Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 4091~4099
e-ISSN: 2548-6861

Benchmarking Random Forest, Support Vector Machine, and XGBoost for Flood Risk Classification Using a Synthetic Dataset: A Case Study of Indramayu Regency

Nur Budi Nugraha [1]*, Rendi [2]*, Yaqutina Marjani Santosa [3]*
* Department of Informatics Engineering, Politeknik Negeri Indramayu
nurbudinugraha@polindra.ac.id [1], rendi@polindra.ac.id [2], yaqutinams@polindra.ac.id [3]

Article Info

Article history:

Received 2026-07-25
Revised 2026-08-01
Accepted 2026-08-13

Keyword:

Random Forest,
Support Vector Machine,
XGBoost,
Flood Risk Classification,
Benchmarking,
Machine Learning, Synthetic Dataset, Probability Calibration

ABSTRACT

Flood risk assessment plays a critical role in disaster mitigation planning, yet flood classification studies in Indonesia have largely relied on a single machine learning algorithm without systematically evaluating alternative classifiers under identical experimental conditions. This study benchmarks Random Forest (RF), Support Vector Machine (SVM), and XGBoost for three-class flood risk classification (Low, Medium, and High) in Indramayu Regency, Indonesia, using a literature-informed synthetic dataset of 4,500 samples generated from seven flood-related features. To ensure a fair comparison, all models were trained and evaluated under an identical preprocessing, hyperparameter optimization, and validation framework, with Logistic Regression and Decision Tree included as baseline classifiers. Experimental results show that SVM achieved the highest predictive performance with an accuracy of 88.56% and a macro F1-score of 86.83%, followed closely by XGBoost (88.44% accuracy, 86.76% macro F1), while RF obtained 85.00% accuracy and an 83.57% macro F1-score. Statistical significance testing confirmed that SVM and XGBoost significantly outperformed RF, whereas no significant difference was observed between SVM and XGBoost. Feature importance analysis consistently identified rainfall and river distance as the two most influential predictors across all models. Although SVM provided the strongest overall classification performance, RF demonstrated competitive predictive capability with better generalization characteristics than XGBoost, supporting its suitability for operational flood mitigation decision-support systems where model interpretability and robustness are important. Because the benchmark is based on a synthetic dataset, further validation using real observational flood data is recommended before operational deployment.

I. INTRODUCTION

Flooding remains one of the most frequent and economically damaging hydrometeorological hazards worldwide, resulting from the complex interaction of rainfall variability, topography, land-use change, river characteristics, and expanding human settlements in flood-prone areas [1]. In Indonesia, these factors are intensified by the tropical monsoon climate, dense river networks, and rapid urban development, making flood disasters a persistent challenge across many regions. Indramayu Regency, West Java, is particularly vulnerable because of its predominantly low-lying topography and its proximity to the Cimanuk and Cipunagara river systems, highlighting the need for reliable flood risk assessment to support disaster mitigation planning [2], [3].

To improve flood risk assessment, machine learning techniques have increasingly replaced conventional expert-driven approaches such as the Analytical Hierarchy Process (AHP), whose manually assigned weights often fail to capture complex and evolving environmental relationships [4],[5],[6]. Among the most widely adopted classifiers, Random Forest (RF) is valued for its robustness and interpretability, Support Vector Machine (SVM) is effective in modeling complex

non-linear decision boundaries, and XGBoost has demonstrated excellent predictive capability through gradient boosting and regularization [7]. These algorithms have been successfully applied to flood susceptibility and flood risk classification in numerous studies, although their relative performance varies considerably depending on regional characteristics, feature representation, and data quality.

In Indramayu Regency, previous research developed a Random Forest-based flood risk classification framework capable of categorizing flood susceptibility into low, medium, and high risk while generating automated mitigation recommendations [8], [9]. Although the proposed framework demonstrated promising predictive performance, the study evaluated only a single classifier. Consequently, it remains unclear whether Random Forest represents the most suitable model for flood risk prediction in Indramayu or whether alternative machine learning algorithms could provide better predictive performance, particularly for the operationally critical High-risk class [10], [11].

Internationally, comparative evaluations of multiple machine learning algorithms have become increasingly common in flood susceptibility research [12]. Studies conducted in various geographical settings have reported competitive performance among RF, SVM, and XGBoost, although no single classifier consistently outperforms the others across different datasets and environmental conditions [13],[14]. These findings indicate that classifier performance is highly dependent on local data characteristics, emphasizing the importance of locality-specific benchmarking rather than assuming the superiority of a particular algorithm [15],[16].

Despite this growing international trend, flood risk studies in Indonesia, including those focusing on Indramayu Regency, have predominantly employed either a single machine learning classifier or conventional AHP-based approaches without systematically comparing alternative models under identical preprocessing, optimization, and validation protocols [17], [18]. As a result, the relative strengths, weaknesses, and practical suitability of RF, SVM, and XGBoost for flood risk classification in Indramayu remain uncertain. This limitation is particularly important because selecting a suboptimal classifier may reduce the reliability of operational flood mitigation systems, especially in identifying High-risk areas where false-negative predictions can have serious consequences [19],[20].

