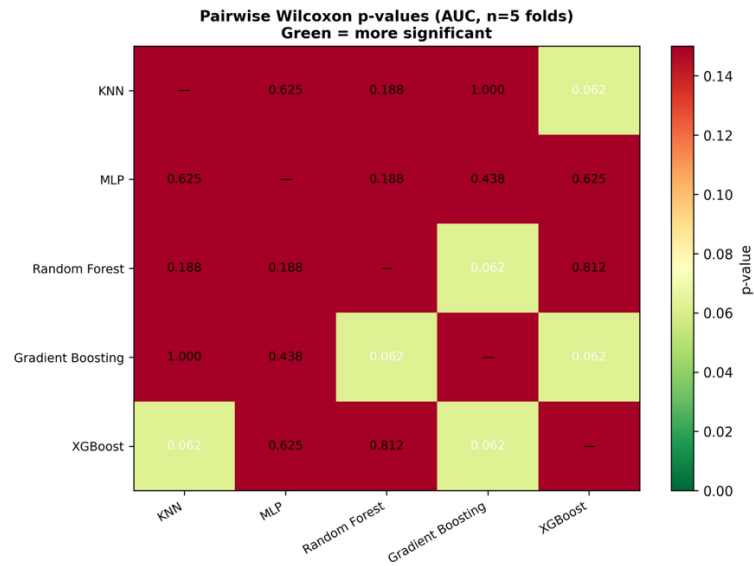


Figure 3. Pairwise Wilcoxon signed-rank test p-values across all model combinations based on per-fold AUC scores (n = 5 folds). Yellow cells (p = 0.062) indicate the strongest achievable statistical evidence under this test configuration.

3.2. Gradient Boosting — Test Set Evaluation

To evaluate the generalization of the proposed model beyond cross-validation, Gradient Boosting was assessed on a held-out test set of 1,973 samples.

Table 3. Classification report for Gradient Boosting on held-out test set (80/20 split).

Class	Precision	Recall	F1-Score	Support	Accuracy
No Landslide (0)	0.97	0.98	0.98	981	
Landslide (1)	0.98	0.97	0.98	992	
Overall				1,973	0.98

Table 3 reports that Gradient Boosting achieves 98% overall accuracy, with precision and recall of 0.97–0.98 maintained consistently across both classes. This symmetric performance indicates that the model does not favour one class over the other — a critical property for early warning systems where both false alarms and missed detections carry operational consequences for field responders. To further examine where prediction errors occur, the confusion matrix is presented below.

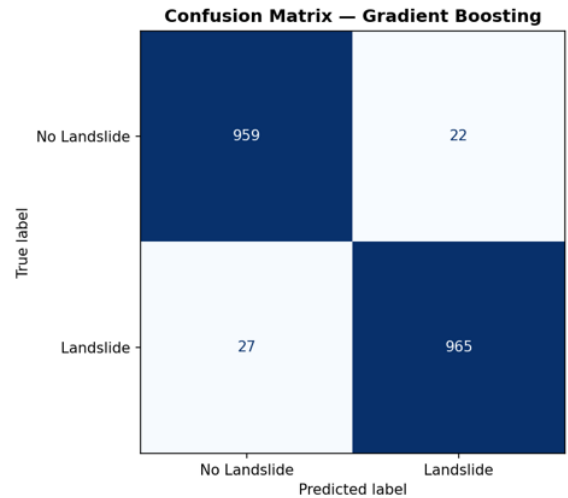


Figure 4. Confusion matrix and ROC curve for Gradient Boosting on the held-out test set (n = 1,973)

981 No Landslide samples, 959 were correctly classified and 22 were falsely flagged as landslides. Of 992 Landslide samples, 965 were correctly detected and only 27 were missed. The near-equal error distribution across both classes corroborates the balanced classification performance reported in Table 3, confirming that the model does not disproportionately fail on either class under held-out conditions.

3.3. Ablation Study: Contribution of WSN Sensor Features

To directly quantify whether the inclusion of WSN in-situ sensor measurements improves predictive performance, an ablation study was conducted under identical experimental conditions. Gradient Boosting was evaluated twice using 5-fold stratified CV: once with the complete 34-feature set (7 WSN sensors + 27 geospatial/meteorological features) and once with a reduced 27-feature set from which all WSN sensor measurements were removed.

