Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 14 Issue 03s, 2025

Flood AI: Data-Driven Prediction with C4.5 Decision Trees

[1]Nilesh Dhannaseth, [2]Khushi Manjare, [3]Kshitij Deshpande, [4]Shruti Chimote
[1,2,3,4] Information Technology St. Vincent Pallotti College of Engineering & Technology
Nagpur, India
Email: [1]ndhannaseth@stvincent.edu.in, [2]khushimanjare65@gmail.com, [3]kshitijdeshpande2003@gmail.com, [4]chimoteshruti@gmail.com

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

Flood Prediction, Flood Detection, DataDriven System, Decision Tree, C4.5 Algorithm, Machine Learning, Environmental

Abstract

Flooding is one of the most severe natural disasters caused when heavy rain and swollen rivers destroy lives, homes, and the surrounding environment. It is also a major source of economic loss worldwide. The ability to predict floods in advance and respond on time plays a vital role in reducing the damage they cause. However, traditional forecasting methods often rely on complicated models and large resources, which are not always practical or accessible in vulnerable regions. In recent years, data-driven approaches such as machine learning have shown promising results in improving flood detection and prediction. This study explores the use of the C4.5 decision tree algorithm to predict floods in three regions of India. By analyzing environmental data such as rainfall, temperature, humidity, and river water levels, the model was able to make accurate predictions about flooding events. The findings show that data-based solutions can not only support existing forecasting systems but also provide quicker responses and better preparedness for communities at risk.

Introduction

Flooding is one of the most damaging and widespread natural disasters, leaving behind longlasting consequences for people, the environment, agriculture, and entire economies. Globally, floods account for nearly 40% of all natural disasters and affect millions of lives each year, making them a serious challenge to sustainable development and community resilience. In India, with its vast geography and diverse climatic conditions, floods are a recurring hazard that disrupt lives on a massive scale. Regions such as Mumbai, the Assam–Brahmaputra Valley, and Chennai are some of the most disaster-prone areas, regularly experiencing heavy flooding that causes economic losses, displaces populations, and brings everyday activity to a halt.

Over the past few decades, the intensity and frequency of floods have only worsened. Climate change, coupled with rapid urbanization, has amplified extreme rainfall events and placed additional pressure on already fragile urban infrastructure and drainage networks. Projections suggest that such severe flood events will increase further in the future, highlighting an urgent demand for prediction systems that are reliable, adaptive, and timely. These systems are especially critical in flood-vulnerable regions, where local communities often lack the resources to recover quickly from disasters. Traditionally, flood forecasting has been carried out using hydrological and meteorological models. These methods depend on rainfall patterns, river discharge, and related environmental indicators. While they have proven effective in many cases, they come with significant limitations: they require large volumes of data, involve complex analysis, and demand considerable cost and technical

expertise. As a result, applying such models widely in resource-constrained areas is often impractical.In response to these challenges, researchers are turning toward machine learning (ML) approaches. ML techniques can examine very large datasets and automatically identify hidden patterns that signal potential floods. This helps make predictions faster, more reliable, and more accessible. Among the various methods, the C4.5 decision tree algorithm stands out for its simplicity and effectiveness. It can handle both numerical and categorical data and generates clear, interpretable rules that authorities and communities can trust. By combining accuracy with transparency, decision tree-based models emerge as a practical and scalable solution for improving flood prediction in India and similar regions across the world.

Literature Survey

Flood prediction and detection has gained significant attention in recent years due to increasing climate variability and the integration of machine learning (ML) and IoT technologies into disaster management systems. Traditional hydrological models, although reliable, often involve high computational costs and require large volumes of historical data and expert calibration. In contrast, modern ML-based and sensor-driven approaches provide faster, more flexible, and more scalable alternatives. Several studies emphasize that combining data-driven methods with meteorological, environmental, and geospatial information significantly improves forecasting accuracy and disaster preparedness.

