Early Detection of Flood using Weather Data

Lukmanul Hakeem M , Sajan Muhammad

Abstract

This report focuses on the importance of flood prediction in Kerala, considering its geographical features and susceptibility to heavy monsoon rains. To address this, a flood prediction model was developed for Alappuzha using binary classification with various machine learning algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Gaussian Naive Bayes, and XGBoost Classifier. The study utilized a dataset obtained from the open-source Open Meteo platform, which consisted of features such as maximum temperature, minimum temperature, and rainfall. Through comparative analysis, it was determined that the XGBoost Classifier exhibited superior performance in flood prediction. This research highlights the significance of accurate flood prediction methods and the potential of machine learning algorithms to aid in timely evacuation measures, resource allocation, and effective disaster management, thereby minimizing the loss of life and property in flood-prone regions like Kerala.

1. Introduction

Climate change has become an urgent global concern, with rising greenhouse gas emissions leading to a dramatic increase in extreme weather events across the planet. In his message for the International Day, Secretary-General António Guterres highlighted how climate disasters are causing unprecedented harm to countries and economies. The consequences, including loss of life, extensive damage, and financial burdens, are more severe than ever before. This report aims to shed light on the impacts of climate change, particularly in relation to frequent floods, and emphasizes the critical need for understanding and addressing this global crisis.

As the average global temperature rises and extreme weather events become more frequent, ecosystems worldwide are undergoing significant transformations, and entire species of plants and animals are at risk. The exploitation of natural resources, driven by the pursuit of improved living standards for populations around the world, has placed enormous pressure on the planet. However, the Earth's capacity to cope with these demands is being strained. Forests are experiencing drought and degradation, rainfall patterns are shifting, wildfires are becoming more prevalent, and the glaciers in both the North and South Poles are rapidly shrinking. One alarming statistic shared by the Secretary-General is that climate disasters now displace three times more people than armed conflicts. The danger zone already encompasses half of humanity, highlighting the widespread impact of climate change on communities and societies across the globe.

This project report aims to provide a comprehensive understanding of the consequences of climate change, specifically focusing on the recurring issue of floods in various regions. By examining the causes and impacts of climate change-induced flooding, this report seeks to contribute to a deeper comprehension of the crisis and emphasize the urgency for proactive measures. To effectively respond and adapt to climate change, it is crucial to first grasp the intricate dynamics of this global phenomenon.

Considering the significant susceptibility of Alappuzha district to frequent floods, we have chosen to concentrate our attention on Alappuzha itself. In this report, we will explore the utilization of machine learning (ML) algorithms for predicting and classifying disasters, which offers great potential. Our work's primary contribution lies in the comparison of different machine learning algorithms for the prediction task, using minimal features. The applications of ML in this context are immense, and we aim to further improve our work to enable real-time prediction and response.

To address the challenge effectively, we explore the application of machine learning (ML) algorithms for predicting and classifying disasters. This approach holds immense potential in improving response strategies. A major contribution of our work lies in the comparison of various ML algorithms for the prediction task, utilizing minimal

features. We aim to further enhance our research to enable real-time prediction and response.

By providing a comprehensive understanding of climate change-induced flooding and evaluating the efficacy of ML algorithms, we hope to contribute to the ongoing efforts in mitigating the impact of floods and fostering proactive adaptation measures.

2. Literature Survey

Kim and Barros [1]: They modified an ANN model for improved flood forecasting by considering atmospheric conditions. The ANN model showed higher accuracy compared to statistical models. Reference [2]: They developed an ANN forecast model for hourly lead time using meteorological and hydrodynamic parameters of three typhoons. The ANN model showed promising results for 5-hour lead time. Danso-Amoako [3]: They provided a rapid system for predicting floods using an ANN. The ANN model demonstrated high generalization ability with an R2 value of 0.70. Panda, Pramanik, and Bala [4]: They compared the accuracy of ANN with FFANN (Feedforward Artificial Neural Network) for short-term water level prediction. FFANN performed faster and relatively more accurately than the ANN model. Kourgialas, Dokou, and Karatzas [5]: They created a modeling system based on ANNs for predicting extreme flow events ahead of floods. The ANN model showed high effectiveness compared to conventional hydrological models. Lohani, Goel, and Bhatia [6]: They improved real-time rainfall-runoff forecasting and compared the results with T-S fuzzy model and TSC-T-S fuzzy model. The fuzzy models provided more accurate predictions with longer lead time. Pereira Filho and dos Santos [7]: They compared autoregressive (AR) models with ANN in simulating forecast stage level and streamflow. The ANN model performed better and was proposed as a better alternative to the AR model. Ahmad and Simonovic [8]: They used a Backpropagation Neural Network (BPNN) for predicting peak flow using causal meteorological parameters. BPNN proved to be a fast and accurate approach with generalization ability. Reference [9]: They used division-based backpropagation with BPNN to improve the simulation of daily streamflow. The BPNN model showed promising results for short-term flood prediction. Reference [10]: They applied BPNN for

assessing flash floods using measured data. The ANN prediction model showed the ability to deal with noisy datasets. Ghose [11]: They predicted daily runoff using a BPNN prediction model. The BPNN model demonstrated high accuracy and efficiency for flood prediction.

