Interactive Flood Inundation Probability Mapping: A Web-Based Application with Machine Learning Models

Anubhav Kumar
Computer Science
KIET Group of Institutions
Abdul Kalam Technical University
Ghaziabad, Uttar Pradesh, India
anubhav.2125cs1142@kiet.edu

Prajjwal Dwivedi
Computer Science
KIET Group of Institutions
Abdul Kalam Technical University
Ghaziabad, Uttar Pradesh, India
prajjwal.2125cs1221@kiet.edu

Ritik Sharma
Computer Science
KIET Group of Institutions
Abdul Kalam Technical University
Ghaziabad, Uttar Pradesh, India
ritik.2125cs1210@kiet.edu

Adarsh Chaudhary
Computer Science
KIET Group of Institutions
Abdul Kalam Technical University
Ghaziabad, Uttar Pradesh, India
adarsh.2125cs1218@kiet.edu

Anurag Mishra
Computer Science
KIET Group of Institutions
Abdul Kalam Technical University
Ghaziabad, Uttar Pradesh, India
anurag.mishra@kiet.edu

***Abstract*— The foremost frequent normal fiasco within the world, flooding influences hundreds of millions of individuals and murders between 6,000 and 18,000 individuals every year. Flood forecast is an basic viewpoint of calamity chance decrease and administration. This paper looks at the application of machine learning methods, particularly classification based model logistic regression and Gradient Boosting models like XGBoost, CatBoost, to anticipate flood likelihood. Two datasets with particular highlights are dissected. Utilizing two unmistakable datasets consolidating natural and geographical components, a vigorous prescient model is created to estimate flood events. The paper highlights the potential of machine learning methods to upgrade flood hazard appraisal while utilizing web applications for the common users.**

Keywords*— Classification based, Regression based, Gradient Boosting models, XGBoost model, CatBoost Model, Logistic Regression, flood hazard, APIs, ANN, Categorical factors, encoding, standardized, sklearn, confusion matrix, Precison, Recall, F1-score, MSE , R^2 score, web application, Normalization, Scaling, learning rate, estimators, Hyperparameters, DEM records, Hydrological modelling, Predicted values, dataset-1 , dataset-2, Accuracy, multilayer perceptron, Susceptibility maps, Synthetic Aperture Radar, one-hot encoding, Label encoding, n_estimators, max_depth, loss function, RMSE, CatBoostRegressor, XGBoostRegressor.

I. INTRODUCTION

Flooding is the foremost visit and obliterating normal catastrophe around the world, influencing millions of lives and causing gigantic financial misfortunes each year. Between 6,000 and 20,000 fatalities are detailed yearly due to floods, whereas incalculable others confront relocation, infrastructure harm, and disturbances to jobs. The expanding recurrence and seriousness of floods are closely connected to climate alter, urbanization, and deforestation, which worsen helplessness in both urban and country zones. As a result, anticipating and mitigating the affect of floods has gotten to be a basic need for governments, communities, and catastrophe administration agencies. Traditional flood expectation strategies depend on hydrological and meteorological information, but they regularly drop short due to the dynamic nature of flooding and the expanding complexity of natural variables. Machine learning (ML), with its capacity to analyze tremendous datasets and reveal hidden patterns, offers a transformative approach to flood forecast. ML models can integrate different inputs, including precipitation, topography, land utilize, and infrastructure conditions, to produce precise and timely predictions. This research centers on leveraging machine learning models like logistic regression, gradient boosting. Utilizing two datasets, this model incorporates different features such as monsoon intensity, drainage quality, urbanization levels, and more , to evaluate flood probabilities successfully. Furthermore, a user-friendly web application is proposed to empower real-time flood chance evaluation based on geological information fetched from APIs. By giving noteworthy bits of knowledge, this work points to back disaster readiness, minimize misfortunes and improve community versatility to flood dangers.

Overall, this research presents a novel approach to predict flood occurrence with utilizing web applications that can significantly improve the efficiency and accuracy of the disaster management.