[1] A. Ullah et al. (2024) conducted a sensitivity analysis of a two-dimensional flood inundation model for Tous Dam and showed that flood predictions can vary considerably based on boundary and terrain conditions. Their research highlighted the importance of optimizing hydraulic parameters to improve inundation mapping accuracy.[2] K. Aatif et al. (2024) developed a deep-learning-based flood forecasting model for the Chenab River in Pakistan using temporal hydrological data. Their ConvLSTM-driven system successfully learned seasonal flood behavior and produced accurate long-term prediction results, demonstrating the usefulness of deep learning in river basin flood analysis.[3] J. Nanditha et al. (2023) analyzed the catastrophic 2022 Pakistan floods and identified rainfall anomalies, glacial lake outbursts, and infrastructure failure as major causes. Their work provided strong evidence that climate variability and poor urban planning greatly magnify flood severity.[4] M. Sajjad et al. (2023) examined disaster resilience trends in Pakistan using geospatial intelligence and catastrophe progression models. Their findings revealed that although early warning frameworks had improved, preparedness and post-flood recovery remained inconsistent.[5] H. B. Waseem and I. A. Rana (2023) reviewed flood patterns and response systems in Pakistan, emphasizing the need for technology-driven risk management strategies such as sensor networks and satellite-assisted models.[6] H. Hamidifar and M. Nones (2023) studied seventy years of flood-related fatalities globally and concluded that early detection and community alert mechanisms play a crucial role in reducing loss of life.[7] C. Prakash et al. (2023) developed FLOODWALL, an IoT-based real-time system that integrates ultrasonic sensors and cloud processing for flash-flood detection. Their work demonstrated that high-frequency sensor networks enhance situational awareness in urban flood environments.[8] Z. Manzoor et al. (2022) reviewed socioeconomic flood impacts and stressed that disaster preparedness planning should prioritize technological support systems for vulnerable populations, especially in developing countries.[9] M. Moishin et al. (2021) designed a hybrid ConvLSTM-based flood forecasting framework which significantly improved spatial and temporal prediction quality. Their system effectively modeled rainfall-runoff relationships.[10] Singh and Kumar (2021) reviewed flash-flood prediction methods and reported that data fusion and multi-sensor integration greatly reduce model uncertainty.[11] M. A. U. R. Tariq et al. (2020) critically reviewed flood risk mitigation strategies and proposed a selection framework that helps planners choose between structural and non-structural measures.[12] S. Puttinaovarat and P. Horkaew (2020) integrated satellite, weather, and crowd-sourced information with machine learning to improve flood predictions and noted that combining multiple data sources improves realtime performance.[13] Khalaf et al. (2020) applied ensemble learning and sensor networks for early warning systems and achieved over 80% prediction accuracy, but also highlighted challenges related to costs and scalability.[14] Amir Mosavi et al. (2018) surveyed over 180 flood-prediction studies and concluded that ensemble and hybrid systems outperform single-model approaches, particularly in complex climate zones.

[15] M. Khan et al. (2018) applied MLP networks to reduce false flash-flood alerts and demonstrated that data pre- processing significantly affects output accuracy.[16] S. Mojaddadi et al. (2017) conducted GIS-based

flood susceptibility mapping using ensemble machine learning and highlighted that remote sensing adds spatial intelligence to flood analysis.[17] R. Mousa et al. (2016) developed an urban flood detection system using infrared and ultrasonic technology and confirmed that sensor fusion improves detection reliability.[18] P. Mitra et al. (2016) introduced one of the earliest IoT-enabled ANN flood models, validating the role of embedded systems in disaster management platforms.

Problem Statement

Flooding is a natural disaster that affects millions of people every year, particularly in India, where states like Assam and cities such as Mumbai and Chennai face it regularly. It leads to extensive damage to farmland, infrastructure, and homes, while also displacing entire communities. In Assam, the Brahmaputra River overflows nearly every year, submerging villages and cutting off transport routes. In Mumbai, heavy monsoon rains often overwhelm the drainage system, causing flash floods and disrupting daily life. In Chennai, rivers such as the Adyar and Cooum flood during the Northeast Monsoon, inundating neighborhoods and critical infrastructure. Floods are a major source of loss of life, property damage, and economic setbacks. Therefore, accurate prediction and timely warning are essential. Without these, floods can inflict severe destruction. Early detection and effective forecasting systems are crucial to protect people and minimize the long- term impact of this disaster.

