Flood Prediction Using Machine Learning Models: A Case Study of Kebbi State Nigeria

Zaharaddeen Karami Lawal
Department of Computer Science
Federal University Dutse
Dutse, Nigeria
deenklawal13@gmail.com

Hayati Yassin
Faculty of Integrated Technolgy
Universiti Brunei Darussalam
Brunei Darussalam
hayati.yassin@ubd.edu.bn

Rufai Yusuf Zakari
Department of Computer Science
University of Electronic Science and Technology of China
Chengdu, China
rufaig6@gmail.com

2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) | 978-1-6654-9552-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/CSDE53843.2021.9718497

***Abstract*— Machine Learning (ML) models for flood prediction can be beneficial for flood alerts and flood reduction or prevention. To that end, machine-learning (ML) techniques have gained popularity due to their low computational requirements and reliance mostly on observational data. This study aimed to create a machine learning model that can predict floods in Kebbi state based on historical rainfall dataset of thirty-three years (33), so that it can be used in other Nigerian states with high flood risk. In this article, the Accuracy, Recall, and Receiver Operating Characteristics (ROC) scores of three machine learning algorithms, namely Decision Tree, Logistic Regression, and Support Vector Classification (SVR), were evaluated and compared. Logistic Regression, when compared with the other two algorithms, gives more accurate results and provides high performance accuracy and recall. In addition, the Decision Tree outperformed the Support Vector Classifier. Decision Tree performed reasonably well due to its above-average accuracy and below-average recall scores. We discovered that Support Vector Classification performed poorly with a small size of dataset, with a recall score of 0, below average accuracy score and a distinctly average roc score.**

***Keywords*— Flood Prediction, Machine Learning, Decision Tree, Logistic Regression, Support Vector Regression.**

I. INTRODUCTION

Natural disasters have always been a part of human history, no matter where they occurred. Terrorist attacks, chemical, biological, radiological, and nuclear threats, as well as human-caused calamities, represent a threat to natural life and human beings. One of the major concerns of authorities as well as individual members of society is the reality of potential risks and disasters. Natural calamities are unavoidable, as we all know. Pre-alarming systems and good management, on the other hand, can reduce their severity and impact. [1]. Early identification of natural disasters, such as floods, can substantially aid humans in minimizing the damage caused by such calamities. [2]. Flood prediction models play a significant role in hazard assessment and extreme event management. Water resource management techniques, policy ideas and analyses, and future evacuation modeling all benefit from robust and accurate prediction. [3]. Many flood control measures are being put in place to help mitigate the damage caused by it. However, in places prone to flash floods, these flood control methods have not been as effective. Machine Learning (ML) has aided in the prevention of numerous natural disasters such as floods due to its capacity to forecast future events. [4]. Machine learning allows you to learn from previous data. It also develops models for future prediction based on historical data. [5].

Flooding is a common environmental issue in Nigeria, and it occurs when a body of water flows over and above an area of land that is not ordinarily inundated. It could alternatively be defined as the temporary flooding of an area that is not regularly flooded by a temporary rise in the level of a stream, river, lake, or sea. [6]. Kebbi state is in the North West geopolitical zone of Nigeria[7]. The main source of income for the majority people in this area is agriculture, which is dependent on rainfall and other factors. Due to the River Niger's pathway from the Benin Republic down to Niger and Kwara states, as well as its link with the River Benue in Kogi state, the majority of Kebbi state's regions are prone to floods. For decades, flooding in Kebbi and other Nigerian regions has killed hundreds of people and ruined property worth millions of dollars. The goal of this study is to create a machine learning model that can predict floods in Kebbi state based on historical rainfall data, so that it can be used in other Nigerian cities with high flood risk.

Predicting rainfall is an application of science and technology for predicting the amount of rain over an area[8]. Machine learning applications can help predict and detect floods, which is a problem that has to be addressed. Furthermore, it is an unavoidable task to withstand the flood's devastation provided there is a feasible means to inform the populace living in the area in a timely and appropriate manner [1][9][10].

Although flood prediction systems have advanced in recent years, many other emerging technologies, which are severely limited in developing countries, will not be able to accurately predict flood conditions unless a predictive model with a high level of precision is established and tested. If a model is successful in predicting the occurrence of a flood, it is vital to evaluate its predictive capacity over a range of historical periods. To choosing the appropriate prediction model from a variety of supervised machine learning techniques. This is the subject of the following research paper.

