International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

xx

COPY RIGHT

IJIEMR Transactions, online available on 23th May 2024. Link
https://www.ijiemr.org/downloads/Volume-13/ISSUE-5

10.48047/IJIEMR/V13/ISSUE 05/43

TITLE: FLOOD PREDICTION USING MACHINE LEARNING

Volume 13, ISSUE 05, Pages: 413-421

Paper Authors **Sumanta Saha, GONGADA NAGARAJU**

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

FLOOD PREDICTION USING MACHINE LEARNING

Sumanta Saha, GONGADA NAGARAJU

assistant professor, Department of IT, Tripura University, Tripura, India, sumantasaha@gmail.com
Department of it, Tripura University, Tripura, India, nagarajugongada100@gmail.com

Abstract -Flooding can be really bad for towns and cities around the world. This thesis is about using super smart computer tools to predict floods and warn people in advance. We use a lot of weather data from different locations and try different computer methods to improve our flood predictions. We want to make sure that our predictions are not only accurate but also help communities prepare. We start by looking really closely at our weather data, fixing any problems like missing information or changes in weather. Then, we use special techniques to select the best weather features that help us make good flood predictions. We try many computer methods like K-Nearest Neighbors, Logistic Regression and others to make them work as best as possible.We check how good our predictions are by using numbers like precision and recall. Also, we think it's important to tell people about floods in an easy way, so we figured out how to communicate this information effectively. By doing all this, our goal is to create smart tools that improve flood prediction, help communities be prepared, and keep everyone safe.

Keywords: - *Flood Prediction, Machine Learning, ML Models,SVC, RF, KNN, LR, DT*

I. INTRODUCTION

In recent years, the increasing frequency and severity of floods has posed significant challenges to communities around the world, necessitating the development of advanced and reliable flood prediction systems. Traditional methods often struggle to provide timely and accurate forecasts due to the complex and dynamic nature of environmental factors influencing flood events.As we enter the era of technological advancements, integration of machine learning techniques into hydrological modeling is emerging as a promising opportunity to enhance flood prediction capabilities.

This thesis focuses on the application of machine learning algorithms to predict floods, leveraging data-driven insights and the power of computational models. The primary objective is to develop a robust and accurate flood prediction system that is capable of analyzing historical and real-time data to forecast potential flood events. Through the use of diverse machine learning models including KNearest Neighbors, Logistic Regression, Decision Trees, Support Vector Machines, Random Forests, and Ensemble Learning, this research aims to determine the effectiveness of these algorithms in predicting floods in a given geographic area. Have to find out. The purpose of this thesis is to advance the field of flood forecasting by utilizing the capabilities of machine learning. By blending scientific rigor with

technological innovation, we aspire to pave the way for a more flexible and adaptive approach to managing the challenges posed by flood events.

Predicting and managing floods is crucial for minimizing their devastating impact on communities and infrastructure. Over the past decade, there has been a significant shift towards leveraging machine learning (ML) models to enhance flood prediction accuracy and efficiency. This paper presents a comprehensive review of recent advancements in flood prediction using ML techniques. One key aspect of this review is the integration of various datasets and technologies to improve prediction models. For instance, Anisha et al. (2018) combined survey data and ML models to predict flood damages in Kerala. Similarly, Fitri Yakub et al. (2022) reviewed hybrid ML models for flood prediction, highlighting the importance of dataset selection. These studies demonstrate the significance of data in enhancing the accuracy of flood prediction models. Furthermore, researchers have explored the integration of block chain technology with ML for flood prediction (K. M. et al., 2016), showing promising results. Additionally, DinhKha Dang et al. (2023) integrated ML with hydrodynamic modeling to address the extrapolation problem in flood depth estimation, further enhancing prediction accuracy. Overall, this review highlights the growing importance of ML in improving flood prediction models, with researchers continually exploring innovative approaches to enhance accuracy and efficiency in flood prediction and management.

II. LITERATURE SURVEY

A literature survey on flood prediction using machine learning models reveals a diverse range of approaches and methodologies employed by researchers worldwide. This survey aims to summarize key findings and trends from selected studies in the field.

One of the earliest works in this area is the study by al. [4], which proposes an integrated approach for flood prediction using blockchain network and machine learning. This work highlights the importance of incorporating emerging technologies to enhance the accuracy and reliability of flood prediction models.

Another significant contribution is the study by Dinh Kha DangHuu Duy Nguyen et al. [5], which focuses on integrating machine learning and hydrodynamic modeling to solve the extrapolation problem in flood depth estimation. This approach demonstrates the potential of combining different computational techniques to improve the performance of flood prediction models.

