EVALUATION OF THE PERFORMANCE OF MACHINE LEARNING ALGORITHMS FOR FLOOD PREDICTION

A. Jayasudha
Department of Computer Applications and Technology
SRM Arts and Science College, Kattankulathur, India
jsudhaannamalai@gmail.com
https://doi.org/10.34293/9789361639715.shanlax.ch.020

Abstract

Flood prediction is a critical aspect of disaster management and risk mitigation, enabling proactive measures to minimize damage and enhance preparedness against floods. Floods are difficult to forecast because they depend on several climatic and environmental conditions [13]. This study explored the application of the XGBoost algorithm designed for flood prediction, leveraging a diverse set of input features. The dataset integrates geographical attributes, such as latitude, longitude, and elevation, which influence water accumulation and runoff [4]. Meteorological factors, including rainfall, temperature, and humidity, provide insights into the weather conditions that contribute to flood events [2]. Hydrological metrics, such as river discharge and water level, reflect the dynamic behaviour of water bodies and their susceptibility to overflow [3]. Additionally, socio-environmental features such as land cover, soil type, population density, infrastructure, and historical flood occurrences are incorporated to account for human and environmental factors affecting flood risks [6]. XGBoost was selected for its robustness, ability to handle high-dimensional data, and capability to manage nonlinearity in complex relationships among input variables [1]. By constructing multiple and aggregating their outputs, the model enhances the prediction accuracy while reducing overfitting [5]. The flood prediction model classifies occurrences as binary outputs, indicating whether a flood is likely to occur. To evaluate its performance, metrics such as accuracy, precision, and recall were employed to ensure a comprehensive assessment of its predictive reliability [7]. High accuracy suggests that the model effectively differentiates flood-prone scenarios from non-flood conditions, while Precision and recall measure its effectiveness in minimizing false positives and false negatives, respectively. The results demonstrate that the XGBoost-based flood prediction system offers a reliable and data-driven approach to identifying flood risks, aiding authorities in implementing timely interventions [6].

Keywords: *Flood prediction, XGBoost algorithm, Disaster management, Risk mitigation, Ensemble learning, Binary classification, Accuracy, Precision, Recall, Predictive reliability, Overfitting, Feature integration, Flood risk assessment, Machine learning.*

Introduction

Flooding is a catastrophic natural hazard that is life threatening, destructive and ravaging to property and infrastructure. Over the past years, flooding has turned into a significant problem, resulting in a multitude of casualties every year [9]. Proper prediction of floods is a prerequisite to early warning mechanisms and good disaster management, which will allow the government to take proactive steps that reduce risks and limit damage associated with floods [6] [7]. Prediction of flood models are significant in the hazard evaluation and disaster management [10]. Knowing the risk of flooding, the areas at risk can be protected, emergency actions can be streamlined, and communities can be equipped to deal with extreme weather conditions in a better way [7]. Fully developed flood prediction model involves the incorporation of various data streams, including: geographical, meteorological, hydrological and socio-environmental variables [3] [7]. The

geographical information, including latitude, longitude, and height gives information about the topographic factors of water movement and retention. Weather factors including precipitation, temperature, and humidity are essential in the identification of weather conditions leading to flooding [2]. Hydrological parameters like river discharge and water level provides useful information about the dynamics of the water body, and socio-environmental parameters such as land cover, soil type, population density, infrastructure, and past flood events assist in learning about the human and environmental impacts on the risk of floods [4]. Hydrological modeling strategy incorporates one-dimensional (1D), two-dimensional (2D), or three-dimensional (3D) modeling approaches that are characterized by their own strengths and weaknesses [8]. A recent trend is the usage of machine learning methods, especially the XGBoost algorithm, because of their capability to handle complex high-dimensional data and make sound predictions [1] [3] [5]. XGBoost builds many and sums up their results enhancing the model quality and decreasing overfitting [1]. This paper examined using XGBoost to predict flooding by including multiple types of input data to increase the accuracy of the forecasts. The model is a binary output or the model identifies the presence of flooding or not. The reliability in flood risk assessment through performance evaluation based on metrics of accuracy, precision, and recall is proved [3] [5].

