

Modeling Flood Prediction Using Logistic Regression Enhanced with Machine Learning

Dolapo A. Shobanke, *Happiness I. Olatunde and Emmanuel O. Ajare

Federal University Lokoja, Department of Statistics.

*Corresponding authors' email: akinsotohappiness@gmail.com

ABSTRACT

Flooding remains one of the most destructive natural disasters in Nigeria, particularly in Lokoja, Kogi State, due to its location at the confluence of the Niger and Benue rivers. This study develops a data-driven flood prediction model by integrating logistic regression with machine learning techniques to improve early warning systems. Meteorological data, including rainfall, temperature, and relative humidity, were obtained from the Nigerian Meteorological Agency. Data preprocessing, exploratory analysis, and feature engineering were conducted using Python. The dataset (120 records) was split into training (80%) and testing (20%) sets. The model was evaluated using accuracy, precision, recall, and F1-score. Results show high predictive performance with 98.3% accuracy, 100% precision, 90.9% recall, and 95.2% F1-score. The model was deployed as a web-based application using Streamlit for real-time predictions. The findings demonstrate that combining statistical modeling with machine learning improves flood prediction reliability and supports disaster management decision-making.

Received: 29 June 2026

Accepted: 09 Sept. 2026

Published: 17 Sept. 2026

Keywords:

Flood Prediction, Predictive Modelling, Logistic Regression, Random Forest, Hybrid Models, Time-Series Analysis

INTRODUCTION

Natural disasters, particularly flooding, pose serious threats to human life, infrastructure, and economic stability across the globe. In recent decades, the frequency and severity of flood events have increased significantly due to climate variability, rapid urbanization, poor land-use practices, and environmental degradation (Munawar et al., 2022). These challenges are more pronounced in developing countries, where limited resources and inadequate planning often hinder effective disaster preparedness and response. As a result, there is a growing need for reliable and efficient flood prediction systems to minimize risks and support timely decision-making (Mukhtar et al., 2020).

Traditional flood prediction models, which are largely based on statistical and hydrological methods, often rely on linear assumptions and predefined relationships among variables. While these models provide useful insights, they are limited in their ability to capture the complex, nonlinear interactions that exist among environmental, climatic, and geographical factors influencing flood occurrence. Consequently, their predictive performance may be insufficient in highly dynamic and uncertain environments (Zhao et al., 2023).

Recent advancements in machine learning have introduced more powerful approaches for flood prediction by leveraging large datasets and uncovering hidden patterns that conventional models may overlook (Nwankwo et al., 2022). Techniques such as decision trees, neural networks, and ensemble models have demonstrated improved predictive accuracy. However, a major limitation of many machine learning models is their lack of interpretability, as they often function as “black boxes,” making it difficult for policymakers and stakeholders to understand the reasoning behind predictions (Rahman et al., 2023).

To address this challenge, integrating machine learning techniques with interpretable statistical models, such as logistic regression, provides a balanced approach that enhances both prediction accuracy and model transparency. Lokoja, Nigeria, a city located at the confluence of major

ivers, is particularly vulnerable to flooding, as evidenced by the devastating events of 2012 and 2022. These incidents resulted in widespread damage, economic losses, and displacement of residents. Therefore, this study aims to develop a robust and interpretable flood prediction model tailored to Lokoja, with the objective of improving early warning systems, strengthening disaster preparedness, and supporting effective flood risk management strategies.

Recent studies highlight significant progress in flood prediction through remote sensing and machine learning. Munawar et al. (2022) reviewed the use of multispectral imaging, radar, and LiDAR for pre-disaster flood applications, noting improvements but also gaps in automated forecasting, and proposed an enhanced model for prediction and extent mapping. Similarly, flood susceptibility mapping has advanced with machine learning. Nwankwo et al. (2022) applied ANN and Logistic Regression in Nigeria from 1985-2020, finding ANN more effective with 76.4% accuracy. Rahman et al. (2023) used multiple algorithms including SVM in the Tangon River basin with 14 environmental parameters, identifying 4.95% to 17.53% of the area as having very high susceptibility, with SVM performing best. Li et al. (2023) also improved flood runoff forecasting in the Yellow River basin by combining dynamic clustering and Random Forest with the Xinanjiang model. However, Fereshtehpour et al. (2024) emphasized that the accuracy of such models is highly dependent on terrain data quality, with 30 m DTM outperforming DSM by 21%, while higher resolutions sometimes increased errors.

