"Enhanced Flood Prediction and Risk Assessment Using Machine Learning Models”

Nishant Asmit Parida Binaya Nanda

[nishantasmit@gmail.com](mailto:nishantasmit@gmail.com) nbinaya56@gmail.com

Odisha University of Technology and Research, Bhubaneswar, Odisha, India

**Abstract**

The increasing frequency and severity of floods driven by climate change, unregulated urbanisation, and deforestation have made floods one of the most destructive natural disasters globally. These disasters affect millions annually, causing widespread infrastructure damage, displacement, and public health crises, especially in developing countries like India. Traditional flood prediction models rely on hydrological and hydraulic simulations that use physical parameters such as rainfall-runoff, river discharge, and watershed characteristics. While scientifically grounded, these models often fall short due to limitations like the need for extensive calibration, sensitivity to data quality, and poor adaptability to dynamic urban landscapes. Previous studies have also struggled to account for multi-dimensional environmental variability and data scarcity, highlighting the need for alternative approaches. This paper leverages machine learning techniques to address these challenges. Using a comprehensive dataset of 50,000 records with 21 environmental and socio-economic features-such as monsoon intensity, urbanisation, and river management- we evaluate the performance of four regression algorithms: K-Nearest Neighbors (KNN), Gradient Boosting, XGBoost, and LightGBM. Models were optimized using RandomizedSearchCV, and visualizations like scatter plots and heatmaps aided in interpretation. The results show that ensemble models, particularly XGBoost, achieved superior performance in predicting flood probabilities. This suggests that machine learning offers a practical and scalable solution for flood early warning systems, capable of delivering real-time, high-accuracy predictions to support disaster mitigation efforts.

**Keywords**: flood prediction, machine learning, XGBoost, ensemble models, climate change, early warning systems.

**1. Introduction**

Floods are among the most destructive natural disasters, typically resulting from extreme precipitation, excessive surface runoff, or overflow in river channels. These events can lead to widespread devastation by damaging infrastructure, disrupting essential services, and posing serious risks to human life and safety.[[1]](https://hess.copernicus.org/articles/26/4345/2022/) Although flooding has been a natural part of Earth’s ecological processes for millions of years processes that human societies have historically adapted to recent decades have seen a marked increase in both the frequency and intensity of floods.[[2]](https://www.wri.org/insights/worlds-15-countries-most-people-exposed-river-floods) These factors increase surface runoff and overwhelm drainage systems, making areas more prone to flooding. As the scale and severity of floods continue to grow, there is an urgent need for efficient and proactive flood risk management strategies. In countries like India, floods affect an estimated 4.84 million people annually, causing major economic, infrastructural, and health-related losses. The rising frequency of extreme weather events highlights the importance of adopting predictive, data-driven approaches to enhance flood preparedness.[[3,](https://pubmed.ncbi.nlm.nih.gov/20534054/)[4]](https://education.nationalgeographic.org/resource/flood/)

Traditionally, flood prediction has relied on hydrological and hydraulic models, which simulate water system behaviour based on physical parameters such as rainfall-runoff relationships, river discharge, and catchment characteristics. While these models are grounded in scientific principles and have been widely used, they come with several limitations. They often require extensive calibration, are sensitive to data quality, and struggle to adapt to the rapidly changing urban environment.[[5]](https://www.sciencedirect.com/science/article/pii/S2212420923002315) In addition, they generally operate under fixed assumptions and are less capable of incorporating real-time, heterogeneous data from multiple sources. As a result, they may fail to capture the highly dynamic and nonlinear interactions between environmental and human-induced factors that influence flooding.[[6]](https://www.theguardian.com/environment/2020/apr/23/flooding-double-number-people-worldwide-2030#:~:text=The%20number%20of%20people%20harmed,people%20just%2010%20years%20ago.) In cities where infrastructure growth outpaces drainage planning, inadequate waste management systems contribute to frequent urban pluvial floods. These floods not only cause immediate disruptions but also create long-term health and environmental concerns due to contaminated water supplies and the increased risk of water-borne diseases [[7]](https://issi.org.pk/issue-brief-on-urban-flooding-in-pakistan/#:~:text=Urban%20flooding%20is%20an%20issue,urban%20flooding%20in%20various%20cities.). Moreover, conventional forecasting systems are challenged by incomplete historical data and regional environmental variability, limiting their effectiveness in real-world scenarios. These systems often lack the adaptability needed to respond to rapid urban development and changing climate patterns. Recognizing this, India has implemented various flood mitigation infrastructures and is increasingly incorporating advanced technologies such as satellite imaging, real-time meteorological inputs, and water-level sensors into its early warning systems.[[8,9](https://www.mdpi.com/2071-1050/12/19/7865)[]](https://ui.adsabs.harvard.edu/abs/2020JHyd..59125216P/abstract)