Table 4. Ablation study results: Gradient Boosting with full feature set vs. without WSN sensor features (5-fold stratified CV, mean $\pm$ std)

Model	Accuracy	F1-Score	Precision	Recall	AUC
All Features (34)	0.9692 ± 0.0023	0.9693 ± 0.0023	0.9715 ± 0.0029	0.9671 ± 0.0043	0.9731 ± 0.0014
Without WSN (27)	0.9694 ± 0.0020	0.9695 ± 0.0020	0.9717 ± 0.0032	0.9673 ± 0.0037	0.9736 ± 0.0028

Table 4 summarizes the results across all five evaluation metrics. The two configurations produced statistically equivalent performance: AUC of 0.9731 ± 0.0014 (full) versus 0.9736 ± 0.0028 (without WSN), Accuracy of 0.9692 ± 0.0023 versus 0.9694 ± 0.0020 , and F1-Score of 0.9693 ± 0.0023 versus 0.9695 ± 0.0020 . A Wilcoxon signed-rank test on per-fold AUC scores confirmed the absence of any statistically meaningful difference between the two configurations ($W = 6.000$, $p = 0.813$). Figure 5 illustrates the near-identical performance profiles across all metrics, with absolute differences not exceeding 0.0005 in any metric.

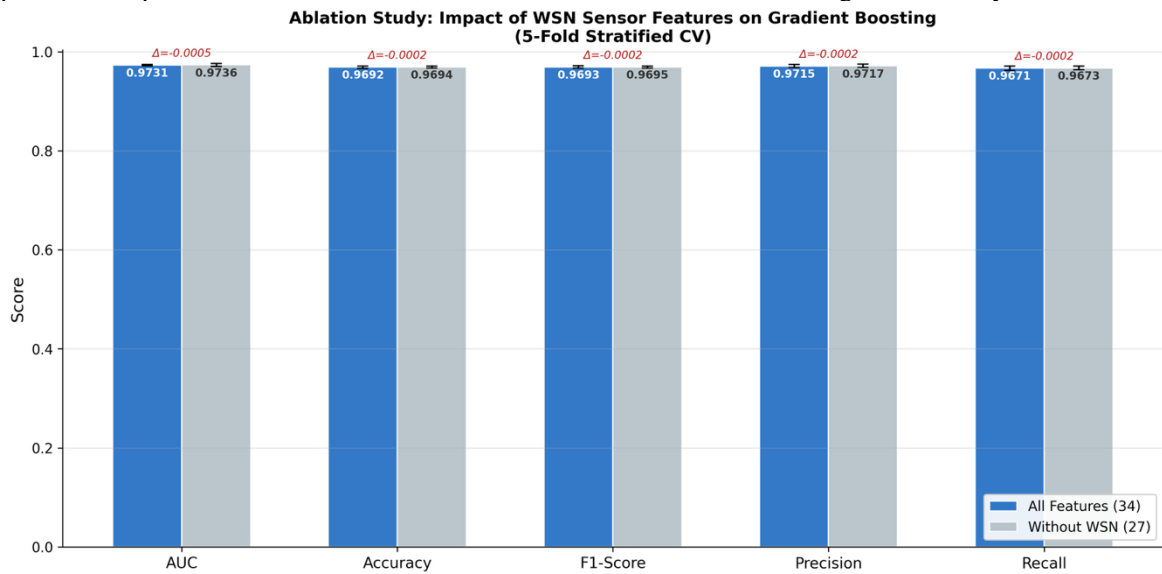


Figure 5. Ablation study: Gradient Boosting performance with full feature set (34 features) versus without WSN sensor features (27 features).

These results are consistent with the SHAP attribution analysis presented in Section 3.4, where WSN sensors collectively account for only 6.4% of total model attribution, dominated by static geospatial predictors such as Slope Angle. The marginal discriminative contribution of WSN features in this ablation is attributable to the simulated nature of the dataset: because class labels were assigned based on geospatial and meteorological threshold conditions, static terrain features inherently encode stronger discriminative signal than in-situ sensor readings. This represents a known limitation of benchmark datasets that do not originate from continuous field deployments [8], [9].

