Explorematicsjournal of Innovative Engineering and Technology

Volume: 05 No: 02 | August -2024
ISSN (Online) 2636 – 590
ISSN (Print) 2636 - 591X

MACHINE LEARNING-BASED FLOOD PREDICTION IN DELTA STATE, NIGERIA: A DATA-DRIVEN APPROACH

Obonyano K. N*[1], Okoye. F. A. [2], Nwobodo-Nzeribe H. N. [2], Oleka. C. V.[2]
1 Department of Computer Engineering, Delta State University of Science and Technology
2 Department of Computer Engineering, Enugu State University of Science and Technology
Author for Correspondence: Obonyano K. N; **Email:** kingdomnelson73@gmail.com

Abstract - The problem addressed in this research is the issue of flood early prediction and response in Nigeria. This study aims at modeling a data-driven flood prediction system using machine learning technique. The methodology used for this research is a mixed approach that involves expert consultation, observation, and simulation approaches respectively. The expert consultation was applied through interaction with domain experts in flood management-related agencies such as the Nigerian Meteorological Agency (NIMET) and the Nigerian Hydrological Service Agency (NIHSA). The observation method was applied for data collection, while the simulation method was used to demonstrate the research, evaluate the model, and then discuss the results. In line with this methodology, the research method is dynamic flood modeling which was achieved using mathematical methods. Then the flood prediction and response system was developed with machine learning techniques, specifically neural network algorithms, and then implemented with Python programming and MATLAB classification software. According to the result of the system validation, the various performances show that the model achieved an average Positive Predicted Value (PPV) of 99.22%, and an average True Positive Rate (TPR) of 98.61%. An average False Negative Rate (FNR) of 1.31% was attained and then, the average False Discovery Rate (FDR) obtained by the system was identified to be 0.97%. Finally, the average accuracy of the model was reported to be 98.88%, this indicates that the overall effectiveness of the model in predicting correctly both positive and negative cases is high.
Keywords: Flood; Delta State; Machine Learning; Artificial Neural Network; Data-Driven

1. Introduction

Flood is characterized as an excess of water that submerges and then gradually disperses (Panchal et al., 2019). According to Matthew (2023), flood is defined as a substantial amount of water covering land under the European Union (EU) Floods Directive. Flooding can occur when water overflows from bodies of water, like rivers or lakes, and breaks or overtops, allowing some of the water to outflow its normal confines. In a similar vein, it might happen because of rainwater gathering on saturated ground after a flood.

According to (Panchal et al., 2019), flood occurs when a large body of water overflows and submerges land, also known as a deluge. When the expression "flowing water" is used, it refers to the tide's inflow rather than its outflow. It usually results from the water flowing or sitting outside of the body's normal boundary, and from the amount of water within a body of water, like a river or lake, exceeding the body's complete capacity. It can also happen in rivers when the water is so strong that it flows in the proper direction, commonly around bends or curves. Naturally, this does not apply in situations like flooding in the sea.

One of the most frequent issues in Nigeria is flooding, particularly in the Niger Delta region in the south-south due to the yearly typhoons that inevitably hit the nation. When the ground is saturated and water cannot discharge or cannot discharge rapidly enough to stop accumulating, flooding can occur in level or low-lying places. As water rushes into nearby rivers and streams from the flood plain, this could be followed by a river flood (Okechukwu et al. 2023).

Because flood has many different aspects, alerting populations to imminent calamities can become a complicated task. The difficulty boils down to foreseeing the disaster, informing the appropriate authorities of that forecast, alerting the impacted communities, and evacuating those communities. According to Fortune (2023), each of the steps stated is further divided into a specific set of duties and issues. These are addressed below.

To provide people ample time to evacuate, the early warning system must first predict the disaster rather than just detect it. River flooding requires the system to predict the flood many hours in advance because water can flow down a huge river like ours in a matter of hours, leaving little time to notify the authorities, much less evacuate the community (Tang et al., 2021). An understanding of the pertinent variables this model requires as input and the predicted output of the model, physical measurements of these variables, communication of this data to the computation location or locations, and a computational system to run the variables through the model are all necessary for prediction (Utkarsh and Shangjia, 2022).

