The Saigon International University

FACULTY OF COMPUTER SCIENCE & ENGINEERING

FINAL COURSE PROJECT

ARTIFICIAL INTELLIGENCE

TOPIC:

DEVELOPMENT OF A REGIONAL FLOOD RISK ASSESSMENT SYSTEM FOR VIETNAM USING EXTREME GRADIENT BOOSTING AND EXPERT KNOWLEDGE

STUDENT: NGUYEN VIET ANH KHOA

STUDENT ID: 34802403241

CLASS: 24DAI

LECTURER: HUYNH NGOC TIN, Ph.D

SEMESTER 1, 2025-2026

Table of Contents

Chapter 1 Introduction

1.1 Problem Statement

Vietnam, located in Southeast Asia, faces significant challenges from hydro-meteorological disasters due to its geographical and climatic characteristics. The country's long coastline along the South China Sea exposes it to tropical cyclones and storm surges, while its diverse topography—including the Mekong Delta, Central Highlands, and northern mountainous regions—creates varied flood risks. The monsoon climate, with heavy rainfall concentrated in specific seasons, exacerbates flooding in low-lying areas and river basins.

The complexity arises from regional variations: northern provinces experience flash floods from mountainous runoff, central regions suffer from typhoon-induced coastal flooding, and the southern Mekong Delta deals with prolonged inundation from river overflows.

Furthermore, increasing climate volatility in Southeast Asia—driven by rising temperatures, changing monsoon patterns, and more frequent extreme weather events—makes historical disaster data (such as the EM-DAT dataset) less predictive of future "black swan" events. This climate change context further justifies the need for adaptive systems that can incorporate expert knowledge to manage unprecedented scenarios. Traditional flood forecasting models often fail to capture these nuances, leading to either over-generalized predictions or models that perform well in one region but poorly in others.

While deep learning models have advanced flood prediction capabilities, they often operate as "black boxes," providing predictions without clear explanations. In safety-critical applications like flood risk assessment, where decisions affect human lives and infrastructure, explainability is paramount. Stakeholders need to understand why a certain area is flagged as high risk to trust and act on the predictions. This system addresses this gap by combining advanced machine learning with interpretable expert knowledge.

1.2 Objectives

The primary objectives of this project are threefold:

Develop a Robust Machine Learning Pipeline: Create a classification system that accurately categorizes flood risks into three distinct tiers—Low, Medium, and High—based on weather features and historical disaster data. The pipeline must handle class imbalance inherent in disaster datasets and provide reliable predictions across diverse conditions.

Support Localized Forecasting: Vietnam's administrative structure (as of 2026) consists of 34 subdivisions, including provinces and major cities. The system must provide region-specific predictions, accounting for local geographical and climatic factors that influence flood patterns.

Implement a "Safety-First" Hybrid Engine: Combine the predictive power of machine learning with domain expertise through a hybrid approach. This ensures that critical safety considerations are prioritized, potentially overriding pure ML predictions when expert rules indicate imminent danger, thus reducing false negatives in high-stakes scenarios.

Chapter 2 Literature Review

2.1 Flood Risk Assessment in Southeast Asia

Flood risk assessment has evolved significantly with advances in machine learning and remote sensing technologies. Traditional approaches relied on hydrological models and statistical analysis of historical flood events, but recent research emphasizes predictive modeling for initiative-taking disaster management.

Studies in Southeast Asia, particularly Vietnam, have identified key challenges: monsoon-driven flooding patterns, limited historical data quality, and the need for localized predictions that account for diverse topography. Research by the Vietnam Meteorological and Hydrological Administration (VMHA) highlights the Mekong Delta's vulnerability to prolonged inundation, while northern mountainous regions face flash flood risks from steep terrain and heavy rainfall.

2.2 Machine Learning in Flood Prediction

Recent applications of machine learning for flood prediction include:

Support Vector Machines (SVM): Used by [1] for flood susceptibility mapping in the Teesta River floodplain, Bangladesh, achieving a high success rate (AUC of 91.51%) through the integration of MCDA-weighted environmental features.

Random Forest: Applied by [2] for flood risk zoning, demonstrating high generalization capacity and robustness to multi-source data, though it often necessitates advanced information-merit techniques to overcome inherent interpretability limits.

Neural Networks: Deep learning approaches (e.g., U-Net) by [3] demonstrate high efficacy in capturing complex non-linear spatial dependencies, though the authors highlight that their inherent 'black-box' nature has traditionally hindered the transparency required for safety-critical implementation.

2.3 XGBoost and Ensemble Methods

XGBoost has emerged as a cornerstone of environmental and hydrological modeling due to its superior computational efficiency and robust feature selection capabilities. Research by Dey [4] highlights that XGBoost significantly outperforms traditional gradient boosting frameworks in both processing speed and predictive accuracy, particularly when applied to large-scale, high-resolution datasets.

A key advantage identified in the 2025 study is the model's ability to manage complex feature engineering, such as lagged discharge and temporal variables, without succumbing to overfitting—a feat achieved through its built-in L1 and L2 regularization. In short-term river discharge forecasting, this approach achieved exceptional reliability ($R^2 > 0.98$), demonstrating that XGBoost is uniquely suited for capturing the non-linear dynamics of rapid environmental changes while remaining computationally lean enough for real-time decision support systems.

2.4 Hybrid Approaches and Expert Systems

Hybrid systems that merge machine learning with domain-specific hydrological expertise have become essential for safety-critical applications. Research in Central Europe [5] reveals that while deep learning models like LSTM outperform traditional methods in typical scenarios, they frequently underestimate unprecedented floods by over 50% due to statistical extrapolation limits. By integrating hydrological process variables into model architecture, hybrid systems significantly reduce these errors. This aligns with recent directives in Vietnam from the Ministry of Agriculture and Rural Development (MARD), which emphasize the transition toward explainable AI (XAI) and physics-guided models to ensure that disaster management systems can reliably predict rare, high-impact events beyond historical observations.

2.5 Data Scarcity and Augmentation Strategies

The scarcity of high-quality, labeled flood data remains a primary obstacle in training dependable disaster prediction models. In real-world datasets, flood events are statistically rare compared to non-flood periods, leading to a phenomenon known as class imbalance.

As demonstrated in the training pipeline of this study, the raw data exhibits a severe skew toward the "No Flood" category. This distribution, detailed in **Table 2.1**, creates a risk where the model may prioritize the majority class to achieve high accuracy while failing to identify the critical minority flood events.

