International Journal of Applied Engineering & Technology

# ENHANCING FLOOD RESILIENCE IN MUMBAI: A MACHINE LEARNING-BASED EARLY WARNING SYSTEM

**Becky Nadar[1], Brita Nadar[2], Seema Yadav[3], Prof. Prachi Patil[4] and Prof. Monali Shetty[5]**
[1, 2, 3, 4,5]Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Mumbai, Maharashtra, India
beckynadar06@gmail.com , britanadar073@gmail.com ,
seemadyadav.2025@gmail.com , prachi@fragnel.edu.in and monali_shetty@fragnel.edu.in[5]

## ABSTRACT

*Floods pose a recurring challenge for the Mumbai district and certain regions in Maharashtra, India, leading to devastating impacts on life, infrastructure, and the economy. This research study aims to predict floods on a monthly basis in Mumbai and on a yearly basis in select districts of Maharashtra. The study proposes an integrated approach that leverages machine-learning techniques to enhance decision-making and resilience in flood management.*

*The research involves the training and testing of various machine-learning models using historical flood data, weather information, and other relevant factors to achieve accurate flood prediction. By developing precise flood prediction models, the study seeks to improve disaster preparedness, minimize the detrimental effects of floods, and mitigate economic losses. Through the integration of advanced technology, the research aims to establish an early warning system specifically tailored to Mumbai's needs. This system will play a crucial role in enhancing the city's resilience by providing timely flood detection and predictions, particularly during periods of intense rainfall. Satellite data, including historical daily rainfall records from previous years, will be utilized to gather vital information for flood prediction. The research employs a training approach where historical flood data and satellite information are divided into 80% training and 20% testing subsets. Machine learning models like SVM, and Random Forest are trained using the training data and fine-tuned to assess their analysis and prediction capabilities. These models are then evaluated for generalization and predictive accuracy using independent testing data. By providing accurate flood inundation estimates well in advance, the system will offer crucial time for residents and authorities to respond, reducing the loss of life and property devastation caused by floods. In conclusion, this study's relevance lies in its potential to significantly improve flood prediction accuracy and enable efficient flood management, making Mumbai and the surrounding regions safer and more sustainable in the face of recurrent floods.*

*Keywords: Flood prediction, Mumbai district, Sustainable City, Machine Learning, Disaster Management, flood forecasting*

## 1. INTRODUCTION

Floods in Maharashtra have become a recurrent phenomenon that has serious repercussions including loss of life, destruction of infrastructure and public utilities, and severe economic consequences. Notably, the severity of the floods on 26th July 2005 resulted in widespread devastation across the state, particularly in Mumbai, causing thousands of deaths, injuries, and significant economic and physical impacts. Similarly, a series of floods occurred on 28th July 2021, resulting in a significant loss of life and ongoing missing persons. The region experiences an average of five days per year when educational institutions and offices are forced to shut down due to flooding during the monsoon season. As Mumbai serves as the commercial capital of India and is situated along the coast of Maharashtra, this study aims to develop machine learning models to predict floods in the Mumbai district on month-wise and many regions in the state of Maharashtra year-wise basis.

This study investigates the efficacy prediction of machine learning models and identifies key factors contributing to floods. By developing accurate flood prediction models, the study seeks to enhance disaster preparedness, minimize the devastating impacts of floods, and mitigate economic losses. The broad scope of the project encompasses various aspects of flood prediction and management, including data gathering and analysis. Comprehensive datasets, including historical flood data and other relevant flood-related elements, were collected and analyzed to provide substantial acumen for informed decision cognition and the making of more accurate flood probability models. The relevance of this study lies in its potential to improve flood prediction accuracy and enable effective flood management.

## 2. LITERATURE REVIEW

In the face of rising natural disasters and heavy rainfall, flood impact is one of the most significant disasters in the world. It is becoming more complex, resulting in loss of lives and damaged properties affecting the environment. This section discusses research related to flood disaster management, open street maps, and techniques used to manage flood risk areas.

The scrutiny highlights the growing use of AI techniques in research, offering enhanced capabilities for data analysis and problem-solving while emphasizing the need for ethical consideration with remote sensing techniques for flood prediction and a categorization substructure for flood identification and mapping [1]. A deep learning neural network method is used to formulate reports concerning the identification of floods from an input image and envisage future instances by learning from big data gathered from multiple citations including historic flood events, social media posts, Google and satellite images, ensuring a comprehensive and diverse dataset for analysis and the addition of a Digital Elevation Model (DEM) module enhances flood assessment by accurately determining the intensity of flooding in each area using advanced elevation data and algorithms. This research is mostly focused on disaster predictions and responses.

Another work, [2] provides the impact of flooding in the coastal area. This research has observed Iran is foreseen to be at a greater exposure to flooding when climate change-induced sea-level rise (SLR) is integrated with episodic rises in sea level. A GIS-based SMCDA procedure, utilizing the AHP model, is implemented to construct a risk index for coastal flooding and generate spatial maps, assimilating multiple criteria to enhance the assessment and decision-making for coastal flood risk management. Core risk components are flood and social vulnerability. The research work specifically examines the coastal flooding risk analysis of Bandar Abbas City by employing a Spatial Multi-Criteria Decision Analysis (SMCDA) framework. This framework integrates a wide span of temporal and spatial data to yield a comprehensive study of coastal flood risks in the area. The framework incorporates a vast amount of temporal and spatial data in a unidirectional progression, enabling comprehensive risk analysis for effective coastal flood management strategies.

