International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Rainfall Based Flood Prediction in Kerala Using Machine Learning

Vidya S[1*], Gayathri S[2], Dilsha M[3], Krithika S[4]

Submitted: 08/12/2023 **Revised**: 15/01/2024 **Accepted**: 29/01/2024

Abstract: Among the most damaging natural disasters are floods, which are very difficult to model. Flood prediction is a critical task that involves forecasting the likelihood of floods in a given area, allowing people to take necessary precautions to minimise damages and prevent loss of life. Machine learning (ML) algorithms have shown great potential in flood prediction, as they can analyse large amounts of data from multiple sources to provide accurate and timely predictions. The objective is to find a prediction model that is more accurate and efficient by incorporating new machine learning techniques and hybridising current ones. Both hydrologists and climate scientists can use this model as a guidance for selecting the appropriate machine learning technique for a given prediction problem. The output of the ML-based flood prediction system can also be integrated with existing flood warning systems, enabling authorities to send out alerts in a timely manner and take necessary precautions to minimise the effects of flooding.

Keywords: *Deep Learning, Flood Prediction, Hydrological model, Machine Learning, Statistical analysis.*

1. Introduction

Flood prediction has its roots in the scientific study of hydrology, meteorology, and climatology. The development of flood prediction methods and techniques has evolved over time as our understanding of these fields has improved [1]. Flood prediction models allow authorities to take proactive measures to mitigate the impact of floods, such as issuing warnings and evacuation orders, building flood protection structures, and mobilising rescue operations.

Today, flood prediction models are an essential tool in managing flood risks and protecting communities and infrastructure from the devastating impacts of flooding. With the ongoing advancements in technology and data science, flood prediction models are becoming more sophisticated and accurate, allowing for better planning and preparedness for future flood events [2].

The development of flood prediction models has also been driven by the need to understand and manage the complex interactions between various factors that contribute to flooding, such as rainfall, river flow, topography, soil moisture, and land use. By analysing and modelling these factors, flood prediction models can provide accurate and reliable forecasts of flood events, enabling authorities to make informed decisions and take appropriate actions [3]. This study explores the application of machine learning techniques to predict floods in the state of Kerala, India, using historical rainfall data spanning the years 1901 to 2018.

2. Methodology

The proposed system has the dataset from the preceding years in several areas of Kerala that are prone to flooding, offering details on several factors such as the usual amount of rainfall, the duration of the rainfall, etc. Our dataset contains the monthly rainfall index and annual rainfall index of Kerala from 1900-2018. The methodology is designed to create robust models for predicting flood occurrences in Kerala through a systematic process. It includes crucial steps such as data cleaning, feature engineering, exploratory data analysis, and algorithm selection. The algorithms used are:

2.1 LSTM based RNN

The response is provided by recurrent neural networks (RNNs), yet this research chose to use long short-term memory (LSTM) because RNNs have shorter memories [4]. As shown in Fig. 1, with its ability to manage long-range dependency, LSTM is primarily designed to be applied to temporal modelling sequences. As learning long-term dependencies, one of the main mathematical challenges is the vanishing gradient problem, which arises as the gradients to the first few input points approach zero [5]. The LSTM activation function handles this, preventing the backpropagated gradient from exploding or disappearing. Because of feedback loops, recurrent neural networks are known to be built for temporal sequences like those in rainfall prediction, which involve dynamic series data sequences [6]. These feedback loops work as memory units by generating feedback from the output and passing it back as input for additional processing. This information relates to the prior activation function. The recurrent neural network can handle temporal and dynamic sequential data thanks to this repetitive process. Memory blocks with gates that control the network's internal operations take the place of the hidden units in LSTM [7].