II. LITERATURE SURVEY

Several data-driven early flood prediction systems have been developed over the years. Machine learning is a popular branch of artificial intelligence that arose from the advancement of self-learning algorithms to extract knowledge from data in order to make predictions. In this section, we'll look at a few of these studies.

[8] To predict floods, it employs machine learning and neural networks. Based on past rainfall data, the best of the two approaches is chosen for prediction..[11] To predict the occurrence of floods caused by rainfall, a prediction model was built utilizing rainfall data. Based on the rainfall range for certain places, the model forecasts whether or not a "flood" would occur. The prediction model is based on rainfall data from Indian districts. Linear Regression, K-Nearest Neighbor, Support Vector Machine, and Multilayer Perceptron are among the techniques used to train the dataset. Among these, the MLP algorithm fared well, with a precision of 97.40 %. [1] Proposed an innovative and robust model that is a real-time flood detection system based on Machine Learning and Deep Learning; Random Forest, Naive Bayes J48, and Convolutional Neural Networks that can detect water level and assess floods with potential humanitarian effects before they occur.. [12] The neural network output data is compared to the original or actual target data during the prediction process. Error calculation is performed, which provides a measurement of how well the network can learn, allowing for easy comparison of new patterns. [13] Summarizes the ML approaches for flood forecasting proposed in this special issue, as well as their considerable benefits. Then, using cutting-edge machine learning, visualization, and system development techniques, it creates an intelligent hydro informatics integration platform (IHIP) to create a user-friendly web interface system for boosting online forecast capabilities and flood risk management. [14] Their research aimed to see how well ML models could estimate flood stages at a crucial gauge station using mostly upstream stage observations, while downstream levels should be provided to account for backwater if it exists. The lower Parma River (Italy) was chosen as the case study for this investigation, and the forecast horizon was extended to 9 hours. The accuracy and computational time of three machine learning techniques, namely Support Vector Regression (SVR), Multi-Layer Perceptron (MLP), and Long Short-term Memory (LSTM), were compared. [15] Their study introduces the most promising long-term and short-term flood prediction approaches. The important developments in enhancing the quality of flood prediction models are also looked into. Hybridization, data decomposition, algorithm ensemble, and model optimization are among the most effective tactics for improving ML approaches, according to the researchers. [16] The goal of this study was to develop an effective flood simulation framework that could be used in large-scale simulations. Using a hybrid of hydraulic model and machine learning methods, the framework provides an innovative, rapid, efficient, and extensible approach for identifying flooded areas and flood depth. To do so, two Machine learning models were trained to forecast river depth over the domain for an arbitrary discharge using a two-dimensional hydraulic model (iRIC) calibrated by measured water surface elevation data.[17] They used a non-tuned machine learning technique called self-adaptive evolutionary extreme learning machine (SaE-ELM) to create an expert prediction model in the chapter they presented. The SaE-primary ELM's use is the prediction of river water levels. In water resources management and flood prediction, developing such water level prediction and monitoring models is a critical optimization problem..[18] Introduced RF as a viable alternative to SVM, which consistently outperforms SVM in flood prediction models. [19] In general applications to floods, researchers compared the performance of ANN, SVM, and RF, finding that RF produced the best results.[20] For flood forecasting, the Support Vector Machine (SVM) was used. They did note, however, that while their objective model outperformed the benchmark models, it took a significant amount of effort to secure the objective model's superior performance in the absence of advanced flood monitoring equipment.

.

III. METHODOLOGY

A. PROBLEMS AND MOTIVATION

Due to the misleading and creeping nature of flood occurrences, designing flood prediction systems for early flood alerts is difficult. This is owing to the fact that early warning system design necessitates a thorough understanding of a variety of technologies. [2] [21]. It's evident that this will provide a greater difficulty for emerging countries like Nigeria. As a result, a cost-effective solution that involves only a little investment in such technology is required for flood prediction. The enhanced flood modeling method presented in this paper will overcome these difficulties. A data-driven model that employs only monthly rainfall data to deliver an accurate result is meant to be a cost-effective choice for countries with limited resources where technological breakthroughs have not yet penetrated. The recurrent destruction caused by flood events in the current study area over many years is another driving element behind our research.

By allowing the general public and rescue groups to be better prepared for future flood disasters, the new flood prediction model can be used to minimize the negative impact of floods not only in Kebbi state, but all over the country. This implies that the study will focused on flood disaster preparedness and prevention thru the implementation of a prediction model.

