IJERST

International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 2, 2026

Peer Reviewed, Referred & Indexed Journal

DISASTER PREDICTION USING MACHINE LEARNING

1. V. Ramya , Asst .prof CSE dept, Gokula Krishna College of Engineering, M. Tech (CSE), Sullurpet, Tirupati District, AP

2. K. Venkata Ramana, Y. Sushma, B. Udaya Sri, K. Prasanna , K. Mohan , B.Tech UG Scholars, Gokula Krishna College of Engineering, Sullurpet, Tirupati District, AP

ABSTRACT

This study presents an advanced disaster prediction framework that enhances traditional machine learning approaches through an intelligent hybrid mechanism integrating multi-source data processing and adaptive learning techniques. Unlike conventional systems that rely heavily on static historical datasets, the proposed model incorporates heterogeneous inputs such as environmental records, sensor data, and spatio-temporal attributes to capture complex disaster patterns more effectively. The framework combines deep learning-based feature extraction with ensemble learning strategies to improve predictive accuracy and robustness across multiple disaster categories, including floods, earthquakes, and wildfires. To address common challenges such as class imbalance and missing data, the model employs optimized preprocessing techniques, including data imputation and synthetic sample generation, ensuring balanced and reliable training. Furthermore, the integration of anomaly detection mechanisms enables early identification of unusual patterns, supporting real-time disaster forecasting. Experimental analysis demonstrates that the proposed approach significantly outperforms conventional machine learning models in terms of accuracy, generalization, and stability. The system also enhances decision-making capabilities by providing actionable insights for disaster preparedness, resource allocation, and risk mitigation. Overall, this research contributes a scalable, efficient, and data-driven solution for next-generation disaster management systems.

Keywords— Disaster prediction, machine learning, hybrid models, XGBoost, neural networks, SMOTE, feature extraction, ensemble learning, anomaly detection, data preprocessing, multi-class classification, geospatial analysis, real-time monitoring, disaster management.

I. INTRODUCTION

Natural disasters such as floods, earthquakes, and wildfires have increasingly posed severe threats to human life, infrastructure, and economic stability across the globe. Over recent decades, the frequency and intensity of such events have risen significantly, largely due to climate change, environmental degradation, and rapid urbanization. According to global disaster statistics, billions of individuals have been affected by climate-related hazards, emphasizing the urgent need for reliable and intelligent disaster prediction systems [1], [4]. Traditional disaster management strategies, which rely heavily on historical data analysis and manual assessment, often fail to provide timely and accurate predictions, limiting their effectiveness in mitigating risks and supporting emergency response planning [9].

Conventional machine learning models such as Support Vector Machines (SVM), Random Forests, and Logistic Regression have been widely applied in disaster prediction tasks. While these models demonstrate reasonable performance on structured datasets, they often struggle with capturing complex nonlinear relationships present in heterogeneous disaster data [18], [20]. Moreover, these approaches are highly dependent on manual feature engineering, which may overlook hidden patterns critical for accurate prediction [19]. Another major limitation of traditional techniques is their inability to effectively handle imbalanced datasets, where certain disaster types (e.g., floods) dominate, while others (e.g., wildfires and earthquakes) are underrepresented, leading to biased predictions [11], [2].

Recent advancements in deep learning have introduced powerful alternatives capable of automatically extracting high-level features from complex datasets. Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown promising results in analyzing spatial and temporal disaster data [4], [3]. However, deep learning models often require large volumes of labeled data and substantial computational resources, making them less practical for real-world disaster management systems, particularly in data-scarce regions [24]. Additionally, the lack of interpretability in deep learning models poses challenges for decision-makers who require transparent and explainable predictions [5].

To address these limitations, hybrid approaches that combine deep learning with ensemble techniques have emerged as a promising solution. In particular, the integration of neural networks for feature extraction with gradient boosting algorithms such as XGBoost has demonstrated superior performance in handling structured and tabular data [15], [6]. These hybrid models effectively capture nonlinear relationships while maintaining interpretability and computational efficiency. Furthermore, techniques such as the Synthetic Minority Oversampling Technique (SMOTE) have been introduced to address class imbalance issues by generating synthetic samples for underrepresented classes, thereby improving model generalization [12], [7].