Traditional flood prediction methods often struggle with data complexity, non-linearity, and limited adaptability to dynamic environmental conditions. These models typically rely on static thresholds and limited variables, making them less effective in accurately forecasting floods in diverse and evolving scenarios.[[10]](https://www.ijraset.com/research-paper/flood-prediction-using-machine-learning) To address these challenges, this study employs a machine learning-based approach utilizing four regression models: K-Nearest Neighbors (KNN), Gradient Boosting, XGBoost, and LightGBM. These models were trained on key environmental and socio-economic indicators such as Monsoon Intensity, Urbanization, Deforestation, and River Management, with hyperparameter tuning conducted using RandomizedSearchCV. The contribution of this work lies in demonstrating the effectiveness of ensemble and boosting techniques in capturing complex data interactions and improving predictive performance for real-time flood forecasting in data-scarce and high-risk regions.[[11]](https://www.mdpi.com/2073-4441/10/11/1536)

This paper focuses on developing a flood prediction model by leveraging advanced machine learning techniques, including KNN, Gradient Boosting, XGBoost, and LGBM, to analyze environmental and socio-economic factors influencing flood probability. The model's performance will then be evaluated using R² score, MSE, RMSE, and MAE to ensure accurate predictions for effective disaster management.

**2. Related Works**

Flood prediction has become a very vital purpose to tackle the environmental setbacks and damage to the public and private property, hence its necessary emphasize major developments in forecasting techniques based on machine learning. It discusses the main breakthroughs in data-driven flood modeling while also spelling out the constraints and future opportunities toward improving prediction accuracy, real-time response, and human-computer interaction for disaster preparedness and risk management.

M. M. A. Syeed [[12]](https://arxiv.org/abs/2208.01234) have used logistic regression, KNN, SVM, and Decision tree (DT) under different conditions for flood prediction. The ranges of monthly rainfall data between 1980 and 2020 were employed in the paper where logistic regression yielded a maximum accuracy rate of 85% with precision and recall scores of 75% and 55%, respectively. Q. Ke et al. [[13]](https://www.sciencedirect.com/science/article/pii/S0309170819311601) used conventional and ML techniques on flood prediction according to the rainfall intensity. These have several ML algorithms in their paper; the functionalities accuracy of their model was between 94% and 96%, & the performance of ML models in comparison to conventional models was quite better.

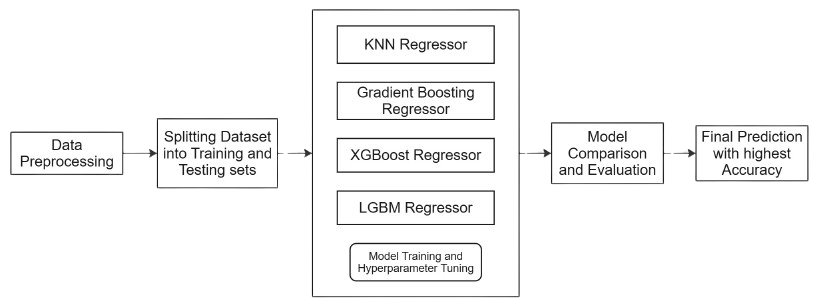
Balamurugan and Manojkumar [[14]](https://www.researchgate.net/publication/336831403_Study_of_short_term_rain_forecasting_using_machine_learning_based_approach) done a comparative study to analyse the performance of statistical methods and ML techniques in rainfall prediction and found ML models were better at predicting the rainfall that others. Kadiyala and Woo [[15]](https://www.researchgate.net/publication/357824114_Flood_Prediction_and_Analysis_on_the_Relevance_of_Features_using_Explainable_Artificial_Intelligence) investigated the mechanisms through which background determinants affecting prediction choices are successfully learned, and it further expands on this avenue by incorporating explainable artificial intelligence (XAI) modules to understand decision making process.