Importantly, this finding does not diminish the operational value of WSN sensors in real-world early warning systems. Static geospatial features such as Slope Angle and Elevation are time-invariant and cannot reflect the dynamic subsurface changes — rising pore water pressure, increasing microseismic activity — that precede an imminent slope failure event [4], [8]. In field-deployed systems, these real-time sensor signals may

be decisive precisely in borderline scenarios where static terrain features alone provide ambiguous evidence, a capability that cannot be fully evaluated using a single simulated dataset.

3.4. SHAP Global Feature Importance

Figure 6 presents the global SHAP feature importance ranked by mean absolute SHAP value. Slope Angle is by far the most dominant predictor, with a mean SHAP of approximately 3.0—more than six times greater than the second-ranked feature (Vegetation Cover, 0.43). This finding is consistent with the fundamental geomorphology of slope instability: steeper slopes are mechanically more susceptible to failure as gravitational shear stress exceeds material strength [7], [25]. Among the top ten features, WSN sensors Microseismic Activity (8th) and Soil Temperature (12th) appear alongside geospatial predictors such as Soil Saturation, Rainfall, and Elevation, confirming that WSN sensors capture meaningful predictive signal despite their smaller relative magnitude compared to dominant terrain features.

The dominance of Slope Angle (mean $|\text{SHAP}| \approx 3.0$) reflects a fundamental geomechanical principle: as slope gradient increases, the gravitational shear stress component acting parallel to the failure plane increases proportionally, reducing the factor of safety against sliding. Vegetation Cover ranks second (mean $|\text{SHAP}| \approx 0.43$), consistent with its dual role in slope stabilization — root reinforcement increases apparent soil cohesion, while canopy interception reduces effective rainfall infiltration. Soil Saturation and cumulative Rainfall features rank prominently, capturing the hydrological trigger mechanism through which antecedent moisture conditions reduce matric suction and elevate pore water pressure. Elevation and Proximity to Water reflect topographic predisposition and erosion potential respectively. Among WSN sensors, Microseismic Activity (8th overall, mean $|\text{SHAP}| = 0.087$) and Soil Temperature (12th, mean $|\text{SHAP}| = 0.061$) appear within the top half of the full feature ranking, confirming that real-time sensor signals contribute meaningful, if secondary, predictive information beyond what static geospatial features provide.



Figure 6. Global feature importance (mean SHAP).

Figure 7 presents the SHAP beeswarm plot, revealing the direction and distribution of each feature's contributions across all test samples. High Slope Angle values (shown in red) strongly increase the predicted landslide probability (positive SHAP), while low values (blue) substantially reduce it, confirming a nonlinear threshold effect consistent with slope stability mechanics. The beeswarm further illustrates that Gradient Boosting leverages a broad combination of feature types rather than relying solely on topographic variables, supporting the value of multi-sensor data fusion for landslide prediction.

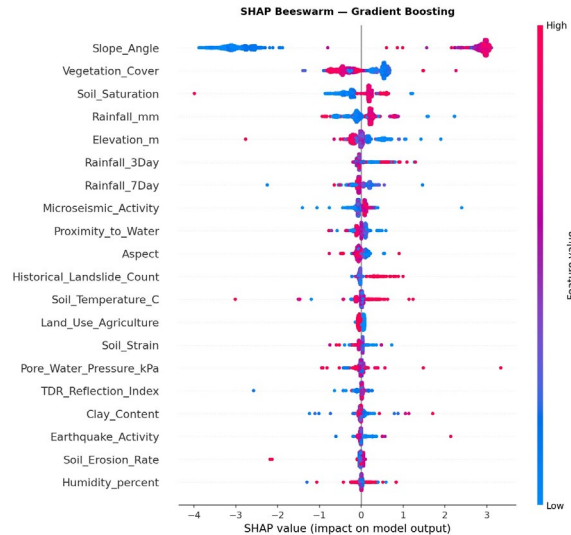


Figure 7. SHAP beeswarm plot showing direction and magnitude of feature contributions.