A class of machine learning techniques known as artificial neural networks (ANNs) is motivated by the composition and operation of the human brain (Kabari and Mazi, 2020). Artificial neural networks (ANNs) are made up of layers of interconnected nodes, or artificial neurons, that may be trained to identify patterns, anticipate outcomes, and resolve challenging issues. The three main components are the input layer, which gets the initial data, hidden layers, which use weighted connections to process the data, and an output layer, which generates the finished product. To enable the network to learn and modify its parameters in response to the given data, each link has a weight that is changed throughout the training phase (Lavanya, 2019).

Zhang et al. (2022) classified floods using the cyclone global navigation satellite system by employing the spatial interpolation method, which is predicated on previously observed behaviour. Likewise, XGboost was merged with cuckoo search, bacterial foraging, and artificial bee colony (Yongqing et al. 2023); network theory-based detection model (Utkarsh and Dong, 2022); dynamic drainage prediction model (Haocheng et al. 2023); Despite their effectiveness, distributed hydrological models and High-Resolution Rapid Refresh (HRRR) have been used to manage flood (Gustavo et al., 2022). However, the behaviour of flood is extremely nonlinear and varies depending on the environment and geographic location. Therefore, creating a flood prediction system that can accurately forecast an impending flood in Nigeria has significant obstacles such as a lack of understanding of the dynamic behaviour/nature of flood data. This study aims at modeling a data-driven flood prediction system using machine learning technique.

2. Methodology

This study employed a mixed methodology that included expert consultation, observation, and simulation techniques, in that order. Through interactions with subject matter experts in agencies connected to flood control, such as the Nigerian Meteorological Agency (NIMET) and the Nigerian Hydrological Service Agency (NIHSA), the expert consultation was put into practice. The data was gathered using the observation technique which uses hydrological sensors, drones, Geographical Information System (GIS) and doppler radar, model evaluation, and discussion of the findings were done using the simulation approach. According to this methodology, the research techniques include the mathematically achieved dynamic flood modelling. Next, using MATLAB classification software and Python programming, the flood prediction and response system was created using machine learning techniques, notably neural network algorithms.

2.1 The Flood Prediction and Response System

The second objective, which involved applying machine learning to create a flood prediction and response system, was addressed in this portion. Data collection, artificial neural network, model training, and finally the flood

prediction and response system, as shown in Figure 1, are the techniques used.

Figure 1: Flood prediction and response system

3 Data Collection and Processing

The Nigerian Meteorological Agency (NIMET) provided flood data for Delta State from January to December 2011 to 2021. The Nigerian Hydrological Services Agency (NIHSA) provided additional flood data from 2015 to 2021 as a secondary source of data collection. All of this data was combined and imported into Excel software to create a new flood data model. The attributes taken into consideration for the data collection were months, sub-division, annual rainfall intensity, and flood outcome. The data source is rainfall data. These features were picked in order to gain a thorough understanding of the patterns and traits of the rainfall that caused flooding in Delta State within the given time range. Furthermore, information on river discharge obtained from NIMET was integrated to evaluate the total amount of water flowing through the area. As seen in Figure 2, the gathered data underwent thorough processing to yield insightful conclusions.