In a similar vein, Miguel de Castro Neto Marcel Motta [7] propose a mixed approach for urban flood prediction using machine learning and GIS. This study emphasizes the importance of spatial data analysis and modeling in predicting flood patterns in urban areas.

Moving on to more specific applications, Fitri Yakub Ainaa Hanis Zuhairi [2] conduct a review of flood prediction hybrid machine learning models using datasets. This review provides a comprehensive overview of the different machine learning techniques used in flood prediction and their relative performance.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

A case study by Hayati Yassin Zaharaddeen Karami Lawal [9] focuses on flood prediction using machine learning models in Kebbi State, Nigeria. This study highlights the importance of tailoring prediction models to specific geographic regions to improve their accuracy and applicability.

On a different note, M. A. Habib, J. O., and Salauddin, M. [6] conduct a systematic review on the prediction of wave overtopping characteristics at coastal flood defenses using machine learning algorithms. This review underscores the significance of considering environmental factors in flood prediction models for coastal areas.

In conclusion, the literature survey highlights the diverse approaches and methodologies employed in flood prediction using machine learning models. From integrating emerging technologies to tailoring models for specific geographic regions, researchers are constantly innovating to improve the accuracy and reliability of flood prediction models. Future research in this area could focus on further refining these models and integrating them into real-time flood forecasting systems for more effective disaster management.

III. METHODOLOGY

Modules:

- Importing required Packages
- Exploring the dataset
- Data Processing - Using Pandas Data frame
- Visualization using seaborn & matplotlib
- Label Encoding using Label Encoder
- Feature Selection
- Train & Test Split
- Training and Building the model - Support Vector Classifier, KNN, Logistic Regression, Decision Tree Classifier, Random Forest.
- Trained model is used for prediction
- Final outcome is displayed through front-end

A) System Architecture

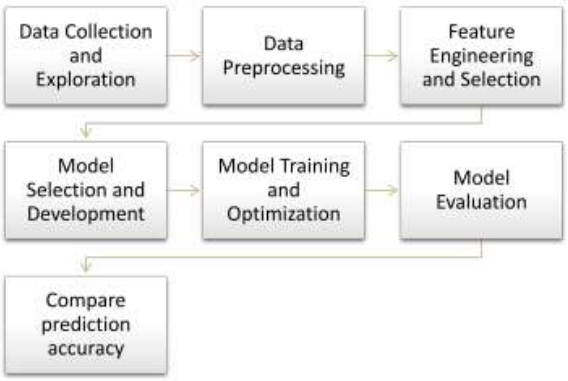


Fig 1: System Architecture

Proposed work

The introduction to this thesis outlines the paramount importance of accurate flood forecasting in ensuring community safety. Floods, as severe natural disasters, pose significant threats, causing widespread destruction, loss of life and profound economic impact. The unpredictability and sudden onset of floods makes accurate prediction critical for the development of effective mitigation strategies and timely response. Addressing the limitations of traditional forecasting methods, the introduction highlights the need to incorporate advanced technologies, particularly machine learning, to enhance forecasting capabilities. Traditional methods,

while valuable, may struggle to deal with the complexity and dynamic nature of the weather patterns that cause floods. Machine learning, with its ability to analyze extensive datasets, identify patterns, and adapt to emerging conditions, emerges as a promising solution to increase the accuracy and reliability of flood predictions.

B) Dataset Collection

The dataset used in this thesis comprises various types of data relevant to flood prediction and management. It includes historical flood data, which provides information on past flood events such as location, severity, and duration. Hydrological data, such as river flow rates and water levels, is also included to understand the behavior of water systems. Additionally, land-use patterns are considered to analyze how human activities impact flood risk. Satellite imagery is incorporated to provide spatial information and identify land cover changes that may affect flooding. For dynamic modeling, seasonal variability data is included to capture the influence of seasonal changes on flood patterns. This data may include information on seasonal precipitation, temperature, and weather patterns. Community engagement data is another component, encompassing surveys, interviews, or social media data to understand public perception and response to flood risk information. Real-time prediction and early warning system data involve real-time weather data, water level sensors, and other monitoring devices to enable timely flood predictions. Interdisciplinary collaboration data includes collaboration records between machine learning experts, hydrologists, meteorologists, and social scientists, indicating the extent of interdisciplinary research efforts. Impact assessment and adaptation strategies data consist of socioeconomic data related to flood impacts, such as property damage, displacement of populations, and economic losses. Continuous model updating data involves methods and records of updating models based on new environmental and climate data, ensuring the models remain accurate over time.

