PAPER NAME

**paper Flood detection or prediction using machine learning test**

WORD COUNT

**6119 Words**

CHARACTER COUNT

**38743 Characters**

PAGE COUNT

**19 Pages**

FILE SIZE

**408.7KB**

SUBMISSION DATE

**Apr 9, 2025 11:39 AM UTC**

REPORT DATE

**Apr 9, 2025 11:41 AM UTC**

## 8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 4% Internet database
- 4% Publications database
- Crossref database
- Crossref Posted Content database
- 5% Submitted Works database

## Excluded from Similarity Report

- Bibliographic material
- Quoted material

# AI-Driven Flood Risk Assessment and Forecasting with Remote Sensing Data

Author Name [1*] , Author Name[2]

[1]*Designation, Department Name, Collage Name, City*

[2]*Designation, Department Name, Collage Name, City*

## Abstract

**Floods are among the most devastating disasters, presenting significant dangers to human life, property, and infrastructure. Accurate and prompt flood detection and forecasting are essential for disaster preparation, early warning systems, as well as risk management. Traditional methods often depend on both physical & statistical models, which may be constrained by data complexity and fluctuating environmental circumstances. A powerful substitute in recent years, "machine learning (ML)" is skilled in spotting intricate, nonlinear patterns in large datasets. This paper examines the use of many machine learning approaches, such as Decision Trees, Random Forests, "Support Vector Machines", Neural Networks, "Gradient Boosting", & Logistic Regression, in flood prediction and detection. It emphasises the data sources, preprocessing techniques, assessment measures, as well as integration with "Internet of Things" (IoT) as well as edge computing technologies. The research also examines the strengths, limitations, as well as practical applicability of various models. The study continues by highlighting future research avenues, underscoring the promise of hybrid models, deep learning, as well as explainable AI in improving both the reliability and accuracy of flood prediction systems.**

***Keywords:*** *Flood Prediction, Flood Detection, Flood Forecasting, Machine Learning, Early Warning System, Data-Driven Models, Rainfall Forecasting, Disaster Management, Hydrology.*

## I. INTRODUCTION

Flooding is a highly devastating natural event, resulting in a huge loss of life and property. Conventional flood forecasting techniques often depend on intricate and time-consuming hydraulic and hydrological models. The advent of "machine learning (ML)" has revolutionised this field by providing improved efficiency and accuracy in predictive capabilities. Machine learning algorithms can analyse large datasets from several sources, including historical flood records, real-time meteorological data, as well as satellite imagery, to forecast flood events with higher accuracy. [1]

Flood prediction systems have been significantly improved in recent years by combining machine learning with remote sensing as well as Geographic Information Systems (GIS). These tools provide geographical analysis as well as visualisation of flood-prone areas, aiding authorities in disseminating early warnings and organising disaster responses. Methods include "Artificial Neural Networks" (ANN), "Support Vector Machines" (SVM), as well as Random Forests have shown significant accuracy in forecasting flood occurrences when trained on high-quality datasets. Furthermore, real-time data collection via Internet of Things (IoT) sensors and meteorological stations enhances the dynamic updating of models, hence augmenting the timeliness as well as reliability of predictions. [2]

In the whole world India has the greatest yearly risk of floods. Water logging usually occurs in low-lying regions of major cities. In these places, flood forecasting is very important. In recent years, several areas, including "Assam, Bihar, Goa, Orissa, Pune, Maharashtra, Tamil Nadu, Karnataka, Kerala, and Gujarat," have experienced flooding.

Chennai saw 1049 millimetres (mm) of rain in November 2015. The precipitation total of 1088 mm was the highest since 1918. The Kanchipuram district receives an average of 64 cm of rainfall between October and December. It had the highest amount of precipitation, 181.5 cm, which is 183% higher than normal. Although 146 cm of rain was reported, the Tiruvallur area typically receives 59 cm.

Despite extensive research on flood prediction, few methodologies provide accurate forecasts. When analysing flood predictions, machine learning is frequently used. Machine learning provides many ways for enhanced problem prediction. This study proposes estimating flash floods to avoid flood-prone areas. Establishing the ML algorithm model is the tactic. For more precise short-term projections in metropolitan areas, it takes the flood component into account. Information retrieval may take hours, depending on the data transfer mode. Nevertheless, streams of photos and videos may provide valuable information for a range of uses. An early warning system for managing and avoiding flooding is one use for the visual sensing approach. Visual sensing systems rely on image processing, which involves using algorithms to derive useful information from digital images. [3]

## II. RELATED WORK

[4] Frequent emergencies are a major issue for mankind. The world has experienced countless natural as well as man-made disasters. More people have been impacted by floods than any other disaster in the 21st century, according to a new analysis. Africa was the second most vulnerable continent after Asia, with 43% of disasters in 2019. Any nation's leadership, especially in flood-prone areas, must manage floods to preserve the environment and the population. Predictive machine learning and analytics may improve risk management. Since flood risk cannot be eradicated, scientific mitigation is essential. The research recommends using machine learning to identify and forecast pluvial floods. Fuzzy rule classification will be used to assess the machine learning approach on pluvial flood conditioning factors.

[5] Climate change causes complex global floods. Certain gauging stations predict floods, although their accuracy is low. Unexpected floods damage lives and infrastructure. Our study is focused on developing a deep learning-based system that can accurately identify and monitor floods in real-time. This study recommends wireless sensor networking for flood detection due to its reliability, low power, and large area connectivity. Our convolutional neural networks also discover flood-affected organisms.

[6] Floods, one of nature's worst calamities, destroy infrastructure, crops, lives, and the economy. Flood forecasting and catastrophe management have been extensively studied. Floods are hard to forecast in real time. Predicting water levels and velocities over a large area requires data and computationally expensive flood propagation models. This study uses numerous machine learning models to forecast floods and provide policy suggestions to reduce their risk. This study will use Decision Tree Classifier, "Binary Logistic Regression", K-"Nearest Neighbour (KNN)", and Support Vector Classifier (SVC) to make accurate predictions. A comparative study will use the findings to evaluate whether model is more accurate.

[7] Floods are one of worst natural calamities, destroying infrastructure, crops, people, and the economy. Flood forecasting and catastrophe management have been extensively studied. Floods are hard to forecast in real time. Predicting water levels and velocities over a large area requires computationally complex flood propagation models and data. In order to lessen the likelihood of this natural catastrophe and provide legislative remedies, this study forecasts floods using several machine learning techniques. "Decision Tree Classifier", "Support Vector Classifier" (the SVC), K-Nearest Neighbour (KNN), and Binary Logistic Regression will all be used in our study to make correct predictions. In order to determine which model is more correct, the results will be compared.

[8] Flooding, a complicated phenomenon that happens all across the world, is the ultimate result of climate change. While certain gauging stations are employed to predict the probability of floods, their accuracy is low. In addition to killing people, unexpected floods are damaging important infrastructure. Our study's overarching goal is to design a deep learning-based system for accurate, real-time flood warning and monitoring. This paper proposes a method for flood detection using "wireless sensor networking technology" that is dependable, uses little power, and communicates over large areas. We also search for any live organisms affected by the flood using a convolutional neural network.