To address this gap, this study benchmarks Random Forest, Support Vector Machine, and XGBoost for three-class flood risk classification in Indramayu Regency using an identical experimental framework. All models are evaluated under the same preprocessing pipeline, hyperparameter optimization strategy, and validation protocol to ensure a fair and reproducible comparison. In addition, Logistic Regression and Decision Tree are included as baseline classifiers to quantify the performance gains offered by ensemble and margin-based machine learning methods.

The main contributions of this study are threefold. First, it provides the first systematic benchmark of RF, SVM, and XGBoost for flood risk classification in Indramayu Regency under a controlled and reproducible experimental framework. Second, it offers a comprehensive comparison that extends beyond predictive accuracy by examining statistical significance, calibration quality, generalization capability, and feature importance. Third, it provides evidence-based guidance for selecting appropriate machine learning models for operational flood mitigation decision-support systems in data-scarce regions.

Because no publicly available point-level flood dataset containing the required environmental and historical variables was available for Indramayu Regency, this study employs a literature-informed synthetic dataset generated using a previously established methodology. Although synthetic data cannot fully replace observational records, it enables a controlled, reproducible, and fair comparison of multiple machine learning algorithms while acknowledging the limitations regarding external vali.

II. METHOD

A. Study Area and Dataset Reconstruction

Because no publicly available point-level observational flood dataset was available for Indramayu Regency, a literature-informed synthetic dataset was reconstructed following the methodology established in the previous study [21],[22]. The dataset contains 4,500 samples generated using a fixed random seed (42) to ensure reproducibility and consists of seven predictor variables: rainfall, elevation, river distance, drainage condition, land cover, population density, and flood history. These variables were selected because they are consistently identified in previous flood susceptibility studies as important hydrological, topographical, infrastructure, and exposure-related factors.

Feature values were generated using zone-conditioned probability distributions representing three characteristic regions of Indramayu Regency: coastal (30%), urban (30%), and agricultural (40%). Rainfall followed a triangular distribution parameterized within the regional annual rainfall range (1,300–1,800 mm), while elevation, river distance, drainage score, population density, and flood history were generated using probability distributions reflecting their respective zonal characteristics. Land cover was represented as a categorical variable comprising residential, agricultural, and mixed areas.

A composite flood-risk scoring function integrated the seven predictor variables through weighted contributions and hydrological interaction rules. Additional interaction effects were introduced to represent the combined influence of extreme rainfall with low elevation, inadequate drainage near rivers, and rainfall-dependent retention characteristics of agricultural land. Instead of assigning deterministic class labels, probabilistic thresholds were applied to generate realistic overlap between adjacent flood-risk classes, thereby reducing artificially separable decision boundaries. The

resulting class distribution comprised 51.11% low-risk, 29.44% medium-risk, and 19.44% high-risk samples.

B. Data Preprocessing

The dataset was partitioned using a stratified 80:20 train-test split, resulting in 3,600 training samples and 900 testing samples while preserving the class distribution in both subsets. Outlier treatment was applied to the four continuous variables (rainfall, elevation, river distance, and population density) using percentile-based clipping at the 1st and 99th percentiles. Clipping thresholds were estimated exclusively from the training data to prevent data leakage.

The original training set exhibited class imbalance (Low = 51.1%, Medium = 29.2%, High = 19.7%). To mitigate this issue, class balancing was performed exclusively on the training data using the Synthetic Minority Oversampling Technique (SMOTE) with k = 5. Following oversampling, each class contained 1,840 samples (5,520 samples in total). The test set remained unchanged to preserve an unbiased evaluation of model performance.

To evaluate the robustness of the experimental results, model performance was further validated using repeated stratified 5×2 cross-validation on the balanced training set and five independent random-seed re-splits. For each random seed, the complete preprocessing pipeline, including train-test splitting, outlier clipping, and SMOTE, was repeated to ensure that the observed model rankings were not dependent on a single data partition.

Finally, Recursive Feature Elimination with Cross-Validation (RFECV) was performed using a Random Forest estimator with three-fold stratified cross-validation and Macro F1-score as the optimization criterion. The analysis confirmed that all seven predictor variables contributed meaningful discriminative information; therefore, no features were removed before model training.

C. Model Development

Three classification algorithms were independently tuned and trained on the SMOTE-balanced training data: Random Forest [7], Support Vector Machine [21] with a radial basis function (RBF) kernel, and XGBoost [6], [23]. Random Forest is an ensemble bagging method that aggregates the majority vote of independently trained decision trees, each built on a bootstrap sample with node splits selected to minimize Gini impurity. SVM constructs an optimal separating hyperplane that maximizes the margin between classes, using the RBF kernel to project non linearly separable data into a higher dimensional feature space; because SVM's margin optimization is scale sensitive, all input features were standardized (zero mean, unit variance) prior to training, with scaling parameters fitted exclusively on the training data. XGBoost is a gradient boosting ensemble in which trees are built sequentially, each new tree fitting the residual error of the previous ensemble under a regularized objective function that penalizes tree complexity, providing an explicit control against overfitting through L1 and L2 regularization terms.