To predict floods, K. Vamshi et al., [[16]](https://www.irjet.net/volume8-issue03) implemented machine learning and neural network-based systems. Based on past rainfall data, the best approach among the two is chosen for prediction. Utilizing rainfall data, a prediction model was built by A. O. Hashi et al., [[17]](https://www.researchgate.net/publication/352242859_A_Real-Time_Flood_Detection_System_Based_on_Machine_Learning_Algorithms_with_Emphasis_on_Deep_Learning) the model forecasts if a flood would occur. The prediction model is for Indian districts based on rainfall data. Linear Regression, KNN, SVM, MLP are used to train the dataset. A. Vinothini et al., [[18]](https://www.researchgate.net/publication/363766106_Prediction_of_Flash_Flood_using_Rainfall_by_MLP_Classifier) utilizes algorithms such as Random Forest, Naive Bayes, J48, and Convolutional Neural Networks to detect water levels and assess floods with potential humanitarian effects before they occur.

F. Chang et al. [[19]](https://www.mdpi.com/books/reprint/1151-flood-forecasting-using-machine-learning-methods) adopted an intelligent hydro-informatics integration platform (IHIP). This platform enhances online forecast capabilities and flood risk management through a web interface system. S. Dazzi et al., [[20]](https://www.mdpi.com/2073-4441/13/12/1612) evaluated how accurately machine learning models predict flood stages at key gauge stations based on upstream water level observations, while incorporating downstream influences caused by backwater effects.

Previous studies employed very basic models like linear regression or decision trees, which had difficulty in capturing complex patterns in flood-related data. Many were not rigorous in applying proper feature selection and tuning, which resulted in poor performance. This paper employed advanced models such as KNN, Gradient Boosting, XGBoost, and LightGBM that were able to handle high-dimensional data and enhance performance through ensemble learning and hyperparameter tuning.

3. Model Framework



**Fig 1: Model Framework for flood detection**

Fig. 1 illustrates the proposed model framework for flood prediction. The starting stage is a data pipeline that imports preprocessed input data. This data is then split into test and training datasets to ensure rigorous evaluation of the models. Four ML models, which are KNN Regressor, Gradient Boosting Regressor, XGBoost Regressor, and LightGBM Regressor, are all trained on the dataset. Models are evaluated and fine-tuned for their predictive performance through hyperparameter optimization techniques. Finally, the top-performing model in accuracy is selected for the final prediction, to ensure optimal reliability and precision in predicting flood probabilities.