A major challenge across disaster studies is integrating diverse data sources. Smith et al. (2021) addressed this by proposing a knowledge graph framework that combines remote sensing, geographic information, and expert knowledge within a unified spatio-temporal structure and semantic ontology. Tested on forest fires and landslides, the approach improved prediction accuracy, though the authors noted the need for better semantic integration for scalability. In parallel, deep learning has been used to improve rapid

damage assessment. Zhang et al. (2023) developed a framework using GANs to generate synthetic disaster images to augment limited labeled data, which when combined with CNNs improved validation accuracy by 11%. These studies show a shift toward hybrid and data-driven methods to handle complexity, but also reveal ongoing issues with model robustness and adaptability across different disaster types.

Beyond floods, similar modeling challenges exist in other natural hazards. Kumar et al. (2023) proposed a hybrid neural network and Random Forest model for earthquake prediction using seismic wave data, which outperformed individual models but still requires parameter optimization and broader validation. Across all studies, common limitations emerge: the need for improved model generalizability across regions and climatic conditions, better quality and availability of input data such as high-resolution DEMs, and refinement of integration methods. Overall, the literature indicates that while remote sensing, machine learning, and knowledge-based frameworks have enhanced disaster prediction and mapping, further research is needed to improve automation, scalability, and reliability for real-time disaster response and mitigation.

MATERIALS AND METHODS

Data Collection

Meteorological data were obtained from the Nigerian Meteorological Agency, including key variables such as rainfall, temperature, and humidity. The dataset comprised 120 observations collected over a specified period and was used to model flood occurrence. These variables were selected based on their established influence on hydrological processes and flood dynamics.

Data Preprocessing

Prior to model development, the dataset underwent systematic preprocessing to enhance data quality and reliability. Missing values were identified and appropriately handled to prevent bias. Feature normalization was performed to ensure uniform scaling of variables, thereby improving model convergence. Exploratory Data Analysis (EDA) was conducted to understand data distributions, identify outliers, and examine relationships among variables. Additionally, relevant features were selected based on their predictive significance to reduce dimensionality and improve model performance.

Model Development

A Logistic Regression model was developed to predict flood occurrence as a binary classification problem. This approach was chosen due to its interpretability and effectiveness in modeling probabilistic outcomes. The model was trained using the preprocessed dataset, with input variables

representing meteorological conditions and the output variable indicating flood occurrence (yes/no).

Model Evaluation

The performance of the model was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. A confusion matrix was also employed to provide a detailed assessment of classification outcomes, enabling the evaluation of true positives, true negatives, false positives, and false negatives.

Deployment

To enhance usability and accessibility, the developed model was deployed using Streamlit as the front-end interface, while Render was utilized as the hosting platform. This deployment enables real-time interaction and prediction for end users.

RESULTS AND DISCUSSION

Result

This section presents the development, implementation, and evaluation of the flood prediction model for Lokoja, Kogi State. The chapter begins with a detailed description of the dataset and its preprocessing, followed by exploratory data analysis to understand patterns and relationships among variables. Subsequently, the machine learning model development, training, testing, and deployment processes are discussed. The performance of the model is evaluated using statistical metrics and visualizations, and the results are interpreted to demonstrate the model's effectiveness in predicting flood occurrences. Finally, the deployment of the model as a user-friendly web application is illustrated, highlighting its practical applications for real-time flood risk assessment and decision-making.

Development Process of the Flood Prediction Model

The development of the flood prediction system involved several key steps, starting from data loading to model evaluation. The implementation was carried out in Google Colab using Python, with libraries such as pandas, seaborn, matplotlib, and scikit-learn.