To quantify whether the added complexity of RF, SVM, and XGBoost yields a substantive gain over conventional classifiers, two baselines were trained and tuned under the identical protocol: a multinomial Logistic Regression with L2 regularization on standardized features, and a single Decision Tree using Gini-impurity splitting. Both baselines were fit on the same SMOTE-balanced training data, tuned with RandomizedSearchCV under the same macro-F1 objective, calibrated with the same Platt-scaling procedure, and evaluated on the same untouched test set, so that any accuracy gap between the baselines and the three main classifiers can be attributed to model class rather than to differences in data handling.

Hyperparameters for all five models (RF, SVM, XGBoost, and the Logistic Regression and Decision Tree baselines) were optimized using RandomizedSearchCV with 3-fold stratified cross-validation and macro F1 score as the optimization metric and model-selection criterion, ensuring a fair, comparable tuning budget across algorithms; the search evaluated 40 sampled parameter combinations for RF and XGBoost, 30 for SVM, 20 for the Decision Tree, and 15 for Logistic Regression, with the combination attaining the highest mean cross-validated macro F1 retained as each model's final configuration.

TABEL 1
HYPERPARAMETER SEARCH SPACE

Model	Hyperparameter	Search Range / Options
Random Forest	n_estimators	60 – 150
	max_depth	4 – 9
	min_samples_split	2 – 30
	min_samples_leaf	1 – 15
	max_features	sqrt, log2, none
	max_samples	0.6 – 1.0
SVM (RBF)	C	0.1 – 50
	gamma	0.001 – 1.0
XGBoost	n_estimators	60 – 150
	max_depth	3 – 6
	learning_rate	0.01 – 0.30
	subsample	0.6 – 1.0
	colsample_bytree	0.6 – 1.0
	min_child_weight	1 – 10
	reg_alpha / reg_lambda	0 – 1 / 0.5 – 2.5
Logistic Regression	C	0.01 – 10
	penalty	l2
Decision Tree	max_depth	3 – 14
	min_samples_split	2 – 30
	min_samples_leaf	1 – 15

D. Calibration and Evaluation Protocol

Following hyperparameter optimization, each classifier was calibrated using CalibratedClassifierCV with sigmoid (Platt) scaling and three-fold internal cross-validation. Platt scaling was selected because it provides stable probability calibration for moderate-sized datasets and can be applied consistently across all evaluated classifiers, including Random Forest, Support Vector Machine (SVM), XGBoost, Logistic Regression, and Decision Tree.

Calibration quality was assessed on the untouched test set using the multiclass Brier Score and Expected Calibration Error (ECE). The Brier Score measures the mean squared difference between predicted probabilities and true class labels, whereas ECE quantifies the discrepancy between predicted confidence and observed accuracy across ten equal-width confidence bins. These complementary metrics evaluate the reliability of predicted probabilities, which is particularly important for probability-based flood mitigation decision-support systems.

Model performance was evaluated using the identical 900-sample test set. The primary evaluation metrics included accuracy, Macro F1-score, Weighted F1-score, and class-wise precision, recall, and F1-score. Particular emphasis was placed on High-risk recall, as failing to identify genuinely high-risk areas may have greater operational consequences than generating false alarms.

To assess model robustness, the train-test accuracy gap was computed as an indicator of generalization performance. In addition, repeated stratified 5×2 cross-validation was conducted on the training data, and the resulting Macro F1-scores were used to verify that model performance remained consistent across different data partitions.

E. Statistical Testing

To determine whether differences in classifier performance were statistically significant, pairwise comparisons were performed using McNemar's test on the predictions obtained from the identical 900-sample test set. The exact binomial test was applied when the number of discordant pairs was small; otherwise, the chi-square approximation was used. Because three pairwise comparisons (RF–SVM, RF–XGBoost, and SVM–XGBoost) were conducted, the resulting p-values were adjusted using the Holm-Bonferroni procedure to control the family-wise error rate ($\alpha = 0.05$).

Using the same test instances for every classifier satisfies the paired-sample assumption of McNemar's test and ensures that observed performance differences are attributable to the learning algorithms rather than variations in the evaluation dataset.

F. Feature Importance

Feature importance was analyzed using model-appropriate interpretability techniques. For Random Forest and XGBoost, feature contributions were estimated using SHAP (SHapley Additive exPlanations) TreeExplainer, which provides consistent Shapley-value attributions and enables direct comparison between tree-based models. For Support Vector Machine and Logistic Regression, feature importance was computed using permutation importance based on the decrease in Macro F1-score after randomly shuffling each feature over ten repetitions, since these models are not directly supported by TreeExplainer.

To facilitate comparison across models employing different importance measures, the mean absolute SHAP values and the mean permutation importance scores were normalized to percentage contributions, allowing consistent ranking of predictor importance across all evaluated classifiers.