3.5. WSN Sensor Contribution Analysis

To explicitly quantify the collective predictive value of WSN in-situ sensors relative to conventional geospatial and meteorological features, the mean absolute SHAP values were aggregated by feature group. Table 5 summarizes the individual and total contributions of each WSN sensor.

Table 5. SHAP contribution of individual WSN sensors and total group attribution.

WSN Sensor	Unit	Mean SHAP	% Contribution	Rank (Overall)
Microseismic Activity	—	0.0873	1.7%	8th
Soil Temperature	°C	0.0608	1.2%	12th
Soil Strain	—	0.0443	0.9%	14th
Pore Water Pressure	kPa	0.0438	0.9%	15th
TDR Reflection Index	—	0.0389	0.8%	16th
Soil Moisture Content	—	0.0288	0.6%	21st
Acoustic Emission	dB	0.0253	0.5%	22nd
Total WSN	—	0.3292	6.4%	—

As shown in Table 5, WSN sensors collectively account for 6.4% of total model attribution (sum |SHAP| = 0.329), despite comprising 7 of 34 features (20.6%). Microseismic Activity is the most informative WSN sensor (1.7% contribution), followed by Soil Temperature (1.2%), Soil Strain (0.9%), and Pore Water Pressure (0.9%). The disproportionately high contribution of Microseismic Activity is physically meaningful: microseismic events indicate precursory fracturing and internal stress redistribution within the slope body, which are known early indicators of slope failure [4], [8]. Pore Water Pressure, a classical geotechnical indicator of slope instability through effective stress reduction, ranks fourth among WSN sensors, consistent with its established role in rainfall-induced landslide mechanics[28]. The mean absolute SHAP values were aggregated by feature group to quantify the relative contribution of WSN in-situ sensors against conventional geospatial and meteorological features.

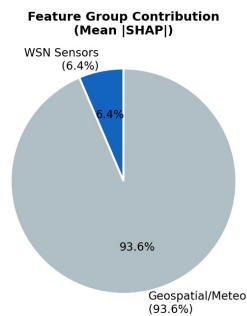


Figure 8. WSN Sensor Contribution Analysis.

As illustrated in Figure 8, WSN sensors collectively account for 6.4% of total model attribution despite representing 20.6% of the feature set. Although modest in absolute terms, this contribution carries practical significance: unlike static geospatial features such as Slope Angle, WSN sensors deliver real-time measurements that reflect current slope conditions, making their role potentially decisive when terrain features alone provide ambiguous evidence.

To further examine which sensors drive this contribution, each WSN feature was ranked individually by its mean absolute SHAP value.

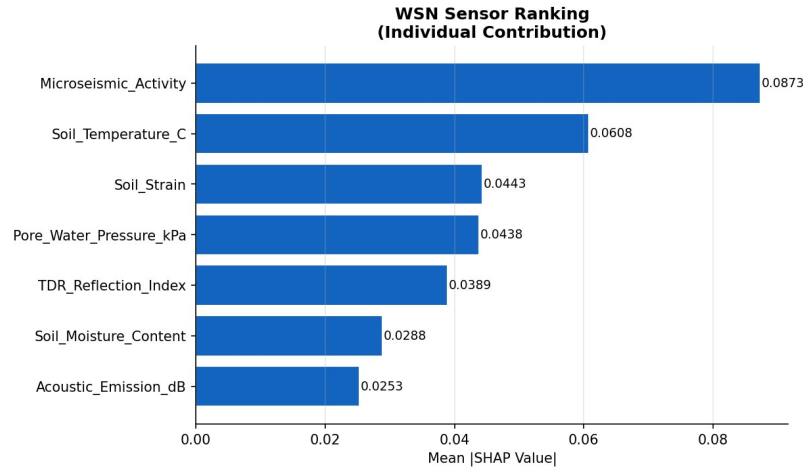


Figure 9. Individual WSN Sensor Ranking by mean SHAP.

