National College of Ireland

Efficiency of Flash Flood Prediction by XGBoost and Random Forest using 15 minutes & 1 hour time period sensor data.

MSc Research Project
Cloud Computing

Ghiridhar Iyer
Student ID: X18183468

School of Computing
National College of Ireland

Supervisor: Dr. Manuel Tova-Izquierdo

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Ghiridhar Iyer
Student ID:	X18183468
Programme:	Cloud Computing
Year:	2020
Module:	MSc Research Project
Supervisor:	Dr. Manuel Tova-Izquierdo
Submission Due Date:	17/8/2020
Project Title:	Efficiency of Flash Flood Prediction by XGBoost and Random Forest using 15 minutes & 1 hour time period sensor data.
Word Count:	8158
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

I agree to an electronic copy of my thesis being made publicly available on TRAP the National College of Ireland's Institutional Repository for consultation.

Signature:	
Date:	28th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Efficiency of Flash Flood Prediction by XGBoost and Random Forest using 15 minutes & 1 hour time period sensor data.

Ghiridhar Iyer
X18183468

Abstract

Floods are one of the costliest and deadliest Natural Disasters known to mankind. Due to the inconsistent nature of rain, estimation of flood becomes complex. Most of the previous works have focused on forecasting floods but limited research has been done on flash flood prediction also known as nowcasting. Since Flash Floods manifest in a matter of hours, people remain unaware of the disaster leading to loss of lives. Many previous works have highlighted the time period (time difference between successive rows) of the dataset as the limitation to predict flash floods. By foreseeing the disaster as well as assessing its threat in real-time would ensure timely actions which can avoid loss of life. This paper predicts flash floods using XGBoost and Random Forest based on UK Sensor Data. This paper also examines the effect of the time period of the dataset on the performance of the prediction model. AWS Platform was used to host the application. GAN was utilised to mimic the dataset and increase the number of records. Algorithms were scripted and were provided to the Sagemaker ML endpoint for training and prediction. Both the algorithms successfully predicted flash floods and river level for about 3 days. The PASS evaluation technique has been adopted for assessing the performance of algorithms. XGBoost outperformed Random Forest in all evaluation aspects and hence saves time and lives of the people. Implementation and performance assessment of Neural Networks is yet to be performed.

1 Introduction

1.1 Background and Motivation

Water forms an integral part of the survival of any living being. Sources of water are underground, lakes, rivers, etc. which get replenished by rain. Rain tends to be inconsistent and in the case of a persistent downpour, can lead to a flood. Natural Disasters never arrive with a prior warning but can be detected & analyzed to plan and prepare the evacuations to save human lives. Floods are one of the costliest and deadliest natural disasters known to mankind. The ability of flood to prove fatal has been underestimated by the youth and elderly which has been the main reason for their deaths due to inundation (Ashley & Ashley (2008)). There was a survey conducted on the damage caused to property by floods. It shows a steady rise in the cost of damage not only to infrastructure but also to human lives on every occurrence of a flood. This is because the density

of constructed structures keeps increasing for accommodating people and setting up the enterprise workplace (Ahmadalipour & Moradkhani (2019)). Amenities like Electricity, Network Connectivity are heavily affected. This conveys that the extent of awareness in people regarding the threat from the flood is lacking.

There are many domains wherein predictions play a vital role. These can be for business purposes aiming for profit or for disaster management purposes aiming at saving lives. Machine Learning has been the prime motivator for creating Prediction models. Cloud is an ideal environment to have an on-demand and scalable application. Research shows Data Processing and Cloud are becoming an integral component for applications designed for providing predictions and insights (Najmurrokhman et al. (2019) Limousin et al. (2019) Furquim et al. (2018) Afzaal & Zafar (2016)).

1.2 Problem

Analyzing the effect of floods can help in understanding the extent of preparedness. Areas with a higher density of people should be targeted first for evacuation. Physical modifications can influence (positively as well as negatively) the possibility and extent of flood in the future (Patrick et al. (2019) Hasan et al. (2019)). Some focused on generating a Flood Map, which is a Geo-Spatial data, representing the areas under threat of being inundated in case of floods. These techniques are mainly used after the occurrence of the flood, as it informs the authorities about the current risk and damage in all the areas. Flood Maps are static content, which gets outdated with time as the change in sea levels, physical modifications in the region and changing pattern of rainfall creates new and challenging scenarios to tackle (Hasan et al. (2019) Patrick et al. (2019)).

Real-time or near real-time prediction is essential to avoid any delay in evacuating people. Flash floods manifest in 4 to 6 hours due to perennial rainfall. Threat to life due to flash floods is high since no one is aware of the upcoming disaster (Ahmadalipour & Moradkhani (2019) Furquim et al. (2014) Ashley & Ashley (2008) Du et al. (2019) Morán-Tejeda et al. (2019)). Most of the previous researches have forecasted the chances of flood beyond 24 hours up to 4 days based on datasets with a time period of 6 hours to 1 day. Since majority of flash floods manifest within this period, prediction of flash flood is not possible by these applications (Du et al. (2019) Hagen et al. (2020) Herman & Schumacher (2018)).

1.3 Research Question

Can Flood Nowcasting and Flash Flood notification be accurately performed by XGBoost than the state-of-the-art techniques like Random Forest using 15 minute and 1 hour time period sensor data?

1.4 Objectives and Contribution

The main objective of this paper is to examine the efficiency of XGBoost and Random Forest in performing Flood Nowcasting and Triggering Flood Warning based on the predicted river level. Flood Nowcasting is predicting the possibility of flood within 24 hours. The paper emphasizes on predicting Flash Floods. Since the emergence of flash floods is possible in 4 to 6 hours, a reasonable time period is expected. The time period is the time difference between two successive rows. In order to understand the trend,

sufficient number of data is required (Du et al. (2019) Herman & Schumacher (2018) Furquim et al. (2014) Furquim et al. (2018)). The accuracy of the algorithms is verified using graphs and sum of error in prediction. The efficiency of the algorithms is verified by analyzing, how many flood warnings are triggered and missed. Incorrect Flood warning Triggers are also examined.

The secondary objective of the paper is to examine the effect of the time period on the accuracy of the prediction. The dataset has a time period of 15 minutes. Aggregation of data to 1 hour would be carried out to assess the prediction accuracy for both time periods - 15 minutes and 1 hour.

Efficient prediction of flash floods would ensure prompt measures to safeguard and in some cases evacuate people to safety. The issue of flash floods is not restricted to certain locations, but is prevalent everywhere and is becoming more frequent. Unlike Flood Forecasting, Flood Nowcasting and flash flood prediction has room for research and exploration which would benefit people across the globe.

1.5 Limitations

The time period and predictor variables like rain, temperature and wind play a pivotal role in this research. Since only one month of sensor data was available, the implementation of Neural Networks becomes non-viable. Forecasting using Neural Networks without a sizable amount of data could lead to impartial predictions. Also, even a small change in values of hyperparameters have a significant effect on the performance. Due to the timing and complexity of Neural Networks, it was not implemented for flash flood prediction.

The paper is structured in the following manner. Section 2 consists of the Literature Review of the State of the Art, to gain clarity regarding Flood Prediction and Flood Monitoring. Section 3 provides an overview of the Methodology. Section 4 deals with the Architecture Design and Section 5 deals with the implementation. Section 6 analyzes the results in detail and section 7 concludes the paper and provides future work.

2 Related Work

There are severe effects of Natural Disaster on the infrastructure and the people living in the affected region. Disasters like Earthquake are devastating depending on their intensity which can be measured using Seismometers. The following section describes the effects of floods.