B. MACHINE LEARNING METHODS

- **Supervised**: Supervised learning is a machine learning activity that involves learning a function that maps an input to an output based on sample input-output pairs**.** [22][23]. To infer a function, it uses labeled training data and a set of training examples. When certain goals are determined to be achieved from a specific set of inputs, supervised learning is used. [22] [24], i.e., a task-oriented strategy. The most typical supervised tasks are data

separation (classification) and data fitting (regression). [22].

- **Unsupervised**: Unsupervised learning is a data-driven method that examines unlabeled datasets without the need for human intervention[22][23]. This is commonly used for generating feature extraction, discovering important trends and structures, groupings in results, and experimental reasons. Clustering, density estimation, feature learning, dimensionality reduction, identifying association rules, anomaly detection, and other unsupervised learning tasks are the most prevalent [22].

According to the literature review, flood prediction models were built using a combination of supervised and unsupervised machine learning methodologies. The following is a brief review of a few of them.

- **Decision Tree (DT):** One of the contributors to predictive modeling is DT's machine learning approach, which has a wide range of applications in flood simulation. A decision tree is used throughout DT, from the branches to the leaf target values. In classification trees (CT), the final variables in a DT are a discrete collection of values, with leaves representing class labels and branches indicating feature label conjunctions. [15]. A regression tree is used when the goal variable in a DT has continuous values and an ensemble of trees is used (RT) [15][25]. Regression and classification trees share some similarities and differences.

- **K-nearest neighbors (KNN):** K-Nearest Neighbors (KNN) [22][26] is a non-generalizing learning or "instance-based learning" algorithm, often known as a "lazy learning" algorithm. Rather of building a broad internal model, it keeps all instances corresponding to training data in n-dimensional space. KNN is a machine learning algorithm that uses data to classify new data points using similarity measurements (such as the Euclidean distance function).)[22]. A simple majority vote of each point's k nearest neighbors is used to classify it. It is relatively unaffected by noisy training data, and accuracy is dependent on data quality. The most difficult aspect of KNN is determining the ideal number of neighbors to consider. The KNN can be used for both classification and regression.

- **Logistic regression (LR):** is a probabilistic-based statistical model that is commonly used to address classification problems in machine learning (LR)[22]. To estimate the probabilities, logistic regression commonly employs a logistic function, which is also known as the mathematically defined sigmoid function.

It works well when the dataset can be divided linearly and can over fit high-dimensional datasets. L1 and L2 regularization approaches [22] In such cases, can be employed to avoid over-fitting. A key disadvantage of Logistic Regression is the assumption of linearity between the dependent and independent variables. It can be used to solve both classification and regression problems, however classification is the most typical application.

- **Support vector machine (SVM):** A support vector machine is a machine learning technology that can be used for classification, regression, and other applications (SVM)[27]. A support vector machine creates a hyper-plane or set of hyper-planes in high- or infinite-dimensional space. Intuitively, the hyper-plane with the greatest distance from the nearest training data points in each class obtains a significant separation since, in general, the larger the margin, the smaller the generalization error of the classifier. It works well in high-dimensional spaces and might act differently depending on the kernel, which is a set of mathematical functions. Linear, polynomial, radial basis function (RBF), sigmoid, and other terms are used to describe many types of functions[22].

- **Random Forests (RF):** is a technique for supervised learning. It creates a "forest" out of an ensemble of decision trees, which are commonly trained using the "bagging" method. The bagging method's basic premise is that combining several learning models improves the overall output. Simply put, random forest combines many decision trees to produce a more accurate and stable prediction. [28].

In this research, a combination of supervised and unsupervised machine learning algorithms would be applied. These are: Decision Tree (DT), Logistic Regression (LR), and Support Vector Classifier (SVC**).**

IV. EXPERIMENT

A. Study Area

Birnin Kebbi city, as well as other towns in Kebbi state, are the focus of this research. Kebbi is a Nigerian state in the northwestern part of the country, covering 36,800 km2 (14,200 sq mi). During this time, the rainy season in Kebbi is normally between April and October. From June through September, there is a lot of rain. A regular flood phenomenon around flood-prone locations in Kebbi state may occur during this time.

B. Dataset

The daily rainfall data for Kebbi state for thirty-three years, from 1st January 1981 to 31st December 2013, was successfully acquired from the Nigerian Meteorological Agency (NiMet) Yelwa station in Kebbi state. 33 years rainfall data consists of (12,053 data points at daily time-steps). The available data was analyzed prior to data pre-processing.