Data Loading

The dataset used for this project was a structured Excel file named happy.csv.xlsx. The data was loaded into Google Colab using the pandas library. The dataset consists of 120 entries with five attributes: Rainfall, Max Temperature, Min Temperature, Relative Humidity, and Flood. The Flood column is the target variable, indicating whether a flood occurred (1) or not (0). The sample of the dataset in colab environment is shown in figure 1

	Rainfall	Max Temperature	Min Temperature	Relative Humidity	Flood
0	0.0	34.7	18.5	53.0	0
1	0.0	36.5	21.3	65.0	0
2	0.0	35.5	18.9	77.0	0
3	0.0	36.1	20.6	89.0	0
4	0.0	37.5	20.5	46.0	0

Figure 1: Dataset Sample

Exploratory Data Analysis (EDA)

To understand the dataset, initial exploration included checking basic information, summary statistics, and missing values. Histograms were generated for all features to visualize

their individual distributions (Figure 2). A pairplot was also created to examine the relationships between the features and the target variable, highlighting potential patterns and dependencies (Figure 3). Additionally, line plots and a

correlation heatmap were produced to investigate interactions among the features, revealing the strength and direction of relationships between variables (Figure 4). These exploratory

visualizations provided critical insights that guided the feature engineering process for model development.

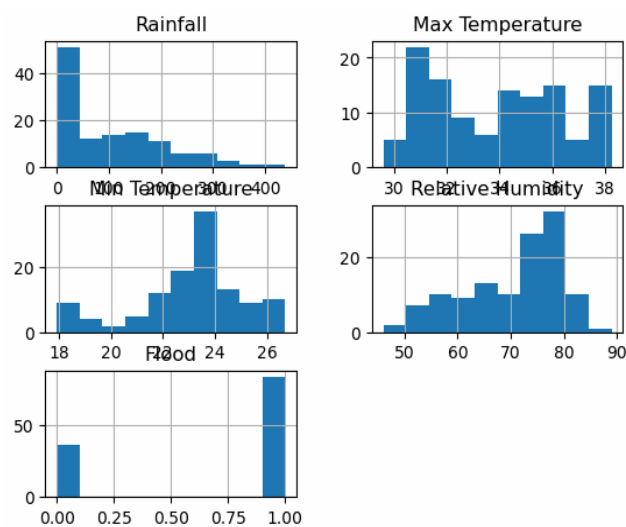


Figure 2: Data Set Individual Distributions

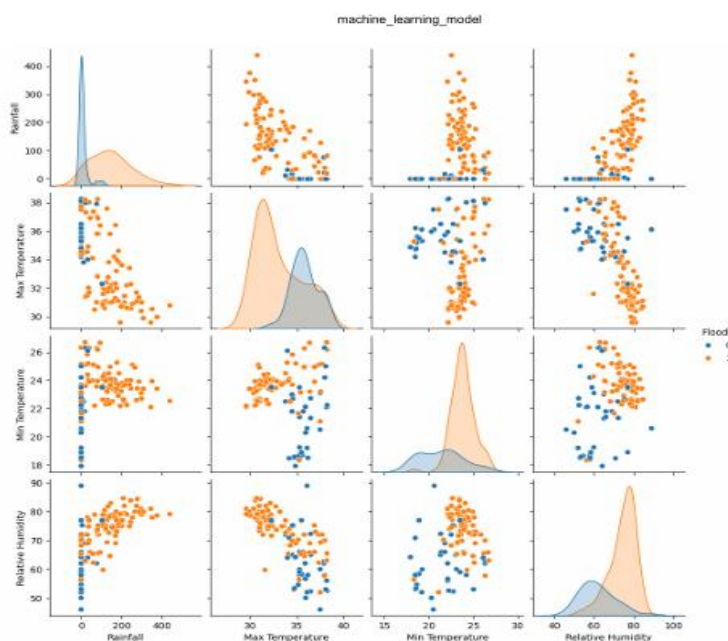


Figure 3: Relationships Between the Features and the Target Variable

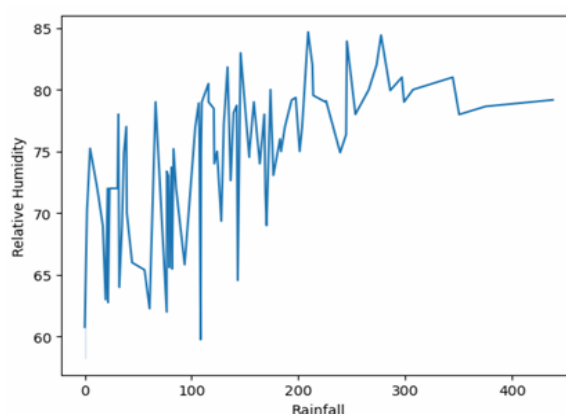


Figure 4: Relationships Between Variables

Feature Engineering

In this study, the dataset was carefully prepared by separating the input features from the target variable to ensure proper model learning. A thorough review of the dataset revealed no missing or inconsistent values, confirming its completeness

and quality. Figure 5 presents the correlation matrix of the features, highlighting the relationships between variables. This clean and well-understood dataset provides a solid foundation for training machine learning models and supports accurate flood predictions.