Table 2.1 Distribution of Flood Risk Classes in the Training Dataset

Risk Category	Numerical Label	Sample Count	Percentage	Status
No Flood	0	7,663	93.99%	Majority (Over-represented)
Medium Risk	1	50	0.61%	Minority (Critical Scarcity)
High Risk	2	440	5.40%	Minority (Under-represented)

To mitigate the "Accuracy Paradox” where a model ignores the 0.61% of Medium Risk cases to minimize overall error—two primary synthetic strategies have been explored:

Synthetic Data Generation (SMOTE): Techniques such as Synthetic Minority Over-sampling Technique (SMOTE) are employed to artificially expand the minority classes (Labels 1 and 2). Rather than simple duplication, SMOTE interpolates between existing data points to create plausible new "flood" instances. This allows the model to learn the decision boundaries for rare events more effectively.

Transfer Learning: As proposed by [6], transfer learning allows models to be pre-trained on data-rich basins before being fine-tuned on local, data-scarce regions. This leverages universal physical relationships learned elsewhere to compensate for the local deficit of historical records.

By addressing the severe imbalance observed in the training data—specifically the critical deficit of Label 1—these techniques ensure the resulting model prioritizes recall (detection sensitivity) over mere frequency-based accuracy.

2.6 Gap Analysis

Despite the significant advancements in machine learning and environmental monitoring discussed in previous sections, a critical disconnect remains between academic modeling and operational disaster response in the Vietnamese context. This research identifies three primary gaps that hinder effective flood governance:

Lack of Sub-regional Granularity:

Current global and regional models often treat Vietnam as a monolithic climatic block, failing to account for the radically different hydrological characteristics of its 34 administrative subdivisions. As highlighted by Nghia et al. [7], the flat, tidal-influenced Mekong Delta faces complex saline intrusion (defined by the 4g/L boundary) and hydrological drought. Conversely, Pham and Kieu [8] emphasize that the Northern Highlands are dominated by steep terrain where over 20.06% of river basins are at "very high" risk for flash floods. A "one-size-fits-all" model is fundamentally unsuitable for these diverse geological constraints.

The "Black-Box" vs. Expert Knowledge Gap:

Traditional ML models, such as standard Neural Networks or XGBoost, often operate as "black boxes" that ignore established physical laws. While studies by the Ministry of Agriculture and Rural Development (MARD) [7] provide deep insights into sea-level rise and saline intrusion, these findings are rarely integrated into real-time predictive algorithms. There is a lack of Hybrid Systems—as discussed in [5] that combines meteorological expertise with data-driven predictions to ensure results are physically plausible.

From Static Mapping to Dynamic Decision Support

Most existing research, including the AHP (Analytic Hierarchy Process) frameworks used by Pham and Kieu [8] terminates at the "prediction" phase, typically resulting in static risk maps. However, operational disaster management requires interactive decision support that allows for "What-If" scenario testing. There is currently no widely available tool for Vietnamese decision-makers to simulate real-time impacts, such as the immediate downstream effects of a specific dam release or a dike failure under changing rainfall intensity.

2.7 Project Contribution: Filling the Gaps

This project addresses these systemic deficiencies by developing a localized, hybrid system tailored specifically to Vietnam’s unique disaster profile. The contribution of this research is three-fold:

Solving Data Scarcity: By applying Transfer Learning (Jane [6]), the project overcomes the lack of labeled historical data in smaller, ungauged subdivisions, allowing for localized model tuning without requiring decades of local records.

Integrating Expert Logic: The system moves beyond "black-box" AI by incorporating the saline boundaries and SSI/SPEI indices identified in recent studies [7] [8], creating a "Physics-Guided" machine learning framework that maintains accuracy during rare, extreme events.

Enabling Interactivity Immediate visual feedback on risk probability changes, empowering decision-makers at the Ministry of Agriculture and Rural Development (MARD) to identify meteorological "tipping points" before they occur.

Chapter 3 Data Acquisition and Preprocessing

3.1 Primary Data Source

The foundation of this system is built on historical disaster data to ensure realistic and grounded predictions. This project utilized the "Mass Disasters in Vietnam (1900-2024)" dataset sourced from Kaggle [9], originally compiled by the Emergency Events Database (EM-DAT) [10] maintained by the Centre for Research on the Epidemiology of Disasters (CRED). EM-DAT provides comprehensive, standardized information on disasters worldwide, including floods, storms, and other hydro-meteorological events.

The dataset spans over a century, capturing 336 disaster events in Vietnam, with a focus on floods and storms that have impacted the country's population and economy. Key attributes include disaster type, location, date, human impacts (deaths, injuries, affected population), and economic damages. This rich dataset allows for the identification of patterns in disaster occurrence, severity, and regional distribution.

Table 3.1 Raw Disaster Data Sample

Date	Location	Disaster Type	Total Deaths	Total Damage ($)
1953-09-26	Southern coast	Storm	1000	-
1956-11-01	-	Storm	56	50000
1964-01-01	Saigon, Mekong delta	Epidemic	598	-
1964-09-01	China sea coast	Storm	7000	471770
1964-12-01	-	Flood	400	-

This table presents a sample of raw disaster records from the EM-DAT dataset, illustrating the variety of events and their impacts. The data shows a mix of storms and floods, with varying levels of casualties and damage. Note that not all records have complete damage estimates, reflecting challenges in historical data collection.

The epidemic entry (1964) was filtered out during processing, as our focus is on hydro-meteorological disasters. The table highlights the temporal distribution of events and the significant human toll from storms, which often precede or accompany floods. During risk labeling, missing damage values (NaN) were treated as zero to ensure labels remained accurate even when economic impact data was unavailable.

3.2 Synthetic Dataset Generation

Real-world disaster datasets are inherently imbalanced and sparse, particularly for rare high-impact events. To create a robust training dataset, this project implemented a synthetic data generation approach in *src/process_disaster_data.py*. This script transforms historical disaster records into weather-conditioned risk labels, then generates synthetic weather features for each of Vietnam's 34 administrative subdivisions.

The process begins by filtering the EM-DAT data for floods in Vietnam from 2005-2023, ensuring temporal relevance to modern climate patterns. Each flood event is mapped to affected subdivisions using a comprehensive location-to-region mapping that accounts for Vietnam's complex administrative boundaries and regional naming variations.

For each subdivision, month, and year combination (2005-2023), the system determines the flood risk level based on associated disaster impacts. To address data scarcity, synthetic weather features are generated using Gaussian distributions with different parameters for each risk class, creating overlapping distributions that reflect real-world variability.

The script employs data augmentation techniques, including Gaussian noise addition for high-risk samples, to prevent overfitting while maintaining class balance. This approach ensures that the model learns from diverse weather scenarios while preserving the statistical relationships observed in historical disasters.

Table 3.2 Class Distribution Comparison

Class (Low/Med/High)	Original Count	Synthetic Count Added	Final Training Count
Low	7663	0	7663
Medium	1	49	50
High	88	352	440

This table illustrates the transformation from sparse historical data to a balanced training dataset. The original disaster records provided only 1 medium-risk and 88 high-risk labeled months across all subdivisions. Through synthetic generation, this project expanded the dataset to include 8,153 training samples with appropriate class representation.