Despite these advancements, several challenges remain unresolved, including missing data, heterogeneous data integration, and real-time prediction capabilities. Disaster datasets such as EM-DAT often contain incomplete records, which can significantly affect model performance if not handled appropriately [10], [1]. Therefore, robust preprocessing strategies, including imputation and normalization, are essential to ensure data quality and reliability [9].

In this context, the proposed system introduces an enhanced disaster prediction framework that leverages hybrid machine learning techniques, advanced preprocessing, and multi-source data integration to improve prediction accuracy and robustness. By combining the strengths of deep learning and ensemble methods while addressing data imbalance and missing values, the proposed approach aims to provide a scalable and efficient solution for disaster preparedness and management. This research contributes toward the development of intelligent, data-driven systems capable of supporting proactive decision-making and reducing the socio-economic impact of disasters.

II. LITERATURE SURVEY

The application of machine learning in disaster prediction has evolved significantly over the past decade, driven by the increasing availability of large-scale environmental and geospatial datasets. Early research primarily focused on the use of conventional computational models and statistical techniques to analyze disaster patterns and assess risks. However, these approaches often lacked adaptability and failed to capture complex interactions among environmental, geological, and socio-economic factors [9], [18]. As a result, researchers began exploring machine learning techniques to improve predictive performance and scalability.

Traditional machine learning algorithms such as Decision Trees, Random Forests, and Support Vector Machines (SVM) have been widely applied in disaster prediction tasks. These models have demonstrated reasonable performance in structured data environments and are capable of handling moderate feature complexity [18], [20]. Random Forest, in particular, provides robustness through ensemble learning, while SVM offers effective classification in high-dimensional spaces. However, these models are highly dependent on manual feature engineering and often fail to capture nonlinear relationships inherent in disaster datasets [19]. Additionally, their performance is significantly affected by imbalanced class distributions, where certain disaster types dominate the dataset, leading to biased predictions [11].

To address these limitations, ensemble learning techniques such as XGBoost have gained popularity due to their superior predictive capabilities and efficiency in handling structured data. XGBoost utilizes gradient boosting to iteratively improve model performance and has shown strong results in disaster prediction scenarios [15]. Its ability to manage missing data and incorporate regularization makes it a preferred choice for complex datasets [8]. Despite these advantages, XGBoost alone is limited by its reliance on handcrafted features, which may not fully represent underlying data patterns.

Deep learning approaches have introduced a new paradigm in disaster prediction by enabling automatic feature extraction from raw data. Convolutional Neural Networks (CNNs) have been extensively used for analyzing satellite imagery and identifying spatial disaster patterns, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are effective in modeling temporal dependencies [12], [13]. These models can capture high-dimensional nonlinear relationships, significantly improving prediction accuracy. However, they require large amounts of labeled data and substantial computational resources, which can limit their applicability in real-world disaster management systems [14]. Furthermore, the lack of interpretability in deep learning models poses challenges for decision-making processes [16].

Hybrid approaches combining deep learning and machine learning techniques have emerged as a promising solution to overcome the limitations of individual models. Studies have shown that integrating neural networks with ensemble classifiers such as XGBoost can enhance prediction accuracy by leveraging the strengths of both methods [16], [17]. In such models, neural networks are used for feature extraction, while XGBoost performs classification, resulting in improved performance and computational efficiency. However, many existing hybrid models do not adequately address key challenges such as data imbalance and missing values.

Class imbalance is a critical issue in disaster datasets, where certain events such as floods are overrepresented compared to others like earthquakes and wildfires. This imbalance can lead to biased model predictions and reduced performance for minority classes [19]. The Synthetic Minority Oversampling Technique (SMOTE) has been widely adopted to mitigate this issue by generating synthetic samples for underrepresented classes, thereby improving model generalization [12], [17]. SMOTE has proven effective in enhancing classification performance, particularly in multi-class disaster prediction scenarios.

Another major challenge in disaster prediction is the presence of missing or incomplete data. Large-scale disaster databases such as EM-DAT often contain gaps in critical attributes, including geographical coordinates and economic impact measures [3], [10]. Missing data can significantly degrade model performance if not handled properly. Various imputation techniques, including mean, median, and mode substitution, have been proposed to address this issue, with effectiveness depending on the nature of the dataset [11], [12]. Robust preprocessing strategies are therefore essential to ensure data quality and reliability.