II. Related Work

In a study by Amir Mosavi, Pinar Ozturk & Kwok-wing chau [2018] [1], the authors proposed a paper that introduces the most promising prediction methods for both long-term and short-term floods, emphasizing the role of machine learning models in flood prediction.

A study by Zaharaddeen Karami Lawal, Hayati Yassin & Rufai Zakari Yusuf [2021] [2]. 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 a study by Nguyen, D.L., Chou, T. V. and chen, M. H. [3], the authors proposed a paper that aims to construct flood susceptibility maps in the Huong Khe district using three machine learning algorithms, namely the K - Nearest Neighbour (KNN), the Support Vector Machine (SVM) and Artificial Neural Network (ANN).

In a study by Sliman Hitouri, Meriame Mohajane, Meriam Lahsaini, Sk Ajim Ali, Tadesual Asamin Setargie, Gaurav Tripathi, Paola D'Antonio, Suraj Kumar Singh and Antonietta Varasano [4], the authors proposed a paper that a novel approach for flood susceptibility mapping by integrating Synthetic Aperture Radar (SAR) data with machine learning techniques.

In a study by Yogesh Bhattarai, Sunil Duwal, & Akitoshi Hanazawa[2024][5], authors proposed a paper that uses machine learning, including ANN, SVM, LSTM, and RF, for flood susceptibility mapping in Laos, highlighting the potential of open-source data in flood risk assessment.

In a study by Adel Rajab, Hira Farman, Norman Islam, Darakhshan Syed, M.A. Elmagzoub, Asadullah Shaikh, Muhammad Akram & Mesfer Alrizq[2023][6], This research focuses on leveraging historical meteorological data to find trends using machine learning and deep learning approaches to estimate rainfall.

In a study by Rahma Khalid & Usman T. Khan [2024][7], this paper explores the influence of model development and validation in FSM using Artificial Neural Networks (ANNs).

In a study by Mohammad Ahmadlou, A'kif Al-Fugara, Abdel Rahman Al-Shabeeb, Aman Arora, Rida Al-Adamat, Quoc Bao Pham, Nadir Al-Ansari, Ngyuen Thi Thuy Linh & Hedieh Sajedi[8], they proposed a novel hybrid model combining the multilayer perceptron (MLP) and autoencoder models to produce susceptibility maps for two study areas, enhancing flood risk assessment.

In a study by Md Hasanuzzaman, Aznarul Islam, Biswajit Bera, Pravat Kumar Shit [2022][9], they proposed a paper about a comparison of performance measures of three machine learning algorithms for flood susceptibility mapping of river Silabati (tropical river, India).

III. Methodology

The idea proposed for this research paper is to make a user-friendly web application where user can chose a location and get the flood probability. The probability are calculated using Machine Learning models.Here are the modules for the proposed approach:

1-Data collection and Preprocessing: The essential objective was to accumulate significant information that seem impact flood probabilities. This included topographical, climatic, and socio-economic factors. Two datasets were utilized. Missing values were dealt with utilizing imputation strategies or by dropping incomplete records. Categorical factors were encoded utilizing one-hot encoding or label encoding. Numerical features were standardized to guarantee uniform scaling. The datasets were part into preparing (80%) and testing (20%) sets to assess demonstrate execution. Confusion metrices were utilized of both dataset.

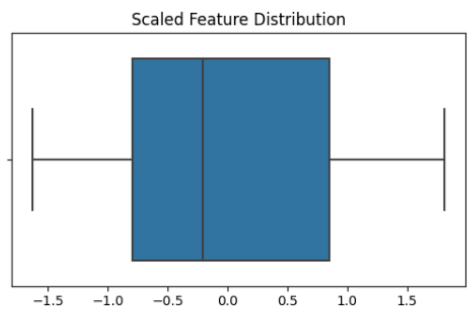

Figure 1: Feature Scaling Effectiveness (Dataset1):