[9] Floods are our century's most catastrophic disaster. Unreliable flood predictions has caused severe infrastructure and human losses. This reinforces the need for flood prediction. This project aims to produce the best flood model. AI calculations and a comprehensive, effective, as well as accurate flood expectation framework will enable residents and the government obtain the relief they need. Thus, the Decision Tree Model is being created. This method performs dataset calculations with varying precision. The software forecasts floods using AI and notifies local and governmental institutions via Android. Gradient Boost, Random Forest, as well as Decision Tree are being compared. This approach seeks to improve prediction by using more complicated data and a sophisticated algorithm.

[10] Floods, one of the greatest natural calamities, must be dealt swiftly. Satellite imaging monitors and assesses flood damage. By evaluating remote sensing data, effective and scalable machine learning and deep learning may improve flood control and detection. Floods are predicted using Sentinel-1 SAR data and Random Forest (RF) and "Histogram-based Gradient Boosting Decision Tree" machine learning algorithms. HHO and ACO metaheuristics optimised hyperparameters. To improve model performance and flooded pixel identification, the pre-trained VGG-16 "Neural Network" was employed as the deep feature extractor. After that, the "four ensemble flood detection models" (RF-HHO, RF-ACO, HGBDT-HHO, and HGBDT-ACO) were compared using statistical measures. The "four ensemble flood detection models" were all successfully validated and tested. Validation accuracy for the models was over 95%, while testing accuracy was above 97%. Top model HGBDT-ACO identified flood pixels with the greatest accuracy as well as lowest error rate. HGBDT models did better than RF models because they learnt quicker and performed similarly. Thus, they were more efficient and computationally complex.

[11] Flooding happens when a water body surges above its capacity, engulfing dry land. Heavy rain and runoff are the main causes of floods, which immerse neighbouring land regions in water and destroy buildings, bridges, power supply, transportation, and the economy. Sensor technologies and active parameter monitoring have been used to predict flood warnings for years. Many future-use data sets were created. Data analytics methods have emerged as a result of the renaissance of machine learning and the concept of intelligent machines. These approaches enable computers to learn from data and create a model equation that can be used to forecast future occurrences. To maximise the utility of the datasets we'll be using to train a Decision Tree-based weak prediction model, we propose a Flood Detection approach that makes use of the "Gradient Boost Algorithm" for data type and regression. To prevent the issue, the results could be presented to the appropriate authorities. This method produces accurate predictions and was developed specifically for these kinds of uses. Accurate datasets for training models are provided by means of sensor and remote sensing technologies.

[12] Flooding is water spilling onto dry land or a considerable rise in water that impacts human existence. As one of the most prevalent natural events, it causes major financial issues for assets and products and affects human lives. Residents would benefit from preventing such floods so they may flee flood-prone areas before they happen. Academics have proposed several flood solutions, including prediction models and infrastructure. All recommended economic cures fail in Somalia. In order to anticipate the humanitarian consequences of flooding, this project seeks to develop a powerful "real-time flood detection system" using machine learning algorithms such as Random-Forest, Naive-Bayes, and J48. The studies conducted using this strategy will address the fourth issue and demonstrate how a combination of Arduino as well as GSM modems might imitate a one-of-a-kind system for monitoring water levels. Classification accuracy for Random Forest is 98.7 percent, while that of J48 and Naive Bayes are 88.1 and 84.2 percent, respectively, according to the study. The recommended flood prevention method enhances research by using artificial intelligence and data mining.

[13] Besides damaging lifelines and infrastructure, urban flooding endangers people and autos. To understand how much flooding an area may face at any one time, cities and local authorities need

faster flood detection techniques. Thus, flood control orders are issued quickly, allowing individuals as well as motorists to avoid flooded areas. In order to create and train "machine-learning models" for the flood detection in San Diego, CA, USA, this study makes use of remotely sensed satellite photographs and records of road closures maintained by the police department. To identify flooded pixels in photos and assess their efficacy, Sentinel 1 flood data is input into supervised and unsupervised machine learning models such as Maximum Likelihood Classifier (MLC), "Support Vector Machine" (SVM), as well as Random Forest (RF). We also build an unsupervised ML system that uses iso-clustering, fuzzy rules, and the Otsu algorithm to identify urban floods. Change detection (CD) is used. Accuracy measurements were 0.69, 0.87, 0.83, and 0.87, while F1-scores were 0.67, 0.85, 0.79, as well as 0.85, and recall values were 0.9, 0.85, & 0.9, according to the performance assessment of the RF, the SVM, MLC, as well as CD models. Precision values were 0.53, 0.85, 0.75, as well as 0.81, respectively. The novel unsupervised flood picture classification and detection method uses less data and processing time to get higher results, allowing for speedier flood mapping. If other flood-prone cities adopt this methodological approach, they may be able to use satellite photos to better identify nuisance floods and mitigate the risks associated with building and planning for urban infrastructure and transportation.

[14] Flooding is the worst land calamity. Flooding will worsen during storms. Risk reduction, preparation, saving lives, and property damage reduction are all goals of the research into flood prediction models. Artificial intelligence has improved the consistency and execution of financial planning and forecasting in the last two years. This activity may prioritise all participants' sentiments. Flood warning, mitigation, and forecast need AI models. Machine learning techniques are widely used since they need a lot of computing power for little data. Representative vectors with the highest scores may come from a tiny data set. The precise and high-scoring tree won. In order to compare and evaluate different approaches, this flood forecast makes use of machine learning techniques such as logistic regression and decision trees. Logistic regression may beat other methods in accuracy, efficiency, and improvement. One of the worst natural disasters, flooding may inflict lifelong devastation and misery. Farmers are the saddest people when their hard effort fails. A safe exposure model is required to monitor water level and velocity over a vast region. These models enhance projections in numerous ways. Additionally, these models properly predict floods in a particular year but don't provide detailed answers.

## III. THEORETICAL BACKGROUND

### A. Limitations of Traditional Flood Prediction Models

Traditional flood prediction techniques primarily depend on hydrological as well as hydraulic models that replicate "natural water systems" using physical as well as mathematical equations. Parameters like "rainfall-runoff relationships," infiltration rates, "terrain topography", as well as "river geometry" are crucial to models like the Soil and Water Assessment Tool (SWAT) and the Hydrologic Engineering Center's River Analysis System (HEC-RAS). Although successful in several organised settings, they need substantial domain-specific expertise, manual adjustment, and significant computing resources.

A significant limitation is their restricted adaptability in managing high-dimensional & non-linear data, typical of dynamic environmental systems. Moreover, their reliance on fixed thresholds as

well as assumptions often results in diminished predictive accuracy amid swiftly changing climatic conditions, urban expansion, and unpredictable hydrometeorological occurrences. These models also encounter difficulties in real-time applications because to delay in data processing and their incapacity to adapt to changing data streams. The need for scalable and accurate flood forecasts has led to a critical shortage of adaptive, data-driven approaches.