C) Pre-processing

Handling missing values and addressing seasonal variability are crucial steps in data preprocessing for climate change analysis. Missing values can occur due to various reasons, such as sensor errors or data collection issues. Imputing missing values using techniques like mean imputation, regression imputation, or using advanced methods like K-nearest neighbors can help maintain data integrity and consistency. Seasonal variability in climate data can be managed by using techniques such as seasonal decomposition, where the data is decomposed into trend, seasonal, and residual components. This allows for a clearer understanding of long-term trends and seasonal patterns in the data. Feature engineering plays a vital role in enhancing the representation of climate change in datasets. Creating new features based on domain knowledge, such as calculating temperature anomalies or aggregating data over different time periods, can provide valuable insights into climate change patterns. Employing feature selection techniques, such as recursive feature elimination or feature importance ranking, helps identify the most relevant predictors for climate change. This can improve model performance and interpretability by focusing on the most influential

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

features. In summary, addressing missing values, managing seasonal variability, and employing feature engineering and selection techniques are essential steps in preparing climate data for analysis, ensuring data integrity, and enhancing the representation of climate change patterns.

D) **Training & Testing**

To implement machine learning models for flood prediction, we first train and optimize each model using a prepared dataset. For K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR), we conduct hyper-parameter tuning to enhance their performance.

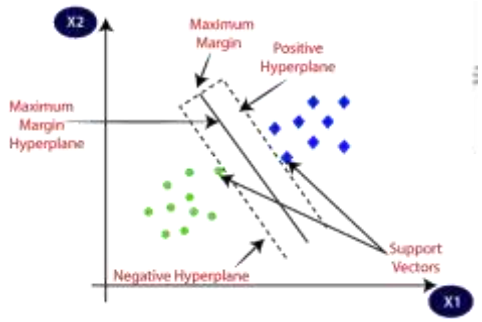
Next, we evaluate and validate these models using cross-validation techniques to ensure their robustness. Performance metrics such as precision, recall, and the Area Under the Receiver Operating Characteristic curve (AUC-ROC) are employed for evaluation.

Additionally, we compare the performance of these machine learning models against traditional flood forecasting methods to validate their effectiveness. By following this approach, we aim to build accurate and reliable models for flood prediction, which can potentially improve the efficiency of flood forecasting and mitigation efforts.

E) Algorithms.

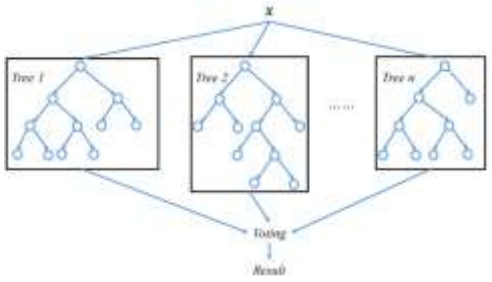
Support Vector Classifier:

A Support Vector Classifier (SVC) is a machine learning model that finds the best possible boundary (hyperplane) to separate different classes of data while maximizing the margin between them. It identifies key support vectors to make accurate classifications, making it effective for both binary and multi-class classification tasks.



Random Forest:

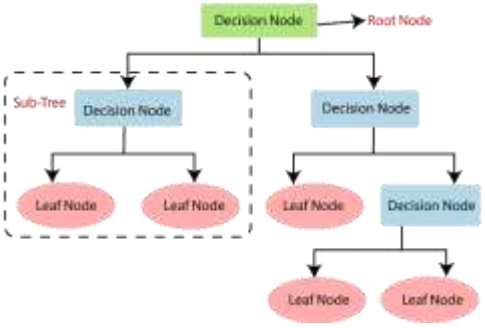
Random Forest [4] is an ensemble learning method that combines multiple Decision Tree [4]s to make predictions. It works by training a collection of Decision Tree [4]s on random subsets of the data and then averaging their predictions. This ensemble approach enhances accuracy, reduces overfitting, and provides robust performance for both classification and regression tasks.