3. Motivation

Alappuzha district, located in Kerala, India, has a population of approximately 2,127,789 people, making it comparable to the nation of Namibia or the US state of New Mexico. The district ranks 216th in India in terms of population. It has a population density of 1,504 inhabitants per square kilometer.

The city of Alappuzha is situated at a specific latitude and longitude (specific values are needed to provide accurate information). The region is known for its scenic backwaters, which include prominent lakes like Vembanad Lake, Punnamada Lake, and Pathiramanal Lake. These water bodies contribute to the district's unique landscape and attract tourists from around the world. In terms of demographics, Alappuzha has a sex ratio of 1100 females for every 1000 males, indicating a slightly higher female population. The district also boasts a high literacy rate of 95.72%, showcasing the emphasis on education.

Alappuzha is prone to floods, making awareness and early prediction essential for mitigating the impact of such events. The district's topography and extensive network of water bodies make it susceptible to flooding. Therefore, raising awareness about flood-related risks and implementing early prediction systems can significantly contribute to preparedness and minimizing damage caused by floods in the region.

4. Contribution

Open-Meteo is an online service owned by OpenWeather Ltd that offers global weather data through an API. It provides various weather information such as current conditions, forecasts, nowcasts, and historical data for any location worldwide. One notable feature is the minute-by-minute hyperlocal precipitation forecast. The historical weather API is particularly noteworthy as it provides over 60 years of hourly weather

data at a 25-kilometer resolution. Despite utilizing a vast amount of data (approximately 20 TB), users can retrieve 60 years of temperature data almost instantly. In order to enhance our services, we incorporated historical flood incident data from newspapers and other archives, specifically focusing on major flood incidents in Alappuzha. This information was also used for training purposes. Since our task involves classification, we included an additional column of class labels. To accomplish this, we sought assistance from a Kaggle dataset containing flood dates. However, the dataset consists of information for more than 33,000 days, with only a few days reported as floods. To ensure unbiased classification, we employed the SMOTE technique to generate synthetic data and address the scarcity of flood-related instances.