### B. "Emergence of Machine Learning in Flood Prediction"

Machine learning (ML) has arisen as a revolutionary substitute for conventional flood prediction by using historical and current information to reveal intricate correlations among many factors. In contrast to physical models, machine learning algorithms may autonomously discern patterns and connections from extensive datasets without requiring intricate physical representations of the system.

Accurate and timely flood warnings can be generated by integrating diverse and heterogeneous features, including rainfall intensity, rivers water levels, land use data, soil saturation, humidity, as well as weather forecasts. Machine learning models that fall into this category include "Support Vector Machines" (SVM), Decision Trees, "Random Forests", Neural Networks, along with Gradient Boosting algorithms. These models are especially advantageous for non-linear & high-dimensional datasets, making them suitable for flood-prone areas where several interdependent factors affect results.

Furthermore, developments in sensor networks & remote sensing have facilitated more accessible real-time data collecting, enabling machine learning systems to perpetually update and enhance predictions. The amalgamation of machine learning with cloud computing and the Internet of Things significantly augments its ability for real-time flood monitoring as well as decision-making assistance systems.

### C. Role of HistGradientBoostingClassifier

This review focusses on the "Histogram-based Gradient Boosting Classifier (HistGradientBoostingClassifier)" as the principal model for flood detection and prediction. This classifier enhances the conventional gradient boosting framework by including a histogram-based binning method, which significantly decreases memory consumption and expedites computation. The model converts continuous numerical characteristics into discrete bins before to fitting the weak learners (decision trees), facilitating the effective management of large-scale tabular information.

HistGradientBoostingClassifier provides robust regularisation features and inherently accommodates missing values, making it effective in noisy or incomplete datasets often encountered in environmental monitoring. It also addresses unbalanced classification challenges—such as identifying infrequent flood occurrences—more efficiently than other baseline models, owing to its sophisticated splitting techniques and emphasis on difficult-to-predict instances during boosting rounds.

By using hyperparameter adjustment (such as learning rate, number of bins, and maximum depth), the model attains enhanced accuracy, recall, as well as precision in flood classification tasks. Its capacity to encapsulate intricate, non-linear interactions among environmental factors and function well in real-time settings establishes it as a formidable instrument in advanced flood forecasting systems.

## IV. METHODOLGY

The methodology chapter outlines the methodological technique used to achieve the study goals of flood detection and prediction using machine learning. This section establishes the basis for understanding the collection, processing, analysis, as well as modelling of the data to provide precise and timely flood predictions. The intricacy of flood occurrences, shaped by several dynamic parameters like the rainfall intensity, "river water levels", soil saturation, as well as climatic conditions, renders a data-driven method using machine learning a viable alternative to traditional hydrological models.

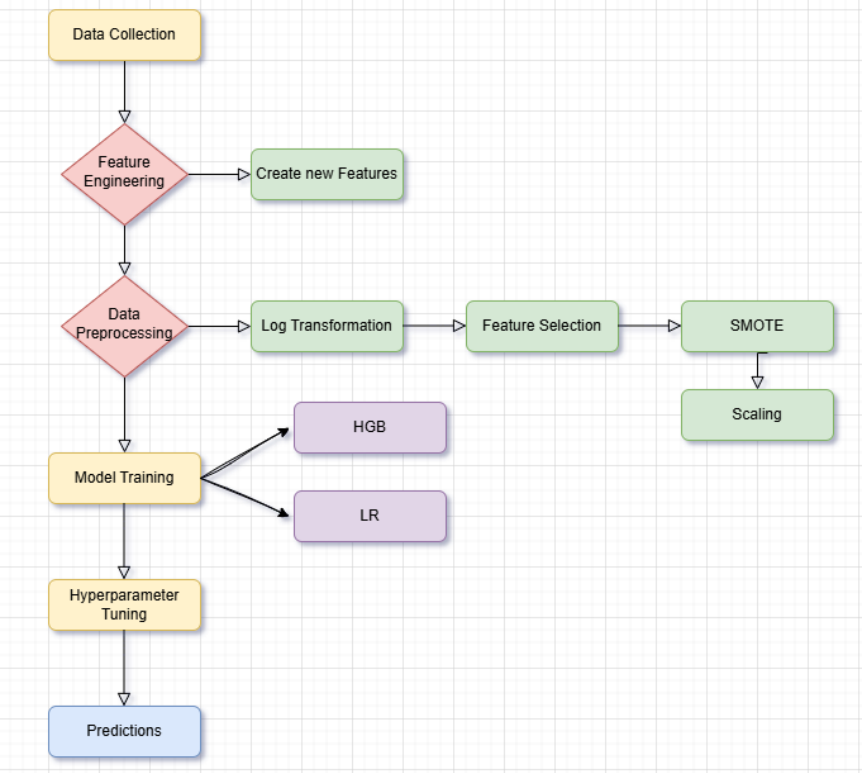


Figure 0-1 Proposed Model Flow

The proposed model for real-time flood prediction is a "Histogram-based Gradient Boosting Classifier (HistGradientBoostingClassifier)". This model is selected for its enhanced accuracy, expedited training time, and capacity to manage intricate, non-linear interactions among variables, which are essential for dependable flood forecasting. Gradient Boosting is the ensemble learning technique that constructs a robust classifier by amalgamating the outputs of several weak learners, usually decision trees. The histogram-based alternative enhances this process by increasing processing performance and managing massive datasets more efficiently.

The study proposes the "Histogram-based Gradient Boosting Classifier (HistGradientBoostingClassifier)" for flood prediction, owing to its efficiency, accuracy, and capability to manage enormous datasets. This section explores the theoretical basis,

implementation procedures, rationale for selection, benefits, and its particular use in flood forecasting.

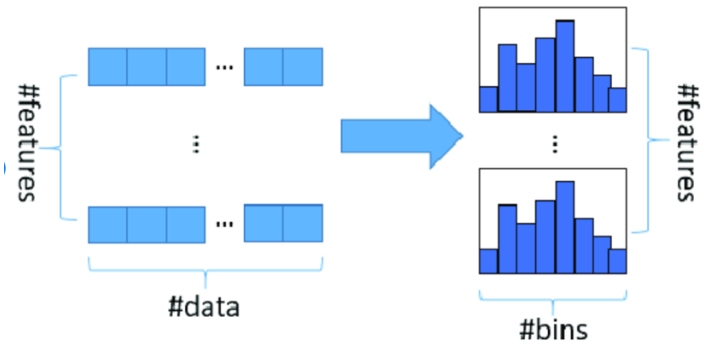


Figure 0-2 F1 scores bar graph Histogram-based Gradient Boosting model

### *A. Data Collection*

The effectiveness of "machine learning-based flood prediction systems" is mostly contingent on the quality, diversity, and “comprehensiveness” of the used datasets. This work employs two unique but complementary datasets sourced from open platforms, that together augment the training, validation, as well as testing of robust predictive models. Despite originating from different sources, both databases provide significant meteorological as well as hydrological information vital to regions susceptible to flooding in India.