2.1 Assessment of Risk due to floods

Ahmadalipour & Moradkhani (2019) affirms that floods inflict drastic damage to biodiversity and infrastructure. The analysis on the frequency, duration and property damage caused by flash floods in the Contiguous United States in the past 22 years provided a clear understanding of the damages due to floods. A Discussion is cited which is based on the property damage caused by flash floods in the United States which concluded Floods as the costliest natural disaster. Although earthquake and Tsunamis cause equivalent or more damage based on its intensity, Floods are becoming very frequent. This ultimately causes more damage to property as well as biodiversity. Ashley & Ashley (2008) has analyzed the flood fatalities from 1959 to 2005 which showed it to be the second deadliest Natural Disaster in the United States. The count of deaths due to drowning outdoors

were the highest as per the findings. This also pinpoints that the awareness regarding the Fatality of Floods is highly unaware among the people. The Age of the dead were from 19 to 30 and above 60. Usually, people between the age group of 30 to 60 are working-class citizens, who either were stuck during the floods at their workplace and/or were aware of the effects and stayed indoors or stay in areas less affected by floods.

Morán-Tejeda et al. (2019) analyzed the water flow in the rivers and rainfall in Spain. The Factorial snow model was used which would compare the snow depth and hence determine the snow meltdown water. Weather Research and Forecasting Model predicted the rainfall. These two models were combined to understand the role of snow meltdown and rainfall in flood incidents. Although ice melt contributed to flooding occurrence, rainfall has been the chief contributor. In about 60 per cent of the flood incidences, rainfall was the major contributor for surface runoff. This paper supports the argument that rainfall plays a major role in the manifestation of floods. Hence an analysis of the behaviour of rain would act as one of the parameters to predict the possibility of flood in the upcoming days.

Orton et al. (2019) finds Flood as one of the deadliest threat due to rise in sea level. This was concluded not only by surveys and expert opinions but also through sea monitoring satellites data. The rising sea level poses a high risk to the coastal cities. Currently, the construction of infrastructure is based on 100-year flood maps. 100-year flood map shows the probability of the number of flood occurrence in a given region, usually a country, in the next 100 years. DEM and projection of sea-level rise are primarily used to create the 100-year flood map. The issue with 100-year flood map is, just one instance of flood in a region changes the 100-year flood probability for that region. 100-year flood maps are subject to change every year. With the rising sea level, the intensity of flood is expected to increase drastically with time.

The wind has been the main influencing factor for the creation of storm and flood. Wind increases the speed and height of tides, which can cause coastal tiding. One point to be noted is, the future protection steps that would be taken by the government will greatly influence the 100-year flood map. Flood Maps can be considered as guidance but cannot be a prime material to assess the safety of the region. It is subject to amendments with the change in environmental factors like rising sea level, changing patterns of rainfall, global warming which causes a rise in humidity and snowmelt, etc. and coastal constructions like dredging, sectional and frame barrier or wall to avoid or at least delay the surface runoff in the region (Patrick et al. (2019)).

The above section clearly states the effects of flood and the need to address it. The frequency of Floods is expected to increase owing to climate change. The contribution by the researchers in the field of Flood majorly dealt with Forecasting. Also, the previous works were based on a specific technology/approach. Each author had focused on a specific technology/approach which they would utilize to develop the solution. For instance, Some authors solely focused on ML while some on Flood Simulators. Hence, analysis of various technologies and approaches in the field of "Forecasting" and "Nowcasting" is obligatory.

Natural Disasters like Floods cannot be prevented but can be predicted through techniques like hydrological models and Machine Learning Techniques. This assists in forecasting the floods before it occurs which assists in Disaster Management. The following section focuses on the functionality of these applications.

2.2 Flood Forecasting

2.2.1 Convolutional Deep Neural Network (CDNN) and Wireless Sensor Area Network (WSAN)

Anbarasan et al. (2020) combined IoT, Big Data and CDNN for predicting the occurrence of floods. The author agrees that IoT can be beneficial in predicting floods due to the continual inflow of data which guides in understanding the behaviour of river water level and rainfall. Historical data has been used to gauge the current behaviour of river level and rainfall. Institute of Environmental Studies has asserted that in the upcoming three decades, 60 per cent of cities would face flood issues. Map-Reduce based Big Data Framework was adopted and the normalisation of the historical and real-time datasets was performed to fit the values into the range of 0 and 1. Normalization was performed because the outcome of the CDNN module is expected to be Logical (Chance or No Chance). It is essential to have an estimated idea of when the flood is most likely to occur which will assist in evacuation planning. This is possible if the outcome is numeric and not logical. Although the approach was creative using CDNN, it will not satisfy the research objective of time series forecasting.

Afzaal & Zafar (2016) proposed a flood detection algorithm wherein the WSAN would be connected to the cloud platform in a form of sensors, gateways or actors. The sensor readings and predefined parameter values determined the actions of the Actuators. Although Cloud is mentioned as a storage and decision-making model, the paper does not discuss in detail its role and influence. Implementation and Validation of the algorithm would provide a better insight into the efficiency of the suggested system.

2.2.2 Flood Simulation Models/Flood Maps

Mai & De Smedt (2017) performed the simulation of a flood using WetSpa application based on hydrological data which provides the behaviour of river level in the river basin. The outcome was an accurate flood map which simulated the possible risk areas and the depth of floodwater in those areas. A general flood map was created to determine the risk areas by simulating flood inundation using DEM and Vegetation map. As it was not based on rainfall data, real-time situations cannot be monitored using this approach. Sanz-Ramos, Marcos et al. (2018) also used meteorological, hydrological as well as hydraulic models to predict flood up to 4 days prior. The High-Resolution Numerical Weather Prediction predicts the extent of precipitation. The data was validated using LiDAR data. Prediction of rainfall based on HR NWP and evaluation of the precipitation collected in the basin was accomplished.

Hasan et al. (2019) utilized the XP Stormwater Management Model (XPSWMM) application to analyze and simulate the flash flood event. The application takes 1-hour rainfall data and performs a hydrological analysis of the river level based on rain frequency, intensity and duration. River water level based on rainfall duration and intensity was obtained. The return period (probability of occurrence in a given year) of rainfall based on intensity was achieved. A flood map describing the areas which can be inundated has been derived. The author firmly states that Construction of structures or alteration of land by removal of sand has been the primary cause of the decrease of groundwater level and increasing impermeable surfaces. The statement holds good for flood occurrence as the alteration of land not only leads to an increased volume of impermeable water but also increased surface runoff of this water into the urban areas causing floods.

There are systems which provide insights on the extent of damage caused by flood after its occurrence. This helps in gauging the vulnerable areas in the region and help strategize the Flood Recovery measures. The following section provides details on the functionality of these applications.

2.2.3 Big Data and IoT in crisis Management

Furquim et al. (2018) combines sensor networks and ML to forecast flood. The Author stresses on combining IoT and cloud. In a real-world scenario, multiple stations would be providing sensor input which needs to analysed and predict chances of flood. For such on-demand and scalable environment with such huge storage, only cloud platform is ideal. With the rise of Edge Computing, Disaster Management especially Floods would be benefited to a great extent. The combined capabilities of concepts will enable gaining real-time insights into the situation. IoT enables capturing parameters which can be utilized for analyzing the status. It is a popular Big Data source. The storage and compute capacity coupled with the resilient and distributed environment makes cloud an ideal platform. The author has cited numerous works which took advantage of the cloud for flood prediction. The author deploys a two-tier WSN, wherein Tier 2 sensors collect data from tier 1 sensors and transmit the data to the cloud. If Tier 2 node fails, Tier 1 temporarily takes up the task. If the cloud is unavailable, the prediction process is done at Tier 2 node.