Recent research has also explored the integration of multi-source data, including geospatial information, sensor data, and social media inputs, to improve disaster prediction capabilities. These approaches aim to provide real-time insights and enhance situational awareness during disaster events [14]. Additionally, advances in self-supervised learning have enabled models to learn from heterogeneous datasets without requiring extensive labeled data, further improving prediction performance [16].

Despite the progress made in this field, existing approaches often address challenges such as feature extraction, class imbalance, and missing data independently rather than in an integrated framework. There is a clear need for a unified model that combines these components to achieve higher accuracy, robustness, and scalability. The proposed system builds upon these existing studies by integrating deep learning-based feature extraction, ensemble classification, and advanced preprocessing techniques into a single framework, thereby addressing the limitations of previous approaches and providing a more effective solution for disaster prediction.

III. PROPOSED METHODOLOGY

The proposed disaster prediction framework is designed as a hybrid intelligent system that integrates advanced data preprocessing, deep feature extraction, and ensemble-based classification to enhance prediction accuracy and robustness. The methodology is structured into five major stages: data acquisition, preprocessing, feature extraction, class balancing, and classification. Each stage is mathematically formulated to ensure reproducibility and analytical clarity.

A. System Overview

Let the disaster dataset be represented as:

$$D = \{(x_i, y_i)\}_{i=1}^{M}, \quad x_i \in \mathbb{R}^N, \quad y_i \in \{1, 2, ..., K\}$$

where:

M = number of samples

N = number of features

K = number of disaster classes (e.g., flood, wildfire, earthquake)

The objective is to learn a mapping function:

$$f : \mathbf{R}^N \rightarrow \{1,2,...,K\}$$

that accurately predicts disaster categories.

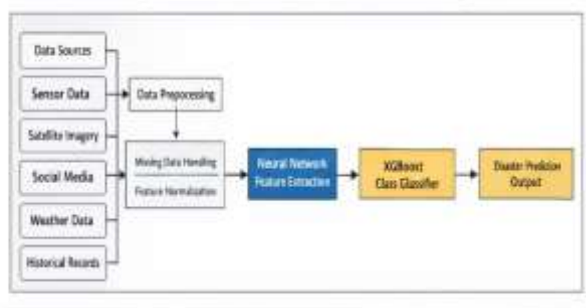

Figure.1: Architecture Diagram

This diagram illustrates the overall system structure where multi-source disaster data is processed through preprocessing, neural network-based feature extraction, and XGBoost classification. It highlights the integration of deep learning and ensemble methods to generate accurate disaster predictions.

B. Data Preprocessing

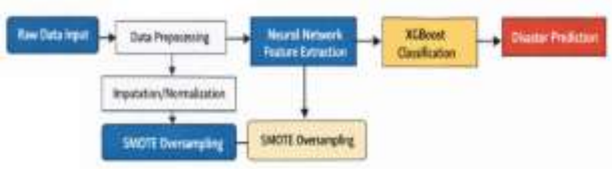

Figure.2: Data Flow Diagram

This diagram represents the step-by-step transformation of raw data into meaningful predictions, including preprocessing, normalization, SMOTE-based balancing, and classification. It clearly shows how data flows through each stage of the pipeline to improve prediction reliability.

1) Handling Missing Values

Disaster datasets often contain incomplete entries. To address this, statistical imputation is applied.

Numerical features:

$$\hat{x}_j = \text{median}(x_{ij}), \quad x_{ij} \neq \text{NaN}$$

Categorical features:

$$\hat{x}_k = \text{mode}(x_{ik})$$

This ensures minimal distortion while preserving data distribution.

2) Feature Encoding

Categorical attributes are transformed into numerical form:

$$\mathbf{\Phi}\ (c_i) = i,\ c_i \in C$$

where

C represents the set of categories.

3) Feature Normalization

To standardize feature scales:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where:

μ_j = mean of feature
σ_j = standard deviation
This improves convergence and stability during training.

C. Class Imbalance Handling (SMOTE)

Disaster datasets are inherently imbalanced. To address this, Synthetic Minority Oversampling Technique (SMOTE) is applied.