For some places in Bangladesh, the data covers the years 1948–2013 and contains monthly averages for things like high temperature, low humidity, wind speed, temperature, cloud cover, and bright sunshine. Includes X and Y coordinates, as well as height and longitude, as well as the weather station numbers.

Having access to high-quality data that is also contextually relevant is crucial for developing a robust and dependable flood prediction model. This research gathered a comprehensive dataset from publicly accessible meteorological and hydrological information, including several parameters that affect flood episodes. These records include many decades and reflect consistent data gathered from several weather monitoring stations in South Asian locations that exhibit similar meteorological and monsoonal traits with flood-prone areas within India.

The collection combines historical atmospheric data with real-time hydrological metrics, offering a comprehensive perspective on the environmental factors that contribute to flood events. This integration facilitates the examination of long-term climate patterns, seasonal precipitation dynamics, and the development of models that can provide precise and prompt flood alerts.

Table 1 Integrated Historical and Real-Time Meteorological Dataset for Flood Prediction

“

| Name | Unit |
|---|---|
| Station Names | |

| | |
|---|---|
| Year | |
| Month | |
| Max Temp | Celsius |
| Min Temp | Celsius |
| Rainfall | cm |
| Relative Humidity | Percentage |
| Wind Speed | meters per second |
| Cloud Coverage | Okta |
| Bright Sunshine | Hours per day |
| Station number | |
| X Coordinate | |
| Y Coordinate | |
| Latitude | |
| Longitude | |
| Altitude | meter |
| Period | |
| Flood? | 1(Yes) |

.”

The presence of the binary target variable denoting flood status facilitates the use of supervised machine learning methods, making the dataset suitable for classification-oriented predictive modelling. The inclusion of real-time data, including river water levels and soil moisture, makes the model better at predicting when floods will happen.

| | Station Na | YEAR | Month | Max Temp | Min Temp | Rainfall | Relative Hu | Wind Spee | Cloud Cov | Bright Sun | Station Nu | X_COR | Y_COR | LATITUDE | LONGITUD | ALT | Period |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Barisal | 1949 | 1 | 29.4 | 12.3 | 0 | 68 | 0.453704 | 0.6 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1949.01 |
| 1 | Barisal | 1950 | 1 | 30 | 14.1 | 0 | 77 | 0.453704 | 0.8 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1950.01 |
| 2 | Barisal | 1951 | 1 | 28.2 | 12.3 | 0 | 77 | 0.453704 | 0.6 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1951.01 |
| 3 | Barisal | 1952 | 1 | 26.6 | 12.3 | 2 | 77 | 0.453704 | 1 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1952.01 |
| 4 | Barisal | 1953 | 1 | 30 | 13.3 | 10 | 75 | 0.453704 | 1.6 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1953.01 |
| 5 | Barisal | 1954 | 1 | 27.8 | 12.7 | 0 | 72 | 0.453704 | 0.5 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1954.01 |
| 6 | Barisal | 1955 | 1 | 26.6 | 12.3 | 2 | 77 | 0.453704 | 1 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1955.01 |
| 7 | Barisal | 1956 | 1 | 29.4 | 14.3 | 17 | 74 | 0.453704 | 0.4 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1956.01 |
| 8 | Barisal | 1957 | 1 | 30.1 | 15.1 | 104 | 80 | 0.453704 | 1 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1957.01 |
| 9 | Barisal | 1958 | 1 | 31.1 | 15.5 | 0 | 77 | 0.453704 | 1.7 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1958.01 |
| 10 | Barisal | 1959 | 1 | 29.9 | 14.5 | 24 | 77 | 0.453704 | 1.4 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1959.01 |
| 11 | Barisal | 1960 | 1 | 27.3 | 12.8 | 0 | 71 | 0.3 | 1.3 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1960.01 |
| 12 | Barisal | 1961 | 1 | 28.8 | 14.5 | 8 | 73 | 0.2 | 1.7 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1961.01 |
| 13 | Barisal | 1962 | 1 | 28.3 | 11.6 | 0 | 74 | 0.3 | 0.5 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1962.01 |
| 14 | Barisal | 1963 | 1 | 27.7 | 12.2 | 0 | 67 | 0.2 | 0.5 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1963.01 |
| 15 | Barisal | 1964 | 1 | 26.6 | 12.3 | 0 | 77 | 0.4 | 0.9 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1964.01 |
| 16 | Barisal | 1965 | 1 | 26.6 | 12.3 | 0 | 80 | 0.1 | 1.5 | 7.831915 | 41950 | 536809.8 | 510151.9 | 22.7 | 90.36 | 4 | 1965.01 |

**Figure 0-3 Data Sample**

The dataset was preprocessed to address issues such as missing values, inconsistent formats, and skewed distributions prior to model building. To improve data quality and model efficacy, methods such as power translation, standardisation, and feature engineering—including polynomial features, interaction terms, and temperature range—were used.

The dataset's richness and variety facilitate the development of time series forecasting models as well as classification algorithms. It establishes a robust basis for developing precise, dependable,

as well as real-time flood prediction systems which may substantially enhance early warning efforts as well as disaster risk reduction initiatives.

This unified dataset enables both "time series forecasting models and categorisation methods". It allows the creation of prediction models which are both accurate and suitable for real-time application in flood-prone areas, hence improving disaster preparedness and risk mitigation strategies.

### *B. "Data Pre-Processing"*

The "flood prediction pipeline" would be incomplete without data pre-processing, which cleans, organises, and otherwise prepares the raw data for training accurate and efficient machine learning models. The relevance and accuracy of the input data greatly determine the predictive system's performance. This project, which focusses on flood prediction using meteorological as well as environmental information, included extensive pre-processing to manage missing values, normalise data, create meaningful features, and rectify class imbalance.

**Handling Missing Values and Data Formatting:** The initial dataset, including rainfall and related meteorological data, often exhibits missing or inconsistent entries attributable to sensor malfunctions or lapses in data collection. To resolve this issue, missing data were imputed by statistical methods such mean substitution, therefore preserving critical information. Additionally, the 'YEAR' column was converted into a standardised datetime format with Python's pandas module to enable time series operations and maintain chronological consistency in modelling.

**Feature Construction and Transformation:** To increase the prediction capability of the "machine learning model", new mathematical features were derived from the current dataset. This included creating derived features that depict monthly as well as seasonal rainfall trends, cumulative rainfall at designated periods, and changes in precipitation patterns. These designed attributes enhance the model's understanding of temporal relationships and environmental changes. Furthermore, the Yeo-Johnson approach was used for power transformation to stabilise variance and normalise the distribution of the skewed data. This modification is very beneficial for machine learning models that are sensitive to data distribution, such gradient boosting as well as vector machines. It guarantees which each feature contributes proportionately throughout the training process.