1Associate Professor, Sri Sai Ram Institute of Technology, Chennai, India
ORCID ID: 0000-0003-0962-8411
2U.G. Student, Sri Sai Ram Institute of Technology, Chennai, India
ORCID ID: 0009-0002-3885-5242
3U.G. Student, Sri Sai Ram Institute of Technology, Chennai, India
ORCID ID: 0009-0007-1808-1514
4U.G. Student, Sri Sai Ram Institute of Technology, Chennai, India
ORCID ID: 0009-0003-8248-5688
**Corresponding Author E-mail: vidya.cse@sairamit.edu.in*

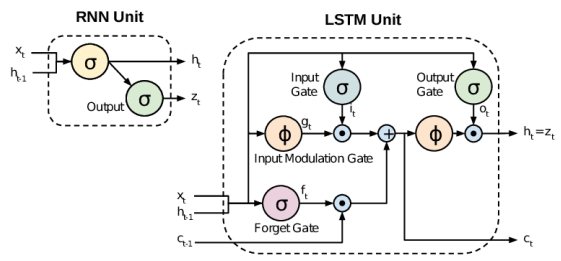

Fig. 1. LSTM based RNN

2.2 Logistic Regression

As shown in Fig. 2, Logistic regression is a binary classification algorithm that is well-suited for problems where the output variable is a binary (0/1) variable. In the case of flood prediction, we are interested in predicting, depending on a collection of input features, whether or not a flood will occur [8]. Logistic regression can be used to build a binary classification model that predicts the occurrence of a flood based on environmental and weather conditions [9]. Logistic regression can be used as a baseline model to evaluate the effectiveness of additional, more advanced machine learning algorithms for flood prediction. It can also be used in combination with other algorithms to improve the overall accuracy of the predictions [10].

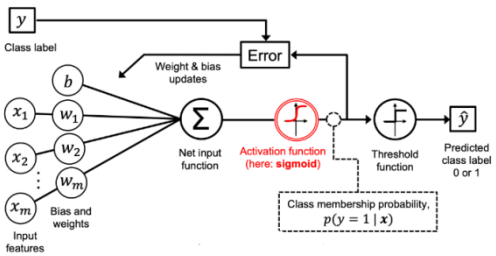

Fig.2. Logistic Regression

2.3 Decision Tree

The decision tree algorithm is a machine learning algorithm that uses a tree-like structure, as shown in Fig. 3, to simulate choices and their possible outcomes. It is used in flood prediction by analysing historical data and relevant features to make predictions about the occurrence or severity of floods. In a dataset, decision trees can be used to determine which variables or characteristics are most significant [11]. By examining the splitting criteria and feature importance measures of a decision tree, one can determine the relative significance of various features and choose the most informative ones for further analysis or modelling. Decision trees can be used for exploratory data analysis and offer a visual picture of the decision-making process. They allow users to understand the relationships between features and their impact on the outcome or class prediction [12]. Decision trees can help uncover patterns, interactions, and decision rules within the data.

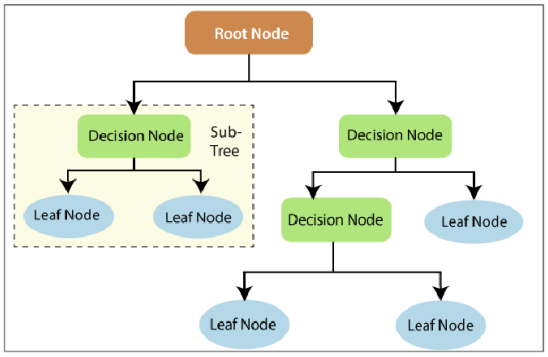

Fig.3. Decision Tree

2.4 Support Vector Machine

Support Vector Machine (SVM), as shown in Fig. 4, is a supervised machine learning approach that may be applied to regression and classification problems. It works especially well when handling complicated datasets and is commonly used in various domains, including flood prediction [13]. Historical data related to floods, including features such as rainfall patterns, river levels, soil moisture, and topography, is collected and pre-processed [14]. Relevant features are selected or engineered to represent the patterns and characteristics of flood events. The dataset is divided into training and testing subsets [15]. The SVM algorithm is applied to the training dataset, using flood occurrence as the target variable. SVM searches for the optimal hyperplane that can separate flood and non-flood instances with the maximum margin, or in the case of non-linear data, it is mapped into a higher-dimensional space using kernel functions [16]. The trained SVM model can be used to predict the likelihood of a flood event for new instances based on their features. The predicted outcomes can be binary (flood or non-flood) or continuous (indicating the severity or probability of flooding) [17].