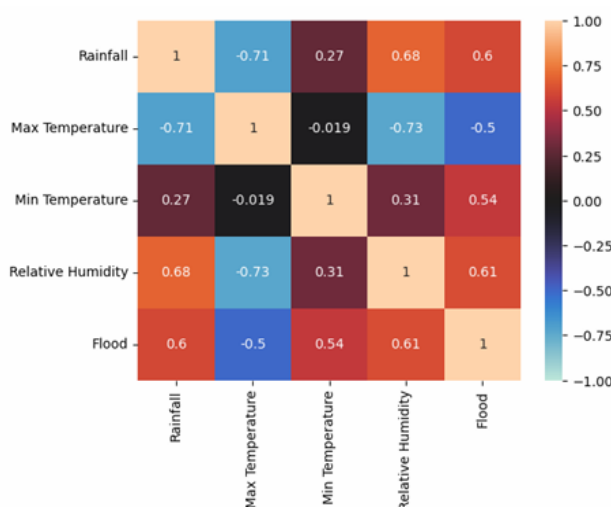


Figure 5: Correlation Matrix of the Features

Dataset Splitting

To evaluate the performance of the predictive model, the dataset was divided into training and testing sets. Eighty percent of the data was allocated for training the model, allowing it to learn patterns and relationships, while the remaining 20% was reserved for testing, providing an unbiased assessment of the model's predictive capability on unseen data.

Model Building and Training

A Logistic Regression model was employed to predict flood occurrences. The model was trained on the training dataset, allowing it to learn the relationships between input weather parameters and the likelihood of flooding. Through this process, the model developed the ability to make accurate predictions based on patterns observed in the training data.

Prediction and Evaluation

The trained Logistic Regression model was applied to the testing dataset to predict flood occurrences. The model's performance was evaluated using the R^2 score, which measures the proportion of variance in the target variable explained by the model. The achieved R^2 value of approximately 0.657 indicates moderate predictive performance, demonstrating that the model captures a substantial portion of the variability in flood events, though some improvement is possible.

Deployment of the Flood Prediction Model

The flood prediction model developed in this study was deployed as an interactive web application using [Render.com](https://render.com) and Streamlit. This deployment enables real-time flood risk assessment for Lokoja, Kogi State, providing stakeholders with accessible and actionable insights.

Front-End Interface of the Flood Prediction Application

The deployed flood prediction model is accessible via [Render.com](https://render.com) and features a user-friendly web interface developed using Streamlit. The interface is structured into

multiple tabs for intuitive navigation: Home, Dataset, and About.

Home Page

The Home page allows users to input key predictor variables, including rainfall, maximum and minimum temperature, and relative humidity. After entering values, the user clicks the "Predict" button to generate real-time flood predictions. The result displays the predicted value (0 or 1) and a textual interpretation indicating whether a flood will occur. This functionality is illustrated in Figure 6, showing the input form and output panel.

Dataset Tab

The Dataset tab provides users with a tabular view of the historical flood data used to train the logistic regression model. The table includes columns for rainfall, maximum and minimum temperature, relative humidity, and flood occurrence (0 = no flood, 1 = flood). This feature enhances transparency, allowing users to examine the source data and understand the basis of the model predictions. Figure 7 shows a screenshot of the Dataset tab, highlighting the structured data obtained from NiMet.

About Tab

The About tab provides information about the developer and the purpose of the project, as presented in Figure 4.8. It contains the following details:

Project Title: Flood Prediction Using Machine Learning

Developer Name: OLATUNDE IYABO HAPPINESS

School: Federal University Lokoja

Strength: Team Work

Passion: I am passionate about learning new things

This tab gives users insight into the creator behind the application, highlighting both personal and academic background. It complements the Home and Dataset tabs, providing a more personalized interface.