5. Methodology

Fig 1. Flow chart of the entire modeling architecture

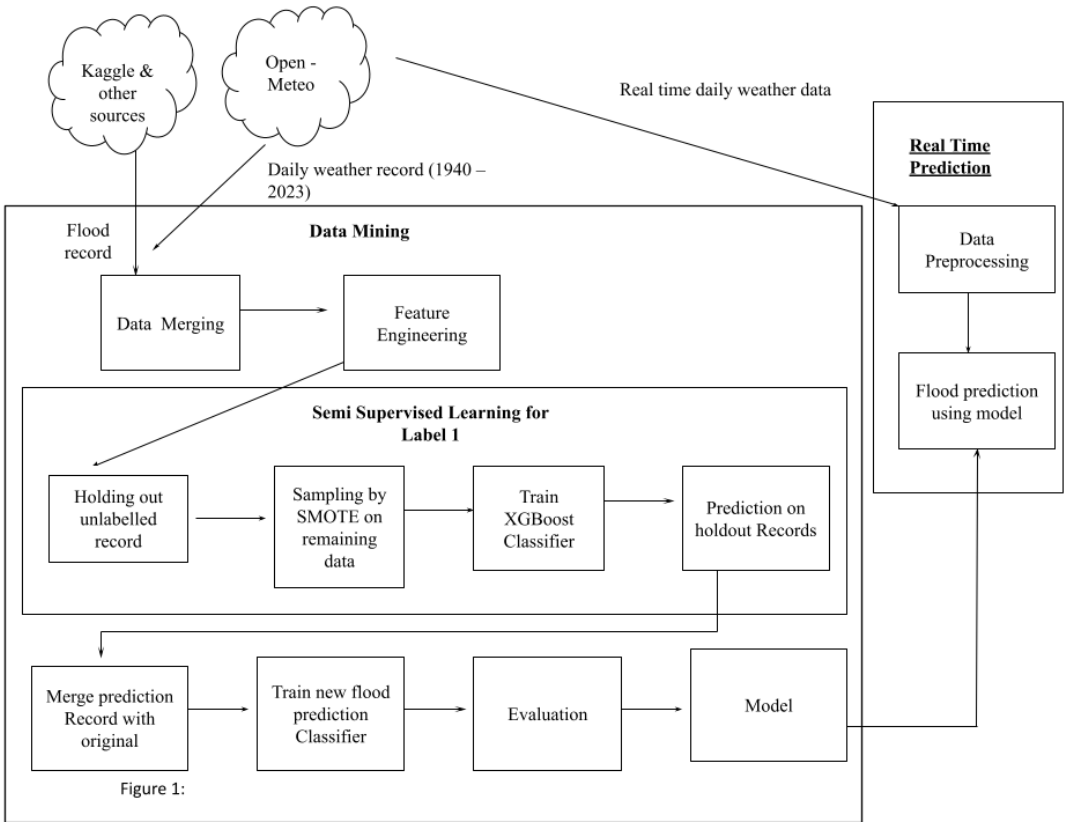

Figure 1:

5.1 Feature selection

We collected historical daily weather data date from 02-01-1940 to 22-03-2023 of Alappuzha district. Data taken from Open-Meteo weather API. There are total 30397 records. Historical weather API consist of about 15 climate measure attribute. Out of 15, we had taken 11 attributes initially. This is used as data features. Selected feature includes:

- Date (Time)
- Max. Temperature (°C)
- Min. Temperature (°C)
- Sunrise Time
- Sunset Time
- Precipitation Sum (mm)
- Rain Sum (mm)
- Snowfall Sum (cm)
- Precipitation Hours (h)
- Max. Windspeed (km/h)
- Max Windgusts (km/h)

Variation in all those features would have hidden relation with the occurance of flood, example: one can say feature Date would have possible relation since most of the flood occurred in Alappuzha was in between Jun and September, day with late sunrise or early sunset can reduce daily evaporation rate from flooded region. This kind of correlation between these features and flood become evident from final model prediction.

In order to prepare suitable data format for model fit, some feature engineering techniques had applied on time related features to convert to sub-features.

- Date feature to Day, Month, Year features
- Sunrise Time feature to Minutes and Hour features (same for Sunset Time)

After above conversion data had 15 input feature and we do not drop any taken features

5.2 ML Algorithm

Inorder to find most suitable algorithms that fit data, we gone through model selection procedure on various classification algorithms. This has been done primarily for selecting algorithm in Semi-Supervised learning to classify on label unknow records. Classification algorithms that tried out are:

- Logistic Regression
- K Nearest Neighbour
- Support Vector Machine Classifier
- Gaussian Naive Bayes
- Decision Tree Classifier
- Ensemble methods: such as Random Forest, BaggingClassifier, AdaBoostClassifier, GradientBoostingClassifier and XGBClassifier

Selection procedure consists estimating cross validation score on all the above algorithms for a fold of 3 splits using Stratified cross validation technique. On each validation, data feed to algorithm through a pipeline to perform SMOTE balancing and scaling. We had figured out XGBoost classifier as best, have good average performance score and better stability in prediction.

Since XGBoost perform well in Semi-Supervised training, we move on with XGBoost algorithm to train on new merged dataset. XGBoost also has low latency compared with others.

6 Experimental Section

In this section, we outline the experimental setup for model training and real-time predictions. We describe the data sources, data preparation steps, model selection, evaluation metrics, and the process of converting the training data mean and variance for real-time predictions.

6.1 Data Sources

We utilized two main data sources for our study: the Open Meteo dataset and a Kaggle dataset. The Open Meteo dataset provided us with daily weather reports for the Alappuzha region, starting from March 1940. This dataset served as the foundation for our weather data, which we used for model training. Additionally, we leveraged the Kaggle dataset to obtain the actual dates of high rainfall events. This dataset included information on red and yellow alerts, allowing us to identify flood possibilities. We added a binary class label column to our weather dataset, assigning a value of 1 to flood possibilities and 0 to other instances.

6.2 Data Preparation

To ensure that our weather dataset contained the necessary class label information, we merged the Open-Meteo and Kaggle datasets. We matched the dates from the Open Meteo dataset with the corresponding flood labels from the Kaggle dataset. However, we encountered missing label information for the flood occurrences in 2005 and 1961. To address this, we employed semi-supervised learning techniques to label the missed information for these years. After combining and labeling the datasets, we proceeded with the data preparation steps, including cleaning the data, handling missing values, and encoding categorical variables.

6.3 Model Selection

To train our model, we evaluated various machine learning algorithms. After comparing their performance, we determined that AG Boost exhibited superior results for our flood prediction task. Therefore, we selected AG Boost as our chosen algorithm for model training and prediction.