During data preprocessing, the following actions were taken for simpler computations and more accurate results. Firstly, the sum of the daily rainfall for each month was taken. Secondly, the sum of rainfall from January-February, March-May, June-September, and October-December of each year,

as well as the annual rainfall of each year was also taken. To fill in the values for missing data points, the calendar mean was employed.

The data was separated into training and testing subsets during the data pre-processing phase. Using the train test split model selection method, 80% of this data were assigned for model training while 20% for testing.

C. Machine Learning Model Design

The Python programming language was used to create a machine learning-based flood prediction model. Python was chosen to develop the predictive models because it provides an efficient environment for machine learning data analysis. Scikit-Learn and other popular Python machine learning tools have also been used to solve machine learning difficulties.[2]-[29]. The two core data structures of pandas, Series (1-dimensional) and Data Frame (2-dimensional), handle the great majority of usual use cases in finance, statistics, social science, and many fields of engineering. Pandas is based on NumPy and is designed to work nicely with a variety of other third-party libraries in a scientific computing environment.[30].

V. RESULT AND DISCUSSION

The results, and discussion of the preprocessed dataset and the performance evaluation of the machine learning model Used to predict future flood situations in Kebbi state are presented in this section.

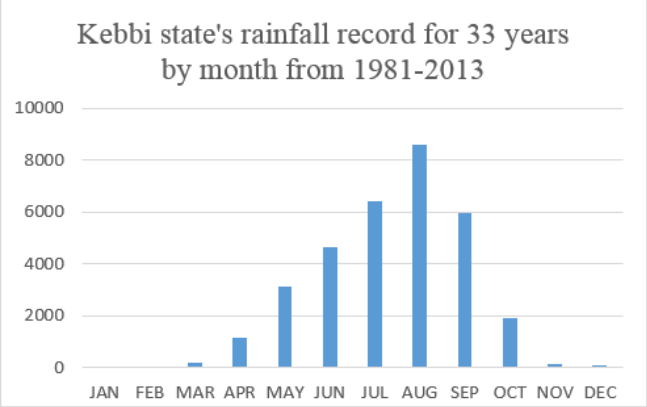

Fig. 1. Rainfall record for 33 years by months

For the past 33 years, the average annual rainfall in Kebbi state has been 977.97mm. The average annual rainfall was used to create a baseline. When the total yearly rainfall is below or equals the average annual rainfall (≤ 977.79mm), we presume that the flood will not occur that year. When the total yearly rainfall exceeds the average rainfall (> 977.79mm), we assume that a flood is likely to occur.

In our dataset table, we added a feature Flood and labeled annual rainfall that was below average as 'NO' and above average as 'YES under the newly added feature. Yes and No were swapped out for 1 and 0 during the data analysis.

Three simple machine learning algorithms to predict the accuracy of the flood occurrence are implemented. The desired algorithm shows the results of occurrence of flood in the upcoming years.

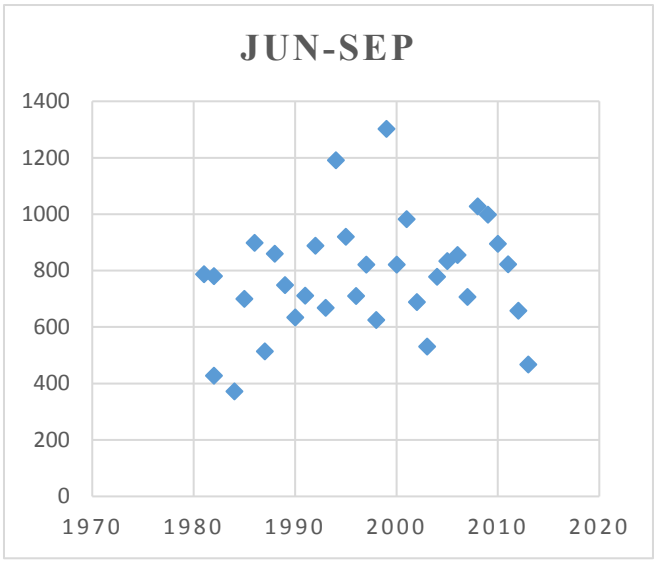

Fig. 2. June – September Rainfall record for 33 years

Model	Accuracy Score	Recall Score	ROC Score
Decision Tree (DT)	57.143	40	70
Logistic Regression (LR)	85.714	100	75
Support Vector Classification (SVC)	28.571	0	50