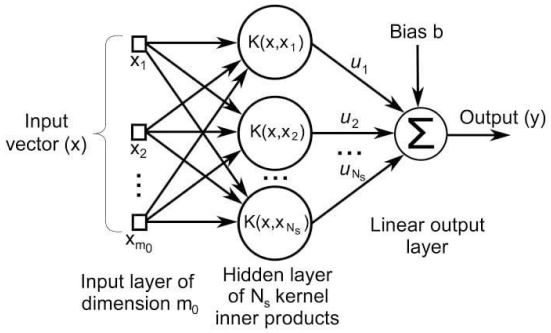

Fig.4. Support Vector Machine

2.5 Random Forest

An ensemble learning technique called the Random Forest algorithm combines several decision trees, as shown in Fig. 5, to develop a forecasting model that is more reliable and accurate [18]. The algorithm introduces randomness in two key ways: (i) Random Sampling: Bootstrapping, sometimes referred to as bagging, is the process of training each tree using a random subset of the training set. This helps to create diversity among the trees. (ii) Random Feature Selection: Only a portion of the features are taken into consideration for splitting at each decision tree node. This decreases connection between trees and increases variety even

more [19]. Based on its internal structure and the features it has taken into consideration, each decision tree in the Random Forest independently predicts the class or value. In classification problems, the Random Forest's ultimate forecast is decided by the individual trees voting together in a majority manner. The expected class is determined by calculating the votes cast in the class. The ensemble of decision trees in Random Forest provides improved accuracy, better generalisation, and reduced over fitting compared to a single decision tree [20].

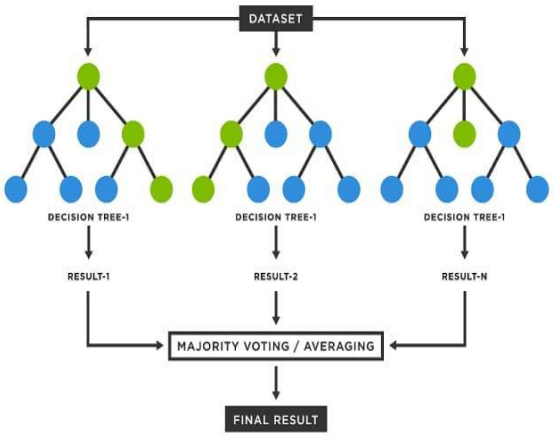

Fig.5. Random Forest

3. Evaluation metrics

3.1 Accuracy

This measures the correctness of a model's predictions. It is frequently applied to classification issues, where the objective is to choose a label from a list of potential labels for each input. The ratio of accurately predicted instances to the total number of instances is used to calculate accuracy. The formula for accuracy is in (1).

$$\text{Accuracy} = \text{Number of Correct Predictions} \div \text{Total Number of Predictions} \quad (1)$$

3.2 Recall

This calculates the proportion of accurately predicted positive occurrences among all positive examples that actually occur. Recall is a measure of the model's accuracy in identifying positive cases. The formula for Recall is in (2).

$$\text{Recall} = \text{TP} \div (\text{TP} + \text{FN}) \quad (2)$$

3.3 F1 Score

This is the model's overall performance measured as the harmonic mean of precision and recall. High recall and precision are indicated by a high F1 score. The formula for F1 score is in (3).

$$F1\ Score = \text{TP} \div (\text{TP} + (1 \div 2\ (\text{FP} + \text{FN}))) \quad (3)$$

3.4 ROC Curve

As shown in Fig. 6, the true positive rate (TPR) at different categorization levels is plotted versus the false positive rate (FPR). It is employed to assess the model's performance across several criteria.

3.5 AUC Score

The ROC curve's area under the curve is used to compare the effectiveness of various models. A higher AUC score denotes a better model overall performance.