The author took 10 minutes of data with an interval of 1 minute to predict five minutes of river level. MLP was used to predict the river level. The author suggests considering only significant data rows, that is, data rows showing a significant increase or decrease in river level rather than being static. Static river level data acts as noise since it portrays lack of influence of predictor variables like rainfall, humidity on river level. The author finds that the prediction accuracy increases when more predictor variables are considered. It is notable that, although the prediction had very less variance between observed and predicted (calculated using r-square) of 0.95, there were significant false positive and false negative values. Predictions are bound to have errors, but accuracy is an important aspect and significant false outcomes show room for improvement.

Several other papers have focused on Big Data to solve counter natural disasters (Cumbane & Gidófalvi (2019) Najmurrokhman et al. (2019) Limousin et al. (2019) Arthur et al. (2018) . Cumbane & Gidófalvi (2019) and Arthur et al. (2018)) focused on social sensing. Social media data was assessed (Arthur et al. (2018)) and ML sentiment analysis was performed on the tweets. Based on the GPS data (based on location tracking setting and/or location mentioned in the tweet), real-time Flood Maps were created. The author admits the issue of significant false positives as well as the inability to detect flash floods. Flash floods happen in hours and in such a scenario, evacuation would occur before tweeting about the flood. Also, the amount of tweets from a region depends on the population density that witnesses it. Cumbane & Gidófalvi (2019) discusses the ability of various Big Data frameworks to process and query spatial data which can enable the creation of real-time flood maps. (Cumbane & Gidófalvi (2019), Najmurrokhman et al. (2019), Limousin et al. (2019)) emphasize on combining ML and IoT for crisis management since IoT is a real-time data source to determine disasters promptly.

2.2.4 Machine Learning (ML)

Furquim et al. (2016) used Machine Learning to predict Floods. He claims that there is a need to continually monitor the river level along with rainfall to examine the trend in the water level. Chaos Theory has been implemented because the river level is subject to change drastically without following a constant pattern as just a small change in one aspect leads to a significant amount of change in another aspect. The experiment showed fluctuation in water level even without any rainfall. One possibility could be precipitation at one of the river's tributaries or opening of a dam.

Furquim et al. (2014) also analysed the performance of different Machine Learning Techniques for now-casting flash floods. Now-casting is the prediction within 24 hours. Forecasting is a prediction beyond today/more than 24 hours. The author assesses how the statistical features of river data influence the performance of the prediction. The author takes into consideration ten minutes of data with time period of 1 minute. Author passes mean, standard deviation and other statistical calculation between the first and last river level value and no dependent feature like rain or temperature. It needs to be noted that all the techniques of ML could still predict the river level accurately based on time and statistical data, although the outcome of each algorithm is different. BFTree Decision Tree could predict t+3 (third interval prediction value) very accurately while Multi-Layer Perceptron could predict first and third river level value accurately.

Du et al. (2019) highlights that about 51 per cent of the natural disasters in 2016 were hydrological (flood, tsunamis, etc) in nature. Among that 51 per cent, about 93 per cent was flooding, causing around 94 per cent of damages and deaths. Based on the above statistics, the author emphasizes on creating a warning system which would enable evacuation and save lives. The author takes into account various parameters like river level, rainfall, temperature, pressure, wind speed, etc. The Author takes 16 years of sensor data with time period of 1 day. Back Propagation Neural Network was used to predict the river level. The application could predict floods up to three days ahead. The author accepts that the prediction of flash floods is not possible with the present time period of data.

Amezquita-Sanchez et al. (2017) discusses the state of the art technologies in the field of natural disaster detection. The author asserts that ML is the most ideal technique irrespective of the disaster. ML and neural networks have proved their efficiency in predicting Floods, Earthquakes as well as Tornadoes.

Hu et al. (2019) uses long short-term memory (LSTM), a feedback-based neural network, for predicting flash floods. The author states that successive data affects the prediction ability of the model and only significant rows contributing to river level change should be considered. Spatial data has been used to predict the river level. Spatial data provides numerous metadata including river level, river depth, etc. The author concludes that considering the records which contribute to the trend of river level would decrease noise and avoid over-fitting of the model.

Hagen et al. (2020) deployed multiple machine learning algorithms to compare the river level prediction and flood warning triggering ability. Multiple datasets were considered with Data ranging between 2 and 30 years and time period ranging between 1 and 24 hours. 50 days were passed as Training data and 10 days as testing data. Most of the features had the time period of 6 hours. Hence the Author aggregated time period of all features (rain, humidity) to 6 hours. Out of the algorithms, Random Forest (91 per cent hit rate) and CDNN (83 per cent hit rate) were most accurate. Around 10 per cent

of flood triggers were false alarms.

Herman & Schumacher (2018) focuses on flash floods in USA. About 11 years of geo-spatial data were used as a dataset with time period of 24 hours to create a probabilistic precipitation forecast model. An accurate one-year rainfall forecast was achieved. Extreme precipitation data was provided to the model along with parameters like wind, moisture, etc. Random Forest was found to be very accurate although there were inaccuracies in regions where high precipitation was uncommon. Such areas had a majority of low or medium precipitation which could have created a biasedness in the data trend.

Hosseiny et al. (2020) creates a hybrid model of Random Forest and MLP to determine the wet nodes(pixels) and compute the water depth. Random Forest is used as a classifier to determine the wet nodes (areas flooded as per satellite imagery) in the geospatial data and MLP is used to compute the depth of the flood water which can aid in understanding the severity. Since geospatial data is a static data unlike IoT but has depth in its metadata, 5 datasets were provided for training. Two decades of water discharge data were also provided. Overall, Random Forest had an accuracy of around 99 per cent and MLP had around 88 per cent.

Kane et al. (2014) demonstrates a time series forecast of Avian Flu using Random Forest and ARIMA models. ARIMA model is usually preferred for time series forecasting. The author highlights that ARMIA assumes a linear relationship between the dependent and predictor variables which is not always the case in time-based scenarios. About 30 weeks of data were provided as training with one week derived as prediction with a time period of 1 day. Random Forest was found to be very accurate. ARIMA although was accurate but had invalid values like negatives values and each record denote the number of infections. Neither of the algorithms could predict accurately large magnitude changes, but Random Forest was reasonably close.

2.2.5 XGBoost

Zhou et al. (2019) proposed a new algorithm CEEDMAN-XGBOOST to predict crude oil prices. CEEDMAN removes the noise from the dataset. Noise in the data hides the trend among the data features. XGBOOST was used to predict the crude oil prices along with Feedforward Neural Network, Support Vector Regression and ARIMA. About 26 years of data were used as training data and 1,3 and 6 years of crude oil price was predicted. XGBOOST had outperformed rest of the algorithms. The author also stated that with an increase in noise in the dataset, the prediction accuracy of all models decreased.

Vanichrujee et al. (2018) predicted taxi demand using XGBOOST, LSTM, Gated Recurrent Unit (Neural Network) and an ensemble of all the three. Only taxi booked from airports, hospitals, residential and educational locations were considered. Also, long trips (duration beyond 100 minutes) were not considered in the study. The data is aggregated to demand/bookings per hour. The author has considered several features apart from time and number of bookings like weather, a national holiday, etc.. 24 hours of data was passed to predict the demand for the upcoming hour. XGBoost predicts the demand better in case of high demand while LSTM performs a bit better when the demand is low.