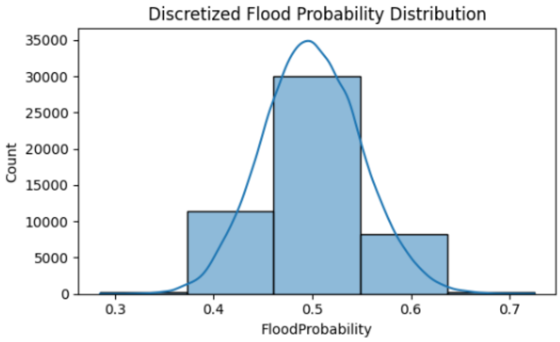

Figure 2: Discretization of Target Variable (Dataset2)

2- Machine Learning Models (Training phase):

(i)-Logistic Regression Model:

A Logistic Regression model was utilized for binary classification of flood chance. It permitted for understanding the relationship between predictor factors and the probability of a flood occasion occurring. The Logistic Regression model from sklearn was designed with parameters like max_iter=200 to adjust training term and model convergence. The target variable was discretized into bins to encourage logistic regression. The model was trained utilizing the preprocessed training data. Performance metrics including precision, accuracy, review, F1 score, and AUC-ROC were utilized to assess the model. A confusion matrix provided a point by point breakdown of the models performance. A classification report was produced to summarize precision, recall, and F1 score for each course during the model learning process.

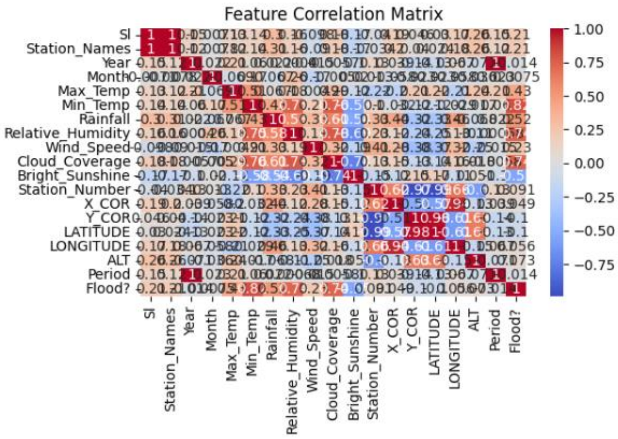

Figure 3: correlation matrix (dataset 1)

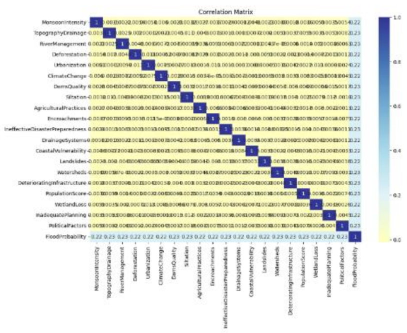

Figure 4: Correlation matrix (dataset - 2)

(ii)-XGBoost Model:

An XGBoost show was utilized for relapse assignments to anticipate ceaseless surge probabilities. This demonstrate is known for its effectiveness and tall expectation accuracy. The XGBRegressor from xgboost was arranged with n_estimators=100, learning_rate=0.1, and max_depth=6 to adjust demonstrate complexity and computational cost. Features were standardized to preserve consistency. The model was prepared on the training information to memorize complex connections between indicator factors and flood probabilities. Performance was evaluated utilizing Mean Squared Error (MSE) and R^2 Score, which gage the models expectation exactness and the variance explained by the model.

Factor	Score
MSE	0.00121
R^2	0.93209

Factor	Score
MSE	0.0002
R^2	0.9049

Figure 5: performance of dataset-1 and dataset-2 respectively

(iii)- CatBoost Model :

A CatBoost demonstrate was utilized for foreseeing persistent surge probabilities, A CatBoost model was utilized for foreseeing continuous flood probabilities, leveraging its strength and high efficiency. The CatBoostRegressor from the catboost library was arranged with iterations=100, learning_rate=0.1, depth=6, and loss_function='RMSE'. Standardization of features was performed to guarantee consistency over models. The model was prepared on the training information to capture connections between indicator factors and flood probabilities. Metrics such as Mean Squared Error (MSE) and R^2 Score were utilized to assess the models performance. During the training phase, Predicted values were compared with genuine values to survey the precision of the models expectations.