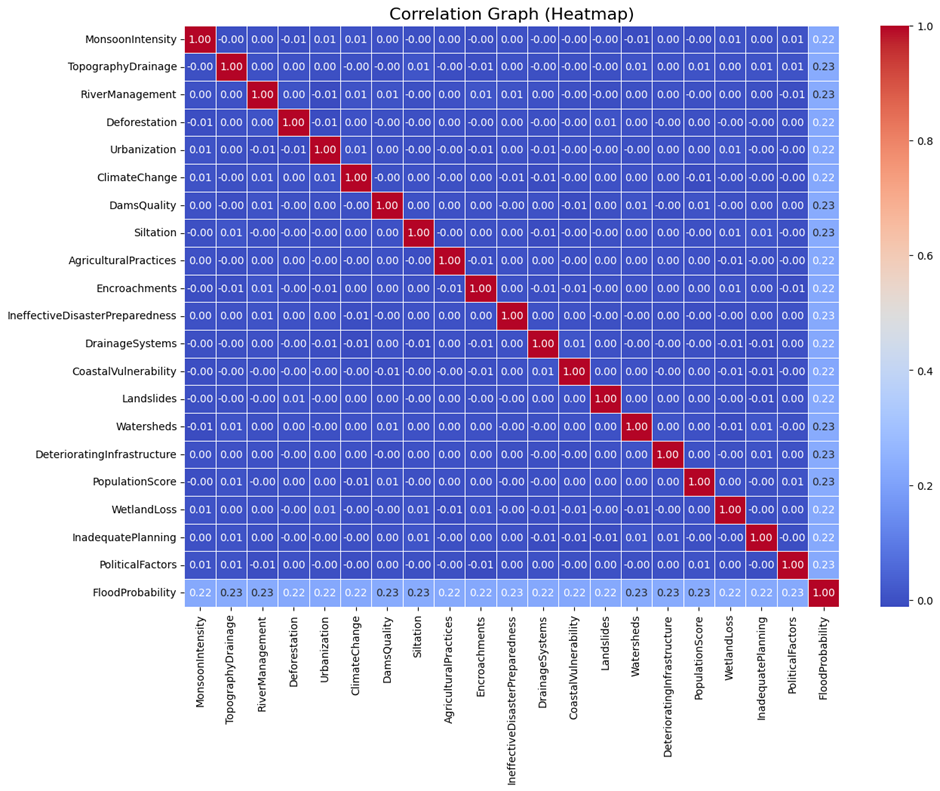
4. Methodology

**4.1. Dataset Description**

An open-source Flood Prediction Dataset from Kaggle (<https://www.kaggle.com/datasets/naiyakhalid/flood-prediction-dataset>) includes environmental and socio-economic factors relevant to flood occurrences. It has been hosted on Kaggle by Naiya Khalid. The dataset includes attributes such as Monsoon Intensity, Topography Drainage, River Management, Deforestation, Urbanization, Climate Change, Dams Quality, Siltation, Agricultural Practices, Encroachments, Ineffective Disaster Preparedness, Drainage Systems, Coastal Vulnerability, Landslides, Watersheds, Deteriorating Infrastructure, Population Score, Wetland Loss, Inadequate Planning, Political Factors, and Flood Probability. This dataset is suitable for developing regression models to predict flood likelihood based on these factors.

**4.2. Data Preprocessing**

After loading the dataset, an initial exploratory analysis was conducted to ensure the data's readiness for further processing and model training. The dataset was first checked for missing values; since none were found, imputation was not required. However, if missing values had been present, standard techniques such as mean or median replacement or row removal would have been considered. To gain deeper insight into the data, various exploratory techniques were applied—these included reviewing column types, checking null counts, examining memory usage, and analyzing summary statistics to detect outliers and understand the overall distribution of values. A sample of data entries was also reviewed to identify any inconsistencies or anomalies. Finally, a correlation heatmap was generated to visualize the relationships between features and identify potential multicollinearity, aiding in feature selection and enhancing model efficiency.



**Fig 2: Correlation Heatmap Plotting**

Fig 2 demonstratescorrelation heatmap using seaborn to examine relationships between different features. High correlation (close to 1 or -1) between independent variables indicates multicollinearity, which can cause redundant information and negatively impact model performance. Features with strong correlations were further analyzed, and if necessary, feature selection techniques like Variance Inflation Factor (VIF) were considered to remove or transform redundant variables, ensuring better model stability and accuracy.

**Feature scaling** was applied to ensure that all numerical features had a similar range, preventing models from being biased towards variables with larger magnitudes. Since different machine learning algorithms are sensitive to feature scales, standardization was performed using Scikit-learn's StandardScaler, which adjusted the data to have a mean of 0 and a standard deviation of 1, making training more efficient and improving model convergence. Scaling was particularly important for models like K-Nearest Neighbors and Gradient Boosting, which rely on distance-based calculations and gradient-based optimization.