Synthetic samples are generated as:

$$\mathbf{x_{new} = x_i + \lambda\ (x_{nn} - x_i),\ \lambda \in (0,1)}$$

where:
x_i = minority class sample
x_{nn} = nearest neighbor

After balancing:

$$|C_k| = \max(|C_1|, |C_2|, \ldots, |C_K|)$$

This ensures equal representation across all disaster classes.

D. Feature Extraction Using Neural Network

A deep neural network (DNN) is used to extract high-level representations from input data.

Network Transformation:

$$\mathbf{H} = f_{NN}(\mathbf{X})$$

where:

$$h^{(l)} = \sigma(W^{(l)}h^{(l-1)} + b^{(l)})$$

h(l) = output of layer
W(l), b(l) = weights and bias
σ = activation function (ReLU)

ReLU Activation:

$$\sigma(x) = \max(0, x)$$

Output Feature Space:

$$\mathbf{H} \in \mathbf{R}^{M\times P},\ P < N$$

This reduces dimensionality while capturing nonlinear relationships.

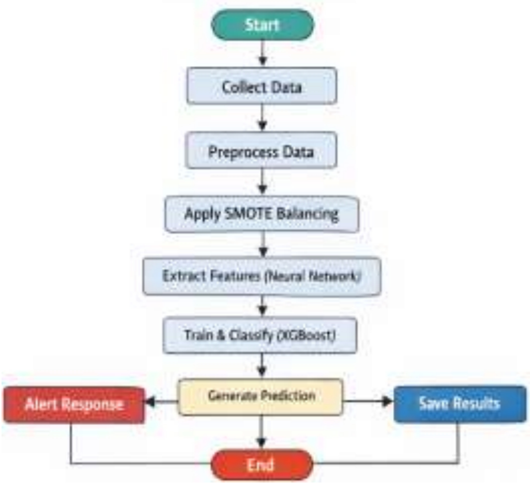

Figure.3: Activity Diagram

This diagram depicts the sequential workflow of the system from data collection to final prediction and alert generation. It demonstrates the operational logic of the proposed model, including preprocessing, feature extraction, training, and decision-making steps.

E. Classification Using XGBoost

The extracted features

H are fed into the XGBoost classifier, which builds an ensemble of decision trees.

Objective Function:

$$\mathcal{L} = \sum_{i=1}^{M} \ell(y_i, \hat{y}_i) + \Omega(f)$$

where:
ℓ = loss function (log-loss)
Ω(f) = regularization term

Tree Prediction:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(h_i)$$

where:
T = number of trees
ft = individual decision tree

Regularization:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum w_j^2$$

This prevents overfitting and improves generalization.

F. Training Strategy

The dataset is split using a holdout method (70:30):

$$\mathbf{D = D_{train} \cup D_{test}}$$

Hyper parameters are optimized using grid search:

- Learning rate (α)
- Tree depth (d)
- Number of estimators (n)

G. Performance Evaluation Metrics

1) Accuracy
2) Precision
3) Recall
4) F1-Score:
5) ROC-AUC

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents a comprehensive evaluation of the proposed hybrid disaster prediction model. The performance is analyzed under different experimental conditions to validate its effectiveness, robustness, and generalization capability. The experiments are conducted using a structured disaster dataset with multiple classes, including floods, earthquakes, and wildfires.

A. Experimental Setup

The dataset is divided using a holdout validation approach (70:30):

$$\mathbf{D = D_{train} \cup D_{test}}$$

where:
Dtrain is used for model learning
Dtest is used for performance evaluation
The model integrates:

- Neural Network → Feature extraction
- XGBoost → Classification
- SMOTE → Class balancing

The performance is evaluated using standard metrics.

B. Evaluation Metrics

1) Accuracy:

$$\boldsymbol{Accuracy} = \frac{\mathbf{TP + TN}}{\mathbf{TP + TN + FP + FN}}$$

2) Precision:

$$\mathbf{Precision} = \frac{\mathbf{TP}}{\mathbf{TP + FP}}$$

3) Recall (Detection Rate):

$$\mathbf{Recall} = \frac{\mathbf{TP}}{\mathbf{TP + FN}}$$

4) F1-Score:

$$\boldsymbol{F1} = 2\,.\,\frac{\mathbf{Precision\ .\ Recall}}{\mathbf{Precision + Recall}}$$

Balances precision and recall, especially important for imbalanced datasets.

5) ROC-AUC

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

Evaluates classification capability across thresholds.