Factor	Score
MSE	0.0008906
R^2	0.9502992

Factor	Score
MSE	0.00017495
R^2	0.929756

Figure 6: performance of dataset-1 and dataset-2 respectively

3- Web Application Development :

To supply an interactive stage where clients can select a area of intrigued, and the trained models foresee the flood likelihood for that location. Clients can select a area by clicking on a map. The chosen locations coordinates are utilized to bring information pertinent to that range from the datasets. Based on the chosen area, the web application permits clients to choose a prepared machine learning model (logistic regression, XGBoost, or CatBoost) to get flood predictions. The application uses the chosen model to handle the location-specific information and produce flood likelihood outputs. The interface was designed to be user-friendly, directing the client through area choice, show choice, and showing forecasts clearly. The web application gets data related to the chosen area from different APIs, preprocesses it if essential, and applies the chosen model to create predictions. The result comes about are at that point displayed on the interface, giving a clear visualization of flood hazard for the chosen area.

4- Comparison Across Models :

The objective was to compare the performance of Logistic Regression, XGBoost, and CatBoost to decide the most viable model for flood prediction. Models were compared using metrics like accuracy, precision, recall, F1 score, MSE, and R^2 Score to provide a comprehensive understanding of their performance.

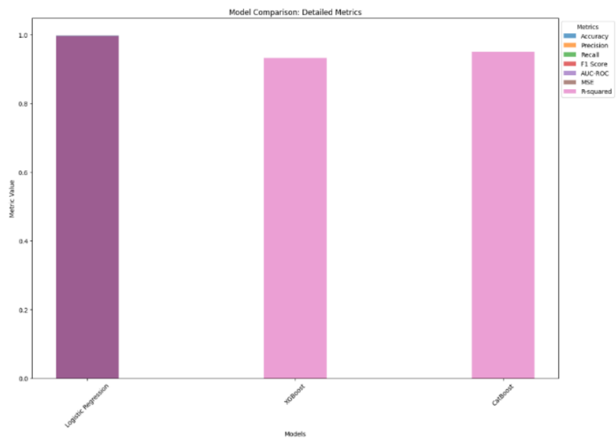

Figure 7: Models performance comparison (dataset 1)

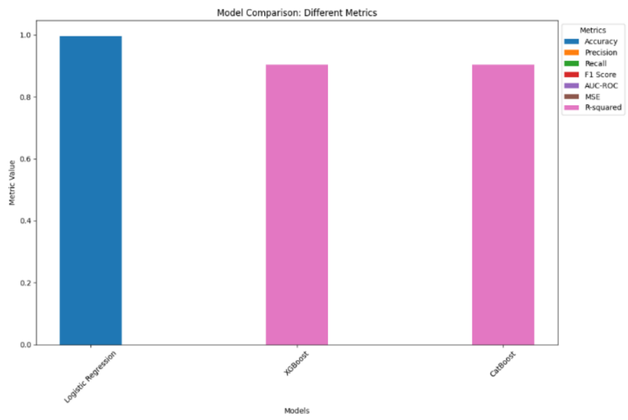

Figure 8 : Models performance comparison (dataset 2)

IV. ALGORITHM

Here is a proposed algorithm for getting flood probabilities for a chosen location using web application and machine learning :

Step 1- Collect the relevant data for the flood prediction model using hydrological modelling , understanding of DEM records, collecting it from various different sources.

Step 2- Analyze the data and apply data preprocessing if required.

Step 3- Train your machine learning model using the dataset.

Step 4- Fine tune the model performance by using different methods like feature engineering, hyperparameters, Normalization, Scaling, categorical encoding (if required).

Step 5- After training the model as per your need, save the model.