Proposed Approach

Floods remain one of the most recurring and devastating natural disasters in India, especially in regions like Assam, Mumbai, and Chennai. Year after year, these areas face heavy losses to life, property, and infrastructure due to unpredictable water levels and extreme rainfall events. To address this urgent challenge, we propose the development of a Secure and Reliable Flood Prediction System that not only predicts floods with accuracy but also ensures the information is accessible and useful for both disaster management authorities and local communities. Unlike conventional flood forecasting systems that often remain static and fail to adapt to changing weather and river dynamics, the proposed system is designed to be responsive and continuously evolving. It integrates two key sources of information: historical flood records and realtime hydrological data. By processing this data through the C4.5 decision tree algorithm, the system aims to generate clear, interpretable forecasts that classify flood threats into three categories: *Flood, Mild Flood, or No Flood*. The major strength of using C4.5 lies in its transparency—unlike blackbox AI models, it produces simple decision rules that authorities can easily understand and trust when making urgent decisions.

The development process begins with the collection of reliable datasets from trusted organizations. Rainfall records will be obtained from sources such as Kaggle repositories and the India Meteorological Department (IMD). Meanwhile, river water-level and discharge information will be sourced from the Central Water Commission (CWC) and India-WRIS stations. To capture the regional variability in flood behavior, additional geographic attributes such as latitude, longitude, and basin-specific identifiers will also be included. Before analysis, the raw data is carefully cleaned and pre-processed. Missing values are corrected using straightforward methods, while new predictive features— such as cumulative rainfall, seasonal averages, and upstream inflows— are generated. Both categorical data (like district codes) and numerical data (such as discharge measurements) are encoded in a way that makes them compatible with the decision tree model. Once trained, the C4.5 algorithm recursively partitions the dataset to identify patterns and thresholds that distinguish between flood intensities. Over time, as new rainfall and river data become available, the model retrains itself. This adaptive mechanism ensures that the system continuously improves, learning from past patterns and evolving climate signals. For disaster management agencies, this provides forecasts that are not only timely but also reliable and backed by clear reasoning. The proposed system not only provides accurate and interpretable flood forecasts but also ensures they are communicated effectively through interactive dashboards, GIS-based maps, and instant mobile alerts for high-risk communities. While beginning with Assam, Mumbai, and Chennai, it is designed to scale across flood-prone regions nationwide. By combining reliable data, a transparent decision tree model, and practical communication tools, the approach has the potential to reduce flood impacts, safeguard lives, and strengthen resilience in vulnerable areas of India.