The research conducted in [3] presents an in-depth analysis conducted in the Maimuna North Divisional Secretariat of Batticaloa, Sri Lanka. This study utilizes an open-source integration to gather field-level data and employs a 3D model to identify areas prone to flood inundation. The study also demonstrates the development of a user-friendly development that integrates open-source tools and Geographic Information System (GIS) technology to create a 3D flood risk model for accurately identifying and estimating flood vulnerability levels. The establishment of the database involves the utilization of JavaOSM and Bing satellite imagery, while exposure data on buildings is collected using free and open-source software. Furthermore, elevation points are extracted from Google Earth to identify flood-prone areas. The research paper employs a combination of qualitative and quantitative modes to collect field-level data, enabling the measurement of flood levels, susceptibility levels, and jeopardy levels within this work.

The authors specifically investigate the roughness parameter and its influence on flood simulation outcomes, particularly examining the impact of different roughness layers [4] . The main purpose of this study is to evaluate multiple roughness maps and assess the effects of various input layers on flood dynamics. The study acknowledges that floodplains and the objects within them exert a significant influence on water flow behavior during flooding events.

The authors utilized Sentinel Synthetic Aperture Radar (SAR) data, specifically VV polarization, known for its effectiveness in flood inundation mapping due to its capability to penetrate cloud cover and rain [5]. The paper's methodology consists of two parts: identification of waterlogged/flooded areas in Mumbai and training flood prediction models utilizing flood conditioning variables and historical flood moments.

The research study [6] offers an overview of flood prediction models, including machine learning (ML) models, physical-based models, numerical models, and statistical models. ML methods have gained popularity among hydrologists for their ability to accurately simulate complex flood processes and provide precise predictions. The study examines various ML algorithms and evaluates their performance in terms of strength, accuracy, potency, and speed. It identifies the most favorable ML models for both long and short-term flood prediction and explores trends for enhancing the quality of these models through hybridization, data disintegration, algorithm ensemble, and model improvement. Additionally, the study discusses physical-based models and numerical models, which have limitations in terms of data requirements and computational intensity for short-term predictions. In flood frequency analysis (FFA), statistical models like autoregressive moving averages (ARMA), multiple linear regression (MLR), and autoregressive integrated moving averages (ARIMA) are widely employed to predict floods by utilizing historical streamflow data and probability distributions, while regional flood frequency analyses (RFFA) have shown improved efficiency over physical models, considering mathematical modeling cost and generalization, by treating floods as stochastic processes. For circumstance, the climatology average method (CLIM), empirical orthogonal function (EOF), multiple linear regressions (MLR), quantile regression techniques (QRT), and Bayesian forecasting models are widely used for predicting major Short-term presumptions of floods are often challenging due to concerns related to the accuracy, complexity, computation cost, and robustness, rendering traditional statistical models unsuitable for such purposes. Advanced data-driven models, particularly ML models, are favored due to their ability to capture flood complexity based only on historical data without relying on detailed knowledge of fundamental physical processes. The objective of this research study is to provide hydrologists and climate scientists with guidance in choosing suitable machine-learning methods tailored to their specific prediction requirements.

Furthermore [7] gives a summary of various machine learning (ML) methods proposed for flood forecasting. The authors introduce an intelligent informatics integration platform (IHIP) that integrates ML, illustration, and system-developing tools to enhance online flood forecasting capabilities and flood hazard mitigation. The IHIP structure consists of five layers (data retrieval, data integration, executor, functional subsystems, and end-user implementation) and a database, enabling efficient handling of flood-related data. Real-time information such as rainfall data and multi-step-ahead zonal flood saturation maps are provided by the IHIP. The integration of Google Maps into the platform improves user accessibility, helps communities make informed decisions regarding flood occurrences, and provides advanced alerts. The IHIP has been deployed in Tainan City, Taiwan, as a case study. The paper also discusses the utilization of the Internet of Things (IoT) and Geographic Information Systems (GIS). The fusion of IoT and ML methods is employed for flood prediction, while GIS enables access to geographical services such as data visualization, interpretation, and querying. The main objective of this analysis is to offer a comprehensive web platform that visualizes hydrological data and provides online regional flood inundation maps, serving as a centralized resource for flood-related information.

## 3. METHODOLOGY

### 3.1 Overview of the System

Floods in India have emerged as one of the deadliest ecological disasters, causing immense human casualties and widespread destruction. Between 1980 and 2017 overall of 235 floods occurred across the country, significantly impacting a population of approximately 1.93 billion people and affecting an estimated loss of 126,286 lives due to these catastrophic events. Floods have led to devasting impacts such as the loss of a living being, structural damage, crop eradication, and livestock loss.

Thus, to address this event again floods and intensify the city's resilience, the system aims to provide an early warning indication to Mumbai City. This system will utilize advanced technology to predict flooding, especially during the interval of heavy rainfall. Therefore, by alerting the individuals and communities in advance, the model will play a crucial role in minimizing the potential impacts of flooding. By providing estimates of flood inundation ahead of time, the system will enable proactive measures to be taken, reducing loss of life and property devastation.