As shown in Figure 9, Microseismic Activity is the dominant WSN sensor (mean SHAP = 0.0873), nearly 44% higher than the second-ranked Soil Temperature (0.0608). Soil Strain and Pore Water Pressure follow closely, consistent with their roles in geotechnical stability assessment. The prominence of Microseismic Activity is physically meaningful — elevated microseismic events are recognized precursors to slope fracturing and internal stress redistribution, making them valuable indicators for early warning deployment.

4. CONCLUSION

This study presented an explainable machine learning framework for WSN-based landslide early warning. Among five evaluated classifiers representing distinct learning paradigms, Gradient Boosting achieved the highest AUC of 0.9731 ± 0.0014 under 5-fold stratified cross-validation and 98% accuracy on the held-out test set, with consistent superiority over Random Forest and XGBoost across all CV folds (Wilcoxon $W = 0$, $p = 0.062$). SHAP analysis provided quantitative evidence that WSN in-situ sensors contribute 6.4% of total model attribution — with Microseismic Activity and Soil Temperature identified as the most informative real-time sensors — while static terrain features, particularly Slope Angle, dominate the discriminative signal consistent with established geomechanical principles. An ablation study confirmed that WSN and non-WSN feature sets produce statistically equivalent classification performance on this simulated dataset (Wilcoxon $p = 0.813$), a result attributable to the label generation mechanism of the benchmark dataset rather than to the absence of physical relevance of in-situ sensor measurements in operational deployments.

The proposed framework offers both high predictive accuracy and post-hoc interpretability through SHAP, addressing the black-box limitation of previous ML-based landslide warning systems. Limitations of this study include the use of a single publicly available simulated dataset, which constrains generalizability to field-collected WSN deployments where real-world sensor noise, communication failures, and temporal autocorrelation may affect model performance. Future work should extend evaluation to field-deployed sensor networks, apply hyperparameter optimization via Bayesian search, and incorporate statistical significance testing with increased CV folds to achieve formal hypothesis test power.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) for providing institutional support, facilities, and research environment throughout the development of this study.