**Data Standardization**: To guarantee that all input characteristics contribute uniformly to the model, data standardisation was executed with StandardScaler from Scikit-learn. Standardisation normalises features to possess a mean of zero and a variance of one, hence avoiding features with higher numeric ranges from overshadowing model training. This is particularly significant for distance-based or gradient-based models, including SVMs and gradient boosting classifiers.

**Feature Selection:** After the creation of more features, the dataset's dimensionality increased, potentially leading to multicollinearity and elevated computational costs. Feature selection was performed using the "ANOVA F-test (SelectKBest) method." This statistical strategy evaluates the association between each parameter as well as the desired variable (flood condition) to discover those with the highest predictive power. The eight most significant features were retained

according to the F-scores. The variables encompassed precipitation, soil moisture, river water levels, humidity, temperature range, and engineering aspects, including the correlation between precipitation and temperature. The removal of unnecessary features improved model performance and reduced overfitting.

**Imbalance Handling:** Flood prediction datasets often demonstrate imbalance, with flood events (label = 1) much less common than non-flood occurrences (label = 0). This discrepancy may result in the model favouring the predominant class, hence diminishing its sensitivity to genuine flood occurrences. The "Synthetic Minority Oversampling Technique" (SMOTE) was used to address this problem. By extrapolating from current samples, SMOTE creates synthetic instances of the minority class, bringing the dataset into balance without sacrificing any data. Improving the model's memory & accuracy for high-risk predictions, this phase was critical in strengthening its capacity to identify real flood occurrences.

**Train-Test Split:** The dataset was partitioned into training and testing sets to facilitate model evaluation. Evaluating the model's generalisability and practical usefulness is made easier by training it on a subset of the data and then testing it on new data. The combination of these pre-processing approaches produced a high-quality dataset suitable for training sophisticated models such as the “Histogram-based Gradient Boosting Classifier”, facilitating dependable and accurate flood prediction.

### *C. Model Building and Hyperparameter Tuning*

The model building stage is crucial to a flood prediction system's capacity to identify trends in the past and forecast future flood events. Several machine learning models were investigated and put into practice in this research to assess how well they predicted flood events. The models were chosen for their ability to classify data, their capacity to withstand noise, and their appropriateness for tabular and unbalanced data.

**Model Selection:** A number of machine learning classification models, such as "Support Vector Machine (SVM), Multilayer Perceptron Classifier (MLPClassifier), AdaBoost Classifier, SGDClassifier, Histogram-based Gradient Boosting Classifier (HistGradientBoostingClassifier), and Logistic Regression," were taken into consideration for the prediction and detection of flood events. When it comes to categorisation jobs, each of these models offers distinct advantages.

Table 2 Machine Learning Models Used for Flood Prediction

“

| Classifier | Description | Use Case |
|---|---|---|
| **SVM (Support Vector Machine)** | Effective in managing high-dimensional spaces and non-linear boundaries using kernel functions. | Classification tasks with complex boundaries. |
| **MLPClassifier** | A type of feedforward neural network suitable for capturing complex, non-linear relationships. | Analyzing flood data and other complex datasets. |

| | | |
|---|---|---|
| **AdaBoost** | An ensemble learning algorithm that improves model performance by combining multiple weak classifiers. | Enhancing accuracy in various classification tasks. |
| **SGDClassifier** | Efficient for large-scale data and online learning, especially with sparse datasets. | Real-time predictions and large datasets. |
| **HistGradientBoostingClassifier** | A faster variant of traditional gradient boosting that discretizes continuous features into histograms. | Reducing memory consumption while maintaining accuracy. |
| **Logistic Regression** | Serves as a baseline model, useful for binary classification tasks. | Binary classification, e.g., flood vs. no flood events. |

."

**Hyperparameter tweaking:** GridSearchCV was used to enhance performance of every "machine learning model". This extensive search approach systematically assesses hyperparameter combinations using cross-validation, determining the optimal configuration for enhanced accuracy and generalisability.

**Table 3 The tuned hyperparameters for each model**

"

| **Model** | **Tuned Hyperparameters** |
|---|---|
| **SVM** | Kernel type (linear, RBF), C (regularization), Gamma |
| **MLPClassifier** | Hidden layer sizes, Activation function (relu, tanh), Learning rate |
| **AdaBoost** | Number of estimators, Learning rate |
| **SGDClassifier** | Loss function (hinge, log), Penalty (l1, l2), Learning rate |
| **HistGradientBoostingClassifier** | Learning rate, Maximum depth, Number of bins |
| **Logistic Regression** | Solver (liblinear, saga), Penalty (l2, none), Regularization strength (C) |

."

**Proposed Model: HistGradientBoostingClassifier**

The HistGradientBoostingClassifier is the efficient gradient boosting solution designed for tabular data. It functions by partitioning continuous input data into discrete bins (histograms), so considerably expediting the training process and reducing memory use. Unlike traditional gradient boosting, which creates trees with precise thresholds for continuous variables, this method use histograms to approximate splits, enabling greater scalability with larger datasets and increased speed without significant loss of accuracy.

The HistGradientBoostingClassifier is the complex ensemble learning method that improves upon the traditional Gradient Boosting framework. By merging many decision trees, which are poor learners, it builds robust prediction models by learning from past mistakes and improving upon

them. The histogram-based variant is distinguished by its novel method of transforming continuous data into discrete histogram bins. This alteration dramatically accelerates model training and reduces memory use without sacrificing performance.

Transforming "continuous input variables" into discrete bins simplifies the identification of splits during tree construction, hence accelerating computation. Despite this optimisation, it maintains the flexibility and capability to depict complex, non-linear interactions. Its tolerance for missing values as well as categorical variables enhances its applicability in actual flood prediction scenarios where data inaccuracies are prevalent.

The objective of using the HistGradientBoostingClassifier in flood prediction is to develop a high-performing and efficient model capable of reliably classifying flood-prone instances by identifying complicated, non-linear patterns within extensive and unbalanced meteorological datasets. This approach improves prediction speed and memory efficiency by transforming continuous data into discrete histograms, it is perfect for use in real-time scenarios that need for accurate and quick flood detection.

## V. “RESULTS AND DISCUSSION”

This section outlines the outcomes obtained from the installation and assessment of various "machine learning models" for flood prediction by employing meteorological data. The models were trained on a preprocessed dataset with sophisticated approaches, including "feature selection (ANOVA F-test), dimensionality reduction (PCA), class balancing (SMOTE), power transformation (Yeo-Johnson), and standardisation" to optimise model performance.

Among the trained models, including "SVM, MLPClassifier, SGDClassifier, AdaBoost, Logistic Regression, as well as HistGradientBoostingClassifier, the Histogram-based Gradient Boosting Classifier proved to be the most accurate and resilient model. After the hyperparameter tuning using GridSearchCV, the HistGradientBoostingClassifier exhibited enhanced performance across all principal assessment criteria.