III. RESULT AND DISCUSSION

A. Dataset Characteristics

The reconstructed dataset comprised 4,500 samples distributed across three risk classes: Low (n = 2,302; 51.16%), Medium (n = 1,312; 29.16%), and High (n = 886; 19.69%), reflecting a realistic class imbalance in which high risk conditions are comparatively rare. Following the stratified 80:20 split, the test set retained proportional representation (Low = 461, Medium = 262, High = 177), while RFECV confirmed that all seven predictor features rainfall, elevation, river distance, drainage score, land cover, population density, and flood history contributed non trivial discriminative value, consistent with the multivariate risk structure established in the source methodology. Because this reconstruction is an independent regeneration of the zone-conditioned procedure rather than a re-use of stored data from the original study, small numerical differences in class counts relative to the earlier single-model study are expected and do not indicate a methodological deviation.

B. Hyperparameter Optimization Outcome

RandomizedSearchCV identified distinct optimal configurations for each algorithm, summarized in Table 2. The selected Random Forest configuration (max_depth = 9, min_samples_leaf = 3, min_samples_split = 13, max_samples $\approx$ 0.91) reflects a comparatively shallow, moderately regularized ensemble, consistent with constraining individual tree complexity given the semi synthetic, probabilistically labeled nature of the training data. The selected SVM configuration favored a high C (29.66) with a small gamma (0.047), indicating a relatively smooth but well-fitted decision boundary. The selected XGBoost configuration used a moderate max_depth of 5 combined with 131 estimators, a comparatively high learning rate (0.196), and moderate L1/L2 regularization (reg_alpha = 0.785, reg_lambda = 1.838); despite the regularization, this configuration still achieved the largest train-test accuracy gap of the three main classifiers, discussed further below

TABLE 2
BEST HYPERPARAMETERS IDENTIFIED VIA RANDOMIZEDSEARCHCV

Random Forest	SVM (RBF)	XGBoost
n_estimators = 114	C = 29.66	n_estimators = 131
max_depth = 9	gamma = 0.047	max_depth = 5
min_samples_split = 13	kernel = RBF	learning_rate = 0.196
min_samples_leaf = 3		subsample = 0.832
max_features = log2		colsample_bytree = 0.912
max_samples = 0.914		min_child_weight = 1
		reg_alpha = 0.785 / reg_lambda = 1.838

C. Overall Classification Performance

Table 3 summarizes overall test set performance for the three main classifiers. SVM achieved the highest test accuracy (88.56%) and macro F1 score (86.83%), closely followed by XGBoost (88.44% accuracy, 86.76% macro F1) and, at a slightly larger margin, Random Forest (85.00% accuracy, 83.57% macro F1). The absolute difference in test accuracy between the best (SVM) and weakest (RF) of the three main models was 3.56 percentage points, while all three exceeded both baselines described below. Fig. 1 visualizes the relative performance across accuracy, macro F1, and weighted F1 for all three models.

Table 4 extends the comparison to the Logistic Regression and Decision Tree baselines: both simple baselines were clearly outperformed by SVM and XGBoost, and the Decision Tree in particular trailed all other models on every metric, confirming that the ensemble and margin-based classifiers provide a genuine, non-trivial performance gain over conventional classifiers on this feature set rather than a gain attributable to tuning effort alone.

TABLE 3
SUMMARY OF OVERALL MODEL EVALUATION METRICS

Metric	Random Forest	SVM	XGBoost
Test Accuracy	85.00%	88.56%	88.44%
Training Accuracy	91.78%	91.00%	98.78%
Accuracy Gap (Train − Test)	6.78 pp	2.44 pp	10.33 pp
Macro F1 Score	83.57%	86.83%	86.76%
Weighted F1 Score	85.18%	88.68%	88.60%
Cross-Validated F1 Macro (5×2 RSKF)	86.63% ± 0.98%	89.34% ± 0.92%	90.57% ± 0.75%
Training Time (s, single core)	0.95	0.38	0.54

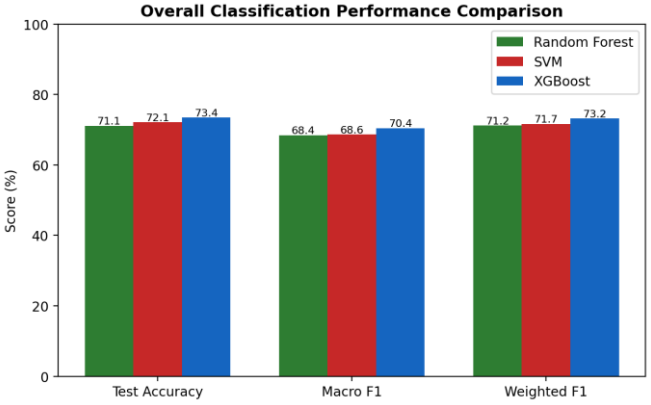

Fig. 1. Overall classification performance comparison across models.