6.4 Model Training

We performed model training using the labeled weather dataset and AG Boost algorithm. The training process involved iteratively adjusting the model's parameters to

minimize a predefined loss function. We optimized the hyperparameters of the AG Boost algorithm through techniques such as cross-validation and grid search. Following training, we evaluated the performance of the trained model using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score.

6.5 Real-Time Prediction

For real-time predictions, we utilized the trained model. However, before making predictions on unseen data, we needed to account for differences in the data distribution. To address this, we converted the mean and variance of the training data to another file. This transformation enabled us to perform standard scaling on the unseen data, ensuring it conformed to the same normal distribution as the training data. Real-time prediction uses real-time daily weather data from Open-Meteo API. This is done using the Open-Meteo python library called Openmeteopy.

6.6 Evaluation Metrics

Our model has a good performance score on validation dataset. Validation carried on input sample of size 6080 records. Out of 6080, only 10 records are positive labels (1) and remaining 6070 are negative labels (0).

Table 1: Model validation score

Classes	**Recall score**	**F1 score**
Flood (Label 1)	0.9	0.95
No Flood (Label 2)	1.0	1.0

Trained XGBoost model able to correctly classify 9 positive out of 10 true positive labels and have one miss classification. Negative classes 6070 records are completely correctly classified. Since our model is for predicting a disaster, we focus more on improving the recall matrix in order to minimize miss-classification of positive

labels as negative. Our model was able to achieve a recall of 0.90. This improvement is achieved by applying balancing on model training data.

7. Conclusion

In this work, we solely performing an simple real-time flood detection based on regional weather data. Right now, we only considered one of the most flood prone district in Kerala, Alappuzha. This kinds of climate disaster modelling had great significance in our lives that further escalated by frequent acceleration on climate change.

As concern increasing, we would have a better climate strategic plan. So, we would like to extend our work further on to predict all other possible Climate Induced Disaster (CID) on a large geographical location. We hope this could be possible by a better weather modelling leaveraging atmospheric weather map images in addition to weather data available.

8. References

1. Kim, G.; Barros, A.P. Quantitative flood forecasting using multisensor data and neural networks. *J. Hydrol.* 2001, *246*, 45–62. [Google Scholar] [CrossRef]
2. Kim, S.; Matsumi, Y.; Pan, S.; Mase, H. A real-time forecast model using artificial neural network for after-runner storm surges on the Tottori Coast, Japan. *Ocean Eng.* 2016, *122*, 44–53. [Google Scholar] [CrossRef]
3. Danso-Amoako, E.; Scholz, M.; Kalimeris, N.; Yang, Q.; Shao, J. Predicting dam failure risk for sustainable flood retention basins: A generic case study for the wider greater manchester area. *Comput. Environ. Urban Syst.* 2012, *36*, 423–433. [Google Scholar] [CrossRef]
4. Panda, R.K.; Pramanik, N.; Bala, B. Simulation of river stage using artificial neural network and mike 11 hydrodynamic model. *Comput. Geosci.* 2010, *36*, 735–745. [Google Scholar] [CrossRef]
5. Kourgialas, N.N.; Dokou, Z.; Karatzas, G.P. Statistical analysis and ann modeling for predicting hydrological extremes under climate change scenarios: The

example of a small Mediterranean Agro-watershed. *J. Environ. Manag.* 2015, *154*, 86–101. [Google Scholar] [CrossRef] [PubMed]

6. Lohani, A.K.; Goel, N.; Bhatia, K. Improving real time flood forecasting using fuzzy inference system. *J. Hydrol.* 2014, *509*, 25–41. [Google Scholar] [CrossRef]
7. Pereira Filho, A.J.; dos Santos, C.C. Modeling a densely urbanized watershed with an artificial neural network, weather radar and telemetric data. *J. Hydrol.* 2006, *317*, 31–48. [Google Scholar] [CrossRef]
8. Ahmad, S.; Simonovic, S.P. An artificial neural network model for generating hydrograph from hydro-meteorological parameters. *J. Hydrol.* 2005, *315*, 236–251. [Google Scholar] [CrossRef]
9. Ju, Q.; Yu, Z.; Hao, Z.; Ou, G.; Zhao, J.; Liu, D. Division-based rainfall-runoff simulations with BP neural networks and Xinanjiang model. *Neurocomputing* 2009, *72*, 2873–2883. [Google Scholar] [CrossRef]
10. Sahoo, G.B.; Ray, C.; De Carlo, E.H. Use of neural network to predict flash flood and attendant water qualities of a mountainous stream on Oahu, Hawaii. *J. Hydrol.* 2006, *327*, 525–538. [Google Scholar] [CrossRef]
11. Ghose, D.K. *Measuring Discharge Using Back-Propagation Neural Network: A Case Study on Brahmani River Basin*; Springer: Singapore, 2018; pp. 591–598. [Google Scholar]