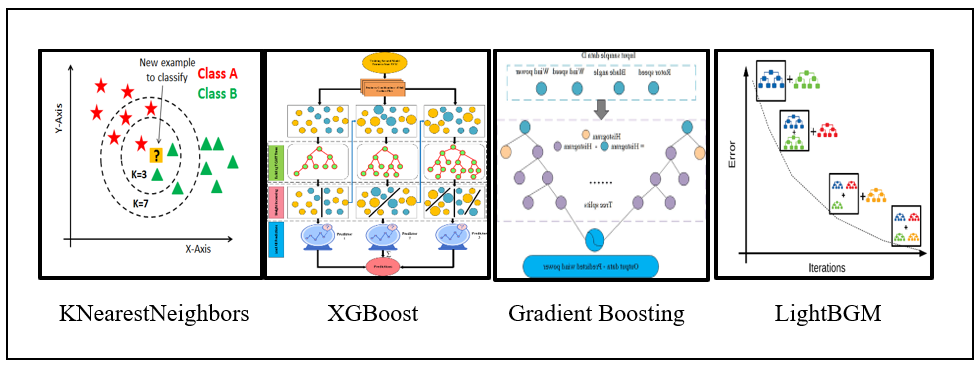
**4.3. Data Splitting**

The dataset consists of 20 independent features (predictors) and one target variable (Flood Probability). For a smooth evaluation, the dataset is split into:

* Training Set (67%) - Used for model training
* Testing Set (33%) - Used for evaluation

The train-test split ensures that the models were trained on a subset of the data and evaluated on unseen portions to minimize overfitting. The random seed (random\_state=42) was set to ensure reproducibility of results. A proper training-testing split is crucial for avoiding overfitting and ensuring the model performs good on real-world predictions.

**4.4. Model Training**



**Fig 3: Models Used**

Fig 3 presents the architecture of the proposed flood prediction model using multiple base regressors—KNN, Gradient Boosting, XGBoost, and LGBM. Each model is independently trained and evaluated to learn different aspects of the dataset. After training, models are compared and the best-performing one is selected based on evaluation metrics such as R-squared, RMSE, and MAE. This model is then optimized through hyperparameter tuning to enhance prediction accuracy. This approach utilizes the complementary strengths of various algorithms to enhance predictive accuracy and ensure more reliable flood occurrence forecasting.

We have Devised a 3-stage process for creation the proposed model:

**Step 1: Machine Learning Model Training**

* **K-Nearest Neighbors (KNN) Regression**: This non-parametric model predicts flood probability by averaging the 'K' most similar historical cases based on Euclidean distance, capturing local patterns in features like rainfall, urbanization, and drainage.

Formula : D(x,y)= ------------(1)

* **Gradient Boosting Regressor (GBR)**: A robust ensemble technique that constructs models in a sequence, where each new model focuses on correcting the errors made by its predecessor, GBR captures complex non-linear relationships in flood prediction using features like topography, monsoon intensity, and deforestation.

Formula: ----------(2)

* **XGBoost Regressor**: An optimized gradient boosting model, XGBoost uses regularization and system enhancements to deliver high accuracy while handling missing data, skewness, and complex feature interactions.

Formula: ----------(3)

* **LightGBM Regressor (LGBM)**: LightGBM is a rapid gradient boosting framework that uses leaf-wise growth and histograms, making it ideal for real-time flood-related environmental and socio-economic data.

Formula: ----------(4)

**Step 2: Hyperparameter Tuning of the Estimators**

Hyperparameter tuning was applied to each of the models to optimize their performance on the dataset. This step ensured that the models were neither underfitting nor overfitting:

* **Random Search Cross Validation**: Randomly sampled combinations of hyperparameters from a defined search space to efficiently find near-optimal model configurations.
* **Grid Search Cross Validation**: Explored all possible combinations of hyperparameter values, albeit at a higher computational cost, to find the best settings for each regressor.
* **Bayesian Optimization**: Used probabilistic modeling to intelligently explore the hyperparameter space and focus on the most promising combinations, balancing exploration and exploitation.

**Step 3: Model Comparison and Evaluation**

Each trained model was evaluated using multiple regression metrics such as R² Score, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The model with the best balance of low error rates and high R² score was selected for final predictions. This multi-model evaluation process allowed for benchmarking and ensured the selection of the most robust and generalizable model for flood prediction.

**5. Result and Discussion**

**5.1 Experimental Setup**

The development of the proposed flood prediction model was carried out using Python, with Google Colab serving as the primary environment for implementation and testing. The libraries NumPy and Pandas were employed for efficient data loading, manipulation, and preprocessing. For visualization and analysis of model performance, Matplotlib and Seaborn were utilized. Machine learning algorithms including K-Nearest Neighbors, Gradient Boosting, XGBoost, and LightGBM were integrated using Scikit-learn and respective libraries to perform regression tasks. RandomizedSearchCV was used for hyperparameter tuning, enabling the selection of optimal configurations for each model. This experimental setup allowed for the A systematic evaluation of model performance enabled the development of precise, data-driven insights for effective flood prediction.