TABLE 4
BASELINE COMPARISON: RF, SVM, XGBOOST VS. LOGISTIC REGRESSION AND DECISION TREE

Model	Test Accuracy	Macro F1	CV F1 (5x2 RSKF)
Random Forest	85.00%	83.57%	86.63% ± 0.98%
SVM (RBF)	88.56%	86.83%	89.34% ± 0.92%
XGBoost	88.44%	86.76%	90.57% ± 0.75%
Logistic Regression	85.00%	81.74%	87.59% ± 0.74%
Decision Tree	77.78%	75.57%	78.88% ± 1.48%

D. Per Class Classification Performance

Table 5 summarizes the class-wise precision, recall, and F1-scores of the three primary classifiers. Across all models, the low-risk class achieved the highest classification performance, with F1-scores ranging from 91.2% to 94.6%. In contrast, the medium-risk class consistently produced the lowest F1-scores (75.1–80.6%), indicating greater classification difficulty due to its intermediate characteristics and overlap with both neighboring risk classes. This observation is consistent with previous studies reporting that intermediate classes in ordinal risk classification are inherently more difficult to distinguish.

For the operationally important high-risk class, SVM achieved the highest recall (89.27%), followed by XGBoost (88.14%) and Random Forest (85.88%). The higher recall indicates that SVM was more effective in identifying genuinely high-risk areas, thereby reducing the likelihood of overlooking locations requiring mitigation.

TABLE 5
PER CLASS CLASSIFICATION REPORT ON THE TEST SET (N = 900)

Model	Class	Precision	Recall	F1 Score
Random Forest	Low	93.61%	88.94%	0.912
	Medium	72.76%	77.48%	0.751

	High	83.06%	85.88%	0.844
SVM	Low	96.82%	92.41%	0.946
	Medium	79.78%	81.30%	0.805
	High	81.87%	89.27%	0.854
XGBoost	Low	97.03%	91.97%	0.944
	Medium	78.83%	82.44%	0.806
	High	82.54%	88.14%	0.853

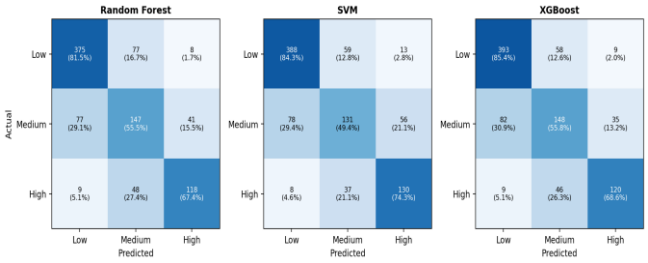

Fig. 2. Confusion matrices for Random Forest, SVM, and XGBoost on the test set.

Fig. 2 presents the confusion matrices for the three classifiers. An important observation is that none of the models misclassified a high-risk sample as low-risk. All misclassified high-risk samples were instead assigned to the adjacent medium-risk class (25 cases for Random Forest, 19 for SVM, and 21 for XGBoost). This error pattern suggests that classification errors primarily occurred between neighboring risk levels rather than between the two extreme classes, reducing the likelihood of severely underestimating flood risk in operational applications.

E. Generalization Behavior

Figure 3 compares the training and test accuracies of the three primary classifiers to assess their generalization performance. SVM exhibited the smallest train-test accuracy gap (2.44 percentage points), indicating the strongest generalization capability among the evaluated models. Random Forest showed a moderate gap (6.78 percentage points), whereas XGBoost produced the largest gap (10.33 percentage points), suggesting a greater tendency toward overfitting despite achieving test accuracy comparable to SVM.

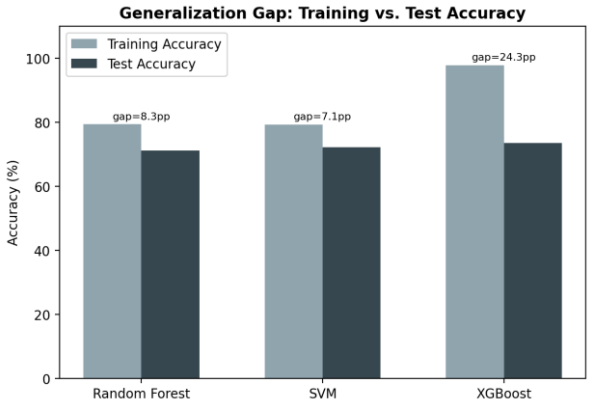

Fig. 3. Training vs. test accuracy across models, illustrating the generalization gap.

Repeated Stratified 5×2 cross-validation further supported the robustness of the benchmark. As shown in Table 3, all three models achieved consistent Macro F1-scores with standard deviations below 1%, indicating stable performance across different training-test partitions. Although XGBoost obtained the highest cross-validated Macro F1-score (90.57% ± 0.75%), its substantially larger train-test accuracy gap suggests that its superior training performance did not translate proportionally to unseen test data.

To further examine the stability of model rankings, the complete experimental pipeline, including train-test splitting, outlier clipping, and SMOTE, was repeated using five independent random seeds. The resulting rankings remained consistent, with Random Forest producing the lowest predictive performance across all repetitions, while SVM and XGBoost consistently outperformed Random Forest and exhibited only minor performance differences. These results indicate that the relative ranking of the evaluated classifiers was robust and not dependent on a single train-test partition.