Through the integration of state-of-the-art techniques and real-time data analysis, the system will deliver reliable flood predictions. Equipping the city with timely and accurate information can empower residents and authorities to take necessary precautions and implement effective disaster response strategies

### 3.2 Architecture of the Proposed System

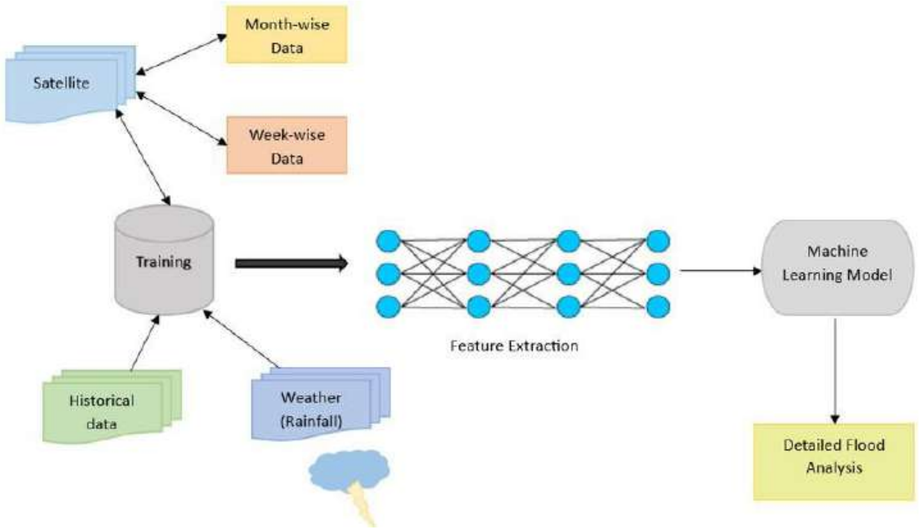


**Fig. 1.** Proposed system.

Component Description

- Satellite Data (Monthly and Weekly): Gathers information from satellites to find the daily rainfall of previous years.
- Training: Historical flood data and satellite data are split into 80% training and 20% testing subsets to train and analyze a machine learning model's analysis and prediction abilities. The model is fine-tuned using the training data and assessed for its generalization and predictive accuracy using independent testing data.
- Historical Data: Offers information on previous floods to shed light on patterns, trends, and prospective danger zones.
- Weather (Rainfall): Uses rainfall information to determine the likelihood of flooding.
- Featured Extraction: To improve the accuracy of flood predictions, feature extraction identifies pertinent information from acquired data, such as precipitation amounts.
- Machine Learning Model: This approach uses gathered data to create predictive models for assessing the frequency and intensity extent of floods.
- Detailed Flood Analysis: Examines predicted flood extent for effective disaster response planning.

The suggested method makes use of historical precipitation data from the trusted and reputable NASA website [8]. By analyzing this data, the system develops an algorithm that takes regional topography, historical trends, and weather conditions into account when determining a threshold rainfall level. An important indicator for locating flood-prone locations is this threshold rainfall.

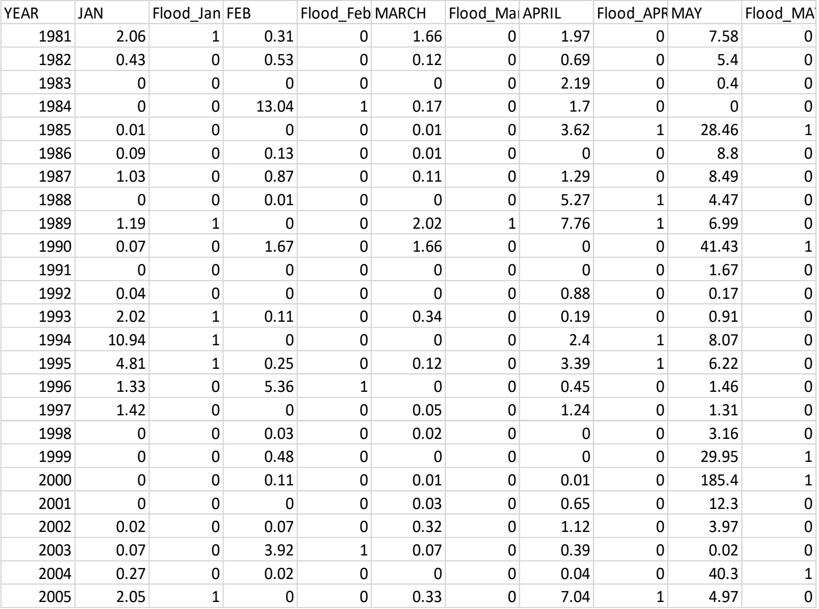
The system determines the region's most likely to experience floods after calculating the threshold rainfall. The system finds the regions where rainfall exceeds the established threshold, indicating a higher danger of flooding. It does this by comparing the most recent rainfall data with the established threshold. To reduce possible risks, additional analysis and intervention actions are prioritized for certain flood-prone locations.

The proposed system makes it possible to analyze and manage flood risk effectively by utilizing NASA's historical precipitation data, a threshold rainfall calculation algorithm. It equips decision-makers with insightful knowledge that enables them to make preventive decisions and allocate resources wisely to lessen the effects of flooding.

### 3.3 Dataset Description

In this study, the datasets employed were obtained from the NASA Power Data Access Viewer, a reputable platform that facilitates researchers' access to satellite data. The latitude and longitude coordinates of the targeted region were extracted meticulously from Google Maps. To ensure the utmost accuracy and authenticity of these coordinates, the validation process involved leveraging the capabilities of the NASA Power Data Access Viewer. Through inputting the coordinates into the viewer and conducting a thorough examination of the corresponding location data and satellite imagery, the veracity of the coordinates was effectively established. By configuring the data selection criteria to encompass crucial variables, including humidity and precipitation, with a specific focus on relative humidity measured at a height of 2 meters, the pertinent dataset essential for conducting rigorous flood prediction analysis was successfully procured.