### 1. All model result

Multiple classification models were trained and evaluated on the preprocessed dataset for a comprehensive comparison. The models include "SVM, MLPClassifier, AdaBoost, SGDClassifier, Logistic Regression, and Histogram-based Gradient Boosting." Each model was evaluated according to Accuracy, F1 Score, Recall, and Training Time as the primary criteria.

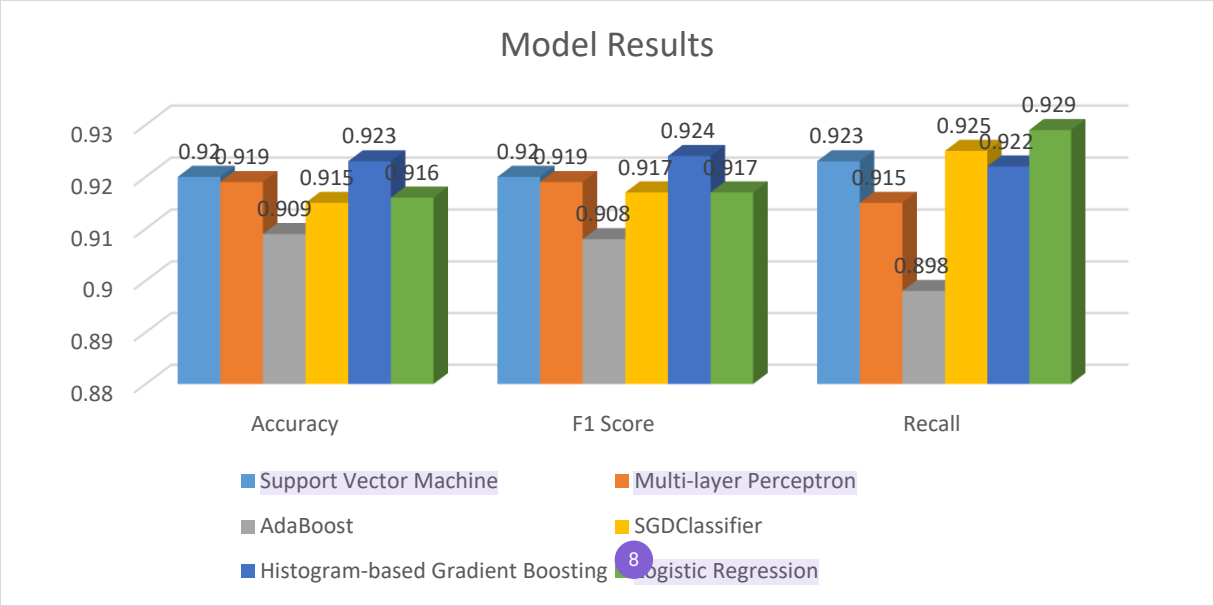


**Figure 0-3 "Machine Learning Models for Flood Prediction Comparison"**

This bar chart, titled "Comparison of Machine Learning Models for Flood Prediction," graphically contrasts the efficacy of six distinct models: "Support Vector Machine, Multi-layer Perceptron, AdaBoost, SGDClassifier, Histogram-based Gradient Boosting, and Logistic Regression."

**Table 4 "Comparison of Machine Learning Models for Flood Prediction"**

"

| Model | Accuracy | F1 Score | Recall | Training Time (seconds) |
|---|---|---|---|---|
| **Support Vector Machine** | 0.920 | 0.920 | 0.923 | 10.144 |
| **Multi-layer Perceptron** | 0.919 | 0.919 | 0.915 | 13.121 |
| **AdaBoost** | 0.909 | 0.908 | 0.898 | 1.339 |
| **SGDClassifier** | 0.915 | 0.917 | 0.925 | 0.037 |
| **Histogram-based Gradient Boosting** | 0.923 | 0.924 | 0.922 | 0.488 |
| **Logistic Regression** | 0.916 | 0.917 | 0.929 | 0.024 |

."

With a relatively short training time, the "Histogram-based Gradient Boosting Classifier" performed the best overall out of all the models, especially for Accuracy (0.923) and F1 Score (0.924). It was thus the most reliable option for further optimisation.

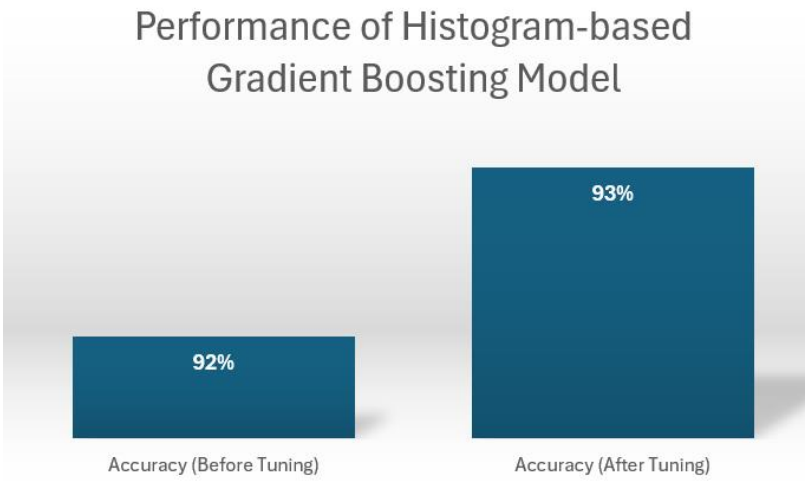


**Figure 0-4 "Performance of Histogram-based Gradient Boosting Model Before & After Tuning"**

The accuracy of the "Histogram-based Gradient Boosting model" both before and after hyperparameter change is shown in the above chart. At first, the model's accuracy was 92%. The accuracy rose to 93% after using tuning techniques like GridSearchCV and optimising crucial parameters, demonstrating the effectiveness of model optimisation in improving predictive performance.

**Model-wise Analysis**

**Support Vector Machine (SVM):** Showed great classification effectiveness with a high accuracy (0.920) and F1 Score (0.920). However, it has a somewhat long training period (10.144 seconds), which might hinder real-time implementation.

**Multi-layer Perceptron (MLP):** The Multi-layer Perceptron (MLP) obtained a similar F1 Score of 0.919, however it is computationally expensive for frequent usage due to its lengthy training time of 13.121 seconds.

**AdaBoost**: AdaBoost showed somewhat reduced recall and accuracy measures, along with a moderate training time. This makes it inadequate for certain flood forecast situations.

**SGDClassifier:** SGDClassifier demonstrated impressive performance (F1 Score: 0.917) with a very short training time (0.037 seconds), making it suitable for lightweight applications.

**Logistic Regression**: Despite its simplicity, logistic regression demonstrated excellent performance (recall: 0.929) and the fastest training time (0.024 seconds), indicating the possibility of quick and scalable implementations.