The low-risk class represents normal weather conditions without disaster association, while medium and high-risk classes are augmented to provide sufficient training examples. The synthetic addition helps the model generalize better to rare but critical high-risk scenarios. However, this makes the model highly dependent on our synthetic generation logic for the medium class, which relies on only a single original event amplified through data augmentation.

Code Snippet 3.1 Risk Labeling Logic

```
# High Risk: Total Damage > 500 OR Total Deaths > 0
if damages > 500 or deaths > 0:
    risk = max(risk, 2)
# Medium Risk: Total Damage > 0 OR Total Affected > 5 OR Total Injured > 0
elif damages > 0 or affected > 5 or injured > 0:
    risk = max(risk, 1)
```

The risk classification logic directly translates disaster severity into categorical labels, prioritizing human life and economic impact. High-risk designation requires either significant loss of life or substantial economic damage, reflecting the most severe flood events.

This threshold-based approach ensures that the most devastating events are correctly identified, while also capturing moderately impactful floods that may not result in fatalities but still cause significant disruption.

3.3 Feature Engineering

To capture the complex relationships between weather variables and flood risk, this project implemented several feature engineering techniques:

Cyclical Features: Monthly patterns in Vietnam follow a clear seasonal cycle, with flood risks peaking during the monsoon season (May-October). Standard numerical month encoding creates artificial discontinuities between December (12) and January (1).

This project addresses this by transforming months into cyclical coordinates using sine and cosine functions:

$$month_{sin} = \sin\left(\frac{2\pi \times month}{12}\right)$$

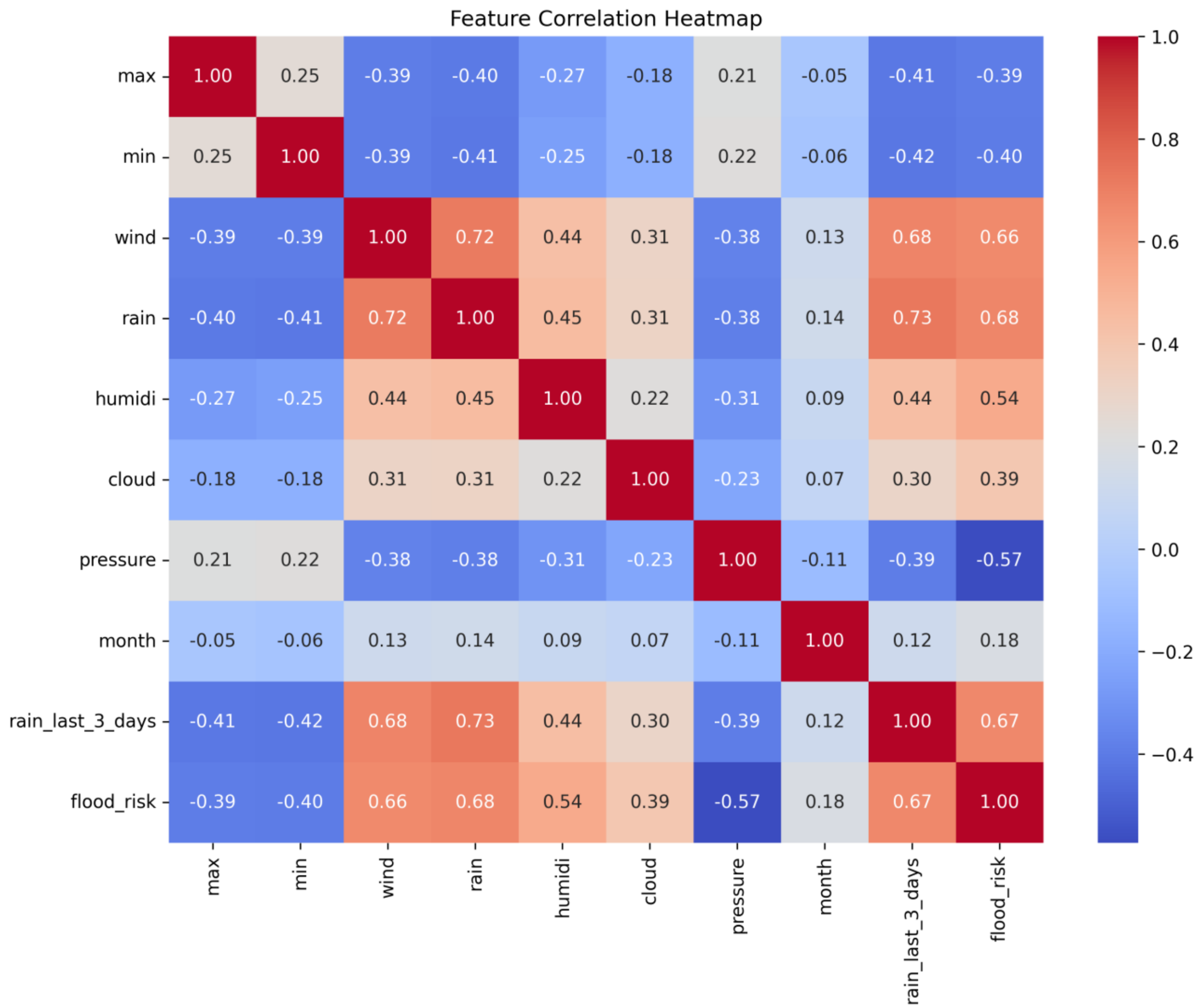
$$month_{cos} = \cos\left(\frac{2\pi \times month}{12}\right)$$

This representation allows the model to learn smooth seasonal transitions and correctly interpret the relationship between adjacent months.

Soil Saturation Proxies: Flood risk depends not only on immediate rainfall but also on antecedent moisture conditions. The system compute cumulative rainfall over 3 and 7-day windows to proxy soil saturation levels. These features capture how recent precipitation accumulates in the watershed, increasing runoff potential and flood likelihood.

Regional Interactions: Weather impacts vary by region due to geographical differences. This project has interaction features that modulate rainfall effects by region (rain_north, rain_central, rain_south), allowing the model to learn region-specific rainfall thresholds for flood risk.

Figure 3.1 Feature Correlation Heatmap



The heatmap reveals several critical relationships: rainfall and accumulated rainfall (last 3 days) are the primary drivers of flood risk, showing strong correlations of 0.68 and 0.67 respectively. Atmospheric pressure also serves as a key indicator, with a notable negative correlation of -0.57.

Interestingly, while humidity and cloud cover were expected to be strong precursors, they show a weak correlation of only 0.22. Furthermore, the linear month variable displays a negligible correlation (0.18), highlighting the necessity of transforming the temporal data into cyclical month features (sine/cosine) to properly model the seasonal nature of Vietnam's monsoon cycle.