Memon et al. (2019) compared the accuracy of XGBoost and ANN for PolSAR Image Classification. PolSAR Image Classification deals with land cover classification in geospatial data. The author has justified the choice by explaining the advantages of both the algorithms. Neural Networks, in general, can learn complex relationships.

The dataset was divided into 80:20 for training and testing respectively. XGBoost not only took less time for training but was more accurate in classification than ANN. ANN took 15 hours for training with around 90 per cent accuracy while XGBoost took around 30 minutes for training with over 92 per cent accuracy. Author highlights that hyperparameter tuning is very crucial in ANN whereas XGBoost's default hyperparameters were robust enough.

Liu et al. (2019) predicted the tourist VOlume for the city of Sanya using XGBoost and ML Graphical Model. Seven years of tourist data were provided for training to predict the two-year tourist volume and income. Many features were taken into consideration like total tourist volume, overnight tourist volume, number domestic and foreign tourists, flight time and type of hotel opted by tourists. One needs to be cautious while consideration of features. Too fewer features will deprive the algorithm of realising the trend in the relationship while too many features can cause overfitting causing deterioration of model performance. Overall, both algorithms have similar accuracy. The author also proposes an ensemble of the above algorithms which has better accuracy than the individual ones.

Krishna et al. (2019) implemented Keystroke based User Identification using XGBoost, Random Forest, Logistic Regression, MLP and Probabilistic Neural Network. A 90:10 approach was used to split the dataset for training and testing respectively. The outcome was not only checked based on accuracy but also the standard deviation. Low Standard deviation means high precision. A prediction can be accurate by the predicted value being close to the actual value but, if the same input fetches three different outputs, the model is not precise. Among the algorithms used, XGBoost had the highest accuracy of 94 per cent and a low standard deviation of 0.37. A statistical t-test was performed to check the significance of the predictions of each model. Only XGBoost was found to be significant. Multi-Layer Perceptron uses back-propagation learning and had an accuracy of around 90 per cent with a standard deviation of 0.61.

2.3 Conclusion

- Due to the increasing frequency, Flood is considered as the costliest and deadliest natural disaster. Flash Floods which manifests in hours or even minutes, is detrimental to the biodiversity of the area since no one is aware of the upcoming disaster (Patrick et al. (2019) Hasan et al. (2019) Du et al. (2019) Ahmadalipour & Moradkhani (2019) Ashley & Ashley (2008) Furquim et al. (2014) Furquim et al. (2018) Hagen et al. (2020)).

- Time Period is crucial for detection of flash floods. Authors have admitted the inability of the application to detect flash floods as the time period used was 6 hours to 24 hours. Due to this the trend of river level during flash flood is unknown to the application (Du et al. (2019) Hagen et al. (2020)).

- Sensor Data which typically has a time period of 1 minute enabled flash flood detection (Furquim et al. (2014) Furquim et al. (2018)).

- The prediction accuracy increases with increasing features/columns (Furquim et al. (2018) Liu et al. (2019) Du et al. (2019) Hagen et al. (2020) Vanichrujee et al. (2018)).

- Noise in Dataset adversely affects the model performance. Even records which has minimal to no change in river level are also considered noisy records (Hu et al. (2019) Hagen et al. (2020)).
- Ensemble Forecast Data aids in overcoming the inaccuracy of the Deterministic Forecast Data (Hagen et al. (2020)).
- XGBoost has outperformed Neural Networks in Classification and Regression. This does not conclude that XGBoost is better than Neural Networks but makes XGBoost a potential and an ideal algorithm to be utilized/implemented for flash flood nowcasting (Memon et al. (2019) Krishna et al. (2019) Zhou et al. (2019) Vanichrujee et al. (2018)). Even Random Forest has very accurate prediction (Herman & Schumacher (2018) Hosseiny et al. (2020) Kane et al. (2014)) and has outperformed CDNN (Hagen et al. (2020)).

3 Methodology

This section describes the methodology followed in this paper to examine the accuracy and efficiency in flash flood prediction using XGBoost and Random Forest based on a suitable time period and all possible features.

3.1 Dataset Source and Credibility Assessment

The dataset of the sensor data was collected from the Environment Agency, UK. The sensor data was available for river level, rainfall, temperature, wind speed and wind direction with a time period of 15 minutes. Dataset was downloaded by the URL link provided in the website which is explained in the manual. Although the volume of historical data available is only one month, the sensor dataset has the parameters and time period deemed essential for flash flood prediction. Permission for public access and utilization data has been declared in the website [1]. The dataset does not have any missing values. This makes the dataset very suitable for this research.

The Research work can be divided into four sections: **Data Processing and Formatting, GAN Creation and Merging, Feature Generation and Prediction, Flood Warning Triggering.**

3.2 Research Methodology

Three stations around River Mersey in Liverpool were considered which were uploaded to S3. Each station provides its region and area names which were used to identify the stations around River Mersey. Athena was used to transform the datasets into a single file with one timestamp column with a time period of 15 minutes and 5 columns each representing a sensor reading. This final dataset was aggregated to 1 hour. Wind speed and direction & temperature were transformed to their respective average values. Since the overall rainfall needs to considered, Sum of Rainfall was performed. River level beyond the safety level, even for a short period is also considered as a flood warning. Since averaging the river level value would distort the findings, Maximum of river level was performed. Two final datasets with a time period of 15 minutes and 1 hour were

[1] Public Sensor Access: `https://environment.data.gov.uk/flood-monitoring/doc/reference`

saved to S3 from Athena. Since Athena saves the output in a compressed (.gz) format, Pandas was used to convert the data to CSV format.

Generative Adversarial Network (GAN) was trained to generate two datasets for a period of one month based on the two S3 datasets. The generated GAN dataset and S3 dataset were merged based on their time period and were saved to S3. The original sensor dataset was for the period - June 12, 2020, to July 11 2020. Since the Summer season in the UK is from June to August and all the parameter values (rainfall, temperature and wind) vary between seasons, only one month of GAN data was generated. The Timestamp in GAN and source dataset was the same. The merged dataset was transmitted to QuickSight for visualization to assess the extent of similarity between them. QuickSight was configured by providing a role to access S3 objects. S3 was configured to block all public access with an exception to QuickSight role.

Boto3 [2] is the AWS based Python SDK to communicate with AWS services. Boto3 was used to stream the file from S3 to Sagemaker and save the predictions from Sagemaker back to S3. Data is encrypted while transit when using Boto3. The sensor dataset of two months was streamed to Sagemaker and GAN timestamp was incremented by one month, making the dataset timestamp from June 12, 2020, to August 11 2020. Since river level is also time-dependent, Pandas was used to create features (columns) based on timestamp value. Custom script was written for each algorithm & time period. Script and dataset were provided to AWS ML EC2 instance for training and deployment of the model [3] [4].

The Environment Agency also provides the river level at which flood warning is triggered. Based on this river level, prediction accuracy and efficiency in triggering the flood warning is evaluated. The actual river level and the predictions of both algorithms were streamed. A visualization between actual river level and predictions by both algorithms were generated in QuickSight. The PASS evaluation technique was adopted for assessment of the algorithms. Statistical Tests like R-Square tests the extent of fit but cannot precisely assess in mission critical situations. The Precision, Accuracy, Specificity and Sensitivity (PASS) are assessed for both the algorithms for both time periods.

4 Design Specification

The architecture is divided into three layers. All the layers are enclosed within the AWS Environment.

Data Storage Layer consists of all the data: Raw Sensor Data, Athena Transformed Data, GAN and sensor merged Data and Prediction Data. Public access to the bucket is blocked and AWS IAM Role is required to access its contents.