**Table I.** Mumbai Region Dataset

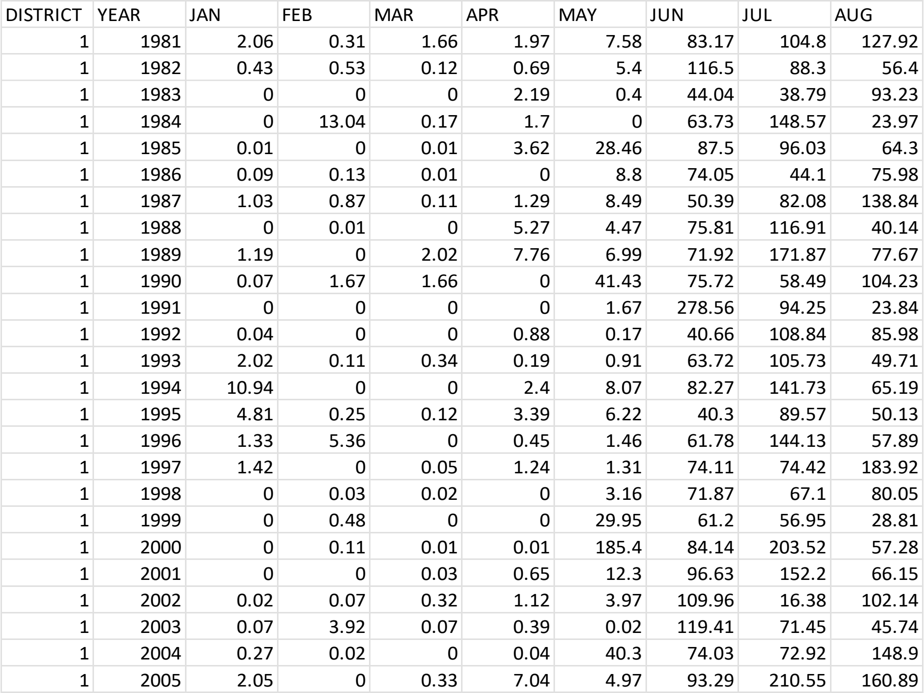


| YEAR | JAN | Flood_Jan | FEB | Flood_Feb | MARCH | Flood_Ma | APRIL | Flood_APR | MAY | Flood_MA |
|---|---|---|---|---|---|---|---|---|---|---|
| 1981 | 2.06 | 1 | 0.31 | 0 | 1.66 | 0 | 1.97 | 0 | 7.58 | 0 |
| 1982 | 0.43 | 0 | 0.53 | 0 | 0.12 | 0 | 0.69 | 0 | 5.4 | 0 |
| 1983 | 0 | 0 | 0 | 0 | 0 | 0 | 2.19 | 0 | 0.4 | 0 |
| 1984 | 0 | 0 | 13.04 | 1 | 0.17 | 0 | 1.7 | 0 | 0 | 0 |
| 1985 | 0.01 | 0 | 0 | 0 | 0.01 | 0 | 3.62 | 1 | 28.46 | 1 |
| 1986 | 0.09 | 0 | 0.13 | 0 | 0.01 | 0 | 0 | 0 | 8.8 | 0 |
| 1987 | 1.03 | 0 | 0.87 | 0 | 0.11 | 0 | 1.29 | 0 | 8.49 | 0 |
| 1988 | 0 | 0 | 0.01 | 0 | 0 | 0 | 5.27 | 1 | 4.47 | 0 |
| 1989 | 1.19 | 1 | 0 | 0 | 2.02 | 1 | 7.76 | 1 | 6.99 | 0 |
| 1990 | 0.07 | 0 | 1.67 | 0 | 1.66 | 0 | 0 | 0 | 41.43 | 1 |
| 1991 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.67 | 0 |
| 1992 | 0.04 | 0 | 0 | 0 | 0 | 0 | 0.88 | 0 | 0.17 | 0 |
| 1993 | 2.02 | 1 | 0.11 | 0 | 0.34 | 0 | 0.19 | 0 | 0.91 | 0 |
| 1994 | 10.94 | 1 | 0 | 0 | 0 | 0 | 2.4 | 1 | 8.07 | 0 |
| 1995 | 4.81 | 1 | 0.25 | 0 | 0.12 | 0 | 3.39 | 1 | 6.22 | 0 |
| 1996 | 1.33 | 0 | 5.36 | 1 | 0 | 0 | 0.45 | 0 | 1.46 | 0 |
| 1997 | 1.42 | 0 | 0 | 0 | 0.05 | 0 | 1.24 | 0 | 1.31 | 0 |
| 1998 | 0 | 0 | 0.03 | 0 | 0.02 | 0 | 0 | 0 | 3.16 | 0 |
| 1999 | 0 | 0 | 0.48 | 0 | 0 | 0 | 0 | 0 | 29.95 | 1 |
| 2000 | 0 | 0 | 0.11 | 0 | 0.01 | 0 | 0.01 | 0 | 185.4 | 1 |
| 2001 | 0 | 0 | 0 | 0 | 0.03 | 0 | 0.65 | 0 | 12.3 | 0 |
| 2002 | 0.02 | 0 | 0.07 | 0 | 0.32 | 0 | 1.12 | 0 | 3.97 | 0 |
| 2003 | 0.07 | 0 | 3.92 | 1 | 0.07 | 0 | 0.39 | 0 | 0.02 | 0 |
| 2004 | 0.27 | 0 | 0.02 | 0 | 0 | 0 | 0.04 | 0 | 40.3 | 1 |
| 2005 | 2.05 | 1 | 0 | 0 | 0.33 | 0 | 7.04 | 1 | 4.97 | 0 |

The first dataset, known as the "Mumbai region dataset" as shown in Table I, is dedicated to capturing monthly rainfall data specifically for the Mumbai region. The dataset is constructed by aggregating daily rainfall measurements to determine the maximum rainfall recorded within each month. This provides a consolidated view of the monthly rainfall patterns.

In addition to the rainfall measurements, the dataset also includes a flood column which indicates the likelihood of flooding on a month-by-month basis. The flood column adopts a binary format where the value 0 signifies the absence of flooding, while the value 1 indicates the occurrence of flooding during that particular month.