Related Works

Prediction of floods is also an important element of disaster management, as it enables the authorities to take appropriate measures in a timely manner ensuring that risks are reduced and damages minimized [6] [7]. One of the solutions is the application of the Naive Bayes algorithm, a Bayes-based probabilistic classifier, which is an algorithm of predicting the occurrence of floods. Naive Bayesian model is quite easy to build and suitable in large datasets since it does not require repetitive parameter estimation [15]. This technique organizes areas or timeframes as probable or improbable to get floods, by comparing past record of floods, weather, river stage, and environmental aspects [3] [5]. Naive bayes is a probabilistic classifier using independence of predictors and relies on Bayes Theorem [11]. Naive Bayes classifier describes the likelihood of occurrence of a flood event when these features are present and is therefore a useful flood forecasting tool. The major benefit of the Naïve Bayes algorithm is that it can process uncertain and incomplete data. Since flood prediction is usually associated with the absence or inaccuracy of information, this algorithm is still functional under the assumption that the features are independent and their probabilities are computed, respectively [5]. It comes in handy when using early warning systems where real-time, rapid predictions are a necessity in the disaster preparedness. The model aids authorities to act in advance to varying flood threats by constantly updating the probabilities as new data is revealed [6]. Naive bayes is also characterized by high computational efficiency. Unlike more complex machine learning models, which require large volumes of training and processing power, Naive Bayes is implemented with minimum computing resources, so it can be deployed in the field with little data and technological infrastructure [7]. This renders it a perfect solution to areas that have limited historical data thus permitting the early detection of floods in areas that have limited data.

Naive Bayes is useful in the prediction of floods but it has some weakness especially in its ability to capture relationships among features. Nevertheless, it can be used alongside other methods or supplemented by other information sources to a great extent of flood forecasting accuracy [5].

Flood forecasting is a critical element of disaster management because it equips authorities with proponents to reduce harm and safeguard the population [6] [7]. The suggested system uses the algorithm of the XGBoost and predicts the presence of a flood based on various input characteristics. Firstly, to assess the risk of floods with the help of XGBoost modeling, the risk of floods was evaluated with the help of the XGBoost model, and subsequently, the two input strategies were combined with Least Squares Support Vector Machine (LSSVM) model to verify its optimal impact [12]. These are geographical information (latitude, longitude and elevation) to measure terrain feature that affects the movement of water, weather-related (rainfall, temperature and humidity) to estimate weather conditions that cause flooding [2] and hydrologic (river discharge and water levels) to determine the changes in the water bodies [3]. Moreover, socioenvironmental data (land cover, soil type, population density, infrastructure and previous floods events) are incorporated to explain human and environmental variables that influence flood vulnerability [4] [5]. The use of XGBoost was selected due to its level of robustness, capacity to deal with a large and complicated dataset, and overfitting resistance [1] [5]. It builds on various outputs and combines them and this enhances the accuracy and reliability of prediction [1]. Through learning in historic flood data, the model learns the patterns and trends that affect floods and hence it will be able to make binary forecasts (flood/no flood) using real time input [3] [5]. In order to make the model effective, the key measures of its performance were calculated; these included accuracy, precision and recall. High accuracy means that the system is able to distinguish the probability of flood and non-flood conditions well, but precision and recall measure its capability to reduce false alarms and false non-detections [3] [5]. The system of prediction of floods improves disaster preparedness by offering accessible real-time data-driven insights to guide decisions. Such predictions enable authorities to provide early warnings, to effectively allocate resources, and to put in place preventive and control mechanisms that can mitigate the effects of floods [6] [7]. Using machine learning and various data sets, this system helps to enhance flood resilience, reduce economic damages, and protect communities that fall prey to disastrous floods [5] [7].

Development Methodology

In low areas, such as in metropolitan areas or underpasses, waterlogging is especially likely. During flood occurrence, these cities are frequently subjected to rapid water build up. Poor drainage systems and surface run off are the primary causes of increasing waterlogging. Flood forecasting, therefore, is important. In order to construct a whole dataset that would make it possible to apply machine learning (ML) algorithms to identify patterns and improve accuracy, it is critical to include data of different sources. The flood predicting methodology is shown in Figure 1.