Data Processing Layer access the data from the Data Storage Layer for four purposes: Raw Data Transformation in AWS Athena, Data formatting, GAN Generation and Merging in AWS Sagemaker (Instance 1), Model Generation and Prediction in AWS Sagemaker (Instance 2) and Assessment of Triggering of Flood Warning by the predictions in AWS Sagemaker (Instance 3). The output is stored in S3. Boto3 streams the

[2]Boto3 Documentation: `https://boto3.amazonaws.com/v1/documentation/api/latest/guide/migrations3.html`

[3]Scripting and Environment Variables in AWS ML Instances: `https://sagemaker.readthedocs.io/en/stable/frameworks/sklearn/using_sklearn.html#create-an-estimator`

[4]XGBoost Containers in Sagemaker: `https://sagemaker.readthedocs.io/en/stable/frameworks/xgboost/using_xgboost.html#use-xgboost-as-a-built-in-algortihm`

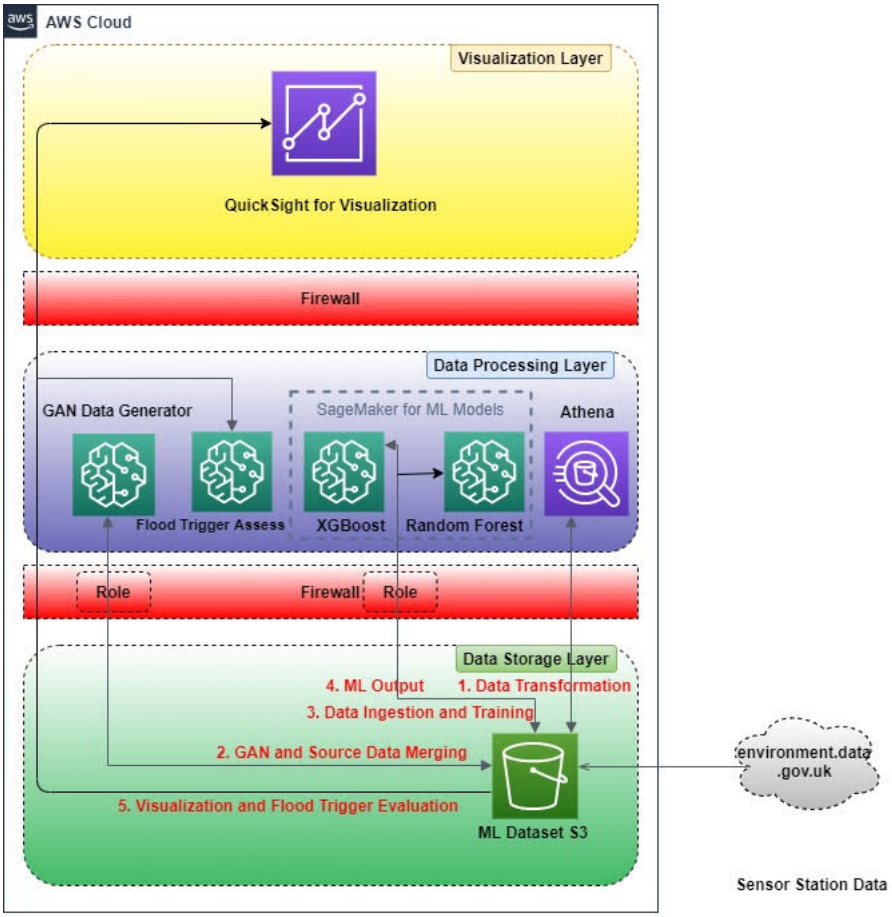

Figure 1: Flash Flood Prediction Architecture

data to and from the S3 bucket.

Visualization Layer consists of QuickSight to generate graphs based on the dataset provided. Two visualizations were produced: Merged Dataset with GAN and sensor data, Actual river level and predictions by both algorithms (15 minutes and 1 hour respectively).

5 Implementation

The implementation process is also illustrated in Figure 1. The implementation process can be divided into five parts.

Raw sensor datasets were processed by Athena for data transformation. Three datasets were transformed into a single dataset based on timestamp value. AWS Athena query S3 objects based on SQL syntax. The transformed dataset had a time period of 15 minutes. Based on this dataset, aggregation of the timestamp to 1 hour was carried out. Average of Wind and Temperature, Sum of Rainfall and Maximum of river level were calculated. The files were saved to S3.

The transformed datasets with both time periods were provided as an input to Sagemaker for GAN Generation (Instance 1). The "CTGAN" library of python was used to generate GAN data Xu et al. (2019). After Training, one month of GAN data was

generated for both time periods. The GAN and sensor data were merged into a single file (referred as "final datasets") and saved to S3.

The Random Forest and XGBoost Algorithms were written in Python. Sagemaker SDK was used to script the algorithm and pass it to Sagemaker ML (Instance 2). The final datasets were provided as training and testing data with GAN data timestamp incremented by one month. The dataset was divided in a ratio of 95:5 for training and testing respectively. River level Prediction was obtained from both algorithms for both time periods. The test dataset and predictions were saved to S3.

The actual values and predictions were provided to AWS Sagemaker (Instance 3) wherein the accuracy, efficiency and reliability of the predictions were assessed based on the flood warning river level provided by the environmental agency. Evaluation in the form of a table was obtained.

QuickSight imports datasets as per the manifest file provided and generates graphs.

6 Evaluation

XGBoost and Random Forest were successful in predicting river levels. A significant advantage was to have a stationary target variable (river level), that is, irrespective of how much the river level value raises, it decreases back to the specific low. This will be clear from Figure 2 which provides the comparison between sensor data and GAN data for both time periods.

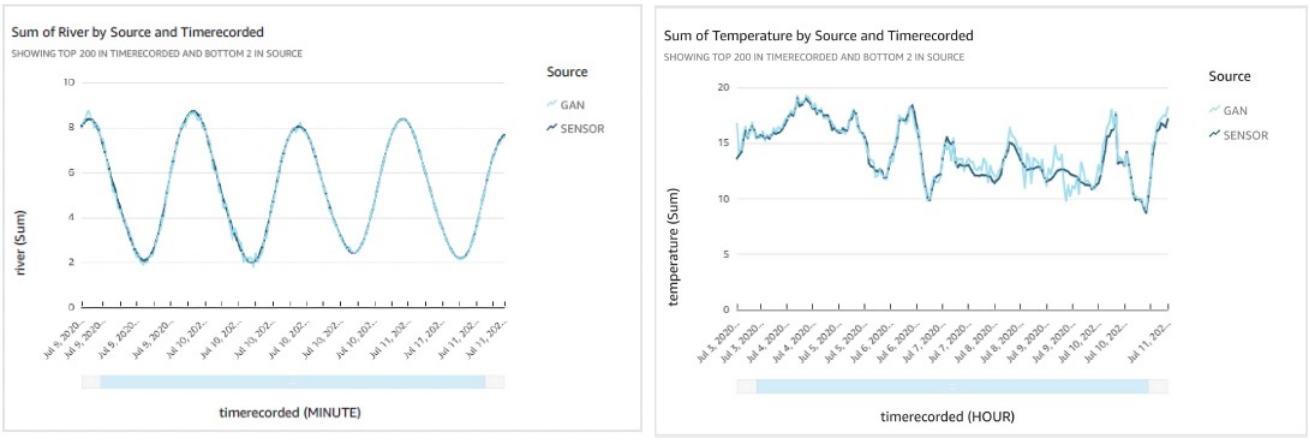

Figure 2: Comparison between Sensor and GAN data of time period 15 minutes (left) and 1 hour (right)

The graph to the left shows the comparison between GAN and sensor river level data (15 minutes). Although the river level raises beyond 9 m, it decreases to around 2 m. This makes decomposing and understanding the trend easily. The graph to the right shows the comparison of GAN and sensor temperature data (1 hour). Both graphs show that GAN and sensor data have a similar distribution of values. The GAN values are not the same but are close enough to resemble the actual sensor data. GAN not only ensures a valid distribution of values in a column but also ensures a valid distribution of values between columns to make sure the dataset seems legitimate.