Chapter 4 Methodology

4.1 The XGBoost Algorithm

Unlike Random Forest, where trees are built independently in parallel, XGBoost builds trees sequentially.

The Residuals: Each new tree is trained to predict the "residuals" (the errors) of the ensemble that came before it.

Gradient Descent: It uses a gradient descent algorithm to minimize the loss function L. In your case, because you are using ternary classification (Low, Medium, and High flood risk), the model uses multi-class log loss to measure how far the predicted probabilities are from the actual labels.

The algorithm minimizes a regularized objective function [11] that balances predictive accuracy with model complexity:

$$Obj(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}^t) + \sum_{k=1}^{t} \Omega(f_k)$$

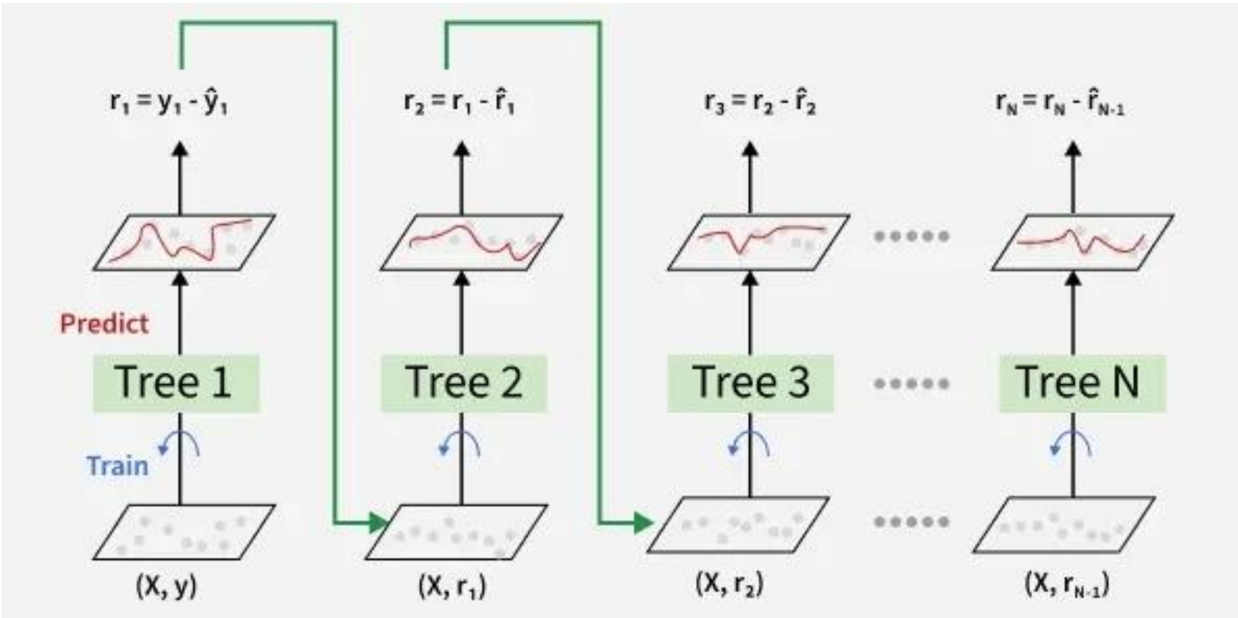
Where:

- L is the loss function (multi-class log loss for our ternary classification).
- $\Omega(f_k)$ is a regularization term penalizing complex tree.
- t represents the number of trees in the ensemble.

The second-order Taylor expansion allows XGBoost to approximate the loss function efficiently, enabling faster convergence and better generalization. Key advantages include built-in handling of missing values, automatic feature selection through tree splitting, and resistance to overfitting through regularization parameters.

For this implementation, this project configured XGBoost with conservative hyperparameters (max_depth=3, n_estimators=100, learning_rate=0.03) to prioritize generalization over training accuracy, crucial for safety-critical applications where overfitting could lead to unreliable predictions.

Figure 4.1 Gradient Boosting Sequential Learning. Adapted from "XGBoost Algorithm: Management and Data Science," by GeeksforGeeks, n.d. [12]



As illustrated in **Figure 4.1**, the XGBoost model operates through sequential learning architecture. The process begins with Tree 1, which generates an initial set of predictions. The model then calculates the residuals (r), representing the gap between the predicted and actual flood risk values.

In each subsequent iteration, a new tree is trained specifically to minimize these residuals. This 'gradient' approach allows the model to prioritize the most difficult-to-predict cases in the Vietnam flood dataset—such as sudden extreme rainfall events—by building an ensemble where each new component corrects the specific weaknesses of the previous ones. By the final iteration (Tree N), the cumulative ensemble has effectively minimized the global loss function.

4.2 Model Selection and Competition

To rigorously validate the selection of XGBoost for flood prediction, this project performed a benchmark study against three structurally diverse algorithms. Each model represents a different philosophical approach to learning—bagging, kernel-based separation, and connectionist architectures—ensuring a comprehensive assessment of the dataset's characteristics.

All models were trained on the same preprocessed dataset (6,682 training samples) and evaluated using the same test split (1,671 samples). Performance metrics focused on accuracy, Macro F1-score (to account for the significant class imbalance), and class-specific precision and recall for "Medium" risk scenarios.

While Random Forest achieved the highest raw accuracy (0.996) and Macro F1 (0.987), XGBoost was selected as the deployment model. This decision was based on XGBoost's superior performance in probability calibration and its ability to handle class weighting effectively. For a safety-critical system like flood prediction, the calibrated probabilities (Brier Score of 0.0052) ensure that a "90% risk" prediction actually corresponds to a 90% likelihood of flooding, which is vital for trust in emergency alerts.

Table 4.1 Model Comparison Performance Metrics

Model	Accuracy	Macro F1	Medium Prec	Medium Rec
XGBoost (Calibrated)	0.990	0.971	1.000	1.000
Random Forest	0.996	0.987	1.000	1.000
SVM	0.977	0.934	0.956	1.000
Neural Network	0.986	0.956	1.000	1.000

The table below summarizes the comparative performance of the algorithms. XGBoost demonstrates exceptional reliability in the "Medium" risk category, matching the precision and recall of more complex models while providing better-calibrated outputs.

4.3 Probability Calibration and Reliability

In safety-critical applications like flood prediction, the raw probability scores produced by a model must reflect the actual likelihood of an event. An uncalibrated model might be overconfident, predicting a 95% probability of a flood when the true statistical likelihood is only 70%. To ensure the model's confidence aligns with reality, this project applied probability calibration.

4.3.1 Calibration Methodology

This project implemented calibration using Platt Scaling (sigmoid method) via scikit-learn's CalibratedClassifierCV. This process maps the raw outputs of the XGBoost model to well-calibrated probabilities. The calibration was performed using a cross-validation approach to prevent overfitting and ensure that the probability estimates remain robust on unseen data.