TABLE I. COMPARING THE PERFORMANCE OF THREE MACHINE LEARNING MODELS

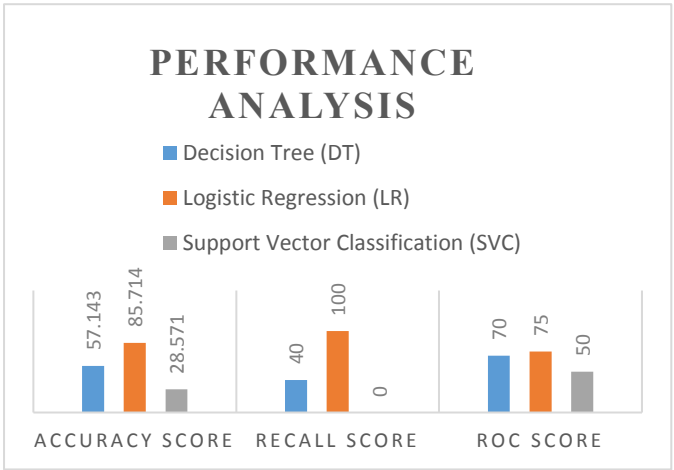

Fig. 3. Performance analysis of the models

VI. Conclusion

In this study, we looked at the case study of Kebbi State, Nigeria, to see if machine-learning algorithms could be used to build a flood prediction model based solely on historical rainfall data. This research focuses on predicting flood disaster preparedness and prevention stages in particular. The prediction models for this case study were created using the historical rainfall data that we have on hand. In terms of accuracy, recall, and roc, we found that logistic regression performed the best of the three methods we used to develop our model.

Although there are certain limitations to this study, the size of our dataset is only 33 years and we employed only two classification and one regression algorithms. We'll need at least one hundred years of historical rainfall data to get the best results. Another limitation of this study is that, aside from Kebbi, several other Nigerian states in nearly all six geopolitical zones are prone to flooding. In the future, we intend to include more states in our research and create our models using advanced machine learning and deep learning algorithms.

Lastly, the authors views and opinions in this article are their own and do not reflect those of the Nigerian government.

References

[1] A. O. Hashi, A. A. Abdirahman, M. A. Elmi, and S. Z. Mohd, "A Real-Time Flood Detection System Based on Machine Learning Algorithms with Emphasis on Deep Learning," vol. 69, no. 5, pp. 249–256, 2021, doi: 10.14445/22315381/IJETT-V69I5P232.

[2] M. Moishin, R. C. Deo, R. Prasad, N. Raj, and S. Abdulla, "Designing deep-based learning flood forecast model with ConvLSTM hybrid algorithm," *IEEE Access*, vol. 9, pp. 50982–50993, 2021, doi: 10.1109/ACCESS.2021.3065939.

[3] A. Mosavi, P. Ozturk, and K. Chau, "Flood Prediction Using Machine Learning Models : Literature Review," no. Idi.

[4] V. B. S and S. Sandhya, "Flood Prediction System using Multilayer Perceptron Classifier and Neural Networks," no. May, pp. 6245–6254, 2020.

[5] N. Ahamed and S. Asha, "Flood prediction forecasting using machine Learning Algorithms," vol. 11, no. 12, pp. 543–546, 2020.

[6] O. Agbonkhese, E. G. Aka, J. Ocholi, and M. Adekunle, "Flood Menace in Nigeria : Impacts , Remedial and Management Strategies," vol. 6, no. 4, pp. 32–41, 2014.

[7] S. E. E. Profile, "Trend analysis of precipitation in Birnin Kebbi , Nigeria," no. June, 2014.

[8] K. Vamshi, S. K. S, B. R. Muralidhar, N. Manjunath, and P. Savitha, "A Review on Rainfall Prediction using Machine Learning and Neural Network," pp. 2763–2769, 2021.

[9] B. Choubin, S. Khalighi-Sigaroodi, A. Malekian, S. Ahmad, and P. Attarod, "Drought forecasting in a semi-arid watershed using climate signals: a neuro-fuzzy modeling approach," *J. Mt. Sci.*, vol. 11, no. 6, pp. 1593–1605, 2014, doi: 10.1007/s11629-014-3020-6.