Random Forest has been utilised widely in the field of flood prediction and has performed efficiently. XGBoost had shown promising prediction ability in many domains and has not yet been implemented for flood prediction. Both algorithms have accurately

predicted around 3 days of river level data. Hence, the accuracy, efficiency and reliability of XGBoost are compared with Random Forest.

6.1 Case Study 1

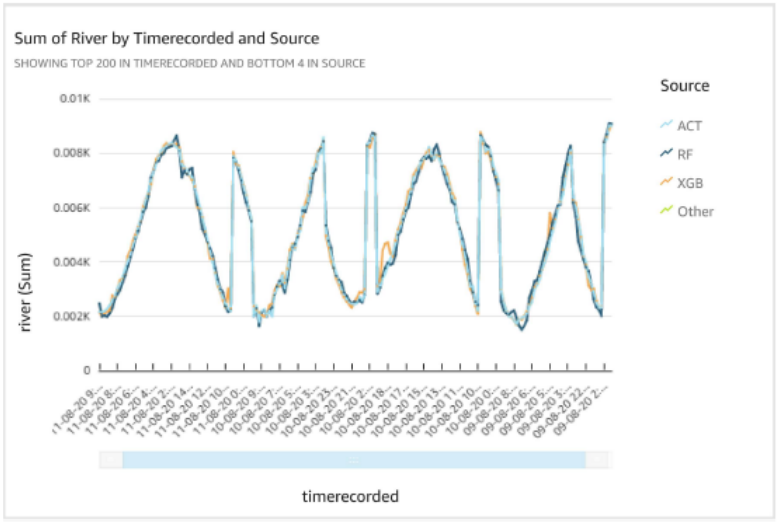

Figure 3: Comparison between Actual river level and predictions by XGBoost and Random Forest of time period 15 minutes

Figure 3 shows the comparison between Actual and predicted river values of both algorithms for 15 minutes time period. The performance of both algorithms seems remarkable. The difference in performance between XGBoost and Random Forest is not evident from the graph. It is assessed in the table below. From the shape of the graph, it is evident that, although GAN could mimic the distribution of the sensor data but could not mimic the trend exactly. The smoothness in the graph is missing. Although it is not an error, GANs ability to mimic the trend smoothly seems to be an area to work on.

6.2 Case Study 2

Figure 4 provides a comparison between actual river level and predictions. Both the algorithms performed well. In 1 hour time period, XGBoost has performed better than Random Forest. The gap between actual values and predictions are more evident in 1 hour time period than in 15 minute time period. In Figure 3 XGBoost and Random Forest graph lines seemed to be overlapping to a greater extent. But in Figure 4, the distinction between them is clear here.

Influenced by the evaluation technique (Hagen et al. (2020) and Furquim et al. (2018)), the R-Square value of both algorithms are calculated. The R-Square value is calculated to find the extent of fit between two arrays - Greater the value better the fit. The Evaluation does not base its conclusion based on R-square test but is just verifying the findings presented by Hagen et al. (2020). As per Hagen et al. (2020), although the statistical tests indicate that the model is a very good fit, but the hit and miss rates of the algorithm denotes room for improvement.

Accuracy is the extent of error in the prediction. This is assessed by finding the sum of the difference between actual and predicted values (can be termed as Prediction Error).

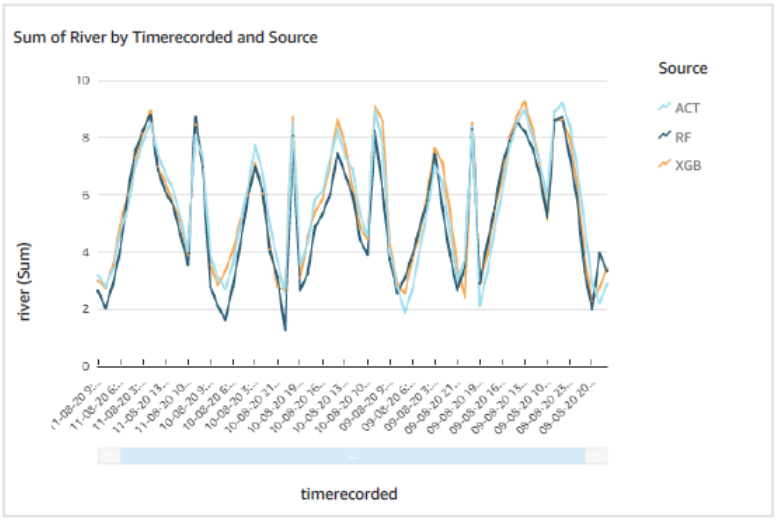

Figure 4: Comparison between Actual river level and predictions by XGBoost and Random Forest of time period 1 hour

Sensitivity/Efficiency in this scenario is assessed based on the number of flood warnings triggered correctly. An algorithm can be efficient to trigger the flood warning but should not be erroneous. Specificity/Reliability is assessed based on the number of erroneous flood warnings triggered. Precision is the number of accurate warnings triggered divided by the actual number of warnings. Influenced by Furquim et al. (2018), the below table summarises these details for both algorithms on both time periods.

Sum of Prediction Error (Accuracy): XGBoost: -7.98 Random Forest: 21.58
Highest Prediction Error: XGBoost: 0.38 Random Forest: 0.74
Total Number of Flood Warning Triggers: 103
Total Correct Flood Warnings Triggered (Sensitivity): XGBoost: 102 Random Forest: 94
Total Flood Warnings missed: XGBoost: 1 Random Forest: 9
Total Erroneous Flood Warnings Triggered (Specificity): XGBoost: 0 Random Forest: 0
Precision: XGBoost: 99.03 Random Forest: 91.26
R-Square Score: XGBoost: 99.4 Random Forest: 98.75

Table 1: 15 Minute Time Period

Sum of Prediction Error (Accuracy): XGBoost: -1.86 Random Forest: 24.79
Highest Prediction Error: XGBoost: 1.32 Random Forest: 1.60
Total Number of Flood Warning Triggers: 29
Total Correct Flood Warnings Triggered (Sensitivity): XGBoost: 27 Random Forest: 21
Total Flood Warnings missed: XGBoost: 2 Random Forest: 8
Total Erroneous Flood Warnings Triggered (Specificity): XGBoost: 1 Random Forest: 1
Precision: XGBoost: 93.10 Random Forest: 72.41
R-Square Score: XGBoost: 94.01 Random Forest: 85.75

Table 2: 1 Hour Time Period

The above statistics convey that: Although Random Forest has an exceptional R-Square score, XGBoost has outperformed Random Forest in Accuracy, Sensitivity, Spe-

cificity and Precision. Statistical Tests are reliable since XGBoost had a higher score than Random Forest. But if one judges the performance of algorithms solely based on R-Square score, it would seem there is no significant difference and both of them have similar prediction with minor differences. But the above tests have clearly distinguished the performance between the algorithms. The evaluation can be known as **PASS Evaluation (Precision, Accuracy, Sensitivity, Specificity)**.