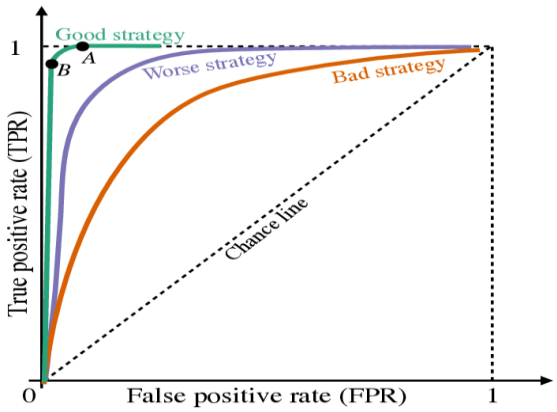

Fig.6. ROC Curve

Comparison of different algorithms using evaluation metrics are shown in Fig. 7 and Fig. 8:

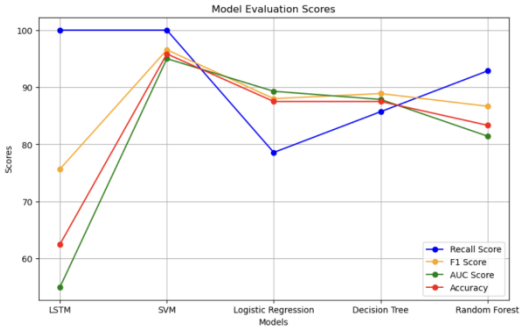

Fig.7. Comparison of different ML algorithms

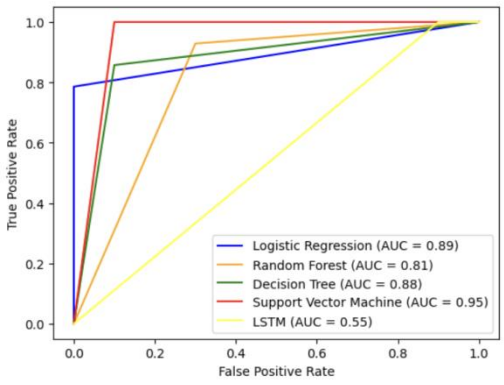

Fig.8. Comparison of ROC Curves for different algorithms

4. Conclusion and Future Directions

In conclusion, this research has undertaken a thorough exploration of machine learning applications for flood prediction in Kerala, employing a diverse set of algorithms such as LSTM-RNN, Random Forest, SVM, Decision Tree, and Logistic Regression. While the machine learning model's flood prediction performance has been promising, there is still a need to enhance the accuracy and reliability of these models. This can be accomplished by combining more historical data, real-time monitoring, and sophisticated algorithms and approaches.

References

[1] Agnibha Sarkar, Dr. Ashutosh M. Kulkarni, Manish R. Khodaskar, Shubhangi Pandurang Tidake, Rahul B. Diwate : "A Predictive Model for Occurrence of Floods Using Machine Learning Techniques", Social Science Journal, vol.13, no.2, January Issue 2023.

[2] Thegeshwar Sivamoorthy, Asif Mohammed Ansari, Dr. B. Sivakumar, V. Nallarasan : "Flood Prediction Using ML Classification Methods on Rainfall Data", International Journal for Research in Applied Science & Engineering Technology (IJRASET), DOI: https://doi.org/10.22214/ijraset.2022.41297

[3] Kishan Kashyap M, Ajay Karthik K, ShivaKumar G.S : "Flood Prediction using Machine Learning", International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 07

[4] Jeba. G. S, P. Chitra, U. M. Rajasekaran : "Time-series analysis and Flood Prediction using a Deep Learning Approach", International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), 2022. https://doi.org/10.1109/WiSPNET54241.2022.9767102

[5] Widiasari. I. R, L. E. Nugoho, R. Efendi : "Context-based Hydrology Time Series Data for A Flood Prediction Model Using LSTM", 5th International Conference on Information Technology, Computer, and Electrical Engineering, 16th Sept.2018. https://doi.org/10.1109/ICITACEE.2018.8576900