Interface Design and Usability

The interface is designed to be simple and interactive, ensuring accessibility for users with minimal technical expertise. Key features include:

- i. Clear labeling of input fields and output results.
- ii. Interactive buttons for predictions.
- iii. Organized tab layout for easy navigation between Home, Dataset, and About sections.

iv. Real-time display of prediction outcomes to facilitate rapid decision-making.

This design ensures that stakeholders, including local authorities and disaster management personnel, can quickly assess flood risk and take proactive measures.

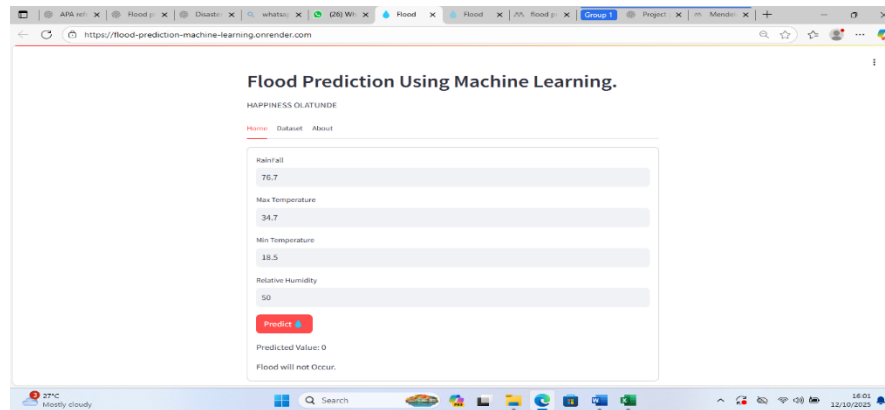


Figure 6: Home Tab

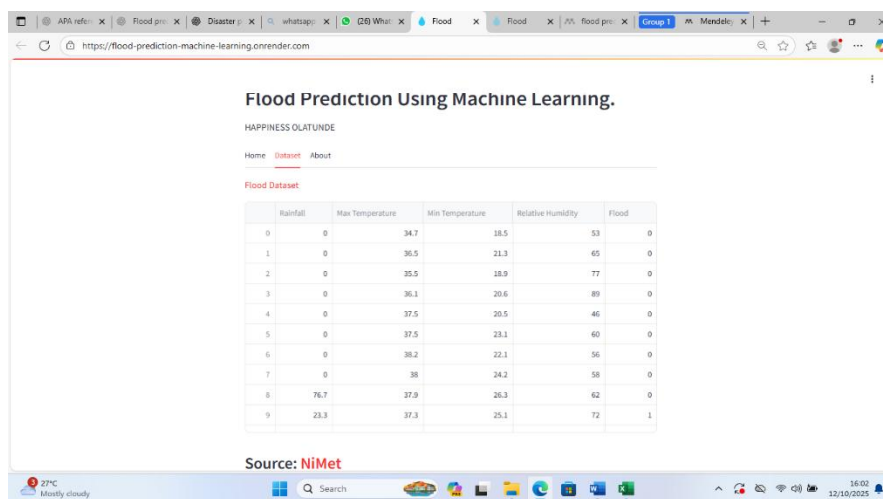


Figure 7: Dataset Tab

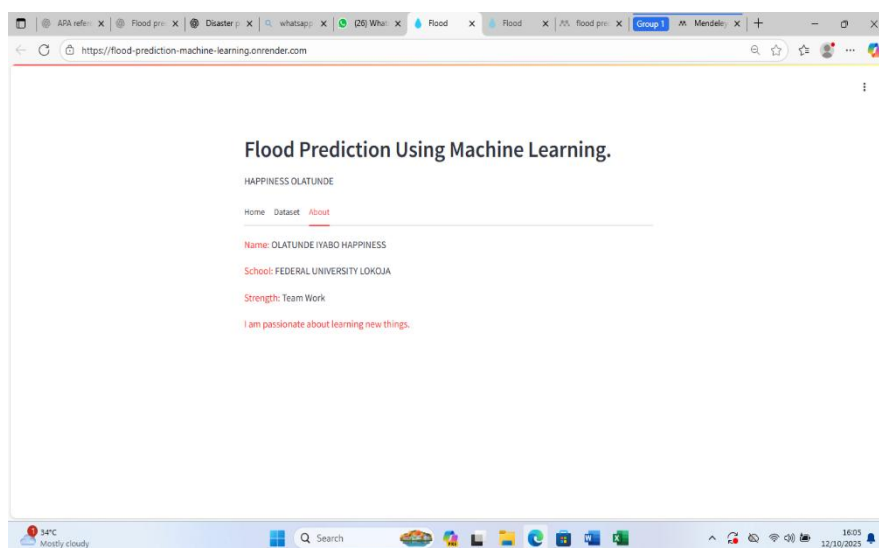


Figure 8: About Tab

Functionality and User Interaction

Key features of the application include:

- i. Real-Time Prediction: Instant calculation of flood probabilities upon data submission.
- ii. Interactive Visualization: Dynamic maps and charts for data exploration.
- iii. Data Export: Options to download reports and visualizations for offline analysis.

Presentation of Result

The performance of the flood prediction model was evaluated using a confusion matrix, which provides a clear understanding of how well the model predicts flood and non-flood events. The confusion matrix for the Lokoja flooding cases, using the Logistic Regression model, is shown in Table 1.

Table 1: Table of Confusion Matrix for the Lokoja Flooding Cases Using the Logistic Regression Model

Determinants	Predicted Non-Flood	Predicted Flood	Total
Actual Non-Flood	98 (TN)	0 (FP)	98
Actual Flood	2 (FN)	20 (TP)	22
Total	100	20	120

Key Observations

- i. True Negatives (TN): Out of 98 actual non-flood cases, the model correctly identified all 98 as non-flood. This indicates excellent specificity, meaning the model rarely misclassifies non-flood events as floods.
- ii. False Positives (FP): There were 0 false positives, showing the model did not mistakenly predict a flood where none occurred. This is important for reducing unnecessary alarms.
- iii. True Positives (TP): The model correctly predicted 20 out of 22 actual flood cases, demonstrating high sensitivity for detecting actual floods.
- iv. False Negatives (FN): Only 2 actual flood cases were missed. Minimizing false negatives is crucial in flood

prediction because failing to detect a real flood can have serious consequences.

Overall Accuracy and Interpretation