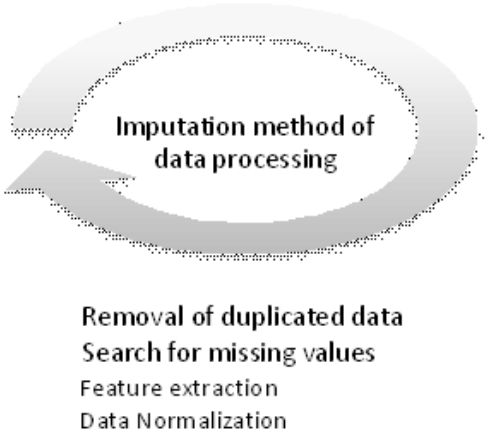

Figure 2: The lifecycle of the data processing steps

The processing stage in Figure 2 used techniques including imputation procedures with Excel software to eliminate all duplicate data. The clustering technique was used for outlier detection, the Fourier Transform algorithm was considered for analyzing seasonal conditions and the Seasonal Normalization approach was applied for data normalization. The search for missing values is an additional data processing step that was used. All missing flood samples in the dataset were automatically replaced. The Analysis of Variance Technique (ANOVA) was used to convert the processed data to geographical disparities and seasonally categorized data. This was accomplished by calculating the precision value of the data and the F-score, which finds the most significant features and computes the balance between the positive and negative ranges of the data values which aligns with the ideas acquired from the domain expert. This was used to evaluate the importance of differences in rainfall patterns and hydrological elements among the several Delta State regions taken into consideration during the data collection process. Prior to training, the features were normalized using the z-normalization approach, which normalises the data points in front of training by utilizing the minimum, maximum, quartile, mean, and standard deviation of the data values. The procedure reduced the dimensionality and preserved the quality of the data by converting the features into a compact feature vector. By concentrating on the most important characteristics for modeling, this technique helps to produce a more concise and easily interpreted feature set that increases the accuracy of flood prediction.

3.1 Proposed flood prediction model

An artificial neural network (ANN) is a network of interconnected massively parallel computational model that simulates human behaviour by processing data from input to output using connection strength (weight), which is acquired by adaptation or learning from a collection of training patterns (Lavanya, 2019). Equation 1 displays the mathematical definition of the ANN process; a computation unit called a neuron receives inputs, processes them, and then outputs the results in a processed form. The weighted sum of the inputs is calculated to obtain the Artificial Neuron's output from the activation function (Pragati, 2021).

$$v_k = \sum_{i=1}^{N} w_{ki} x_i \qquad (1)$$

The neuron's output is obtained by sending the weighted sum v_k as the activation function (Sigmoid activation function) φ input that resolves the output of the specific neuron. $y_k = \varphi(v_k)$. A step function with threshold t can be used to express a simple activation as;

$$\varphi(x) = \begin{cases} 1 & if \quad x \geq t, \\ \quad \vdots & \\ 0 & if \quad x < t \end{cases} \qquad (2)$$

However, bias is most time used instead of a threshold in the network to learn the optimal threshold by itself by adding $x_o = 1$ to every neuron in the network. According to Lavanya (2019), the step activation function for the bias becomes.

$$\varphi(x) = \begin{cases} 1 & if \quad x \geq 0, \\ 0 & if \quad x < 0 \end{cases} \qquad (3)$$

Multiple neurons are employed as a multi-layered network of neurons produced by feeding the output of one neuron to the input of another neuron in order to speed up the learning process and also to enable adaptive learning. Hidden layers are the layers that lie between the input and output layers. A group of neuron nodes representing each input feed with the class of the data set comprise each layer of the multilayer network.

3.2 Training of the model

The optimisation back-propagation approach was used to train the artificial neural network. The neural network automatically divided the file into training, test, and validation sets once the processed data were first inputted. After initialising the neurons, the optimisation back-propagation modifies the weight, bias, learning rate, and momentum of the neurons while observing the gradient loss that transpired throughout the training phase. As the process moves forward, the hyper-parameters which was generated using random search method are updated in light of the gradient loss values that are approaching zero, after which the model is validated and produced as an output. The training process of the model was done in 10 iterations.

3.3 The Flood Prediction and Response Model (FPRM)

This study introduces the FPRM, an approach for improved flood control in Delta State through early flood prediction. The model created by the neural network algorithm's training was used by the FPRM to function. The network of rainfall patterns and features in the Delta State during flood was updated through the use of training data from the flood. When this model receives tested rainfall data as input, it first extracts the features and uses the trained features to classify the data to determine whether or not there is a flood. The model's response component predicts flooding so that residents of the Delta State municipal can evacuate without delay and prepare for any potential flooding. The algorithm for flood prediction and response is described as follows:

Flood Prediction and Response Algorithm

1. Start
2. Initialization of prediction model parameters
3. Load incoming test data
4. Feature extraction
5. Load the initialized trained neural network model
6. Proceed with classification process
7. If
8. Flood features matches = True
9. Return output as flood
10. Else
11. Return the load new data
12. End if
13. End