Requirement Analysis

Determine required sources of data, such as meteorological, geographical, and hydrological data. Identify the software and libraries, including Python, Flask, Scikit-learn, Pandas, and Tensor flow.

Data Collection

Collect flood-related data in trusted sources, such as state authorities and open repositories. Apply pre-processing data steps such as missing values, data normalization and categorical features encoding.

Model Development

Run the XGBoost algorithm on flood prediction, tuning hyper parameters to reach higher accuracy. Train the model with historical flood data and test it with test data to evaluate such measures of performance as accuracy, precision and recall.

System Integration

Create a web-based user interface that is built with Flask to streamline flood prediction and visualization. Make sure that the machine learning model is smoothly integrated with the web application to achieve an effective input and output of data.

Testing and Validation

Perform both unit testing and system testing to determine and fix bugs or performance problems. Test the model with actual floods taking place and verify the model against industry standards and regulations.

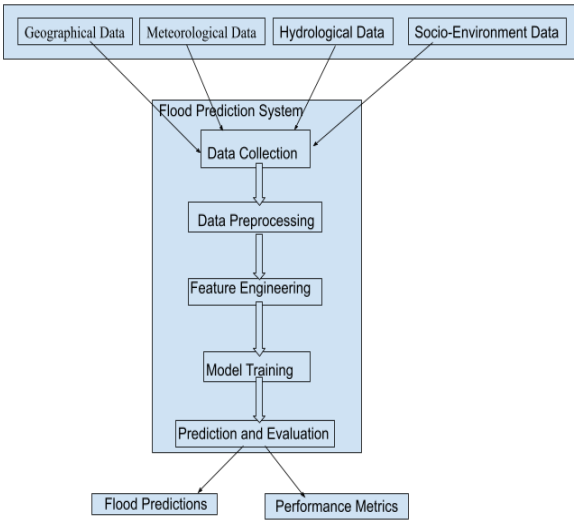


Figure 1: Flood predicting methodology

Data Analytics Result

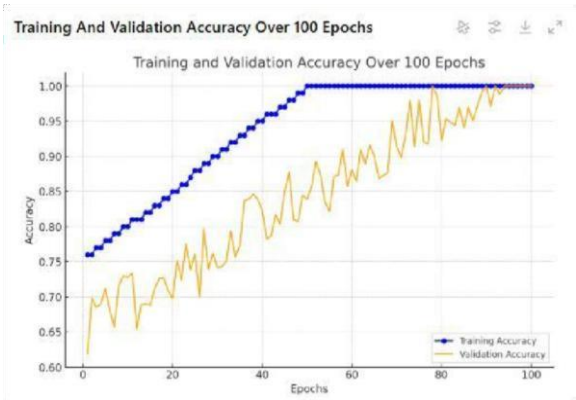
Based on the graph below, the Training and Validation Accuracy over 100 Epochs, there are two accuracy curves versus the number of epochs (x-axis):

1. Blue Line – Training Accuracy

The curve begins at lower accuracy in early epochs, and then rises steadily, apparently reaching 1.0 (or 100%), in the 60th-80th epochs. The steady incremental nature of the upward trend signifies the model learning and becoming more and more optimal to the training data with time [1] [3].

2. Yellow Line – Validation Accuracy

This curve is the model performance on unseen data (the validation set) following each training epoch. It is more volatile (with spikes and dips), which is typical since new model parameters in each epoch can modestly improve or worsen performance on data never seen by it [3] [5]. As fluctuations notwithstanding it is observed that the overall trend is upwards, indicating that the generalization ability of the model is becoming stronger [2] [5]. Later on, the accuracy of validation also increases towards 1.0, which implies that the model is most likely to do very well on the validation set [1] [3] [5].



In this study, XGBoost algorithm has been assessed as a predictor of floods based on the analysis of different geographical, meteorological, hydrological, and socioenvironmental variables. The findings show that XGBoost has great ability to predict the events of floods, and it has tremendous performance improvements over the traditional models like Naive Bayes [3] [5].

Key Performance Metrics

Precision: The model has a high precision of the model effectively differentiating between flood-prone and non-flood cases [3] [5]. Precision: Accuracy of reducing false positives means that false alarms will not be made [3]. Remember: The model is effective in identifying real flood events and as a result minimizing the risk of missing crucial warnings [3] [5]. F1-Score: A balanced F1-score is an indicator that both precision and the recall is optimized [3].