Step 6- Make a web application where user can chose a location on map.

Step 7- For the chosen location, the geodata will be fetched from different authentic APIs. This data will be helpful in predicting the flood probabilities.

Step-8- User can also chose a model they want to use (logistic regression, XGBoost, CatBoost)

Step-9- This Fetched data will be sent to previously trained machine learning model as a input.

Step-10- After this we will be able to get the predicted flood probabilities which will be shown to the user at the developed web application.

This algorithm could be further refined and optimized through machine learning techniques to improve its accuracy and efficiency.

V. CONCLUSION

In conclusion, This research presents a comprehensive system for flood prediction, joining machine learning algorithms with an interactive web application to improve usability and encourage informed decision-making. By utilizing logistic regression, XGBoost, and CatBoost models, the study illustrates the potential of machine learning to accurately and flexibly predict flood probabilities. The incorporation of a web-based interface empowers clients to choose particular geological areas and seamlessly access insights derived from the models, in this manner advancing timely decision-making and effective hazard management.

The proposed framework not only approves the viability of advanced machine learning methods within the domain of natural hazard appraisal but too underscores the significance of user-oriented plan in transforming complex algorithms into commonsense, noteworthy instruments. This technique viably bridges the partition between technical development and its application in real-world scenarios, giving fundamental support for calamity readiness, urban arranging, and climate adjustment strategies.

Future research endeavors may build upon this foundational work by joining real-time information streams, consolidating geospatial analytics, and growing the system's functionalities to figure other normal catastrophes. In advancing the integration of data science and calamity management, this study highlights the significant part of innovation in moderating natural dangers and cultivating versatile communities.

REFERENCES

[1] A. Mosavi, P. Ozturk, and K. Chau, "Flood Prediction Using Machine Learning Models: Literature Review," Water, vol. 10, no. 11, p. 1536, Oct. 2018, doi: https://doi.org/10.3390/w10111536.

[2] Z. K. Lawal, H. Yassin, and R. Y. Zakari, "Flood Prediction Using Machine Learning Models: A Case Study of Kebbi State Nigeria," *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pp. 1–6, Dec. 2021, doi: https://doi.org/10.1109/csde53843.2021.9718497.

[3] D. L. Nguyen, T. Y. Chou, T. V. Hoang, and M. H. Chen, "Flood Susceptibility Mapping Using Machine Learning Algorithms: A Case Study in Huong Khe District, Ha Tinh Province, Vietnam," *International Journal of Geoinformatics*, vol. 19, no. 7, Jul. 2023, doi: https://doi.org/10.52939/ijg.v19i7.2739.

[4] Sliman Hitouri *et al.*, "Flood Susceptibility Mapping Using SAR Data and Machine Learning Algorithms in a Small Watershed in Northwestern Morocco," *Remote Sensing*, vol. 16, no. 5, pp. 858–858, Feb. 2024, doi: https://doi.org/10.3390/rs16050858.

[5] Sackdavong Mangkhaseum, Y. Bhattarai, Sunil Duwal, and Akitoshi Hanazawa, "Flood susceptibility mapping leveraging open-source remote-sensing data and machine learning approaches in Nam Ngum River Basin (NNRB), Lao PDR," *Geomatics Natural Hazards and Risk*, vol. 15, no. 1, May 2024, doi: https://doi.org/10.1080/19475705.2024.2357650.

[6] A. Rajab *et al.*, "Flood Forecasting by Using Machine Learning: A Study Leveraging Historic Climatic Records of Bangladesh," *Water*, vol. 15, no. 22, p. 3970, Jan. 2023, doi: https://doi.org/10.3390/w15223970.

[7] R. Khalid and U. T. Khan, "Flood susceptibility mapping using ANNs: a case study in model generalization and accuracy from Ontario, Canada," *Geocarto international*, vol. 39, no. 1, Jan. 2024, doi: https://doi.org/10.1080/10106049.2024.2316653.