**Table II.** Maharashtra Region Dataset



| DISTRICT | YEAR | JAN | FEB | MAR | APR | MAY | JUN | JUL | AUG |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1981 | 2.06 | 0.31 | 1.66 | 1.97 | 7.58 | 83.17 | 104.8 | 127.92 |
| 1 | 1982 | 0.43 | 0.53 | 0.12 | 0.69 | 5.4 | 116.5 | 88.3 | 56.4 |
| 1 | 1983 | 0 | 0 | 0 | 2.19 | 0.4 | 44.04 | 38.79 | 93.23 |
| 1 | 1984 | 0 | 13.04 | 0.17 | 1.7 | 0 | 63.73 | 148.57 | 23.97 |
| 1 | 1985 | 0.01 | 0 | 0.01 | 3.62 | 28.46 | 87.5 | 96.03 | 64.3 |
| 1 | 1986 | 0.09 | 0.13 | 0.01 | 0 | 8.8 | 74.05 | 44.1 | 75.98 |
| 1 | 1987 | 1.03 | 0.87 | 0.11 | 1.29 | 8.49 | 50.39 | 82.08 | 138.84 |
| 1 | 1988 | 0 | 0.01 | 0 | 5.27 | 4.47 | 75.81 | 116.91 | 40.14 |
| 1 | 1989 | 1.19 | 0 | 2.02 | 7.76 | 6.99 | 71.92 | 171.87 | 77.67 |
| 1 | 1990 | 0.07 | 1.67 | 1.66 | 0 | 41.43 | 75.72 | 58.49 | 104.23 |
| 1 | 1991 | 0 | 0 | 0 | 0 | 1.67 | 278.56 | 94.25 | 23.84 |
| 1 | 1992 | 0.04 | 0 | 0 | 0.88 | 0.17 | 40.66 | 108.84 | 85.98 |
| 1 | 1993 | 2.02 | 0.11 | 0.34 | 0.19 | 0.91 | 63.72 | 105.73 | 49.71 |
| 1 | 1994 | 10.94 | 0 | 0 | 2.4 | 8.07 | 82.27 | 141.73 | 65.19 |
| 1 | 1995 | 4.81 | 0.25 | 0.12 | 3.39 | 6.22 | 40.3 | 89.57 | 50.13 |
| 1 | 1996 | 1.33 | 5.36 | 0 | 0.45 | 1.46 | 61.78 | 144.13 | 57.89 |
| 1 | 1997 | 1.42 | 0 | 0.05 | 1.24 | 1.31 | 74.11 | 74.42 | 183.92 |
| 1 | 1998 | 0 | 0.03 | 0.02 | 0 | 3.16 | 71.87 | 67.1 | 80.05 |
| 1 | 1999 | 0 | 0.48 | 0 | 0 | 29.95 | 61.2 | 56.95 | 28.81 |
| 1 | 2000 | 0 | 0.11 | 0.01 | 0.01 | 185.4 | 84.14 | 203.52 | 57.28 |
| 1 | 2001 | 0 | 0 | 0.03 | 0.65 | 12.3 | 96.63 | 152.2 | 66.15 |
| 1 | 2002 | 0.02 | 0.07 | 0.32 | 1.12 | 3.97 | 109.96 | 16.38 | 102.14 |
| 1 | 2003 | 0.07 | 3.92 | 0.07 | 0.39 | 0.02 | 119.41 | 71.45 | 45.74 |
| 1 | 2004 | 0.27 | 0.02 | 0 | 0.04 | 40.3 | 74.03 | 72.92 | 148.9 |
| 1 | 2005 | 2.05 | 0 | 0.33 | 7.04 | 4.97 | 93.29 | 210.55 | 160.89 |

The second dataset, labeled the "Maharashtra region dataset" as shown in Table II, gives comprehensive detail on monthly rainfall for four regions in the state with a year spanning 40 years. The datasets include district name, monthly rainfall data, and flood column indicating 0's for no occurrence of flood and 1's indicating the possibility of flood on a year-wise basis. The development of the dataset is done by computing daily rainfall measurements to get maximum rainfall records for all 12 months, thus providing a better view of monthly rainfall records.

It is important to note that both datasets have undergone rigorous collection and cleaning processes to ensure accuracy and data consistency. For this research study, these datasets form the fundamental basis, which aims to conduct an in-depth analysis of the correlation between rainfall patterns and the likelihood of flooding in both, Maharashtra and Mumbai regions.

### 3.4 Working of The Proposed Methodology

To accurately forecast future flood events, the following algorithm has been implemented: K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). These algorithms prove their effectiveness in analyzing historical rainfall data to predict floods presenting them the most ideal option for this research.

For accurate flood predictions, each algorithm was evaluated to conclude its strengths, limitations, and suitability. The K-Nearest Neighbors algorithm is a classification method that allocates a new data point to the nearest neighbors for classification in data training. With the parameter K specified is called a number of neighbors. Additionally, Logistic Regression is used for binary classification whereas a Decision Tree divides the data into sub-data based on given requirements. While on the other hand, Random Forest employs an ensemble of Decision Trees to make predictions.

The accuracy of each algorithm is determined using various evaluation metrics including precision, recall, and roc score. A comprehensive understanding of these algorithms was on these metrics. For this study, Random Forest has the highest accuracy and thus emerges as the best model among the four algorithms. Its ability to generalize the unseen data and handle the outliers effectively is one of the crucial roles for the prediction of complex and dynamic rainfall patterns.