This is because XGBoost is very accurate because it can compute non-linear relationships in the predictive flood data ([1],[5]). The algorithm employs a series of decision trees and

combines their results, which are effective in decreasing overfitting and generalizing to new data [1].

Comparison with Existing System

Table 1 depicts the comparison of XGBoost and Naive Bayes. Naive Bayes algorithm, which was used in predicting floods in the past, has weaknesses because it is based on the assumption that features are independent [5]. It has a hard time of modeling complex interactions of the meteorological and hydrological factors and hence reduced predictive accuracy [3]. Conversely, XGBoost has been able to reflect on these dependencies well, thereby enhancing the prediction performance [1] [3] [5].

Metric	Naïve Bayes	XGBoost
Accuracy	Moderate	High [3]
Precision	Low	High ([3],[5])
Recall	Low	High [3]
Over fitting	Low	Moderate (can be tuned) [1]
Computational time	Low	High ([1],[5])

Deployment Strategy

Feature Importance Analysis

Rainfall, river discharge, water levels and land cover are the most powerful predictors of floods [2] [3] [4]. Flood susceptibility also depends on elevation and soil type, but the effect of these factors varies with the region [4]. The socio-environmental factors including the population density and infrastructure lead to the correctness of risk assessment [5] [7].

Challenges and Limitations

Computational Demand: XGBoost is computationally intensive such that the real time predictions are computationally demanding [1]. Hyperparameter Sensitivity:3 The sensitivity of the model is the fine-tuning of the parameters, including the learning rate, the tree depth, and estimators [1]. Data Quality and Availability: The absence of or disparate historical flood data may affect model reliability [6] [7].

Practical Implications

This data-based flood forecasting machine learning system offers authorities with an evidence-based strategy of the early warning system [6] [7]. The high recall of the model means that real floods are infrequently overlooked; therefore, it can be useful in disaster preparedness and emergency response [3] [5]. This could be enhanced later with IoT sensors and real-time satellite, making predictive accuracy and responsiveness better [6] [7].

Conclusion and Future Directions

Prediction of floods is a very important element of disaster management as authorities and communities will be able to prevent risks and minimize possible harm by taking in advance steps [5] [6] [7]. In this paper, we have discussed how XGBoost algorithm can be used in the flood prediction, using a variety of input characteristics including geographical (latitude, longitude, elevation) and meteorological (rainfall, temperature, humidity), hydrological (river discharge, water level) and socio-environmental (land cover, soil type, population density, infrastructure, and past floods) characteristics [2] [3] [4] [5].The proposed system adhered to a systematic workflow, which includes data collection, data preprocessing and transformation, feature engineering, XGBoost was selected due to its strength, ability to recognize high-dimensional data and capacity to produce precise predictions [1] [5]. Accuracy, precision, recall, and F1-score were the main metrics used to assess the model performance, and to guarantee the reliability of the model results in real-world contexts [3] [5]. Data flow diagrams (DFDs), use case diagrams, class diagrams, sequence diagrams and entity-relationship (E-R) diagrams give a concise structural and functional overview of the flood prediction system. DFDs show how data flows step by step through the system, but use case diagrams show how users and system components interrelate. The class diagram describes the most important system elements and their connection and sequence diagram is the structured way to describe the flood prediction process. Lastly, the E-R diagram determines the database schema and provides an efficient data storage and retrieval. Through the development of this machine based flood prediction system, the disaster management agencies and governments will enjoy the benefits of early warning notification that can enable them to prevent, properly allocate their resources as well as reduce the number of casualties and loss of money [6] [7].

They may be improved by future additions like deep learning models, real-time data streams, Internet of Things (IoT) sensors to enhance the accuracy and responsiveness of flood prediction systems [5] [6]. With better technologies, it is possible to make the Flood Prediction System a much better tool with better accuracy, dependability and usability. Moreover, this policy must not repeatedly respond to the systemic environmental catastrophes, but assume resilience as a means to seize the opportunities towards sustainability, economic expansion and transformational development [14].