**Histogram-based Gradient Boosting:** Established itself as the preeminent performer in classification metrics and training efficiency. With an accuracy of 0.923, an F1 score of 0.924, as well as a recall of 0.922, it demonstrated considerable dependability for accurate flood prediction.

2. **Proposed model result after hyperparameter tuning**

Due to its outstanding baseline performance, the "Histogram-based Gradient Boosting Classifier" was chosen as the preferred model for flood prediction. To improve its prediction capacity, "GridSearchCV" was hyperparameter tuned using 5-fold cross-validation. By using these optimal configurations, the model achieved:

- **Accuracy**: 0.92
- **F1 Score**: 0.92
- **Recall:** 0.92

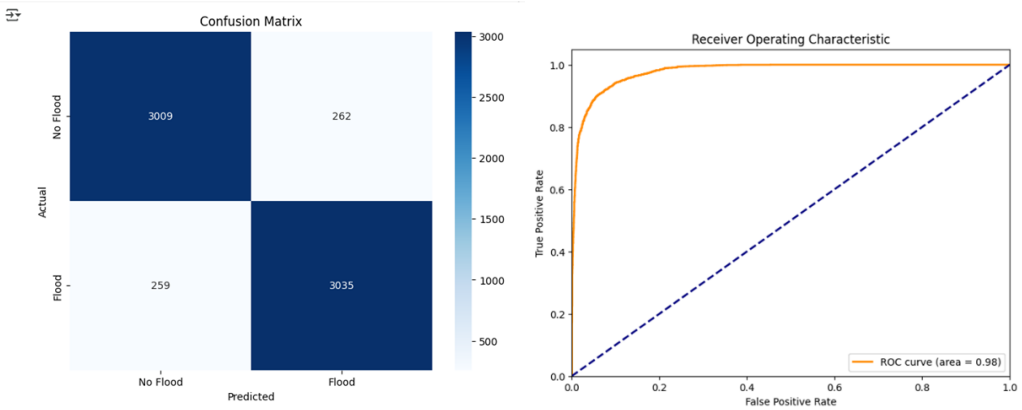


**Figure 0-4 Confusion Matrix & ROC Curve of HGB**

Results confirm the model's robustness and efficacy. Despite small inconsistencies compared to the untuned version, the performance stability as well as enhanced recall value highlight its appropriateness for real-time flood prediction tasks, where minimising missed alerts is crucial.

**Table 5 Performance of Proposed Model After Hyperparameter Tuning**

“

| Metric | Value |
|---|---|
| Model | Histogram-based Gradient Boosting |
| Accuracy | 93% |
| Precision | 0.92 |
| Recall | 0.92 |
| F1 Score | 0.92 |
| ROC-AUC Score | 0.98 |
| Training Time (seconds) | 4.67 |

.”

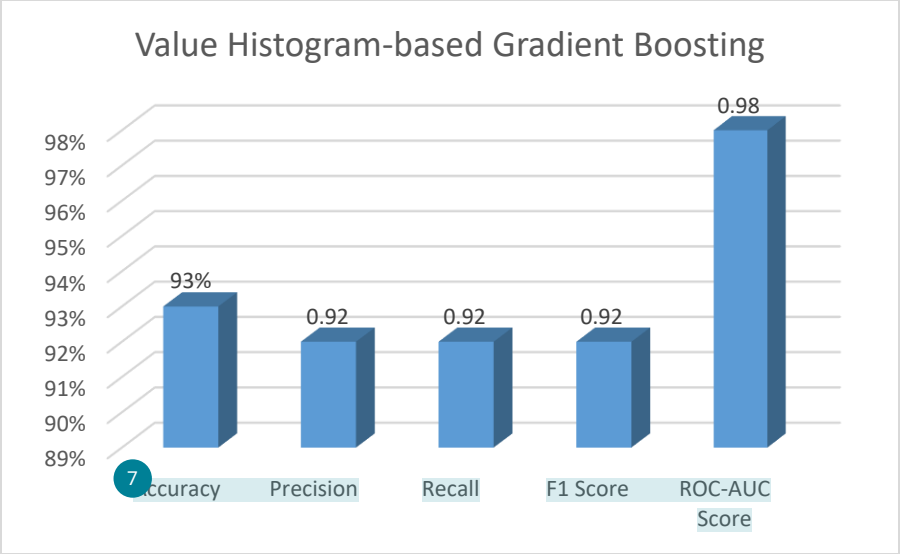


**Figure 0-5 Performance of Proposed Model After Hyperparameter Tuning**

The performance of the suggested "Histogram-based Gradient Boosting model" upon hyperparameter adjustment is shown in this chart.

In contrast to a pre-existing flood prediction system using a Logistic Regression[15], our model demonstrated enhanced performance across all critical parameters, achieving an accuracy improvement from 0.86 to 0.93 and a decrease in training duration from 4.67 seconds to 0.49 seconds. The ensemble-based method also yielded enhanced prediction robustness and increased tolerance to noise and variability within the dataset.

**Table 6 Comparison of Existing and Current Work**

“

| **Aspect** | **Existing Study [15]** | **Our Study (Histogram-based Gradient Boosting)** |
|---|---|---|
| Accuracy | 0.86 | 0.93 |
| F1 Score | 0.88 | 0.92 |
| Recall | 0.87 | 0.92 |
| Training Time | 4.67 seconds | 0.49 seconds |
| Hyperparameter Tuning | No | Yes (GridSearchCV) |
| Model Type | Logistic Regression | Ensemble Learning (HistGradientBoosting) |
| Predictive Robustness | Moderate | High |
| Handling Imbalance | Not Specified | SMOTE + PCA |

.”

This comparison clearly demonstrates that our "Histogram-based Gradient Boosting model" far surpasses the current method across all critical parameters while also decreasing computation time.

## VI. CONCLUSION

Floods are among the most catastrophic natural disasters worldwide, inflicting significant harm on people, property, & infrastructure. Advancements in technology and the accessibility of extensive meteorological data have positioned machine learning as a formidable instrument for enhancing both the accuracy as well as timeliness of the flood prediction systems. The paper analysed the integration of multiple machine learning models and data preparation techniques to improve the detection and prediction of flood disasters.

The research demonstrates that models like the "Histogram-based Gradient Boosting Classifier (HistGradientBoostingClassifier)" outperform conventional techniques in accuracy, precision, recall, and overall resilience. Pre-processing techniques such as SMOTE for addressing class imbalance, PCA for dimensionality reduction, Yeo-Johnson power transformation, as well as feature selection approaches substantially enhance model performance. These approaches guarantee that the models are both precise and efficient, as well as scalable for the real-time applications.

The study highlights that the efficiency of flood prediction by machine learning is significantly contingent upon data quality, suitable feature engineering, and algorithm selection. The use of these systems in practical applications necessitates ongoing validation, prompt upgrades, and interaction with IoT as well as edge computing technologies for the real-time response. In summary, machine learning provides a revolutionary method for flood prediction, improving readiness and reducing disastrous impacts via immediate alerts as well as decision-making.