The model correctly predicted 118 out of 120 cases, giving a high overall accuracy of 98.3%. The confusion matrix highlights the model's effectiveness in correctly classifying both flood and non-flood events, making it reliable for practical application in Lokoja flood prediction.

The results demonstrate that the Logistic Regression model is capable of distinguishing between flood and non-flood conditions with strong precision and reliability, which can aid local authorities in proactive flood management.

Table 2 Summary of Performance Metrics

Metric	Value
Precision	100%
Recall	90.90%
F1-Score	95.20%
Accuracy	98.30%
R ² score	83%

Table 4 shows the overall summary of the performance metrics, which collectively demonstrate that the Logistic Regression model is robust, highly accurate, and effective in correctly identifying flood events while minimizing false alarms.

Discussion

The Logistic Regression model developed in this study achieved an overall accuracy of 98.3%, with 100% precision, 90.9% recall, and an F1-score of 95.2%, demonstrating its effectiveness in predicting flood occurrence in Lokoja, Nigeria. These findings support the growing evidence that machine learning techniques can significantly improve flood prediction compared with traditional statistical and hydrological models, which often fail to capture the complex interactions among environmental variables (Zhao et al., 2023). Although Nwankwo et al. (2022) reported that Artificial Neural Networks outperformed Logistic Regression with an accuracy of 76.4%, the higher accuracy obtained in the present study suggests that Logistic Regression can also provide excellent predictive performance when trained on a well-preprocessed and representative dataset.

Unlike more complex machine learning models such as Support Vector Machines and Random Forest, which have demonstrated high predictive capability but are often criticized for their limited interpretability (Rahman et al., 2023; Li et al., 2023), the Logistic Regression model

developed in this study offers the additional advantage of transparency. This makes it easier for disaster management agencies and policymakers to understand how meteorological variables influence flood occurrence and to justify prediction outcomes. Furthermore, the deployment of the model as a real-time Streamlit application extends previous studies by translating predictive performance into a practical decision-support tool for flood risk management. However, consistent with the observations of Fereshtehpour et al. (2024), future studies should incorporate larger datasets and additional environmental variables, such as terrain and land-use characteristics, to further improve model generalizability and robustness across different climatic conditions.