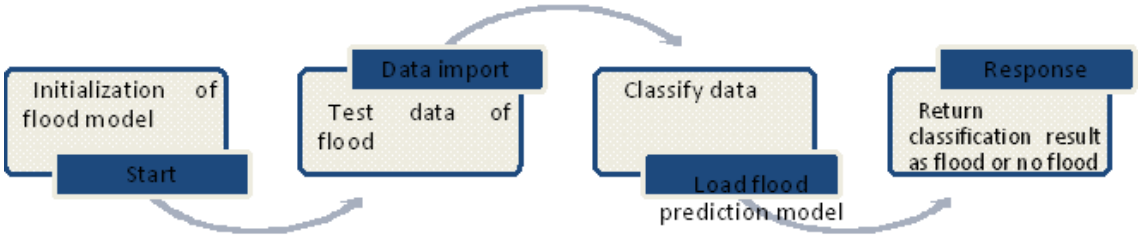

Figure 3: Data flow diagram of the flood prediction and response model

4 Implementation of the model

The Python programming language (Numpy) and MATLAB were used to create the system. The gathered data was loaded into the A.I. programming framework and processed using the Python programming language. To deal with duplicate data issues, the imputation technique was applied. The processed data was then sent to MATLAB's classification learner software, where it was trained.

4.1 Parameters of system evaluation

To properly evaluate the efficacy of the flood prediction model, a multivariate analysis incorporating several important variables must be conducted. The most important of these are loss and accuracy, which are key performance indicators for the model. Accuracy serves as a fundamental measure of the model's capacity to accurately classify cases, offering a percentage representation of correctly predicted outcomes. This statistic is crucial in clarifying the overall precision and dependability of the flood prediction model, revealing insights into its potential to accurately classify between flood and non-flood events. Conversely, loss is a crucial aspect of the assessment procedure since it captures the discrepancy between the model's predictions and the real ground truth. It functions as a numerical measure of how well the model minimises errors during training. A lower loss number highlights the model's ability to learn and generalise patterns within the dataset by showing a tighter match between predicted and actual values. The neural network technique makes use of these two parameters to assess the model and guarantee that the optimal version is produced.

5 Results

This study's flood prediction operation used an ANN model that was trained and tested using the MATLAB neural network classification toolbox. The dataset was divided into training, testing, and validation sets. The model generated during training was then used to test the system using the Delta test dataset, with the results shown in Figure 4 representing the Receiver Operator Characteristics (ROC) result obtained after testing the model.

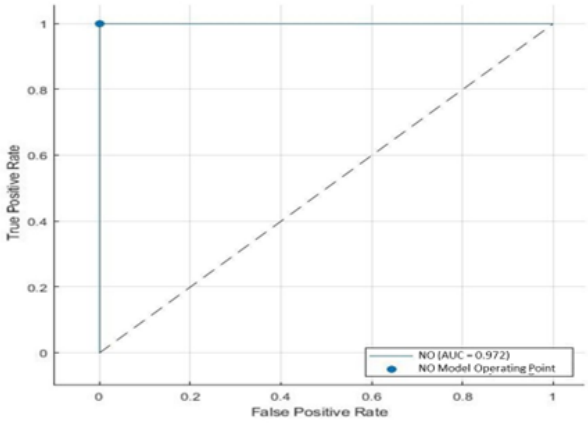

Figure 4: ROC result of the model

Figure 4 presents the flood prediction model's ROC performance, demonstrating the model's capacity for diagnosis. By showing the ratio of accurate positive observations to all actual positives and inaccurate positive observations to all actual negatives, the ROC provides information on true positive and false positive rates. The two-dimensional area beneath the ROC curve is then displayed by the Area Under Curve (AUC) value, which is used to measure it. The AUC of 0.972, as indicated by the ROC analysis, indicates that the ANN model is ideal for flood prediction. Figure 5 displays the model's resultant confusion matrix. Regularization and cross validation technique

was applied to ensure that the proposed ANN model generalizes well to unseen data.

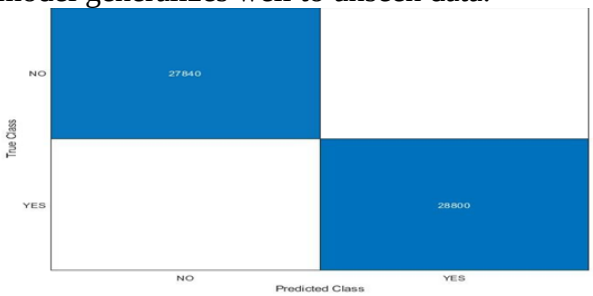

Figure 5: Confusion matrix result of the model

The confusion matrix performance of the model is shown in Figure 5, which visualises a comparison between the actual and predicted classifications. The result of the implementation is shown in Figure 5, where additional examples of the confusion matrix performance of the model are shown. Specifically, 27,840 cases of no flooding condition were determined to be correctly classified as no flood, while 28,800 sets of data were correctly classified as flooding condition.