[6] Babar. M, M. Rani, I. Ali : "A Deep learning-based rainfall prediction for flood management", 17th International Conference on Emerging Technologies (ICET), 2022. https://doi.org/10.1109/ICET56601.2022.10004663

[7] Wang. S, Wang. J : "Research on prediction model of mountain flood level in small watershed based on deep learning", 4th International Conference on Intelligent Control, Measurement and Signal Processing (ICMSP), 2022. https://doi.org/10.1109/ICMSP55950.2022.9859047

[8] Saiesh Naik, Anang Verma, Srushti Ashok Patil, Prof. Anil Hingmire :"Flood Prediction using Logistic Regression for Kerala State", International Journal of Engineering Research & Technology (IJERT), Volume 9, Issue 3

[9] Jaeyeong Lee, Byunghyun Kim : "Scenario-Based Real-Time Flood Prediction with Logistic Regression", Water 2021, 13, 1191, https://doi.org/10.3390/w13091191

[10] Muhammad Ikbal Hidayat Jati, Suroso and Purwanto Bekti Santoso : "Prediction of flood areas using the logistic regression method (case study of the provinces Banten, DKI Jakarta, and West Java)", IOP Publishing, Journal of Physics: Conference Series, DOI 10.1088/1742-6596/1367/1/012087

[11] Naveed Ahamed, S.Asha : "Flood prediction forecasting using machine Learning Algorithms", International Journal of Scientific & Engineering Research Volume 11, Issue 12, December-2020

[12] Mr. B. Samuel John Peter, Mr. N. Thilak Chandhra, Mr. Shaik.Afrid, Mr. N. Prasanna Kumar : "MACHINE LEARNING BASED FLOOD PREDICTION", International Journal of Research in Engineering, IT and Social Sciences, Volume 12 Issue 12, December 2022

[13] Hussein. E, M. Ghaziasgar, C. Thron : "Regional Rainfall Prediction Using Support Vector Machine Classification of Large-Scale Precipitation Maps", IEEE 23rd International Conference on Information Fusion (FUSION), 2020. https://doi.org/10.23919/FUSION45008.2020.9190285

[14] Samantaraya. S, A. Sahoo, A. Agnihotri : "Prediction of Flood Discharge Using Hybrid PSO-SVM Algorithm in Barak River Basin", MethodsX Volume 10, 2023. https://doi.org/10.1016/j.mex.2023.102060

[15] Shada. B, N. R. Chithra, S. G. Thampi : "Hourly Flood Forecasting Using Hybrid Wavelet-SVM", Journal of Soft Computing in Civil Engineering, 2022. https://doi.org/10.22115/SCCE.2022.317761.1383

[16] Nadia Zehra : "Prediction Analysis of Floods Using Machine Learning Algorithms (NARX & SVM)", International Journal of Sciences: Basic and Applied Research (IJSBAR) (2020) Volume 49, No 2, pp 24-34, https://core.ac.uk/download/pdf/287366682.pdf

[17] Shie-Yui Liong, Chandrasekaran Sivapragasam : "FLOOD STAGE FORECASTING WITH SUPPORT VECTOR MACHINES", JAWRA JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION, https://doi.org/10.1111/j.1752-1688.2002.tb01544.x

[18] Grady. F, J. K. Tarigan, J. R. Wahidiyat, A. Prasetyo : "Classification of Flood Alert in Jakarta with Random Forest", IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA), 2022. https://doi.org/10.1109/ICITDA55840.2022.9971411

[19] Choi. C, J. Kim, H. Han, D. Han, H. S. Kim : "Development of Water Level Prediction Models Using Machine Learning in Wetlands: A Case Study of Upo Wetland in South Korea", Water, vol. 12, no. 1, p. 93, Dec. 2019. https://doi.org/10.3390/w12010093

[20] Esfandiari. M, S. Jabari, H. Mcgrath, D. Coleman: "Flood mapping using random forest and identifying the essential conditioning factors; a case study in Fredericton, New Brunswick, Canada", ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., V-3-2020, 609–615, 2020. https://doi.org/10.5194/isprs-annals-V-3-2020-609-2020