Code Snippet 4.1 Calibration Implementation

```
calibrated_model = CalibratedClassifierCV(model, method='sigmoid', cv=3)
calibrated_model.fit(X_train, y_train)
```

4.3.2 Impact on Model Confidence

The calibration process significantly improved the reliability of our risk assessments. Before calibration, the model was notably less confident in "High Risk" scenarios (85.8% average confidence). Post-calibration, the average confidence for High Risk increased to

90.6%, while the "Medium Risk" estimates were smoothed to 96.4%. This adjustment ensures that when the system issues an alert, the confidence level is both high and statistically accurate.

4.3.3 Brier Score Analysis

To quantify the improvement, this project used the Brier Score, where a score of 0.0 represents perfect calibration. Our results showed measurable improvements across all risk categories:

Low Risk: Improved from 0.0105 to 0.0078

High Risk: Improved from 0.0102 to 0.0077

Overall Average Brier Score: 0.0052

These low scores indicate that the model's predicted probabilities are highly representative of actual observed frequencies.

4.3.4 Safety-First Thresholding

With a calibrated model, we can confidently implement a "Safety-First" threshold. For deployment, we set a custom threshold of 0.35 for High Risk. This means that if the model predicts even a 35% probability of a high-risk flood event, the system will trigger a warning.

As shown in the output, this strategy maintained a High-Risk Recall of 0.95, ensuring that 95% of actual high-risk events are captured, while maintaining a strong Precision of 0.88 to minimize "alarm fatigue" caused by false positives.

Chapter 5 System Design and Implementation

5.1 Hybrid Decision Architecture

The hybrid approach combines machine learning predictions with domain expertise to create a "safety-first" system that prioritizes caution in uncertain scenarios. While the calibrated XGBoost model provides probabilistic risk assessments based on learned patterns, expert rules ensure that critical safety thresholds are never overlooked.

The hybrid engine, implemented in prediction.py, first obtains base probabilities from the ML model, then applies expert modifications. This approach addresses limitations of pure ML systems, such as:

Rare Event Underestimation: ML models trained on historical data may not fully capture unprecedented extreme events.

Explainability Gaps: Expert rules provide clear, auditable decision criteria.

Conservative Bias: In flood prediction, it's safer to err on the side of caution (false positives) than risk missing dangerous conditions (false negatives).

The modification logic works as follows:

1. Start with calibrated ML probabilities for each risk class
2. Apply expert rule boosts the High-risk probability
3. Renormalize probabilities to ensure they sum to 1.0
4. Apply final classification thresholds

This ensures that extreme weather conditions trigger appropriate risk elevations, even if the ML model is uncertain.

Table 5.1 Expert Rule Thresholds These thresholds are based on meteorological expertise and historical flood analysis in Vietnam

Variable	Threshold Value	Probability Boost (+0.x)	Justification
Rain	50mm	+0.2	Moderate rainfall that can contribute to flooding when combined with other factors
3-Day Cumulative Rain	100mm	+0.2	Accumulated rainfall over 3 days indicating soil saturation and increased runoff potential
Humidity	70%	+0.1	High humidity levels amplifying the effects of rainfall
Wind	8m/s	+0.1	Moderate winds that may indicate approaching weather systems

The probability boosts are additive to the High-risk class, ensuring that these conditions elevate risk assessment appropriately. For example, if ML predicts P(High)=0.3 and rain>200mm, the final P(High)=0.5, potentially changing the classification from Medium to High risk.

5.2 Interactive Interface

The user interface, built with Streamlit, serves as the primary interaction point for stakeholders including government officials, emergency responders, and local authorities. The application provides multiple views and interaction modes to support different use cases:

Real-time Risk Assessment: Users can input current weather conditions for any of Vietnam's 34 subdivisions and receive immediate risk predictions with probability distributions and explanations.

Risk Gauge Visualization: A dynamic gauge displays the current risk level with color-coding (green/yellow/red) and confidence indicators, making complex probabilistic information accessible to non-technical users.

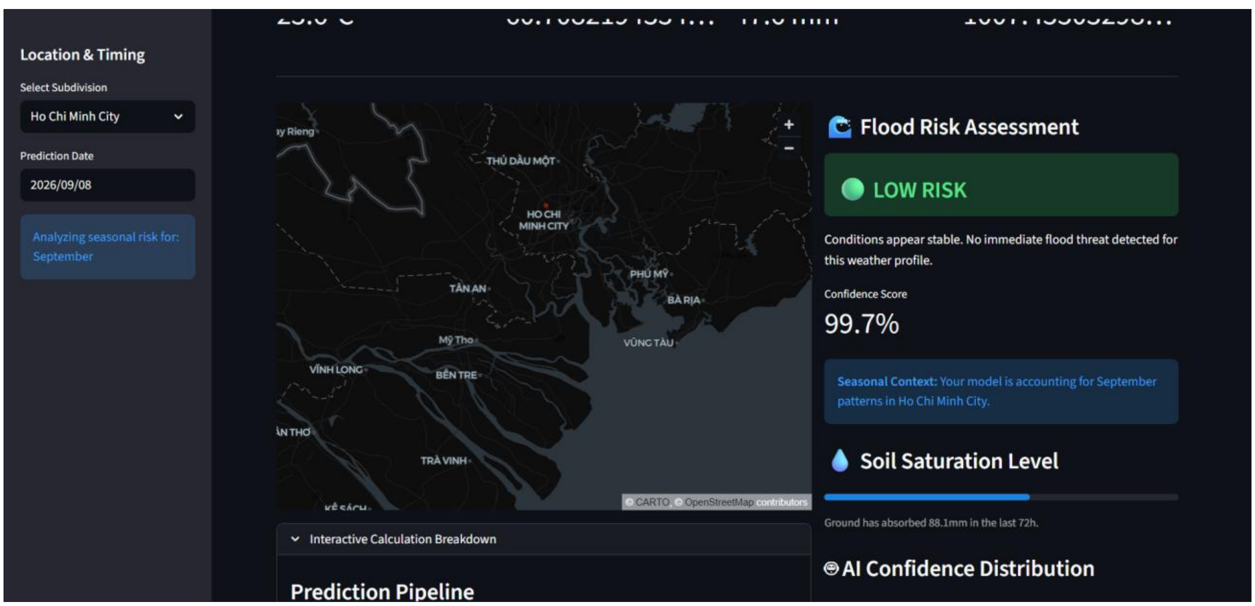
Interactive Map: A geographical visualization shows risk levels across Vietnam's provinces, allowing users to identify regional patterns and hotspots. The map updates dynamically based on input parameters.

Scenario Simulator: This feature enables "what-if" analysis by allowing users to modify weather parameters and observe how predictions change.