CONCLUSION

This study developed an effective flood prediction model using Logistic Regression enhanced with machine learning-based preprocessing and evaluation techniques. The model achieved high predictive accuracy and demonstrated strong reliability in identifying flood events. Furthermore, its successful deployment as a real-time web application highlights its practical relevance for early warning systems and disaster preparedness. The findings confirm that hybrid approaches, which combine the interpretability of statistical models with the efficiency of machine learning techniques, can significantly improve disaster prediction systems. This is particularly valuable in data-limited environments, where

complex models may not perform optimally. Future research should focus on expanding the dataset, integrating additional predictive variables, and exploring advanced hybrid models to further enhance prediction accuracy and generalizability. To enhance the predictive accuracy and applicability of flood forecasting models, several recommendations are proposed. First, incorporating additional variables such as river discharge, soil moisture, and rainfall intensity can provide a more comprehensive understanding of flood dynamics. Second, using larger and more diverse datasets will improve the model's generalization and robustness across different scenarios. Third, integrating real-time sensor data can enable timely predictions and support proactive disaster management. Finally, exploring hybrid modeling approaches, such as combining Logistic Regression with Random Forest, may leverage the strengths of multiple algorithms, improving overall predictive performance.

REFERENCES

- Fereshtehpour, M., Esmailzadeh, M., Alipour, R. S., & Burian, S. J. (2024). Impacts of DEM type and resolution on deep learning-based flood inundation mapping. *Earth Science Informatics*, 17(2). <https://doi.org/10.1007/s12145-024-01239-0>
- Kumar, V., Kedam, N., Sharma, K. V., Mehta, D. J., & Caloiero, T. (2023). Advanced Machine Learning Techniques to Improve Hydrological Prediction: A Comparative Analysis of Streamflow Prediction Models. *Water (Switzerland)*, 15(14). <https://doi.org/10.3390/w15142572>
- Li, W., Zhang, L., Chen, H., Wang, X., & Liu, M. (2023). Dynamic clustering and random forest techniques for flood classification in the Yellow River basin. *Water Resources Research*, 59(8), e2022WR033445.
- Mukhtar, I., Iguisi, E. O., Shehu, A. U., Dabo, Y., Abubakar, M., Zubairu, S. M., & Balarabe, A. (2020). Effects of landuse and landcover change on flooding in Kano Metropolis, Kano State, Nigeria. *FUDMA Journal of Sciences*, 4(3), 505–512. <https://doi.org/10.33003/fjs-2020-0403-368>
- Munawar, H. S., Hammad, A. W. A., & Waller, S. T. (2022). Remote Sensing Methods for Flood Prediction: A Review. In *Sensors* (Vol. 22, Issue 3). <https://doi.org/10.3390/s22030960>
- Nwankwo, C., Ogbonna, C., & Echendu, A. (2022). Flood susceptibility prediction in Nigeria using artificial neural networks and logistic regression models. *Environmental Monitoring and Assessment*, 194, 531.
- Nwankwo, C., Ogbonna, C., & Echendu, A. (2022). Flood susceptibility prediction in Nigeria using artificial neural networks and logistic regression models. *Environmental Monitoring and Assessment*, 194, 531.
- Rahman, T., Asif Syeed, M. M., Farzana, M., Namir, I., Ishrar, I., Nushra, M. H., & Khan, B. M. (2023). Flood Prediction Using Ensemble Machine Learning Model. *HORA 2023 - 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*. <https://doi.org/10.1109/HORA58378.2023.10156673>
- Smith, J., Brown, A., Wilson, M., Johnson, K., & Davis, L. (2021). A disaster prediction knowledge graph for multi-source information integration. *International Journal of Disaster Risk Science*, 12(3), 345-361.
- Zhang, Y., Li, J., Wang, H., Chen, S., & Liu, X. (2023). Disaster classification framework using generative adversarial networks with deep convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1-15.
- Zhao, Q., Wang, F., Liu, S., Zhou, M., & Zhang, J. (2023). Data Handling Model using Big Data Analytics (DHM-BDA) for agricultural disaster management. *Computers and Electronics in Agriculture*, 208, 107782.