Real-time streams offered by IoT sensors, satellites, and weather stations will be integrated, which will allow continuously monitoring the conditions in the environment, which will result in more accurate predictions of floods [6] [7]. Such real-time feeds can supplement previous historical data and the system will be able to pick up early warning signs more efficiently than ever before. Moreover, a deeper architecture, including Long Short-Term Memory (LSTM) networks may sharpen the skill of the system to locate more intricate temporal patterns and long-term relations in flood-related data [5]. LSTMs can be easily applied to sequential data processing and thus will be suitable to predict the occurrence of floods using weather patterns and sensor readings recorded in history [2] [5]. The use of Geographic Information System (GIS) mapping will also give a visual display of the geographically vulnerable regions to floods, which will help prepare and plan disaster

response more effectively [6] [7]. GIS-based maps will also assist authorities and communities in determining areas of high risk and creating the relevant evacuation plans. Moreover, an application on a mobile phone to the system will provide real-time availability to flood warnings and safety measures to enable individuals and organisations to make timely precautionary measures [6]. Another way to improve the effectiveness of the system is to introduce the blockchain technology that will guarantee safe, transparent, and immutable data management of disaster responses. Blockchain is capable of supporting the coordination of government, relief agencies and local communities by ensuring that a permanent record of the flood data, response and resource allocation is maintained [7]. Moreover, increasing the size of the dataset by incorporating more historical data and worldwide floods will improve the prediction models, which will be more fit to different geographical areas [3] [5]. A detailed set of data will enhance generalization of the model and the system will make correct predictions in varied climatic conditions [5].

References

1. L. Breiman, "XGBoosts," Machine Learning, 45(1), 5-32, 2001.
2. M. A. Nayak, and S. Ghosh, "Prediction of extreme rainfall events using weather pattern recognition and machine learning techniques," IEEE Transactions on Geoscience and Remote Sensing, 51(12), 6741-6753, 2013.
3. P. P. Jena, et al, "Machine learning based flood prediction model using hydrological and meteorological data," Environmental Science and Pollution Research, 27, 462-474, 2020.
4. M. Pal, and P. M. Mather, "An assessment of the effectiveness of methods for land cover classification," Remote Sensing of Environment, 86(4), 554-565, 2003.
5. N. Sharma, et al, "A survey on machine learning applications for flood prediction," International Journal of Disaster Risk Reduction, 62, 102412, 2021.
6. "NOAA National Weather Service," Understanding flood risk prediction, 2023, https://www.weather.gov/media/learning
7. "United Nations Office for Disaster Risk Reduction (UNDRR)," The role of predictive analytics in disaster mitigation, 2022.
8. A Fares Hamad Aljohani, B. Ahmad Alkhodre, Adnan Ahamad Abi Sen, Muhammad Sher Ramazan, Bandar Alzahrani, Muhammad Shoaib Siddiqui, "Flood Prediction using Hydrologic and ML-based Modeling: A Systematic Review," Faculty of Computing and Information Technology, Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah KSA2, 6, 2023.
9. N. Zehra, "Prediction analysis of floods using machine learning algorithms (NARX & SVM)," International Journal of Sciences: Basic and Applied Research (IJSBAR), 49(2), 24-34, 2020.
10. A. P Ozturk, K Chau, "Flood Prediction using Machine Learning Models: Literature Review," 2018.

11. J. Linus, V. Tanjaya, A. Kurniawan, "Performance Comparison of Machine Learning Methods for Flood Prediction," Procedia Computer Science, 2024.
12. M. Ma, G. Zhao, B. He, Q. Li, H. Dong, S. Wang, Z. Wang, "XGBoost-based method for flash flood risk assessment," Journal of Hydrology, 2021.
13. H.S. Munawar, A.W.A. "Hammad, Remote sensing methods for flood prediction: A review ST Waller," Sensors, 2022.
14. M. Motta, M. de Castro Neto, P. Sarmento, "A mixed approach for urban flood prediction using Machine Learning and GIS," International journal of disaster ,2021
15. Sankaranarayanan,M. Prabhakar, S. Satish, P. Jain, A. Ramprasad, A. Krishnan, "Flood prediction based on weather parameters using deep learning," Journal of Water and Climate Change, 11 (4), 1766–1783, 2019.