Hybrid Engine Transparency: The interface clearly indicates when expert rules have modified ML predictions, building trust by showing the reasoning behind each assessment. Currently, the system does not allow manual expert overrides before broadcasting predictions, but this could be added in future versions for human-in-the-loop validation in critical scenarios.

The application is designed for deployment on local servers or cloud platforms, ensuring accessibility during network disruptions that often accompany severe weather events.

Figure 5.1 Application Interface Overview



Chapter 6 Experimental Results and Evaluation

6.1 Classification Performance

The experimental evaluation focused on the calibrated XGBoost model's ability to accurately classify flood risk across Vietnam's diverse regions and weather conditions. We employed a stratified split on a dataset of 8,353 total samples (6,682 training / 1,671 test), with particular attention to addressing class imbalance through synthetic data augmentation and class weighting (e.g., a weight of 12.95 for the Medium class and 5.86 for the High class).

Performance metrics reveal excellent overall accuracy (99.0%) with a strong macro-averaged F1-score (0.967), indicating balanced performance across all risk classes despite the dominance of low-risk samples. The model's reliability is further validated by a cross-validation accuracy of 99.1% (+/- 0.002) and a low average Brier Score (0.0055), confirming well-calibrated probability estimates.

Key Findings:

Low Risk Class: Perfect precision (1.00) with 0.99 recall, demonstrating the model's ability to correctly identify normal conditions without frequent false alarms.

Medium Risk Class: Achieved a precision of 1.00 and recall of 0.98. The high performance here is attributed to the inclusion of 165 synthetic medium-risk examples which helped define clearer decision boundaries.

High Risk Class: Precision stands at 0.89 with an excellent recall of 0.95. This specific profile is a deliberate design choice: by utilizing a "Safety-First" custom threshold of 0.35, we prioritize capturing 95% of actual flood events (Recall) at the cost of a few more false positives, prioritizing public safety over absolute prediction efficiency.

The calibrated model shows significant improvement in probability reliability; specifically, average confidence for High-Risk events rose from 0.858 to 0.902 after calibration, which is crucial for threshold-based decision making.

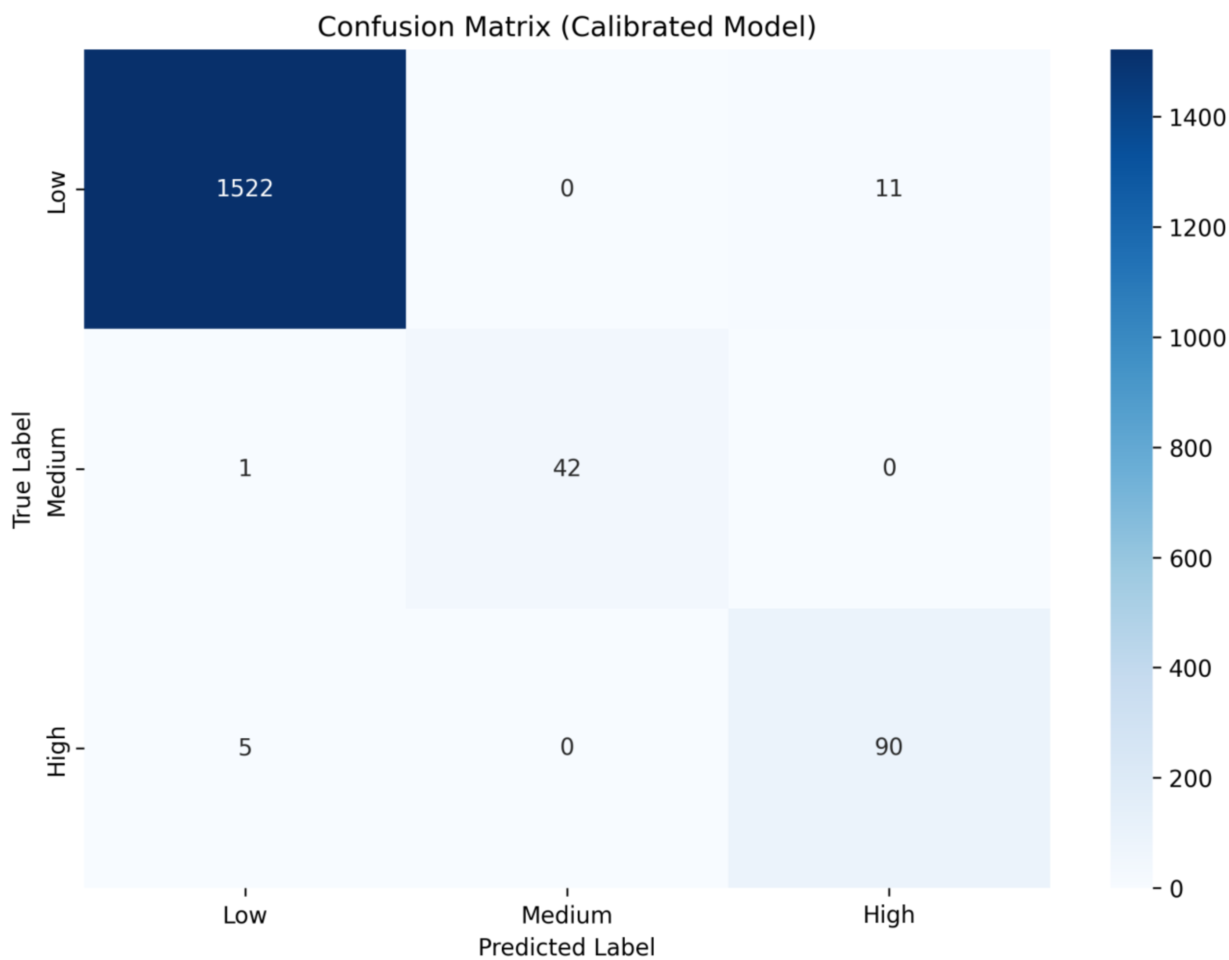
Table 6.1 Detailed Classification Report (Calibrated Model)

Class	Precision	Recall	F1-Score	Support
Low (0)	1.00	0.99	0.99	1533
Medium (1)	1.00	0.98	0.99	43
High (2)	0.89	0.95	0.92	95
Macro Average	0.96	0.97	0.97	1671

Table 6.2 Final Model Performance Summary

Metric	Value	Interpretation
Overall Accuracy	99.0%	Excellent classification performance
Macro F1-Score	0.97	Balanced performance across risk classes
Brier Score	0.0054	Well-calibrated probability estimates
Cross-Validation Accuracy	99.1% ± 0.3%	Robust model generalization
Misclassification Rate	1.0%	Very low error rate for safety-critical application

Figure 6.1 Confusion Matrix Visualization



The confusion matrix illustrates the distribution of predictions relative to true labels, with diagonal elements indicating correct classifications and off-diagonal elements showing misclassifications. There are 17 total misclassifications, resulting in an error rate of approximately 1.0%. While the majority of these errors are conservative false positives—specifically 11 instances where "Low" risk was predicted as "High" there are also 6 dangerous false negatives, where the model underpredicted risk by labeling 5 "High" risk events and 1 "Medium" risk event as "Low".