Predictions are prone to contain errors. It is notable that, the 15-minute prediction did not contain any erroneous triggers whereas 1-hour prediction had 1 erroneous trigger by both the algorithms. It possible reason could be that the trend is more clear in 15 minutes time period than in 1 hour time period - For 2 hours: 1 hour time period has only 2 records but 15 minutes time period has 8 records.

Lastly, Random Forest overestimates while predicting while XGBoost slightly underestimates while predicting (Based on Accuracy factor).

6.3 Discussion

The stationary trend of the river level data assisted in better performance of the algorithms. Although both the algorithms were efficient, XGBoost has outperformed Random Forest. It can be concluded that the design enabled meeting the objectives of the research.

Only Furquim et al. (2014) (Furquim et al. (2018)) and Hu et al. (2019) have worked on flash flood detection. Hu et al. (2019) had used geospatial data and the rest while Furquim et al. (2014) (Furquim et al. (2018)) used numerical dataset. These works had taken only a short period of data for training and hence could predict only a short period of time ahead. Du et al. (2019) had highlighted the use of multiple features and a suitable time period to enable flash flood prediction. But Hagen et al. (2020) used a time period of 6 hours (which is smaller than many of the previous works of 24 hours) and accurately predicted floods for 10 days.

By considering all possible features like rainfall, temperature and wind and implementing XGBoost to predict river level with a time period of 15 minutes, makes the research unique. Evaluation of algorithms based on the PASS technique shows that model efficiency should not be decided solely on the outcome of statistical tests like R-square. The outcome of the Statistical tests was not wrong but the interpretation solely based on them is not sufficient. The Evaluation methodology followed by Furquim et al. (2018) needs to be done in case of gauging the algorithm thoroughly.

But there are potential improvements that can be implemented on the design. Neural Networks are efficient to understand the underlying relationship and predict accurately. The number of Features and hyperparameter values greatly influence the performance of Neural Networks (Memon et al. (2019)). Due to time and complexity issues, Neural Networks were not implemented for predicting flash floods.

It can be clear that: A bigger dataset (greater extent of historical data) will ensure the model can perform efficiently even if the behaviour of river level changes in the future. As informed earlier, this dataset is ideal for flash flood prediction since flash flood warnings were predicted efficiently. This data source was considered for the research since other sources of data either did not contain all these features, had a greater time period (typically 24 hours) or not accessible publicly.

As discussed by Hagen et al. (2020) and Hu et al. (2019), With an increasing number of days with a small time period, there will be instances wherein the change in river level

is minimal. These records distort the relationship between the dependent variable (river level) and predictor variables. Dropping those rows and carrying out Data assimilation on the dataset has been suggested as the solution for this issue. Hagen et al. (2020) utilised Ensemble Forecast Data which enabled him to predict flood levels beyond one week.

Implementation of flash flood prediction with Neural Networks with a bigger dataset is essential future work. It is noteworthy to utilise data assimilation technique and Ensemble forecast data.

Based on the AWS Sagemaker ML endpoint, Lambda can be configured to invoke the endpoint with the input of dependent variables to get the prediction. Hence a website can be developed and configured to get flood prediction. It was considered in the research, but due to a recurrent error during implementation, it was dropped.

7 Conclusion and Future Work

Flash Floods have been a threat to biodiversity due to their unpredictable nature. Due to the changing climate and infrastructure, flash floods are expected to increase in frequency and intensity. Implementation of XGBoost based on a smaller time period to predict flash floods had not been implemented yet. XGBoost and Random Forest were able to perform better than Neural Networks in some instances.

In this paper, XGBoost and Random Forest were implemented to predict flash floods based on 15 minutes and 1 hour time period. Various features like rainfall, temperature and wind were considered. XGBoost has outperformed Random Forest in all aspects although both the algorithms accurately predicted the river level data for about 3 days. This paper also highlights that the Evaluation of algorithm should not be solely based on Statistical Tests like R-Square. Statistical Tests are reliable but the factors considered for evaluation in this paper not only makes the assessment clear but also easier. The GANs ability to resemble the smoothness of the trend seems to be a research area.

"**Flash Flood Prediction based on 50 days of Historical data by considering all dependent features like rainfall,temperature & wind and obtaining 3 days of river level prediction with 15 minutes and 1 hour time period using XGBoost**" marks the uniqueness of the research. Flash flood prediction using Neural Networks with a bigger dataset is an essential future work.

References

Afzaal, H. & Zafar, N. A. (2016), Cloud computing based flood detection and management system using wsans, *in* '2016 International Conference on Emerging Technologies (ICET)', pp. 1–6. Islamabad, Pakistan,Cited by 6, Scopus ID: 57053550400.
URL: *https://doi.org/10.1109/ICET.2016.7813213*

Ahmadalipour, A. & Moradkhani, H. (2019), 'A data-driven analysis of flash flood hazard, fatalities, and damages over the conus during 1996–2017', *Journal of Hydrology* **578**, 124106. Impact Factor=3.727.
URL: *http://www.sciencedirect.com/science/article/pii/S002216941930841 8*

Amezquita-Sanchez, J., Valtierra-Rodriguez, M. & Adeli, H. (2017), 'Current efforts for prediction and assessment of natural disasters: Earthquakes, tsunamis, volcanic erup-

tions, hurricanes, tornados, and floods', *Scientia Iranica* **24**(6), 2645–2664. Impact Factor = 0.5.
URL: *http://scientiairanica.sharif.edu/article*$_4$589*.html*

Anbarasan, M., Muthu, B., Sivaparthipan, C., Sundarasekar, R., Kadry, S., Krishnamoorthy, S., R., D. J. S. & Dasel, A. A. (2020), 'Detection of flood disaster system based on iot, big data and convolutional deep neural network', *Computer Communications* **150**, 150 – 157. Impact Factor=2.613.
URL: *http://www.sciencedirect.com/science/article/pii/S0140366419310357*

Arthur, R., Boulton, C. A., Shotton, H. & Williams, H. T. P. (2018), 'Social sensing of floods in the uk', *PLOS ONE* **13**(1), 1–18. Impact Factor = 2.766.
URL: *https://doi.org/10.1371/journal.pone.0189327*

Ashley, S. T. & Ashley, W. S. (2008), 'Flood fatalities in the united states', *Journal of Applied Meteorology and Climatology* **47**(3), 805–818. Impact Factor=2.236.
URL: *https://doi.org/10.1175/2007JAMC1611.1*

Cumbane, S. P. & Gidófalvi, G. (2019), 'Review of big data and processing frameworks for disaster response applications', *ISPRS International Journal of Geo-Information* **8**(9). Impact Factor=1.723.
URL: *https://www.mdpi.com/2220-9964/8/9/387*

Du, W., Chen, N., Yuan, S., Wang, C., Huang, M. & Shen, H. (2019), 'Sensor web - enabled flood event process detection and instant service', *Environmental Modelling Software* **117**, 29 – 42. Impact factor=4.177.
URL: *http://www.sciencedirect.com/science/article/pii/S1364815218312842*

Furquim, G., Filho, G. P. R., Jalali, R., Pessin, G., Pazzi, R. W. & Ueyama, J. (2018), 'How to improve fault tolerance in disaster predictions: A case study about flash floods using iot, ml and real data', *Sensors* **18**(3). Impact Factor = 2.475.
URL: *https://www.mdpi.com/1424-8220/18/3/907*