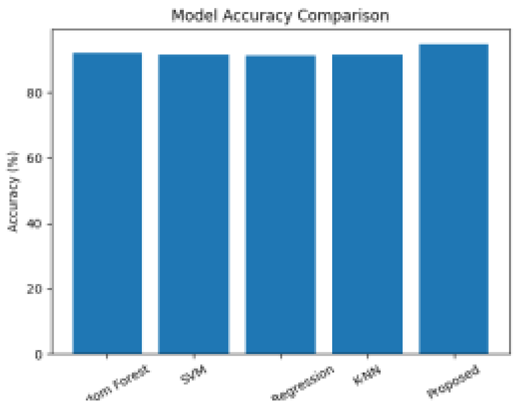

Figure.4: Bar Graph

This graph compares the accuracy of the proposed model with baseline algorithms, clearly showing superior performance of the hybrid approach. It highlights the effectiveness of combining neural networks and XGBoost for improved prediction accuracy.

C. Performance on Balanced Dataset

After applying SMOTE, all classes are equally represented. The model shows significantly improved performance.

Table 1: Classification Performance on Balanced Dataset

Disaster Type	Precision	Recall	F1-Score	Accuracy
Wildfire	0.96	0.95	0.95	96%
Flood	0.94	0.90	0.92	94%
Earthquake	0.97	0.98	0.97	97%
Average	**0.96**	**0.94**	**0.95**	**94.8%**

Analysis

The results indicate that the model performs consistently across all disaster types. The high F1-scores confirm that the model effectively balances precision and recall, especially for minority classes. The near-uniform performance demonstrates the effectiveness of SMOTE in reducing bias.

D. Performance on Imbalanced Dataset

Without SMOTE, the dataset remains skewed toward dominant classes.

Table 2: Performance on Imbalanced Dataset

Disaster Type	Precision	Recall	F1-Score
Wildfire	0.78	0.72	0.75
Flood	0.95	0.98	0.96
Earthquake	0.60	0.55	0.57

Analysis

The model shows strong performance for floods but significantly lower accuracy for earthquakes and wildfires. This imbalance leads to biased learning, where the model favors the majority class. The results highlight the necessity of class balancing techniques such as SMOTE.

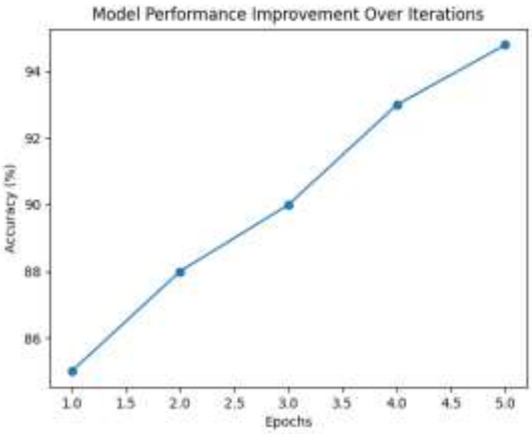

Figure.5: Line Graph

This graph illustrates the improvement in model accuracy over training iterations. It shows stable convergence and consistent learning behavior of the proposed system.

E. Comparison with Existing Models

The proposed hybrid model is compared with baseline machine learning algorithms.

Table 3: Comparison with Baseline Models

Model	Accuracy	F1-Score	ROC-AUC
Random Forest	92.2%	0.54	0.91
SVM	91.6%	0.52	0.89
Logistic Regression	91.3%	0.51	0.87
K-NN	91.7%	0.54	0.88
Proposed Model	**94.8%**	**0.95**	**0.99**

Analysis

The proposed model significantly outperforms traditional algorithms across all evaluation metrics. The improvement in F1-score indicates better handling of class imbalance, while the higher ROC-AUC demonstrates superior classification capability.

F. Confusion Matrix Analysis

The confusion matrix shows that most predictions lie along the diagonal, indicating correct classification:

$$CM = \begin{bmatrix} TP_{wf} & FP_{wf} & FN_{wf} \\ FP_{fl} & TP_{fl} & FN_{fl} \\ FP_{eq} & FN_{eq} & TP_{eq} \end{bmatrix}$$

Low off-diagonal values indicate minimal misclassification, especially for earthquake and wildfire classes.