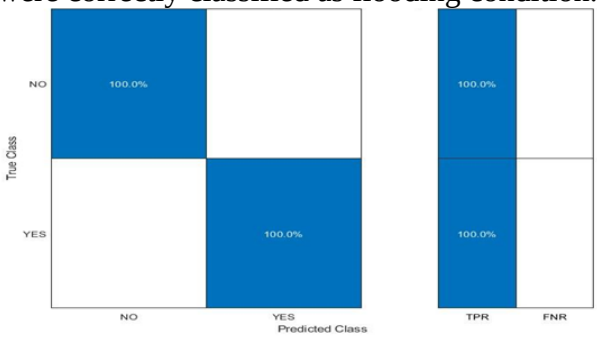

Figure 6: Confusion matrix performance of the model

Figure 6 provides additional examples of the confusion matrix performance, reporting on the true class and predicted class performance of the system model for flood prediction. The model's performance in the true class was 100%, correctly identifying either a flood or no flood, indicating that the system achieved a 100% True Positive Rate (TPR) and 0% False Negative Rate (FNR) in flood prediction. Figure 7 displays the confusion matrix performance, taking into account the model's Positive Predictive Value (PPV) and False Discovery Rate (FDR).

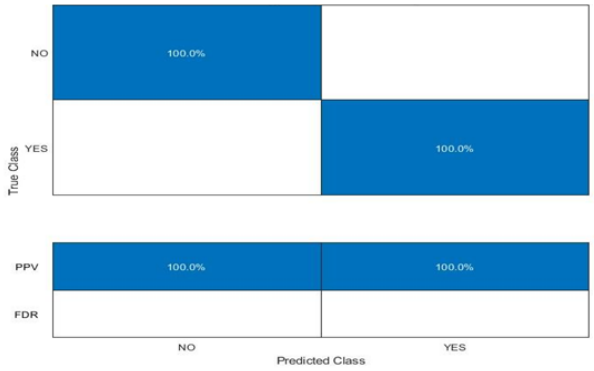

Figure 7: Confusion matrix result of the model

Figure 7 displays the model's confusion matrix performance when PPV and FDR are examined. The PPV result, which describes the model's positive prediction capacity and the proportion of predicted positive cases that are real positives, is included in the result. It also offers information on the model's accuracy in making positive predictions. The false positive errors created by the model in relation to the number of positive predictions made are correlated with the FDR performance, which measures how well the classification model performs in terms of the number of anticipated positives that are false positives. Figure 7 illustrates that the system obtained a PPV result of 100% and an FDR result of 0%, suggesting that the model is ideal.

The different system implementation performances throughout numerous model execution iterations are presented based on the results. The model's average PPV of 99.22% is evident from the data, indicating a high rate of correctly recognised true positives in case prediction and a highly dependable method for positive predictions. With an average TPR of 98.61%, the model demonstrated the system's efficacy in detecting true positive cases and reducing false negatives.

With an average false negative rate (FNR) of 1.31%, the system is good at correctly identifying negatives and has a low false positive rate. The system's average false discovery rate (FDR) was found to be 0.97%, indicating that it can produce high positive predictions that are indeed true positives. Ultimately, the model's average accuracy was reported to be 98.88%, indicating a high level

of overall efficacy in properly forecasting both positive and negative cases.

6 Conclusion

The problem addressed in this study is the issues of flood early prediction and response in Nigeria. While many studies have focused on solving this problem considering the hydrological flood model in their environment; flood as a model is very dynamic and not the same for every region. This means that a model developed after charactering flood in Europe for instance, may not be the best to address similar problem in Nigeria. In order to do this, a solution that best describes the behaviour of floods in Nigeria must be found, and a prediction model must be created. Along with the numerous other issues previously listed in the research background, this problem has led to further issues such as the delayed identification of floods, the loss of lives and property. Many individuals who live in flood-prone areas, especially those in the many southern Nigerian states, will benefit from this study.