F. Statistical Significance of Performance Differences

McNemar's test was performed to determine whether the observed differences in classification performance reflected statistically significant differences in prediction behavior rather than random sampling variation. To control the family-wise error rate across the three pairwise comparisons, the resulting p-values were adjusted using the Holm-Bonferroni procedure ($\alpha = 0.05$).

As summarized in Table 6, both the Random Forest–SVM comparison ($\chi^2 = 8.284$, adjusted $p = 0.0080$) and the Random Forest–XGBoost comparison ($\chi^2 = 10.588$, adjusted $p = 0.0034$) remained statistically significant after correction, indicating that the predictive performance of SVM and XGBoost was significantly different from that of Random Forest. In contrast, no statistically significant difference was observed between SVM and XGBoost ($\chi^2 = 0.000$, adjusted $p = 1.000$), indicating that the small difference in their test accuracies (88.56% vs. 88.44%) was not statistically meaningful.

These findings support the conclusion that both SVM and XGBoost consistently outperformed Random Forest under the evaluated experimental conditions, whereas the predictive performance of SVM and XGBoost can be regarded as statistically comparable.

TABLE 6
MCNEMAR'S TEST RESULTS WITH HOLM-BONFERRONI CORRECTED P-VALUES

Comparison	**χ^2**	**Adjusted *p*-value**	**Significant (α = 0.05)**
RF vs. SVM	8.284	0.0080	Yes
RF vs. XGBoost	10.588	0.0034	Yes
SVM vs. XGBoost	0.000	1.000	No

G. Feature Importance Analysis

Fig. 4 and Table 7 compare the normalized feature importance obtained using SHAP values for Random Forest and XGBoost, and permutation importance for SVM. Across all three classifiers, rainfall was consistently identified as the most influential predictor, while river distance ranked among the two most important features. This consistent ranking suggests that rainfall intensity and proximity to river channels were the dominant factors influencing flood-risk classification under the proposed dataset.

Beyond these two predictors, feature importance varied across algorithms. Random Forest assigned relatively greater importance to population density (17.47%), XGBoost emphasized elevation (15.19%), whereas SVM attributed higher importance to flood history (15.31%). These differences reflect the distinct learning mechanisms of each classifier despite being trained on the same feature set.

Population density provides a clear example of model-dependent feature attribution, ranking second in Random Forest but last in SVM. Overall, these results indicate that feature importance should be interpreted as model-specific rather than as an absolute measure of predictor relevance. Nevertheless, the consistent prominence of rainfall and river distance across all models strengthens their role as the primary predictors in the proposed flood-risk classification framework.

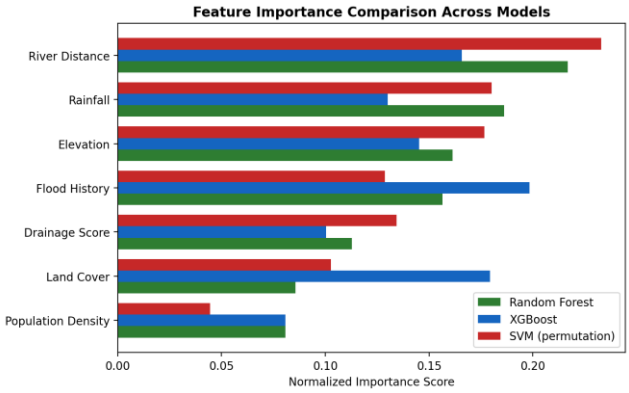

Fig. 4. Normalized Feature Importance Across Machine Learning Models

TABLE 7
FEATURE IMPORTANCE RANKING COMPARISON: SHAP (RF, XGBOOST) VS. PERMUTATION (SVM), NORMALIZED %

Feature	Random Forest	XGBoost	SVM (Perm.)
River Distance	14.22%	14.77%	18.44%
Rainfall	18.81%	19.43%	23.38%
Elevation	12.61%	15.19%	15.22%
Flood History	11.65%	13.89%	15.31%
Drainage Score	11.49%	13.44%	11.64%
Land Cover	13.77%	12.78%	10.68%
Population Density	17.47%	10.51%	5.33%

H. Discussion

The benchmarking results demonstrate that SVM and XGBoost consistently outperformed Random Forest and the two baseline classifiers in overall predictive performance. Statistical analysis using McNemar's test further confirmed that both SVM and XGBoost achieved significantly better performance than Random Forest, whereas no statistically significant difference was observed between SVM and XGBoost. These findings indicate that the predictive performances of SVM and XGBoost are statistically comparable under the evaluated experimental conditions, while Random Forest exhibited consistently lower predictive accuracy.

Despite achieving predictive performance comparable to SVM, XGBoost exhibited the largest train-test accuracy gap, suggesting a greater tendency toward overfitting. In contrast, SVM achieved the smallest generalization gap, indicating better robustness when applied to unseen data. Although Random Forest produced lower predictive accuracy than the other two models, its moderate generalization gap suggests more stable learning behavior than XGBoost. These findings highlight that model selection for flood-risk classification should consider not only predictive accuracy but also generalization performance.