[8] M. Ahmadlou *et al.*, "Flood susceptibility mapping and assessment using a novel deep learning model combining multilayer perceptron and autoencoder neural networks," *Journal of Flood Risk Management*, vol. 14, no. 1, Dec. 2020, doi: https://doi.org/10.1111/jfr3.12683 .

[9] M. Hasanuzzaman, A. Islam, B. Bera, and P. K. Shit, "A comparison of performance measures of three machine learning algorithms for flood susceptibility mapping of river Silabati (tropical river, India)," *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 127, p. 103198, Oct. 2022, doi: https://doi.org/10.1016/j.pce.2022.103198.

[10] Gauhar, Noushin and Das, Sunanda and Moury, and Khadiza Sarwar, "Prediction of Flood in Bangladesh using k-Nearest Neighbors Algorithm," *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, pp. 357–361, 2021. [12]

[11] N. Khalid, "Flood Prediction Dataset," *Kaggle.com*, 2024. https://www.kaggle.com/datasets/naiyakhalid/flood-prediction-dataset [1]

[12] T. Chen and C. Guestrin, "XGBoost: a Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pp. 785–794, 2016, doi: https://doi.org/10.1145/2939672.2939785.

[13] Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin, "CatBoost: unbiased boosting with

categorical features," *arXiv (Cornell University)*, Jun. 2017, doi: https://doi.org/10.48550/arxiv.1706.09516.

[14] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting," *The Journal of Educational Research*, vol. 96, no. 1, pp. 3–14, Sep. 2002, doi: https://doi.org/10.1080/00220670209598786.

[15] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, no. 21, 2013, doi: https://doi.org/10.3389/fnbot.2013.00021.

[16] J. Terven, D. M. Cordova-Esparza, A. Ramirez-Pedraza, and E. A. Chavez-Urbiola, "Loss Functions and Metrics in Deep Learning. A Review," *arXiv.org*, Jul. 05, 2023. https://arxiv.org/abs/2307.02694

[17] D. B. Figueiredo Filho, J. A. Silva Júnior, and E. C. Rocha, "What is R2 all about?," *Leviathan (São Paulo)*, vol. 3, no. 3, pp. 60–68, Nov. 2011, doi: https://doi.org/10.11606/issn.2237-4485.lev.2011.132282.

[18] Scikit-learn, "scikit-learn: Machine Learning in Python," *Scikit-learn.org*. https://scikit-learn.org/stable/

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

[20] Agri Sustainability, Food Research, S. Sahu, and Shobhana Ramteke, "Floods disaster in India, mitigation and their impacts," *Sustainability Agri Food and Environmental Research*, vol. Vol. 12 (2024), no. Special issue: Climate change, Apr. 2023, Accessed: Dec. 18, 2024. [Online]. Available: https://www.researchgate.net/publication/371069486_Floods_disaster_in_India_mitigation_and_their_impacts

[21] "(PDF) Confusion Matrix-based Feature Selection.," *ResearchGate*. https://www.researchgate.net/publication/220833270_Confusion_Matrix-based_Feature_Selection

[22] V. Sharma, "A Study on Data Scaling Methods for Machine Learning," *International Journal for Global Academic & Scientific Research*, vol. 1, no. 1, Feb. 2022, doi: https://doi.org/10.55938/ijgasr.v1i1.4.

[23] S. Simon, N. Kolyada, C. Akiki, M. Potthast, B. Stein, and N. Siegmund, "Exploring Hyperparameter Usage and Tuning in Machine Learning Research," *IEEE Xplore*, May 01, 2023. https://ieeexplore.ieee.org/document/10164726

[24] E. B. and J. Wilber, "Logistic Regression," *MLU-Explain*. https://mlu-explain.github.io/logistic-regression/

[25] J. Wilber, "ROC and AUC," *MLU-Explain*. https://mlu-explain.github.io/roc-auc/

[26] J. Wilber, "Train, Test, and Validation Sets," *MLU-Explain*. https://mlu-explain.github.io/train-test-validation/