### 3.5 Threshold Calculation

**Algorithm 1: Monthly Threshold Rainfall Calculation**

```
Input: start_year, end_year, rainfall_data
Determine the time period: N ← end_year - start_year + 1
max_rainfall[12] ← It is an array that stores the maximum monthly
month
threshold_rainfall[12] ← It is an array that stores the Threshold
mean_max_rainfall ← It represents the mean of the maximum
std_max_rainfall ← It represents the standard deviation of the
for j=1 to N do
    for i=1 to 12 do
        if rainfall > max_rainfall[i], then
            max_rainfall[i] ← rainfall
        end if
    end for
end for
mean_max_rainfall ← sum(max_rainfall) / 12
sum_of_squares ← 0
for i=1 to 12 do
    difference ←max_rainfall[i] + mean_max_rainfall
    square ← difference * difference
    sum_of_squares ← sum_of_squares + square
end for
std_max_rainfall ← square_root(sum_of_squares/12)
for i=1 to 12 do
    threshold_rainfall[i] ← mean_max_rainfall + std_max_rainfall
end for
Output: threshold_rainfall
```

**Algorithm 2: Yearly Threshold Rainfall Calculation**

```
Input: start_year, end_year, rainfall_data
Determine the time period: N ← end_year - start_year + 1
max_rainfall[12] ← It is an array that stores the maximum monthly rainfall values for each
month
threshold_rainfall[12] ← It is an array that stores the Threshold rainfall values for each month
mean_max_rainfall ← It represents the mean of the maximum monthly rainfall value
std_max_rainfall ← It represents the standard deviation of the maximum monthly rainfall values
```

Copyrights @ Roman Science Publications Ins.

```
for j=1 to N do
    for i=1 to 12 do
        if rainfall > max_rainfall[i], then
            max_rainfall[i] ← rainfall
        end if
    end for
end for
mean_max_rainfall ← sum(max_rainfall) / 12
sum_of_squares ← 0
for i=1 to 12 do
    difference ←max_rainfall[i] + mean_max_rainfall
    square ← difference * difference
    sum_of_squares ← sum_of_squares + square
end for
std_max_rainfall ← square_root(sum_of_squares/12)
for i=1 to 12 do
    threshold_rainfall[i] ← mean_max_rainfall + std_max_rainfall
end for
total_threshold_rainfall ← sum(threshold_rainfall)
Output: total_threshold_rainfall
```

## 4 Implementation and Result

### 4.1 Analysis of the Mumbai Region

**Table III**. Performance Evaluation Metrics for the Classification Algorithm

| Evaluation Metrics | LR | DT | RF |
|---|---|---|---|
| **Accuracy** | 82.4% | 85.28% | 86.103% |
| **Recall** | 62.78% | 62.78% | 62.78% |
| **ROC** | 80.79% | 79.6% | 79.6% |
| **F1 score** | 75.23% | 73.84% | 73.84% |
| **Precision** | 95.83% | 93.75% | 93.75% |
| **NPV** | 83.62% | 82.67% | 83.14% |
| **Specificity** | 98.81% | 96.43% | 96.43% |

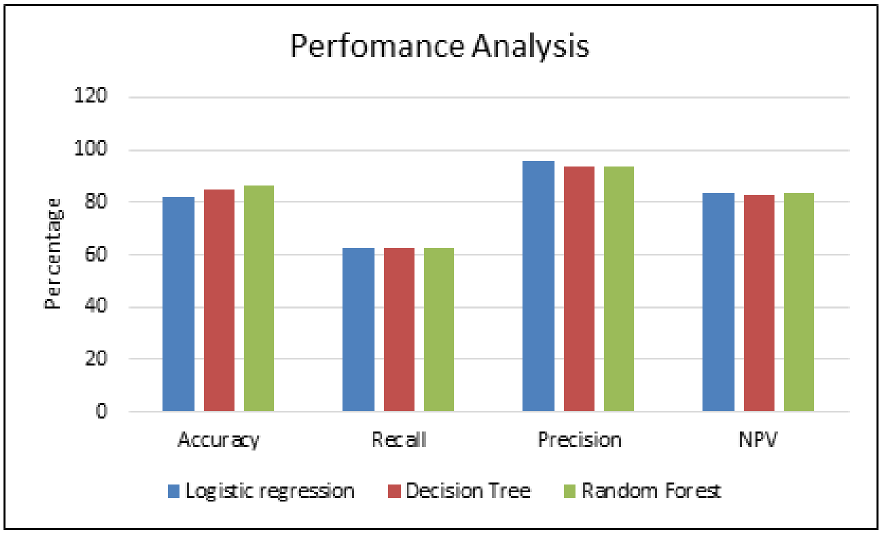


**Fig. 2.** Performance analysis for Mumbai

Table III presents the ML-proposed model's experimental outcomes, illustrating the competence of the Machine Learning model using evaluation metrics Accuracy, a fundamental metric, quantifies the overall correctness of the model's predictions by indicating the percentage of instances correctly classified. Logistic Regression (LR) achieves an average accuracy of 82.4%, Decision Tree (DT) achieves 85.28%, and Random Forest (RF) outperforms both with an accuracy of 86.103%. This indicates that Random Forest has the highest percentage of correct predictions across the dataset. The recall score evaluates the model's capability to correctly identify actual positive instances. It calculates the proportion of true positives relative to the sum of true positives and false negatives. All three algorithms exhibit the same recall score of 62.78%, indicating an equal capacity to correctly identify positive instances. Receiver Operating Characteristic (ROC) curve serves as a graphical representation of a binary classification model's performance. Although Logistic Regression has a slightly higher ROC score of 80.79%, the difference is not statistically significant.