[10] M. Khalaf, A. Hussain, D. Al-Jumeily, P. Fergus, and I. Idowu, "Advance flood detection and notification system based on sensor technology and machine learning algorithm," *2015 Int. Conf. Syst. Signals Image Process.*, pp. 105–108, 2015.

[11] A. Vinothini, L. Kruthiga, and U. Monisha, "Prediction of Flash Flood using Rainfall by MLP Classifier," no. 1, pp. 425–429, 2020, doi: 10.35940/ijrte.F9880.059120.

[12] "PREDICTING FLOOD USING ARTIFICIAL NEURAL NETWORKS," vol. 15, no. 1, pp. 53–57, 2020.

[13] F. Chang, K. Hsu, and L. Chang, *Flood Forecasting Using Machine Learning Methods*. .

[14] S. Dazzi, R. Vacondio, and P. Mignosa, "Flood Stage Forecasting Using Machine-Learning Methods : A Case Study on the Parma River (Italy)," 2021.

[15] A. Mosavi, P. Ozturk, and K. W. Chau, "Flood prediction using machine learning models: Literature review," *Water (Switzerland)*, vol. 10, no. 11, pp. 1–40, 2018, doi: 10.3390/w10111536.

[16] H. Hosseiny, F. Nazari, V. Smith, and C. Nataraj, "OPEN A Framework for Modeling Flood Depth Using a Hybrid of Hydraulics and Machine Learning," *Sci. Rep.*, pp. 1–14, 2020, doi: 10.1038/s41598-020-65232-5.

[17] Z. M. Yaseen and I. Ebtehaj, "Hybrid Data Intelligent Models and Applications for Water Level Prediction," no. 1, pp. 121–139, doi: 10.4018/978-1-5225-4766-2.ch006.

[18] M. S. Tehrany, B. Pradhan, and M. N. Jebur, "Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS," *J. Hydrol.*, vol. 512, pp. 332–343, May 2014, doi: 10.1016/j.jhydrol.2014.03.008.

[19] D. T. Bui, T. A. Tuan, H. Klempe, B. Pradhan, and I. Revhaug, "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree," *Landslides*, vol. 13, pp. 361–378, 2015.

[20] D. Han, L. Chan, and N. Zhu, "Flood forecasting using support vector machines," *J. Hydroinformatics*, vol. 9, no. 4, pp. 267–276, Oct. 2007, doi: 10.2166/hydro.2007.027.

[21] V. V Krzhizhanovskaya *et al.*, "Flood early warning system: design, implementation and computational modules," *Procedia Comput. Sci.*, vol. 4, pp. 106–115, 2011, doi: https://doi.org/10.1016/j.procs.2011.04.012.

[22] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[23] J. Han, M. Kamber, and J. Pei, "8 - Classification: Basic Concepts," in *Data Mining (Third Edition)*, Third Edit., J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 327–391.

[24] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *J. Big Data*, vol. 7, no. 1, p. 41, 2020, doi: 10.1186/s40537-020-00318-5.

[25] G. D ath and K. Fabricius, "CLASSIFICATION AND REGRESSION TREES: A POWERFUL YET SIMPLE TECHNIQUE FOR ECOLOGICAL DATA ANALYSIS," *Ecology*, vol. 81, pp. 3178–3192, 2000.

[26] K. P. Georgakakos and J. A. Sperfslage, "Operational Rainfall and Flow Forecasting for the Panama Canal Watershed BT - The Río Chagres, Panama: A Multidisciplinary Profile of a Tropical Watershed," R. S. Harmon, Ed. Dordrecht: Springer Netherlands, 2005, pp. 325–335.

[27] S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy, "Improvements to Platt's SMO Algorithm for SVM Classifier Design," *Neural Comput.*, vol. 13, pp. 637–649, 2001.

[28] Z. Wang, C. Lai, X. Chen, B. Yang, S. Zhao, and X. Bai, "Flood hazard risk assessment model based on random forest," *J. Hydrol.*, vol. 527, pp. 1130–1141, 2015.

[29] D. K. Barupal and O. Fiehn, "Generating the blood exposome database using a comprehensive text mining and database fusion approach," *Environ. Health Perspect.*, vol. 127, no. 9, pp. 2825–2830, 2019, doi: 10.1289/EHP4713.

[30] W. McKinney and P. D. Team, "Pandas - Powerful Python Data Analysis Toolkit," *Pandas - Powerful Python Data Anal. Toolkit*, p. 1625, 2015