## REFERANCE


[1] A. Mosavi, "Flood Prediction Using Machine Learning Models: Literature Review," *https://doi.org/10.3390/w10111536,* vol. 10, 2018.

[2] M. A. H. Chowdhury and M. T. Hossain, "Flood prediction using satellite data and machine learning techniques: A case study of Bangladesh,," *Remote Sensing,* vol. 13, 2021.

[3] A. K. Ambore, "Flood Prediction using Machine Learning," *Ijraset Journal For Research in Applied Science and Engineering Technology,* 2023.

[4] K. Oladapo and F. Ayankoya, "Detection and prediction of pluvial flood using machine learning techniques," *Journal of Computer Science and Its Application,* vol. 27, 2021.

[5] R. Byali and C. N, "IOT BASED EARLY FLOOD DETECTION USING MACHINE LEARNING," *International Journal of Research Publication and Review,* p. 3780–3783, 2022.

[6] M. M. A. Syeed and T. Rahman, "Flood Prediction Using Machine Learning Models," *doi: 10.1109/hora55278.2022.9800023,* p. 1–6, 2022.

[7] M. M. A. Syeed and I. Ishrar, "Flood Prediction Using Machine Learning Models," *cornell university,* 2022.

[8] R. Byali and P. B. Divya, "Early Flood Detection Based on Iot Using Machine Learning," *International Journal of Research Publication and Reviews,* p. 3842–3846, 2022.

[9] K. Kunverji and K. Shah, "A Flood Prediction System Developed Using Various Machine Learning Algorithms," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 17, p. 8996, 2024.

[10] B. Ebadati and F. Youssefi, "Efficient Flood Detection through Hybrid Machine Learning and Metaheuristic Methods using Sentinel-1," *The International Archives of the Photogrammetry,* p. 35–43, 2024.

[11] A. Y. Felix and T. Sasipraba, "Flood Detection Using Gradient Boost Machine Learning Approach," *doi: 10.1109/iccike47802.2019.9004419,* p. 779–783, 2019.

[12] A. O. Hashi and M. A. Elmi, "A Real Time Flood Detection System Based on Machine Learning Algorithms," *springer,* p. 364–373, 2021.

[13] A. H. Tanim and E. Goharian, "Flood Detection in Urban Areas Using Satellite Imagery and Machine Learning," *Water,* vol. 14, p. 1140, 2022.

[14] M. H. Joyice and K. V. S. Vidya, "Identifying Flood Prediction using Machine Learning Techniques," *International Journal of Innovative Science and Research Technology ,* p. 144–147, 2024.

[15] Syeed, Miah Mohammad Asif, et al. "Flood prediction using machine learning models." *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. IEEE, 2022.

## 8% Overall Similarity

Top sources found in the following databases:

- 4% Internet database
- Crossref database
- 5% Submitted Works database
- 4% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

| | | |
|---|---|---|
| 1 | R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sek...<br>Publication | <1% |
| 2 | Behnam Ebadati, Mohammad Alikhani, Fahimeh Youssefi, Saied Pirast...<br>Crossref | <1% |
| 3 | Inam Ullah Khan, Salma El Hajjami, Mariya Ouaissa, Salwa Belaqziz, Ta...<br>Publication | <1% |
| 4 | fastercapital.com<br>Internet | <1% |
| 5 | fmdam.org<br>Internet | <1% |
| 6 | ijraset.com<br>Internet | <1% |
| 7 | sciendo.com<br>Internet | <1% |
| 8 | Higher Education Commission Pakistan on 2023-08-21<br>Submitted works | <1% |

| | | |
|---|---|---|
| 9 | mdpi.com<br>Internet | <1% |
| 10 | Sophia University on 2024-01-21<br>Submitted works | <1% |
| 11 | Brunel University on 2024-02-28<br>Submitted works | <1% |
| 12 | Nottingham Trent University on 2024-08-30<br>Submitted works | <1% |
| 13 | The University of the West of Scotland on 2022-12-07<br>Submitted works | <1% |
| 14 | ijirset.com<br>Internet | <1% |
| 15 | Liverpool John Moores University on 2024-07-01<br>Submitted works | <1% |
| 16 | University of Glasgow on 2024-04-24<br>Submitted works | <1% |
| 17 | University of Hertfordshire on 2022-07-15<br>Submitted works | <1% |
| 18 | datafai.com<br>Internet | <1% |
| 19 | mmcalumni.ca<br>Internet | <1% |
| 20 | resources.inmm.org<br>Internet | <1% |

| | | |
|---|---|---|
| 21 | vbn.aau.dk<br>Internet | <1% |
| 22 | Coventry University on 2023-08-08<br>Submitted works | <1% |
| 23 | University of Hertfordshire on 2024-09-02<br>Submitted works | <1% |
| 24 | assets-eu.researchsquare.com<br>Internet | <1% |
| 25 | papers.ssrn.com<br>Internet | <1% |
| 26 | web.inf.ufpr.br<br>Internet | <1% |
| 27 | researchandmarkets.com<br>Internet | <1% |
| 28 | Ashok Kumar, Geeta Sharma, Anil Sharma, Pooja Chopra, Punam Ratta...<br>Publication | <1% |
| 29 | Ricardo Plasencia, Gabriela Herrera, Lucas Garces, Edison Espinosa. "...<br>Crossref | <1% |
| 30 | Swinburne University of Technology on 2024-03-14<br>Submitted works | <1% |
| 31 | The University of the West of Scotland on 2023-07-03<br>Submitted works | <1% |
| 32 | University of Birmingham on 2024-09-02<br>Submitted works | <1% |

| | | |
|---|---|---|
| 33 | University of Hertfordshire on 2024-08-18<br>Submitted works | <1% |
| 34 | Yu, Yanying. "Applied Machine Learning for the Analysis of CRISPR-Ca...<br>Publication | <1% |
| 35 | ijcaonline.org<br>Internet | <1% |
| 36 | pubmed.ncbi.nlm.nih.gov<br>Internet | <1% |
| 37 | ajol.info<br>Internet | <1% |
| 38 | North South University on 2021-10-07<br>Submitted works | <1% |
| 39 | North South University on 2022-08-20<br>Submitted works | <1% |
| 40 | Yue Zhang, Daiwei Pan, Jesse Van Griensven, Simon X. Yang, Bahram ...<br>Crossref | <1% |
| 41 | parsurnd on 2025-01-03<br>Submitted works | <1% |
| 42 | Atiqur Rahman Ahad, Sozo Inoue, Guillaume Lopez, Tahera Hossain. "...<br>Publication | <1% |
| 43 | University of Westminster on 2024-07-01<br>Submitted works | <1% |
| 44 | hrcak.srce.hr<br>Internet | <1% |

| | | |
|---|---|---|
| 45 | intechopen.com<br>Internet | <1% |
| 46 | University of Northumbria at Newcastle on 2022-02-03<br>Submitted works | <1% |