The F1 score, a measure of precision and recall, offers a balanced assessment of the model's performance, particularly in imbalanced datasets. The F1 score values reported in the table range from 73.84% to 75.23% across the algorithms. Precision quantifies the accuracy of positive predictions made by the model , The precision values presented range from 93.75% to 95.83% for the different scenarios. NPV (Negative Predictive Value) evaluates the model's ability to accurately identify actual negative instances, measuring the proportion of true negatives out of the sum of true negatives and false negatives. The NPV values for Logistic Regression, Decision Tree, and Random Forest are 83.62%, 82.67%, and 83.14% respectively. Specificity gauges the model's aptitude to correctly identify actual negative instances as negative. It calculates the proportion of true negatives in relation to the sum of true negatives and false positives. The specificity values reported in the table span from 96.43% to 98.81% for the different scenarios.

When considering the importance of accurate predictions, as well as other technical factors such as interpretability, model complexity, robustness to outliers, and computational efficiency, the Random Forest algorithm emerges as the better choice. Random Forest offers a good balance between accuracy and interpretability, handles non-linear relationships effectively, is robust to outliers, and demonstrates acceptable computational efficiency. Therefore, considering the overall performance and the aforementioned factors, Random Forest is the preferred model for the research study. It provides a higher accuracy rate, making it well-suited for accurately predicting the desired outcomes while maintaining interpretability and handling the complexity of the dataset. Thus, it’s the preferred model for the analysis of the Mumbai region.

## 4.2 Analysis of the Maharashtra Region

**Table III.** Performance Evaluation Metrics for the Classification Algorithm

| Evaluation Metrics | LR | DT | RF |
|---|---|---|---|
| **Accuracy** | 90.91% | 78.79% | 96.97% |
| **Recall** | 25% | 50% | 75% |
| **ROC** | 62.50% | 66.38% | 87.50% |
| **F1 score** | 40% | 36.36% | 85.71% |
| **NPV** | 90.62% | 92.30% | 96.67% |

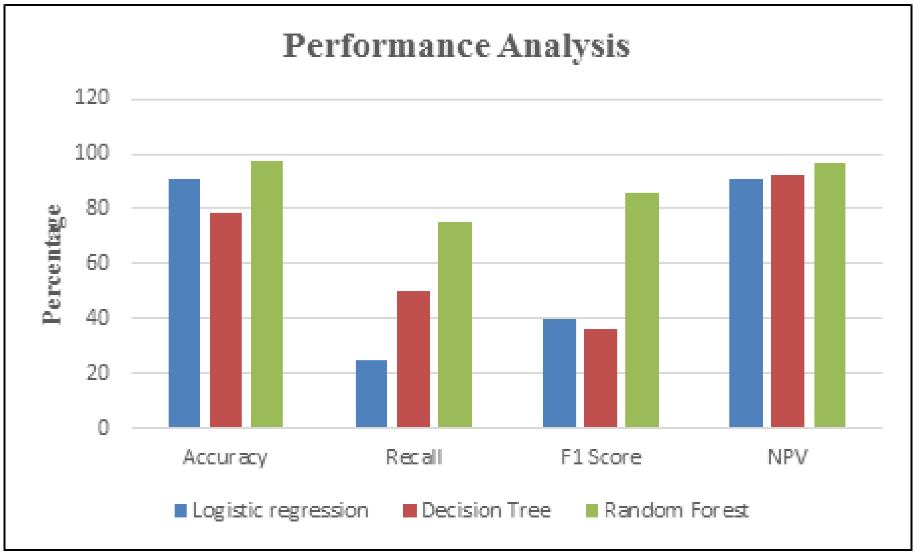


**Fig. 3.** Performance analysis for Maharashtra

Similar to evaluation metrics for the Mumbai region, Table IV presents the evaluation metrics of Maharashtra. Here Logistic Regression (LR) achieves an average accuracy of 90.91%, Decision Tree (DT) achieves 78.79%, and Random Forest (RF) outperforms both with an accuracy of 96.97%. This indicates that Random Forest has the highest percentage of correct predictions across the dataset. Recall, measuring the model's ability to identify positive instances, showed values of 25%, 50%, and 75%. Receiver Operating Characteristic (ROC) ranges from 62.50% to 87.50%. The F1 score was reported as 40%, 36.36%, and 85.71% across the scenarios, offering a balanced assessment of the model's performance. The model's ability to accurately identify actual negative instances was evaluated using NPV values of 90.62%, 92.30%, and 96.67% respectively.

The Random Forest model outperformed other models with the highest accuracy, recall, ROC, F1 score, and NPV, indicating its superior performance in correctly classifying instances and accurately identifying positive and negative cases. Hence,it’s the preferred model for the research study in the region of Maharashtra

## 5 Conclusion And Future Enhancement

### 5.1 CONCLUSION

In conclusion, the research study presented a comprehensive evaluation of three machine learning algorithms - Logistic Regression (LR), Decision Trees (DT), and Random Forest (RF) - for flood prediction in Mumbai and selected districts of Maharashtra. After careful analysis, Random Forest emerged as the most effective model, exhibiting high accuracy rates of 86.103% for Mumbai and an impressive 96.97% for Maharashtra. The suitability of Random Forest for flood prediction in the region is attributed to its interpretability, capability to handle non-linear relationships, resilience to outliers, and computational efficiency.

The successful application of these machine learning algorithms to process and analyse historical flood data revealed their capacity to discern crucial features and patterns within the dataset, leading to accurate flood predictions. These findings hold significant implications for flood management and preparedness, as the integration of real-time data and monitoring systems can enable the development of efficient early warning systems. The research study's outcomes offer valuable insights for stakeholders involved in flood risk mitigation and infrastructure planning. By leveraging the predictive power of machine learning algorithms, authorities can make informed decisions and implement proactive measures to minimize flood damage and protect lives and property.