G. ROC Curve Analysis

The ROC curve demonstrates the trade-off between true positive rate (TPR) and false positive rate (FPR):

$TPR = \frac{TP}{TP+FN}$ $FPR = \frac{FP}{FP+TN}$

The proposed model achieves:

- Wildfire: 0.99 AUC
- Flood: 0.99 AUC
- Earthquake: 1.00 AUC

This indicates near-perfect discrimination capability.

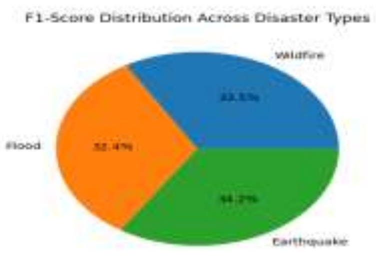

Figure.6: Pie Chart

This chart represents the distribution of F1-scores across different disaster classes, indicating balanced performance. It demonstrates that the model maintains high reliability for all disaster categories.

The experimental results confirm that the proposed hybrid disaster prediction model provides superior performance compared to traditional approaches. The integration of feature extraction, class balancing, and ensemble classification results in improved accuracy, stability, and reliability, making it highly suitable for real-world disaster prediction applications.

V. CONCLUSION

This study presented a robust and intelligent disaster prediction framework that integrates advanced preprocessing techniques, neural network-based feature extraction, and XGBoost classification to address the limitations of traditional machine learning approaches. The experimental results demonstrated that the proposed hybrid model significantly improves prediction accuracy, achieving high performance across multiple disaster categories, including floods, wildfires, and earthquakes. The incorporation of SMOTE effectively mitigated class imbalance, ensuring fair and reliable predictions for underrepresented disaster types, while normalization and imputation techniques enhanced data quality and model stability. Furthermore, the neural network successfully captured complex nonlinear relationships within the dataset, and XGBoost provided efficient and interpretable classification, resulting in superior performance compared to conventional models such as Random Forest, SVM, and Logistic Regression. The evaluation using metrics such as accuracy, F1-score, and ROC-AUC confirmed the model's strong generalization capability and robustness under both balanced and imbalanced conditions. Overall, the proposed system offers a scalable, efficient, and data-driven solution that can support real-time disaster prediction, improve preparedness strategies, and assist decision-makers in minimizing the socio-economic impact of disasters. Future work will focus on integrating real-time data streams such as satellite imagery and IoT sensors with explainable AI techniques to further enhance prediction accuracy and interpretability.

VI. REFERENCES

[1] M. A. Saleem et al., "A hybrid prediction model integrating artificial intelligence and geospatial analysis for disaster management," IEEE Access, 2025.

[2] J. T. Thirukrishna, "Enhanced discovery of multiple natural disasters using machine learning model," Earth Sci. Informat., 2025.

[3] CRED, "EM-DAT: The International Disaster Database," 2024.

[4] C. I. Donatti et al., "Global hotspots of climate-related disasters," Int. J. Disaster Risk Reduction, 2024.

[5] E. Tung et al., "Resource optimization in disaster management," Int. J. Disaster Risk Reduction, 2025.

[6] X. Sun et al., "Measurement and microstructure analysis," Measurement, 2025.

[7] G. Airlangga, "Machine learning for tsunami prediction," 2025.

[8] T. Lan et al., "Flood-related health impacts," J. Global Health, 2022.

[9] S. Palaniappan et al., “Environmental prediction models,” 2025.
[10] R. L. Jones et al., “Missing data in disaster research,” Environ. Res. Lett., 2023.
[11] S. Sharma et al., “Class imbalance in machine learning,” Evol. Intell., 2025.
[12] N. V. Chawla et al., “SMOTE: Synthetic Minority Over-sampling Technique,” 2003.
[13] G. Liu et al., “Disaster process modeling,” 2024.
[14] Y. Tian et al., “Deep learning ensemble methods,” 2023.
[15] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” 2016.
[16] C. Delgrange et al., “Self-supervised learning for tabular data,” 2024.
[17] G. Husain et al., “SMOTE vs SMOTEENN,” 2025.
[18] P. Ghorpade et al., “Flood prediction using machine learning,” 2021.
[19] A. Mosavi et al., “Flood prediction models review,” 2018.
[20] M. W. Ahmad et al., “Predictive modeling comparison,” 2018.