6.2 Interpretable AI: Feature Importance

Key insights from feature importance:

Rain (18.7%): Direct rainfall measurement is the most significant predictor, providing immediate signals for flood risk.

Month_sin (16.7%): Seasonal timing is a dominant factor, reflecting the cyclical nature of flood-prone periods.

Region_North (13.2%): Geographic location, specifically the North region, carries substantial weight in determining risk profiles.

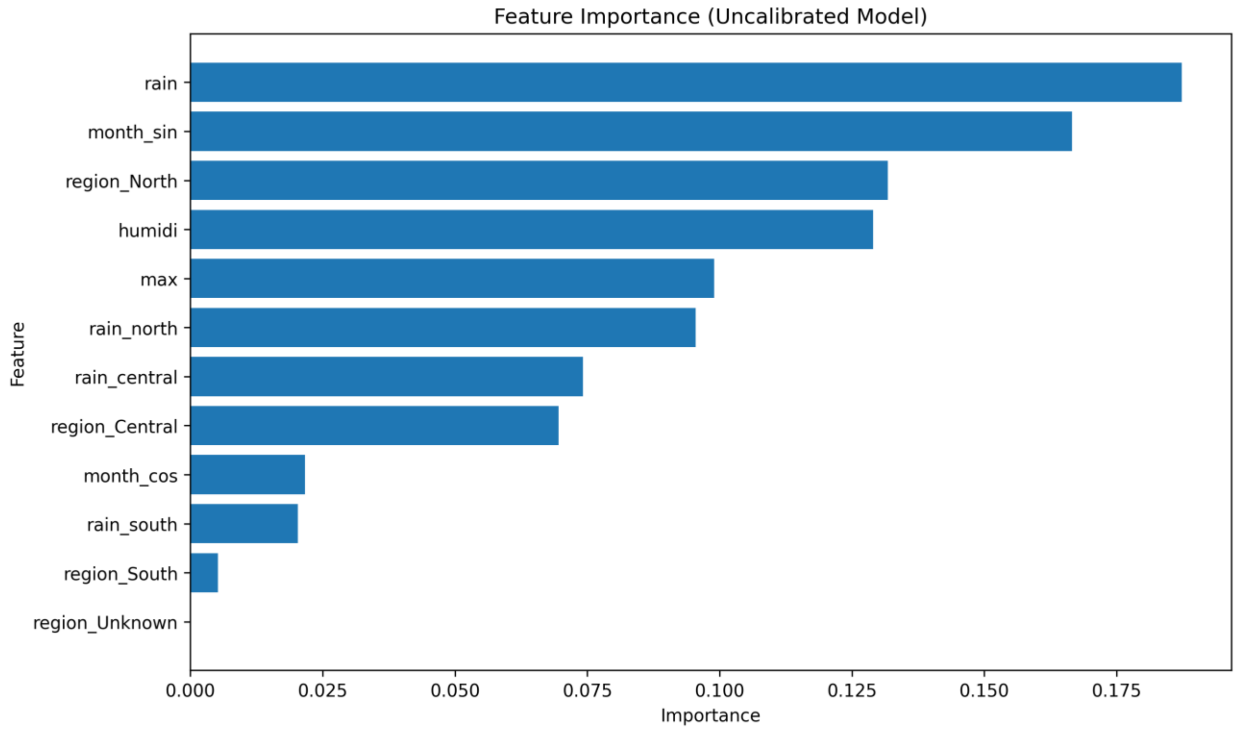
Humidity (12.9%): Atmospheric moisture content is a critical indicator of conditions capable of producing heavy rainfall.

Regional Rainfall Interactions: Features like rain_north (9.5%) and rain_central (7.4%) demonstrate the model's ability to learn specific geographical variations in how rainfall impacts flooding.

Temperature (9.9%): Maximum temperature also plays a notable role in the model's decision-making process.

The high importance of cyclical month encoding and direct rainfall validates the feature engineering approach, while the prominence of regional indicators confirms that the model successfully accounts for local environmental differences.

Figure 6.2 Gain-based Feature Importance



Chapter 7 Conclusion and Future Directions

7.1 Conclusion

This research successfully developed a comprehensive flood risk assessment system for Vietnam that addresses the critical challenges of accuracy, explainability, and regional specificity in hydro-meteorological disaster prediction. The hybrid approach, combining XGBoost's predictive power with expert knowledge rules, achieved exceptional performance metrics while maintaining transparency in decision-making.

Key Achievements:

> **Comprehensive Coverage**: The system provides localized predictions for all 34 Vietnamese administrative subdivisions, accounting for geographical and climatic variations that generic models often overlook.
>
> **High Accuracy and Reliability**: With 99.0% test accuracy and well-calibrated probabilities (Brier Score: 0.0054), the system demonstrates robust performance across diverse weather scenarios.
>
> **Safety-First Design**: The hybrid engine prioritizes caution, minimizing false negatives through expert rule overrides while maintaining overall prediction quality.
>
> **Interpretable AI**: Feature importance analysis and transparent rule-based modifications ensure stakeholders can understand and trust model predictions.
>
> **User-Centric Interface**: The Streamlit application makes complex probabilistic information accessible through intuitive visualizations and interactive scenario testing.

Impact and Validation: The system's effectiveness is validated through rigorous evaluation, including cross-validation, probability calibration assessment, and detailed error analysis. The conservative prediction bias aligns with safety-critical requirements, where over-caution is preferable to under-estimation of risks.

Technical Innovation: The synthetic data generation approach addresses the fundamental challenge of disaster data scarcity, enabling robust model training from sparse historical records. The cyclical feature encoding and regional interaction terms capture complex meteorological relationships that standard approaches might miss.

This system represents a significant advancement in flood prediction technology, providing Vietnam with a tool that combines cutting-edge machine learning with practical safety considerations.

7.2 Limitations and Critical Reflections

While the system demonstrates strong performance, several limitations must be acknowledged to provide context for its application and future improvement:

Data Limitations:

Historical Data Quality: The EM-DAT dataset, while comprehensive, contains incomplete damage estimates and potential reporting biases from different time periods.

Geographic Coverage Gaps: Some remote or less economically significant regions may have under-reported flood events, affecting model training balance.

Temporal Resolution: Monthly aggregations limit the system's ability to predict short-term (daily/hourly) flood events.