Furquim, G., Neto, F., Pessin, G., Ueyama, J., d. Albuquerque, J. P., Clara, M., Mendiondo, E. M., d. Souza, V. C. B., d. Souza, P., Dimitrova, D. & Braun, T. (2014), Combining wireless sensor networks and machine learning for flash flood nowcasting, *in* '2014 28th International Conference on Advanced Information Networking and Applications Workshops', pp. 67–72. Victoria, BC, Canada, Cited by 14, Scopus ID: 56278079600.
URL: *https://doi.org/10.1109/WAINA.2014.21*

Furquim, G., Pessin, G., Faiçal, B. S., Mendiondo, E. M. & Ueyama, J. (2016), 'Improving the accuracy of a flood forecasting model by means of machine learning and chaos theory', *Neural Comput. Appl.* **27**(5), 1129–1141. Impact Factor=4.213.
URL: *https://doi.org/10.1007/s00521-015-1930-z*

Hagen, J. S., Cutler, A., Trambauer, P., Weerts, A., Suarez, P. & Solomatine, D. (2020), 'Development and evaluation of flood forecasting models for forecast-based financing using a novel model suitability matrix', *Progress in Disaster Science* **6**, 100076. Impact Factor = 2.1.
URL: *http://www.sciencedirect.com/science/article/pii/S2590061720300132*

Hasan, H. H., Mohd Razali, S. F., Ahmad Zaki, A. Z. I. & Mohamad Hamzah, F. (2019), 'Integrated hydrological-hydraulic model for flood simulation in tropical urban catchment', *Sustainability* **11**(23). Impact Factor=2.075.
URL: *https://www.mdpi.com/2071-1050/11/23/6700*

Herman, G. R. & Schumacher, R. S. (2018), 'Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests', *Monthly Weather Review* **146**(5), 1571–1600. Impact Factor = 3.25.
URL: *https://doi.org/10.1175/MWR-D-17-0250.1*

Hosseiny, H., Nazari, F., Smith, V. & Nataraj, C. (2020), 'A framework for modeling flood depth using a hybrid of hydraulics and machine learning', *Scientific Reports* **10**(1), 8222. Impact Factor = 4.12.
URL: *https://doi.org/10.1038/s41598-020-65232-5*

Hu, R., Fang, F., Pain, C. & Navon, I. (2019), 'Rapid spatio-temporal flood prediction and uncertainty quantification using a deep learning method', *Journal of Hydrology* **575**, 911 – 920. Impact Factor = 3.73.
URL: *http://www.sciencedirect.com/science/article/pii/S0022169419305323*

Kane, M. J., Price, N., Scotch, M. & Rabinowitz, P. (2014), 'Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks', *BMC Bioinformatics* **15**(1), 276. Impact Factor = 2.21.
URL: *https://doi.org/10.1186/1471-2105-15-276*

Krishna, G. J., Jaiswal, H., Teja, P. S. R. & Ravi, V. (2019), Keystroke based user identification with xgboost, *in* 'TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)', pp. 1369–1374. location=Kochi,India, Core Ranking = C.

Limousin, P., Azzabi, R., Bergé, L., Dubois, H., Truptil, S. & Gall, L. L. (2019), How to build dashboards for collecting and sharing relevant informations to the strategic level of crisis management: an industrial use case, *in* '2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)', pp. 1–8. paris, france, Core Ranking=C.
URL: *https://doi.org/10.1109/ICT-DM47966.2019.9032970*

Liu, X., Chen, Y., Qiu, Z. & Chen, M. (2019), Forecast of the tourist volume of sanya city by xgboost model and gm model, *in* '2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)', pp. 166–173. Guilin, China, Scopus ID: 7406348622.
URL: *https://doi.org/10.1109/CyberC.2019.00038*

Mai, D. T. & De Smedt, F. (2017), 'A combined hydrological and hydraulic model for flood prediction in vietnam applied to the huong river basin as a test case study', *Water* **9**(11). Impact Factor=2.069.
URL: *https://www.mdpi.com/2073-4441/9/11/879*

Memon, N., Patel, S. B. & Patel, D. P. (2019), Comparative analysis of artificial neural network and xgboost algorithm for polsar image classification, *in* B. Deka, P. Maji, S. Mitra, D. K. Bhattacharyya, P. K. Bora & S. K. Pal, eds, 'Pattern Recognition and Machine Intelligence', Springer International Publishing, Cham, pp. 452–460. Scopus

ID: 36457207300, ORCID: https://orcid.org/0000-0002-4280-6446, location=Tezpur, India.
URL: *https://link.springer.com/chapter/10.1007/978-3-030-34869-4*$_4$9

Morán-Tejeda, E., Fassnacht, S. R., Lorenzo-Lacruz, J., López-Moreno, J. I., García, C., Alonso-González, E. & Collados-Lara, A.-J. (2019), 'Hydro-meteorological characterization of major floods in spanish mountain rivers', *Water* **11**(12). Impact Factor=2.069.
URL: *https://www.mdpi.com/2073-4441/11/12/2641*

Najmurrokhman, A., Kusnandar, Komarudin, U., Daelami, A. & Arisandy, R. (2019), Development of internet-of-things based building monitoring system for supporting the disaster mitigation in the city, *in* '2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)', pp. 179–183. Yogyakarta, Indonesia, ORCID: https://orcid.org/0000-0002-2674-6009, Scopus ID: 55919091000.
URL: *https://doi.org/10.1109/ICITISEE48480.2019.9003886*

Orton, P., Lin, N., Gornitz, V., Colle, B., Booth, J., Feng, K., Buchanan, M., Oppenheimer, M. & Patrick, L. (2019), 'New york city panel on climate change 2019 report chapter 4: Coastal flooding', *Annals of the New York Academy of Sciences* **1439**(1), 95–114. Impact Factor=4.277.
URL: *https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.14011*

Patrick, L., Solecki, W., Gornitz, V., Orton, P. & Blumberg, A. (2019), 'New york city panel on climate change 2019 report chapter 5: Mapping climate risk', *Annals of the New York Academy of Sciences* **1439**(1), 115–125. Impact Factor=4.277.
URL: *https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.14015*

Sanz-Ramos, Marcos, Amengual, Arnau, Bladé, Ernest, Romero, Romu & Roux, Hélène (2018), 'Flood forecasting using a coupled hydrological and hydraulic model (based on fvm) and highresolution meteorological model', *E3S Web Conf.* **40**, 06028. Lyon-Villeurbanne, France, Cited by 5, ORCID: https://orcid.org/0000-0003-2534-0039.
URL: *https://doi.org/10.1051/e3sconf/20184006028*

Vanichrujee, U., Horanont, T., Pattara-atikom, W., Theeramunkong, T. & Shinozaki, T. (2018), Taxi demand prediction using ensemble model based on rnns and xgboost, *in* '2018 International Conference on Embedded Systems and Intelligent Technology International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES)', pp. 1–6. Khon Kaen, Thailand, Cited by 4, Scopus ID:36598530800.
URL: *https://doi.org/10.1109/ICESIT-ICICTES.2018.8442063*

Xu, L., Skoularidou, M., Cuesta-Infante, A. & Veeramachaneni, K. (2019), 'Modeling tabular data using conditional GAN', *CoRR* **abs/1907.00503**. Cited by 15, ORCID= https://orcid.org/0000-0002-3328-501X.
URL: *http://arxiv.org/abs/1907.00503*

Zhou, Y., Li, T., Shi, J. & Qian, Z. (2019), 'A ceemdan and xgboost-based approach to forecast crude oil prices', *Complexity* **2019**, 4392785. Impact Factor = 1.83.
URL: *https://doi.org/10.1155/2019/4392785*