Decision Tree:

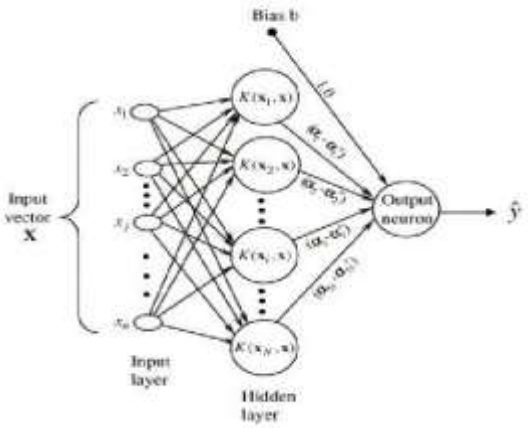
A Decision Tree [4] is a machine learning model that makes decisions by recursively splitting data into subsets based on the most significant feature, aiming to classify or predict outcomes. It creates a tree-like structure where each node represents a feature and each branch represents a possible decision, making it interpretable and effective for various tasks.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org



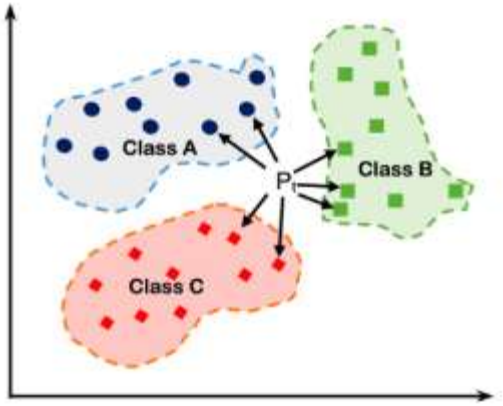
Logistic Regression:

Logistic Regression is a classification algorithm that predicts the probability of an input belonging to a specific category. It employs the sigmoid function to map the input features to a probability score between 0 and 1, and a threshold is applied to classify the input into one of two or more categories based on this probability. The model learns coefficients during training to best fit the data and make accurate classifications.



K-nearest neighbor:-

KNN, or the k-nearest neighbor algorithm, is a machine learning algorithm that uses proximity to compare one data point with a set of data it was trained on and has memorized to make predictions. This instance-based learning affords kNN the 'lazy learning' denomination and enables the algorithm to perform classification or regression problems. kNN works off the assumption that similar points can be found near one another — birds of a feather flock together. As a classification algorithm, kNN assigns a new data point to the majority set within its neighbors. As a regression algorithm, kNN makes a prediction based on the average of the values closest to the query point.



IV. EXPERIMENTAL RESULTS

A) Comparison Graphs → Accuracy, Precision, Recall, f1 score

Accuracy: A test's accuracy is defined as its ability to recognize debilitated and solid examples precisely. To quantify a test's exactness, we should register the negligible part of genuine positive and genuine adverse outcomes in completely examined cases. This might be communicated numerically as:

Accuracy = TP + TN TP + TN + FP + FN.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

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

Precision: Precision measures the proportion of properly categorized occurrences or samples among the positives. As a result, the accuracy may be calculated using the following formula:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: Recall is a machine learning metric that surveys a model's capacity to recognize all pertinent examples of a particular class. It is the proportion of appropriately anticipated positive perceptions to add up to real up-sides, which gives data about a model's capacity to catch instances of a specific class.

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

F1-Score: The F1 score is a machine learning evaluation measurement that evaluates the precision of a model. It consolidates a model's precision and review scores. The precision measurement computes how often a model anticipated accurately over the full dataset.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

B) Performance Evaluation Graph.

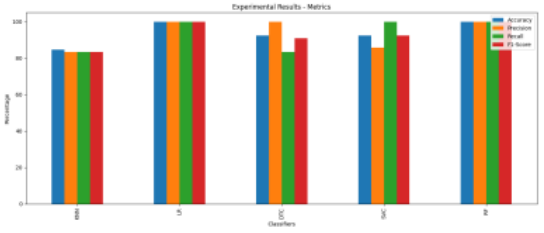


Fig 2: Bar Plot of Result for Five different Classifiers

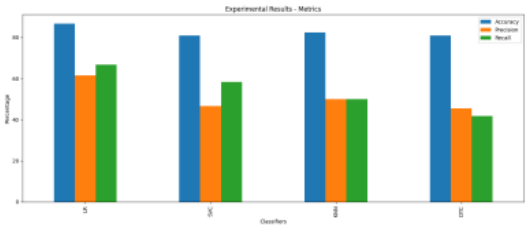


Fig 3: Bar Plot of Base Paper Result for Four different Classifier

C) Observations

In this observational study, we explore how different machine learning tools perform. These tools are like assistants that make predictions or decisions based on data. We are looking at K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree Classifier (DTC), Support Vector Classifier (SVC), and Random Forest (RF). The goal is to understand what these tools are good at, where they may have

difficulty, and when to use them. We'll examine things like how easy it is to understand their predictions, how fast they work on computers, and how well they handle different types of data. By comparing their strengths, weaknesses, and capabilities, we want to help people choose the right tool for their specific data and project goals. This analysis will focus on how easy it is to explain the tool's decisions, how quickly it can process information, and how well it can handle different types of data setups. This way, we can provide useful information to help make smart choices in various machine learning projects.