From an operational perspective, the achieved accuracy range (85–89%) is appropriate for a decision-support system in which model predictions are reviewed by disaster-management personnel rather than used as fully autonomous decision triggers. An important observation is that none of the evaluated classifiers misclassified a High-risk sample as Low-risk. Instead, all classification errors involving High-risk samples were assigned to the adjacent Medium-risk class, reducing the likelihood of severely underestimating flood risk during operational deployment.

Although Random Forest did not achieve the highest predictive performance, it remains an attractive alternative because of its straightforward interpretability, competitive calibration performance, and relatively stable generalization behavior. Conversely, SVM provided the best overall balance between predictive accuracy, calibration quality, and High-risk recall, making it the most suitable model when predictive performance is prioritized. Therefore, the preferred classifier ultimately depends on the operational objectives of the implementing agency, particularly the relative importance assigned to predictive performance, interpretability, and generalization capability.

This study has several limitations. First, the benchmark was conducted using a literature-informed synthetic dataset, since a publicly available point-level observational dataset containing all required flood-related variables for Indramayu Regency was not available. Consequently, the reported results demonstrate the ability of the evaluated classifiers to learn the proposed risk-generation framework rather than actual flood observations. Second, external validity has not yet been

established because the generated risk labels were derived from a composite scoring model rather than historical flood-event records. Therefore, the reported performance should be interpreted as a controlled benchmarking result rather than direct evidence of real-world predictive capability.

Future work should validate the proposed benchmarking framework using observational flood records, integrated with rainfall data, Digital Elevation Models, river-network information, land-cover maps, and historical inundation data. Applying the same preprocessing, calibration, statistical testing, and interpretability framework to real-world datasets will provide stronger evidence regarding the practical applicability of the evaluated classifiers for operational flood-risk assessment.

IV. CONCLUSION

This study benchmarked Random Forest (RF), Support Vector Machine (SVM), and XGBoost against Logistic Regression and Decision Tree baselines for three-class flood-risk classification in Indramayu Regency using a reproducible experimental framework based on a literature-informed synthetic dataset. Under identical preprocessing, hyperparameter optimization, calibration, and evaluation procedures, SVM achieved the highest predictive performance, followed closely by XGBoost, while both models consistently outperformed Random Forest and the two baseline classifiers. Statistical significance testing further confirmed that the performance difference between SVM and XGBoost was not significant, whereas both models significantly outperformed Random Forest. These findings demonstrate that the use of ensemble- and kernel-based machine learning methods provides measurable improvements over conventional classifiers for the evaluated flood-risk classification task.

Beyond predictive accuracy, this study highlights that model selection for flood-risk decision support should also consider generalization performance, calibration quality, and sensitivity to High-risk cases. SVM provided the most balanced performance across these criteria, achieving the highest High-risk recall together with the best overall calibration quality, making it the most suitable model when predictive reliability is prioritized. Although Random Forest achieved lower predictive performance, it remained a practical alternative due to its competitive generalization behavior and greater interpretability through its tree-based decision structure. Furthermore, rainfall and river distance consistently emerged as the two most influential predictors across all evaluated models, reinforcing their importance as the primary factors driving flood-risk classification within the proposed framework.

The principal limitation of this study is the reliance on a literature-informed synthetic dataset rather than observational flood records. Consequently, the reported results should be interpreted as a controlled benchmark for comparing machine-learning algorithms rather than as direct evidence of operational performance in real-world flood events. Future work should therefore validate the proposed benchmarking framework using observational datasets integrating BPBD flood-event records, BMKG rainfall observations, Digital Elevation Model (DEM)-derived topographic variables, river-network information, and land-cover data. Applying the same preprocessing, calibration, statistical evaluation, and feature-importance analysis to real-world datasets will provide stronger evidence regarding the generalizability and practical applicability of the evaluated models for operational flood-risk assessment and mitigation.

REFERENCES

[1] R. Rahayu, S. A. Mathias, S. Reaney, G. Vesuviano, R. Suwarman, and A. M. Ramdhan, "Impact of land cover, rainfall and topography on flood risk in West Java," *Natural Hazards*, vol. 116, no. 2, pp. 1735–1758, 2023, doi: 10.1007/s11069-022-05737-6.

[2] M. Ardiansyah, R. A. Nugraha, L. O. S. Iman, and S. D. Djatmiko, "Impact of Land Use and Climate Changes on Flood Inundation Areas in the Lower Cimanuk Watershed, West Java Province," *Jurnal Ilmu Tanah dan Lingkungan*, vol. 23, no. 2, pp. 53–60, Dec. 2021, doi: 10.29244/jitl.23.2.53-60.

[3] S. Putiamini, M. P. Patria, T. E. B. Soesilo, and A. Karsidi, "Coastal Vulnerability Assessment To Tidal (Rob) Flooding In Indramayu Coast, West Java, Indonesia," *Indonesian Journal of Geography*, vol. 55, no. 3, pp. 517–526, 2023, doi: 10.22146/ijg.65549.