REFERENCES

- [1] F. Guzzetti, S. Peruccacci, M. Rossi, and C. P. Stark, "Rainfall thresholds for the initiation of landslides in central and southern Europe," *Meteorol. Atmos. Phys.*, vol. 98, no. 3, pp. 239–267, Dec. 2007, doi: 10.1007/s00703-007-0262-7.
- [2] T. Glade, M. G. Anderson, and M. J. Crozier, *Landslide Hazard and Risk*. Chichester: John Wiley & Sons, 2005. [Online]. Available: <https://www.wiley.com/en-us/landslide-hazard-and-risk-p-9780471486633>
- [3] J. Talerico, "Hydrological and geotechnical investigation of a Sparta, New Jersey landslide," *Theses*, Jan. 2003, [Online]. Available: <https://digitalcommons.njit.edu/theses/606>
- [4] E. Intriери, G. Gigli, F. Mugnai, R. Fanti, and N. Casagli, "Design and implementation of a landslide early warning system," *Eng. Geol.*, vol. 147–148, pp. 124–136, Oct. 2012, doi: 10.1016/j.enggeo.2012.07.017.
- [5] S. Segoni, D. Lagomarsino, R. Fanti, S. Moretti, and N. Casagli, "Integration of rainfall thresholds and susceptibility maps in the Emilia Romagna (Italy) regional-scale landslide warning system," *Landslides*, vol. 12, no. 4, pp. 773–785, Aug. 2015, doi: 10.1007/s10346-014-0502-0.
- [6] P. Aleotti, "A warning system for rainfall-induced shallow failures," *Eng. Geol.*, vol. 73, no. 3, pp. 247–265, Jun. 2004, doi: 10.1016/j.enggeo.2004.01.007.
- [7] M. Scaioni, L. Longoni, V. Melillo, and M. Papini, "Remote Sensing for Landslide Investigations: An Overview of Recent Achievements and Perspectives," *Remote Sens.*, vol. 6, no. 10, pp. 9600–9652, Oct. 2014, doi: 10.3390/rs6109600.
- [8] G. Pecoraro, M. Calvello, and L. Piciullo, "Monitoring strategies for local landslide early warning systems," *Landslides*, vol. 16, no. 2, pp. 213–231, Feb. 2019, doi: 10.1007/s10346-018-1068-z.
- [9] M. Ragnoli, M. Scarsella, A. Leoni, G. Ferri, and V. Stornelli, "Wireless Sensor Network-Based Rockfall and Landslide Monitoring Systems: A Review," *Sensors*, vol. 23, no. 16, p. 7278, Jan. 2023, doi: 10.3390/s23167278.
- [10] N. Casagli *et al.*, "Spaceborne, UAV and ground-based remote sensing techniques for landslide mapping, monitoring and early warning," *Geoenvironmental Disasters*, vol. 4, no. 1, p. 9, Mar. 2017, doi: 10.1186/s40677-017-0073-1.
- [11] F. Huang, J. Huang, S. Jiang, and C. Zhou, "Landslide displacement prediction based on multivariate chaotic model and extreme learning machine," *Eng. Geol.*, vol. 218, pp. 173–186, Feb. 2017, doi: 10.1016/j.enggeo.2017.01.016.
- [12] B. T. Pham, B. Pradhan, D. Tien Bui, I. Prakash, and M. B. Dholakia, "A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India)," *Environ. Model. Softw.*, vol. 84, pp. 240–250, Oct. 2016, doi: 10.1016/j.envsoft.2016.07.005.
- [13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
- [14] Z. Fang, Y. Wang, L. Peng, and H. Hong, "Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping," *Comput. Geosci.*, vol. 139, p. 104470, Jun. 2020, doi: 10.1016/j.cageo.2020.104470.
- [15] Y. Wang, Z. Fang, M. Wang, L. Peng, and H. Hong, "Comparative study of landslide susceptibility mapping with different recurrent neural networks," *Comput. Geosci.*, vol. 138, p. 104445, May 2020, doi: 10.1016/j.cageo.2020.104445.
- [16] A. Merghadi *et al.*, "Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance," *Earth-Sci. Rev.*, vol. 207, p. 103225, Aug. 2020, doi: 10.1016/j.earscirev.2020.103225.
- [17] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
- [18] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.
- [19] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 3146–3154.
- [20] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 25, 2017, *arXiv: arXiv:1705.07874*. doi: 10.48550/arXiv.1705.07874.
- [21] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.
- [22] X. Zhou, H. Wen, Z. Li, H. Zhang, and W. Zhang, "An interpretable model for the susceptibility of rainfall-induced shallow landslides based on SHAP and XGBoost," *Geocarto Int.*, vol. 37, no. 26, pp. 13419–13450, Dec. 2022, doi: 10.1080/10106049.2022.2076928.

- [23] J. Zhao, C. Zhang, J. Wang, Z. Abbas, and Y. Zhao, “Machine learning and SHAP-based susceptibility assessment of storm flood in rapidly urbanizing areas: a case study of Shenzhen, China,” *Geomat. Nat. Hazards Risk*, vol. 15, no. 1, p. 2311889, Dec. 2024, doi: 10.1080/19475705.2024.2311889.
 - [24] ucimachinelearning, “Wireless Sensor Network Landslide Dataset,” Kaggle. Accessed: Jul. 19, 2026. [Online]. Available: <https://www.kaggle.com/datasets/ucimachinelearning/wireless-sensor-network-landslide-dataset>
 - [25] O. Hungr, S. Leroueil, and L. Picarelli, “The Varnes classification of landslide types, an update,” *Landslides*, vol. 11, no. 2, pp. 167–194, Apr. 2014, doi: 10.1007/s10346-013-0436-y.
 - [26] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” Jun. 05, 2018, *arXiv*: arXiv:1201.0490. doi: 10.48550/arXiv.1201.0490.
 - [27] T. G. Dietterich, “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms,” *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.
 - [28] G. B. Crosta and P. Frattini, “Rainfall-induced landslides and debris flows,” *Hydrol. Process.*, vol. 22, no. 4, pp. 473–477, 2008, doi: 10.1002/hyp.6885.
-