Methodology

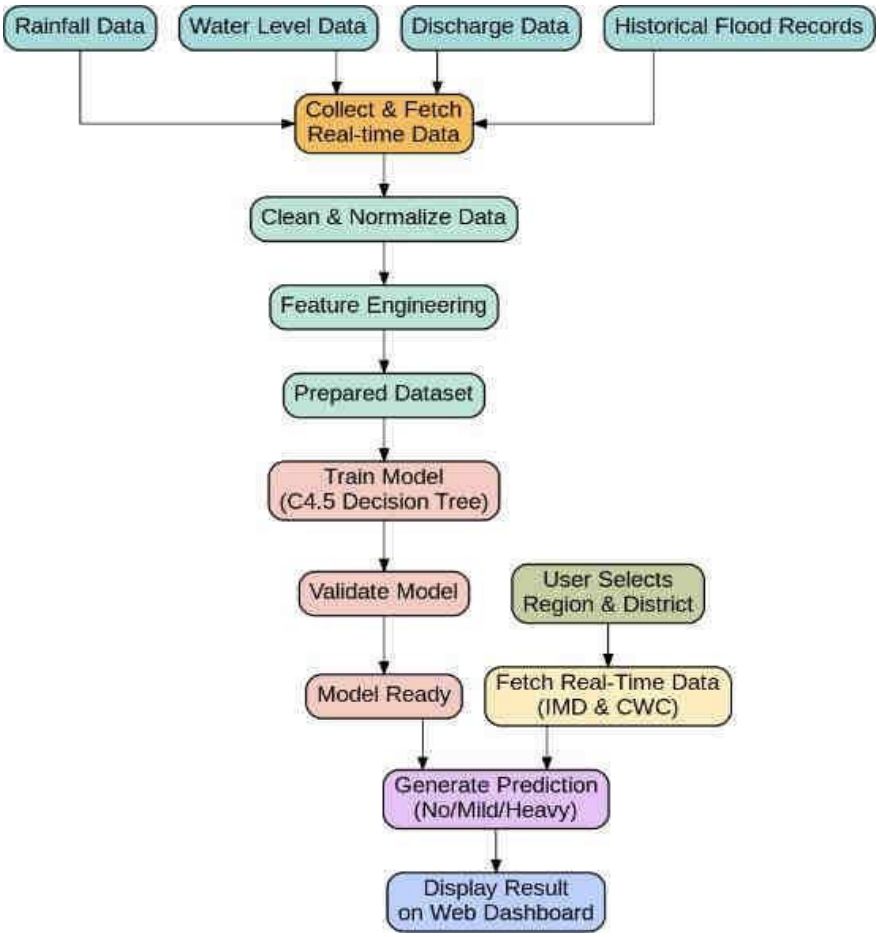

Fig. 1. Flow of Proposed Methodology

The research methodology for the AI-based flood prediction and detection system follows a stepbystep process, beginning with collecting the right data and moving through preparation, model building, and evaluation. The overall aim of this approach is to create a practical and reliable system that can help provide early flood warnings. The stages of the methodology are explained below.

A. Data Handling

Data Collection: Historical and real-time data were gathered from trusted sources. Rainfall data came from Kaggle and the India Meteorological Department (IMD), while river levels and discharge were obtained from the Central Water Commission (CWC) and India-WRIS. Geographic details such as latitude and longitude for flood-prone regions including Assam, Mumbai, and Chennai were also incorporated to enhance spatial analysis and regional prediction accuracy.

Data Pre-processing: A series of preprocessing steps were applied to ensure data quality and compatibility with the C4.5 decision tree model.

Missing Data Handling: Gaps in historical rainfall, river water level, and discharge datasets were addressed using interpolation and nearest neighbor methods to maintain consistency and reliability of input data.

Feature Engineering: Additional features relevant to flood prediction were created, including cumulative rainfall over specific time intervals, seasonal rainfall averages, upstream river inflow, and riverbed saturation levels. These engineered features help the C4.5 algorithm make more accurate and region-specific predictions.

Categorical Encoding: Location-specific features such as district or basin were encoded appropriately to allow the decision tree to handle both numerical and categorical data.

B. AI Model Selection

The C4.5 decision tree algorithm was chosen for its effectiveness in classification tasks, interpretability, and ability to handle both numerical and categorical features. The model uses rainfall, river water level, discharge, and region-specific parameters as input features. Recursive partitioning based on information gain is applied to split the dataset at each node, creating a tree structure that classifies flood

conditions into three categories: Flood, No Flood, and Mild Flood. The decision tree generates clear and interpretable rules, enabling disaster management authorities to understand the rationale behind predictions and take timely preventive actions. Separate models were developed for Assam, Mumbai, and Chennai to capture localized hydrological patterns and improve prediction accuracy for each region.

C. Training the Model

The dataset was split into training (70%), validation (15%), and test (15%) sets. The C4.5 model was trained iteratively using historical rainfall, water level, and discharge data, with parameter tuning for minimum leaf size and pruning thresholds to prevent overfitting. The model is periodically retrained with updated datasets to maintain accuracy and adaptability in flood predictions.

D. Evaluation and Performance Metrics

Accuracy: Measures the proportion of correct predictions (Flood / No Flood / Mild Flood) relative To the total number of predictions.

Precision and Recall: Evaluated to determine the model's effectiveness in identifying flood events while minimizing false alarms.

Interpretability of Decision Rules: Ensures that flood management authorities can understand and trust the predictions for actionable early warning.

Region-Specific Validation: Model performance is assessed separately for Assam, Mumbai, and Chennai to confirm localized effectiveness.