Therefore, there is need for this research which proposes to develop a data-driven model for flood prediction using machine learning technique. In this study, the validation of the result obtained is performed using a 10-fold validation technique. This is done to ensure the model adopted meets with the intended purpose and requirement for implementation and to determine that it operates in its full specifications and operates correctly in the intended environment.

6.1 Recommendations

To enhance this study of early flood prediction and response system in Delta State region of Nigeria, it is recommended that future research works consider the integration of regional real-time data of the locality while considering the geographic and seasonal variability of the zone being considered. Then, it is equally recommended that mobile Application should be integrated to improve the response of the flood prediction system. By implementing these recommendations, this study will contribute significantly to the adaptive and locally tailored flood prediction model for Delta State of Nigeria, which will reduce the impact of flood and improve the livelihood and survivability of the local indigenes.

References

Fortune E. (2023)"Huge Flood Looms In Nigeria As Cameroon Set To Open Lagdo Dam" https://www.vanguardngr.com/2023/08/huge-flood-looms-in-nigeria-as-cameroon-set-to-open-lagdo-dam/

Gustavo de A., Celso M., James L., (2022) "Multiscale and multi event evaluation of short-range real-time flood forecasting in large metropolitan areas" Journal of Hydrology; Volume 612, Part C, 128212

Haocheng H., Xiaohui L., Weihong L., Ziyuan W., Mingshuo Z., Hao W., & Lizhong J., (2023) "Effects analysis and probability forecast (EAPF) of real-time management on urban flooding: A novel bidirectional verification framework" Science of The Total EnvironmentIn Press, Journal Pre-proofWhat's this?

Kabari L., & Mazi Y., (2020) "Rain–Induced Flood Prediction for Niger Delta SubRegion of Nigeria Using Neural Networks", EJERS, European Journal of Engineering Research and Science. DOI: http://dx.doi.org/10.24018/ejers.2020.5.9.2114

Lavanya S. (2019) "Designing your neural networks" https://towardsdatascience.com/designing-your-neural-networks-a5e4617027ed;

Matthew O., (2023)" Over 2 million Nigerians displaced by flood in 2022, says NEMA"https://guardian.ng/news/over-2-million-nigerians-displaced-by-flood-in-2022-says-nema/

Okechukwu Nnodim, John Charles, Collins Sunday, Justin Tyopuusu, Ikenna Obianeri and Chika Otuchikere (2023)"Cameroon dam opening: Evacuate to prevent deaths, FG, states warn flood-prone communities" https://punchng.com/cameroon-dam-opening-evacuate-to-prevent-deaths-fg-states-warn-flood-prone-communities/;

Panchal U., Ajmani H., and Sait Y. (2019), "Flooding Level Classification by Gait

Analysis of Smartphone Sensor Data," IEEE Access, vol. 7, ISSN: 181678–181687; pp. 201-216

Pragati B. (2021) "The essential guide to neural network architectures" https://www.v7labs.com/blog/neural-network-architectures-guide;

Tang W., Zhan W., Jin B., Motagh M., and Xu P. (2021) "Spatial Variability of Relative Sea-Level Rise in Tianjin, China: Insight fromInSAR, GPS, and Tide-Gauge Observations." IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 14, pp. 2621–2633.

Utkarsh G., & Shangjia D., (2022)"Critical facility accessibility rapid failure early-warning detection and redundancy mapping in urban flooding" Reliability Engineering & System Safety"; Volume 224, September 2023, 108555

Yongqing L., Xin L., Brian T., Qin C., & Navid J., (2023) "V-FloodNet: A video segmentation system for urban flood detection and quantification" Environmental Modelling & Software; Volume 160, https://doi.org/10.1016/j.envsoft.2022.105586Get rights and content

Zhang S., Zhongmin M., Qi L., Shengwei H., Yuxuan F., Hebin Z., & Qinyu G., (2022) "POBI interpolation algorithm for CYGNSS near real time flood detection research: A case study of extreme precipitation events in Henan, China in 2021" Advances in Space Research; Volume 71, Issue 6, Pages 2862-2878