V. CONCLUSION

In conclusion, this thesis used machine learning to advance flood prediction methods, using well-prepared datasets of weather-related variables. Models such as K-Nearest Neighbors, Support Vector Classifier, Random Forest, Decision Tree, and Logistic Regression demonstrated their potential in flood prediction through rigorous development and optimization. The evaluation and validation processes provided valuable insights using performance metrics for comparative analysis. The emphasis on model interpretation and effective communication underlined the importance of linking technological advances with community understanding. Validation against traditional methods confirmed the reliability of the proposed machine learning model. Nevertheless, the limitations of the study, particularly in representing complex flood dynamics, require future research to increase accuracy by addressing these challenges and exploring additional contributing factors.

VI. FUTURE SCOPE

This thesis establishes a foundational framework for advancing flood prediction through machine learning, identifying key avenues for future research and development. Integrating additional data sources, such as hydrological data, land-use patterns, and satellite imagery, holds promise for enhancing the generality and accuracy of predictive models. Dynamic modeling techniques capable of adapting to seasonal variability and changing weather patterns are essential for robust predictions under diverse conditions. Community engagement is crucial, and exploring innovative strategies to promote understanding of flood risk information and proactive measures can enhance community resilience. Real-time prediction systems, integrating machine learning models into early warning systems, offer potential for timely and effective responses to flood events. Interdisciplinary collaboration among machine learning experts, hydrologists, meteorologists, and social scientists is essential for developing holistic flood prediction models that consider technical, environmental, and social factors. Impact assessments and adaptation strategies, including optimization strategies based on model results, are necessary to minimize risks and optimize resource allocation. Continuous model updating methods are also vital to adapt to changing environmental and climate conditions, ensuring sustained accuracy over time.

REFERENCES

[1] S.M.ASCEA. Anisha, K. M. K. E. K. D. A. S. A. K. S. R. S. M.-A. and Davis, R. . 2018 kerala flood damage: Survey, identification, and damage

prediction models using machine learning. In American Society of Civil Engineers. National Institute of Technology, Calicut, 2023.

[2] Fitri YakubAinaa Hanis Zuhairi, S. A. Z. M. S. M. A. . Review of flood prediction hybrid machine learning models using datasets. In IOP Conf. Series: Earth and Environmental Science. Malaysia-Japan International Institute of Technology, 2022.

[3] B.Yashwanth SaiD.Keerthi Reddy, S. P. K. D. K. . Flood prediction using machine learning. In International Journal of Multidisciplinary Research in Science, Engineering, Technology Management (IJMRSETM), pages 1395–1402. INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH, 2023.

[4] al, K. M. et . An integrated approach for flood prediction by using block chain network and machine learning. IOP Conf. Series: Materials Science and Engineering, 2016.

[5] Dinh Kha DangHuu Duy Nguyen, Y. N. N. C. P. V. T. T. V. N. Q.-H. N. X. L. N. L. T. P. V. T. P. and Bui, Q.-T. . ntegration of machine learning and hydrodynamic modeling to solve the extrapolation problem in flood depth estimation. In Journal of Water and Climate Change. University of Science, Vietnam National University, 2023.

[6] M. A. Habib, J. O. and Salauddin, M. . Prediction of wave overtopping characteristics at coastal flood defences using machine learning algorithms: A systematic rreview. In IOP Conf. Series: Earth and Environmental Science. UCD Dooge Centre for Water Resources Research, UCD School of Civil Engineering, and UCD Earth Institute, University College Dublin, Dublin, Ireland, 2022.

[7] Miguel de Castro NetoMarcel Motta, P. S. . A mixed approach for urban flood prediction using machine learning and gis. In International Journal of Disaster Risk Reduction, pages 350–402. INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH, 2021.

[8] Maisha FarzanaMiah Mohammad Asif Syeed, I. N. I. I. M. H. N. T. R. . Flood prediction using machine learning models. In ResearchGate, pages 1395–1402. BRAC University, 2022.

[9] Hayati YassinZaharaddeen Karami Lawal, R. Y. Z. . Flood prediction using machine learning models: A case study of kebbi state nigeria. In Conference Paper. Department of Computer Science Federal University Dutse Dutse, Nigeria, 2021.