E. Technology Used

Backend Programming Language: Python

Framework: Flask for API integration and dashboard interface

Machine Learning Libraries: scikit-learn for C4.5 implementation, Pandas and NumPy for preprocessing

Prediction Flow**:** Historical rainfall, river discharge, and water level data are processed and fed into the C4.5 model to generate flood predictions with confidence scores

Frontend Framework: HTML, CSS, JavaScript with GIS mapping libraries (Leaflet.js)

User Interface: Web-based dashboard displaying region- specific flood maps, historical trends, and alerts

Expected Result

The expected outcome of this project is the development of a reliable and data-driven flood prediction system capable of generating medium-term forecasts (1–2 years ahead) for the selected regions by systematically analyzing historical rainfall patterns, river water levels, and discharge data. These predictions will enable the identification of high-risk flood-prone zones, allowing government agencies and disaster management authorities to implement timely preventive and mitigation strategies. By providing flood susceptibility assessments, the system will support informed planning of evacuation routes, reinforcement of embankments, and protection of critical infrastructure such as roads, bridges, hospitals, and power supply systems.

In addition, the system aims to improve community preparedness by delivering understandable and actionable warnings through visualization dashboards and alert mechanisms. This will empower local populations to take early protective measures, ensuring better safety for lives, properties, and livestock. The project also contributes to safeguarding agricultural lands by helping farmers prepare in advance, thereby minimizing crop damage and reducing long-term economic losses. Furthermore, the proposed framework is designed to be scalable and adaptable, allowing it to be extended to other flood-vulnerable regions with minimal modifications. With continued data integration and system enhancement, the model can evolve into a comprehensive decision-support tool for disaster resilience planning, strengthening both short-term response strategies and long-term climate adaptation efforts

Conclusion

The C4.5 decision tree algorithm demonstrates strong potential as a practical and reliable solution for flood prediction in highly vulnerable regions such as Assam, Mumbai, and Chennai, where flooding remains a recurring threat. By incorporating essential hydrological parameters such as rainfall, river water levels, and water discharge, the proposed model is capable of rapidly identifying flood risk levels and generating meaningful rule-based decisions. One of the major strengths of the C4.5 model lies in its interpretability, as it provides transparent decision paths that enable disaster management authorities to clearly understand the reasoning behind each prediction. This enhances trust in the system and allows officials to take confident and timely action during emergency situIn addition to its predictive capability, the model offers scalability and adaptability to different geographic regions and climatic patterns. With the inclusion of region-specific environmental factors, such as soil moisture levels, land- use patterns, reservoir status, and temperature

variations, the accuracy and reliability of the system can be further improved. The integration of real-time data from IoT devices, satellite observations, and meteorological systems can allow for continuous system updates and improved responsiveness to rapidly changing conditions. Furthermore, the use of visualization dashboards and alert-based interfaces can help communicate risk information clearly to both authorities and the public, improving situational awareness and community preparedness.

Although advanced machine learning and deep learning methods continue to evolve in the flood prediction domain, the proposed C4.5-based model offers a balanced solution by combining operational efficiency, simplicity, and analytical effectiveness. Its lower computational cost makes it suitable for deployment in resource-constrained environments, where complex models may be difficult to implement or maintain. With further enhancements and long-term data integration, this model can serve as a valuable decision- support tool in disaster management systems. Overall, the proposed approach contributes to building resilient infrastructure, improving early warning mechanisms, and reducing the socio-economic impact of floods through informed preparedness and proactive risk mitigation strategies.

References

A. Ullah, S. Haider, and R. Farooq, "Sensitivity analysis of a 2D flood inundation model: A case study of Tous Dam," Environmental Earth Sciences, vol. 83, no. 7, p. 213, 2024.

K. Aatif, M. A. Fahiem, and F. Tahir, "Forecasting floods using deep learning models: A longitudinal case study of Chenab River, Pakistan," IEEE Access, vol. 12, pp. 115802–115819, 2024.