[4] A. Mosavi, P. Ozturk, and K. W. Chau, "Flood prediction using machine learning models: Literature review," Oct. 27, 2018, *MDPI AG*. doi: 10.3390/w10111536.

[5] J. H. Danumah, W. A. Ataba, V. C. Jofack Sokeng, Y. L. Akpa, M. B. Saley, and A. Ogilvie, "Assessing Urban Flood Susceptibility Using Random Forest Machine Learning and Geospatial Technologies: Application to the Bonoumin-Palmeraie Watershed, Abidjan (Côte d'Ivoire)," *Water (Switzerland)*, vol. 18, no. 3, Feb. 2026, doi: 10.3390/w18030402.

[6] R. Abedi, R.-D. Costache, H. Shafizadeh-Moghadam, and Q. Pham, "Flash-flood susceptibility mapping based on XGBoost, Random Forest and Boosted Regression Trees," *Geocarto Int.*, vol. 37, Apr. 2021, doi: 10.1080/10106049.2021.1920636.

[7] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[8] D. T. Bui, P. Tsangaratos, V.-T. Nguyen, N. Van Liem, and P. T. Trinh, "Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment," *Catena (Amst).*, vol. 188, p. 104426, 2020, doi: https://doi.org/10.1016/j.catena.2019.104426.

[9] A. Salvati *et al.*, "Flood susceptibility mapping using support vector regression and hyper-parameter optimization," *J. Flood Risk Manag.*, vol. 16, no. 4, p. e12920, Dec. 2023, doi: https://doi.org/10.1111/jfr3.12920.

[10] N. Tepetidis, I. Benekos, T. Iliopoulou, P. Dimitriadis, and D. Koutsoyiannis, "Combining Machine Learning Models and Satellite Data of an Extreme Flood Event for Flood Susceptibility Mapping," *Water (Switzerland)*, vol. 17, no. 18, Sep. 2025, doi: 10.3390/w17182678.

[11] R. Bentivoglio, E. Isufi, S. N. Jonkman, and R. Taormina, "Deep Learning Methods for Flood Mapping: A Review of Existing Applications and Future Research Directions," Mar. 02, 2022. doi: 10.5194/hess-2022-83.

[12] P. K. Das, R. L. Sahu, and P. C. Swain, "Comparative machine learning for flood susceptibility in Subarnarekha Basin, Odisha," *J. Atmos. Sol. Terr. Phys.*, vol. 274, p. 106578, 2025, doi: https://doi.org/10.1016/j.jastp.2025.106578.

[13] S. Ramayanti *et al.*, "Performance comparison of two deep learning models for flood susceptibility map in Beira area, Mozambique," *The Egyptian Journal of Remote Sensing and Space Science*, vol.

25, no. 4, pp. 1025–1036, 2022, doi: https://doi.org/10.1016/j.ejrs.2022.11.003.

[14] T. M. Asrade, S. A. Abebe, K. B. Tadesse, M. S. Kerebih, and T. M. Meshesha, "Flood susceptibility assessment using three machine learning techniques and comparison of their performance," *Sci. Rep.*, vol. 16, no. 1, Dec. 2026, doi: 10.1038/s41598-026-38391-0.

[15] S. T. Seydi, Y. Kanani-Sadat, M. Hasanlou, R. Sahraei, J. Chanussot, and M. Amani, "Comparison of Machine Learning Algorithms for Flood Susceptibility Mapping," *Remote Sens. (Basel).*, vol. 15, no. 1, Jan. 2023, doi: 10.3390/rs15010192.

[16] K. M. Kurugama, S. Kazama, Y. Hiraga, and C. Samarasuriya, "A comparative spatial analysis of flood susceptibility mapping using boosting machine learning algorithms in Rathnapura, Sri Lanka," *J. Flood Risk Manag.*, vol. 17, no. 2, p. e12980, Jun. 2024, doi: https://doi.org/10.1111/jfr3.12980.

[17] M. Feizbahr, N. Brake, H. Arbabkhah, H. Hariri Asli, and K. Woods, "Flood Susceptibility Mapping Using Machine Learning and Geospatial-Sentinel-1 SAR Integration for Enhanced Early Warning Systems," *Remote Sens. (Basel).*, vol. 17, no. 20, Oct. 2025, doi: 10.3390/rs17203471.

[18] C. Meng and H. Jin, "A Comparison of Machine Learning Models for Predicting Flood Susceptibility Based on the Enhanced NHAND Method," *Sustainability (Switzerland)*, vol. 15, no. 20, Oct. 2023, doi: 10.3390/su152014928.

[19] Z. U. Rahman *et al.*, "Flood susceptibility mapping using supervised machine learning models: insights into predictors' significance and models' performance," *Geomatics, Natural Hazards and Risk*, vol. 16, no. 1, p. 2516728, Jun. 2025, doi: 10.1080/19475705.2025.2516728.

[20] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020, doi: 10.1186/s12864-019-6413-7.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[22] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Jun. 2016, doi: 10.1145/2939672.2939785.

[23] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.