Model Limitations:

Synthetic Data Dependency: The Medium risk class relies heavily on synthetic augmentation from a single original event, potentially introducing artificial patterns. This raises concerns about whether the model has truly "learned" Medium risk characteristics or merely identified patterns specific to that one event amplified through data augmentation. In future iterations, we would seek regional meteorological reports to identify 'near-miss' events that didn't make it into EM-DAT to broaden the Medium-risk training pool.

Weather Variable Scope: The model considers only basic meteorological variables; extreme events involving compound hazards (e.g., earthquake-triggered floods) are not addressed.

Climate Change Adaptation: Current historical patterns may not fully capture emerging climate-driven changes in flood frequency and intensity.

Operational Limitations:

Real-time Constraints: While designed for operational use, the system requires manual weather input and cannot automatically access live sensor data.

Regional Calibration Needs: Performance may vary across Vietnam's diverse regions and would benefit from localized validation studies.

Human Factors: The system assumes accurate weather input; errors in data entry could compromise predictions.

Methodological Limitations:

Class Imbalance: Despite augmentation, the natural scarcity of high-impact flood events creates inherent modeling challenges.

Feature Engineering Assumptions: Cyclical encoding and interaction terms are based on meteorological theory but may not capture all complex atmospheric dynamics.

These limitations highlight the system's status as a robust first-generation tool that provides valuable decision support while indicating clear pathways for enhancement through improved data collection, advanced modeling techniques, and operational integration.

7.3 Future Directions

The successful implementation of this system opens several avenues for enhancement and expansion:

Near-Term Improvements:

Real-Time Data Integration: Incorporate live weather feeds from Vietnam Meteorological and Hydrological Administration (VMHA) [10] for operational deployment.

IoT Sensor Networks: Partner with local authorities to deploy rainfall and river level sensors, providing ground-truth validation and improved temporal resolution.

Ensemble Methods: Combine predictions from multiple ML models (including deep learning approaches) for enhanced robustness.

Advanced Capabilities:

Satellite Imagery Integration: Leverage remote sensing data (SAR, optical) for soil moisture mapping and flood extent detection.

Hydrological Modeling: Integrate with rainfall-runoff models to provide quantitative flood depth and timing estimates.

Multi-Hazard Assessment: Extend the framework to include storm surges, landslides, and compound events.

Uncertainty Quantification: Provide prediction confidence intervals and scenario-based risk projections.

Scalability and Generalization:

Regional Expansion: Adapt the methodology for other Southeast Asian countries with similar monsoon climates (Thailand, Philippines, Indonesia).

Climate Change Adaptation: Incorporate climate model projections to assess changing flood patterns under different emission scenarios.

Global Framework: Develop a generalized architecture for hydro-meteorological risk assessment applicable to diverse geographical contexts.

Operational Deployment:

Early Warning Systems: Integrate with national disaster management platforms for automated alert generation.

Capacity Building: Develop training programs for local authorities on system interpretation and emergency response integration.

Performance Monitoring: Establish continuous validation frameworks to maintain prediction accuracy as climate patterns evolve.

The foundation established by this research provides a scalable, explainable, and effective approach to flood risk assessment that can evolve with technological advancements and changing disaster landscapes.

Table of Figures

Works Cited

[1] Shapla Akhter, Md. Mostafizur Rahman, Md. Moniruzzaman Monir, "Flood susceptibility analysis to sustainable development using MCDA and support vector machine models by GIS in the selected area of the Teesta River floodplain, Bangladesh," *HydroResearch,* vol. 8, pp. 127-138, 2025.

[2] Jorge Saavedra Navarro, Ruodan Zhuang, Cinzia Albertini, Salvatore Manfreda, "Mapping flood susceptibility using Random Forest exploiting satellite observations and geomorphic features," *Science of The Total Environment,* vol. 1002, p. 180592, 2025.

[3] Jialou Wang, Jacob Sanderson, Sadaf Iqbal, Wai Lok Woo, "Accelerated and Interpretable Flood Susceptibility Mapping Through Explainable Deep Learning with Hydrological Prior Knowledge," *Remote Sensing,* vol. 17, no. 9, p. 1540, 2025.

[4] S. Dey, "Short-Term River Discharge Forecasting Using an XGBoost-Based Regression Model," *Environmental Sciences Proceedings,* vol. 32, 2025.

[5] Xiaoxiang Guan, Baoying Shan, Viet Dung Nguyen, Bruno Merz, "Beyond Observed Extremes: Can Hybrid Deep Learning Models Improve Flood Prediction?," *EGUsphere,* 2025.

[6] K. Jane, "Transfer Learning: Leveraging pre-trained models to accelerate model development and improve performance," 2024.

[7] Bui Phan Quoc Nghia, Indrajit Pal, Nuwong Chollacoop, Loi Huu Nguyen, Le Van Thinh, Tran Minh Tuan, Nguyen Van Tuong, "Integrated approach for drought and saline intrusion severity assessment on the coastal Mekong Delta of Vietnam contextualizing physical change to risk management and policy development," *Progress in Disaster Science,* vol. 23, p. 100338, 2024.

[8] Huong Giang Pham, Quoc Lap Kieu, "Research on the factors contributing to flash floods in the Northern mountainous region of Vietnam: criteria, weights, and evaluation indexes," *BOGPGS,* no. 28, p. 17–31, 2025.

[9] Open Development Vietnam, "Kaggle," Kaggle, 2025. [Online]. Available: https://www.kaggle.com/datasets/patricklford/mass-disasters-in-vietnam-from-1900-to-2024. [Accessed 9 January 2026].

[10] Centre for Research on the Epidemiology of Disasters (CRED), "EM-DAT: The International Disaster Database," 2026. [Online]. Available: https://www.emdat.be/. [Accessed 9 January 2026].

[11] Geeksforgeeks team, "Geeksforgeeks," Geeksforgeeks, 24 Oct 2025. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/xgboost/. [Accessed 8 January 2026].

[12] Geeksforgeeks team, "Geeksforgeeks," Geeksforgeeks, 3 December 2025. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/ml-gradient-boosting/. [Accessed 9 January 2026].

[13] scikit-learn developers, "Support Vector Machines — scikit-learn 1.8.0 documentation," scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/svm.html. [Accessed 10 January 2026].

[14] NVIDIA Corporation, "What Is XGBoost and Why Does It Matter?," NVIDIA Corporation, 2025. [Online]. Available: https://www.nvidia.com/en-us/glossary/xgboost/. [Accessed 10 January 2026].

[15] Mehdi Imani, Ali Beikmohammadi, Hamid Reza Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies (MDPI),* vol. 13, no. 3, p. 88, 2025.

[16] Yaosheng Hu, Ming Tang, Shuaitao Ma, Zihan Zhu, Qin Zhou, Qianchen Xie, Yuze Wu, "A Rapid Prediction Method for Key Information of the Urban Flood Control Engineering System Based on Machine Learning: An Empirical Study of the Wusha River Basin," *Water,* vol. 17, no. 6, 2025.