Hazarika, I., Khalfan, J., Ahmed, M., Yousif, A., & Hussain, J. (2024). *Role of fintech as an enabler to fulfill HR requirements and attain sustainability.* In A. Hamdan & A. Harraf (Eds.), *Business development via AI and digitalization* (Vol. 537, pp. 59–69). Springer. https://doi.org/10.1007/978-3-031-62106-2_5

J. Nanditha, A. P. Kushwaha, R. Singh, I. Malik, H. Solanki, D. S. Chuphal, S. Dangar, S. S. Mahto, U. Vegad, and V. Mishra, "The Pakistan flood of August 2022: Causes and implications," Earth's Future, vol. 11, no. 3, p. e2022EF003230, 2023.

M. Sajjad, Z. Ali, and M. Waleed, "Has Pakistan learned from disasters over the decades? Dynamic resilience insights based on catastrophe progression and geo-information models," Natural Hazards, vol. 117, no. 3, pp. 3021–3042, 2023.

H. B. Waseem and I. A. Rana, "Floods in Pakistan: A state-of-the-art review," Natural Hazards Research, vol. 3, no. 3, pp. 359–373, 2023.

H. Hamidifar and M. Nones, "Spatiotemporal variations of riverine flood fatalities: 70 years global to regional perspective," River, vol. 2, no. 2, pp. 222–238, 2023.

C. Prakash, A. Barthwal, and D. Acharya, "FLOODWALL: A real-time flash flood monitoring and forecasting system using IoT," IEEE Sensors Journal, vol. 23, no. 1, pp. 787–799, 2023.

Jumde, A., Hazarika, I., & Akre, V. (2023). *Challenges and opportunities in integrating rapidly changing technologies in business curriculum*. In *Proceedings of the 2023 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)* (pp. 203–208). IEEE. https://doi.org/10.1109/ICCIKE58312.2023.10131683

Sharma, B. (2023). *Impact of artificial intelligence on the legal industry: Advantages, challenges, and ethical implications.* BioGecko, 12(2), 3363–3374.

Z. Manzoor, M. Ehsan, M. B. Khan, A. Manzoor, M. M. Akhter, M. T. Sohail, A. Hussain, A. Shafi, T. Abu-Alam, and M. Abioui, "Floods and flood management and its socio- economic impact on Pakistan: A review of the empirical literature," Frontiers in Environmental Science, vol. 10, p. 1021862, 2022.

M. Moishin, R. C. Deo, R. Prasad, N. Raj, and S. Abdulla, "Designing deep-based learning flood forecast model with ConvLSTM hybrid algorithm," IEEE Access, vol. 9, pp. 50982–50993, 2021.

M. A. U. R. Tariq, R. Farooq, and N. Van de Giesen, "A critical review of flood risk management and the selection of suitable measures," Applied Sciences, vol. 10, no. 23, p. 8752, 2020.

S. Puttinaovarat and P. Horkaew, "Flood forecasting system based on integrated big and crowdsource data by using machine learning techniques," IEEE Access, vol. 8, pp. 5885–5905, 2020.

M. Khan et al., "Application of MLP models to study flash floods using soil flux and CO_2 for

reducing false alarms," in Proc. ICSSA, pp. 130–134, 2018.

S. Mojaddadi et al., "Flood risk assessment using ensemble machine learning and GIS with multisensor remote sensing data," Geomatics, Natural Hazards and Risk, vol. 8, no. 2, pp. 1080–1102, 2017.

R. Mousa, Y. Zhang, and C. Claudel, "Flash flood detection in urban areas using ultrasonic and infrared sensor technologies," Sensors Journal, vol. 16, no. 19, pp. 7204– 7216, 2016.

Hazarika, I., Saoji, S., Bhandari, R. B., Jorvekar, G., Rao, P. H., & Porwal, T. (2025). *Mapping resilience pathways: A conceptual framework for portfolio risk management in microenterprise lending during economic shocks.* Enterprise Development and Microfinance, 35(1), 1–20.

Sharma, B. (2025). *Ethical and AI concerns in data privacy: A charismatic dilemma.* International Journal of Multidisciplinary Research and Development, 12(7), 18–32.

P. Mitra et al., "Flood forecasting using Internet of Things and artificial neural networks," in Proc. 2016 IEEE 7th Annual International Conference on IEMCON, pp. 1–5, 2016.