While Random Forest proved to be the optimal model for this study, ongoing research and continuous improvement of the models hold promise for even better flood prediction capabilities. As technology advances and data collection methods improve, the effectiveness of these algorithms will likely increase, leading to further advancements in flood forecasting and disaster resilience. In conclusion, the research underscores the importance of leveraging machine learning techniques to address the recurring challenge of floods in Mumbai and Maharashtra. By harnessing the power of data-driven insights, early warning systems can be fortified, bolstering the region's ability to respond swiftly and effectively to potential flood events. This research contributes to the broader field of disaster management, emphasizing the significance of interdisciplinary approaches that combine technological innovation and scientific analysis to create a safer and more sustainable future.

### 5.2 FUTURE ENHANCEMENT

1. To enhance the correctness of flood prediction, additional factors such as the location of shelters and hospitals can be included in the model as input features This can be achieved by integrating geospatial data with machine learning algorithms.

2. Providing the safest route for users can be done using routing algorithms that take into account the predicted flood levels and road conditions. This requires real-time monitoring of weather conditions and integrating this information into the prediction model.

3. To implement the project on a large scale, hydrological and hydrometeorological data can be collected and used in WRF models to predict flood events. These models can provide high-resolution forecasts of weather conditions and river levels to support flood prediction.

4. Including the BMC wards dataset and watersheds can provide additional insights into the characteristics of the region and the factors that contribute to flooding This can help in making more accurate flood prediction models and better disaster management strategies.

## REFERENCES

[1] H. S. Munawar, A. W. Hammad and S. Waller, "Remote Sensing Methods for Flood Prediction: A Review," Sensors, vol. 22, no. 3, p. 960, 2022.

[2] V. Hadipour, F. Vafaie and K. Deilami, "Coastal Flooding Risk Assessment Using a GIS-Based Spatial Multi-Criteria Decision Analysis Approach," Water, vol. 12, no. 9, p. 2379, 2020.

[3] S.Suthakaran, A. Withanage, M.Gunawardhane and J. , "Flood Risk Assessment based on OpenStreetMap Application: A case study in Manmunai North Divisional Secretariat of Batticaloa, Sri Lanka," in FOSS4G Asia 2018, Moratuwa, 2018.

[4] H. Dorn, M. Vetter and B. Hofle, "GIS-Based Roughness Derivation for Flood Simulations: A Comparison of Orthophotos, LiDAR and Crowdsourced Geodata," Remote Sensing, vol. 6, no. 2, pp. 1739-1759, 2014.

[5] S. Khatri, P. Kokane, V. Kumar and S. Pawar, "Prediction of waterlogged zones under heavy rainfall conditions using machine learning and GIS tools: a case study of Mumbai," GeoJournal, pp. 1-15, 2022.

[6] A. Monsavi, P. Ozturk and K.-w. Chau, "Flood Prediction Using Machine Learning Models: Literature Review," Water, vol. 10, no. 11, p. 1536, 2018.

[7] L.-C. Chang, F.-J. Chang, S.-N. Yang, I.-F. Kao, Y.-Y. Ku, C.-L. Kuo and I. Z. b. Mat Amin, "Building an Intelligent Hydroinformatics Integration Platform for Regional Flood Inundation Warning System," Water, vol. 11, no. 1, p. 9, 2019.

[8] "NASA," [Online]. Available: https://power.larc.nasa.gov/data-access-viewer. [Accessed February 2023].

[9] P. Sudar and K. Subrahmanya, "SpatialMapping of Flood Susceptibility Using Decision Tree-Based Machine Learning Models for the Vembanand Lake System in Kerala, India," Water Resources, vol. 149, no. 10, 2023.

[10] M. Khairudin, N. Mustapha, N. T. Mohd Aris and Z. Maslina, "In-Depth Review On Machine Learning Models For Long-term Flood Forecasting," Journal Of Theoretical and Applied, vol. 100, no. 10, 2022.

[11] R. Balamurugan, K. Choudhary and S. Raja, "Prediction of Flooding Due to Heavy Rainfall in India Using Machine Learning Algorithms: Providing Advanced Warning," IEEE Systems,Man, and Cybernetics Magazine, vol. 8, no. 4, pp. 26-33, 2022.

[12] "About being MeMumbai," MeMumbai Department, [Online]. Available: https://memumbai.com/about-being-memumbai/. [Accessed 5 January 2023].

[13] P. Ashok, J. S. Kavin, N. Sankara and P. Vasanth, "Flood Prediction in the Area of Tamil Nadu and Andhra Pradesh Using Machine Learning," in International Conference on Computational Science and Technology, Chennai, 2022.

[14] N. Razali, S. I. Ismail and A. Mustapha, "Machine Learning approach for flood risks prediction," IAES International Journal of Artificial Intelligence(IJ-AI), vol. 9, no. 1, pp. 73-80, 2020.

[15] S.-j. Park and D.-k. Lee, "Prediction of Coastal Flooding Risk Under Climate Change Impacts In South Korea Using Machine Learning Algorithms," Environmental Research Letters, vol. 15, no. 9, 2020.

[16] "Humanitarian OpenStreetMap Team," [Online]. Available: https://www.hotosm.org/updates/osm-kerala-the-past-present-and-future/. [Accessed 7 December 2022].

[17] "Sahana Foundation," [Online]. Available: https://sahanafoundation.org/. [Accessed 7 December 2023].

[18] M. Hasanuzzaman, A. Islam, B. Bera and P. K. Shit, "A Comparison of performance measures of three machine learning algorithms for flood susceptibility mapping of river Silabati(tropical river,India," Physics and Chemistry of the Earth, Parts A/B/C, vol. 127, p. 103198, 2022.