**5.2 R2 Score, MAE, MSE, RMSE**

The evaluation of regression models was conducted using four key performance metrics: R² Score, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Table 1: Final Regression Metrices of all models

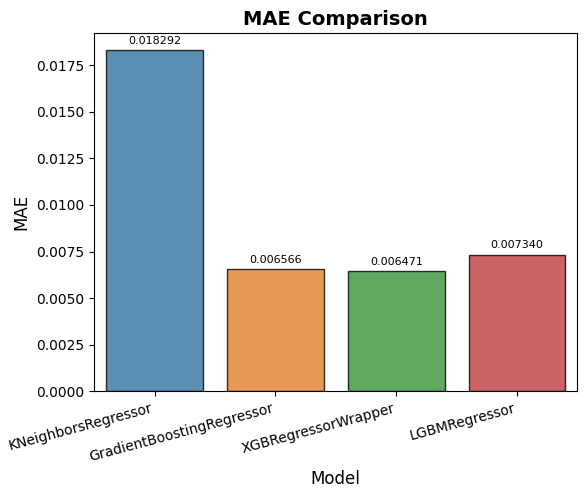
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R² Score** | **MAE** | **MSE** | **RMSE** |
| KNNRegressor | 0.784627 | 0.018292 | 0.000546 | 0.023360 |
| GradientBoostingRegressor | 0.972805 | 0.006566 | 0.000069 | 0.008301 |
| XGBRegressor | 0.973470 | 0.006471 | 0.000067 | 0.008199 |
| LGBMRegressor | 0.965043 | 0.007340 | 0.000089 | 0.009411 |

Table 1 summarizes the performance comparison of four regression models using key evaluation metrics. It highlights the effectiveness of ensemble-based methods over simpler models. The ensemble models, particularly the gradient boosting variants, demonstrate superior predictive accuracy and lower error rates across all metrics. In contrast, the KNN Regressor underperforms relative to the others, indicating its limitations in capturing complex relationships in the data. Overall, the results suggest that advanced ensemble methods are more suitable for precise and reliable stock market forecasting in this context.

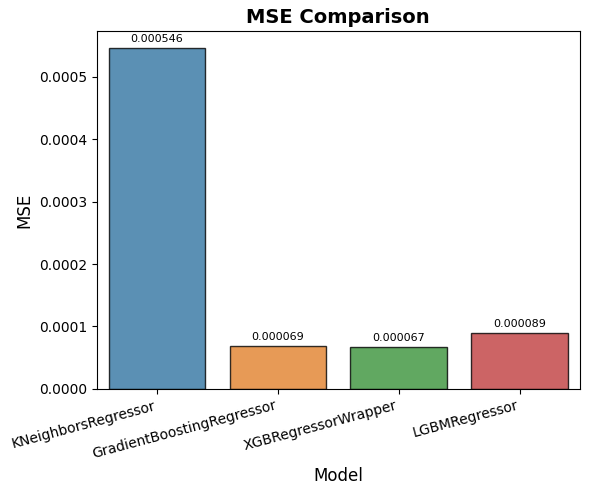
Figure 4 displays the R² Score for all four regression models. Both XGBoost and Gradient Boosting achieve the highest R² values, indicating a strong ability to explain the variance in the target variable and demonstrating high prediction accuracy in estimating flood probabilities. Figure 5 shows the Mean Absolute Error (MAE) comparison. Gradient Boosting records the lowest MAE, which means it consistently makes predictions that are closest to the actual values on average, highlighting its precision in flood risk estimation. Figure 6 presents the Mean Squared Error (MSE) across the models. Lower MSE values for XGBoost and Gradient Boosting suggest that these models produce fewer large errors, making them more reliable in scenarios where larger deviations are critical. Figure 7 illustrates the Root Mean Squared Error (RMSE) comparison. Both XGBoost and Gradient Boosting models achieve lower RMSE values, reaffirming their overall predictive accuracy and stability in forecasting flood occurrence.



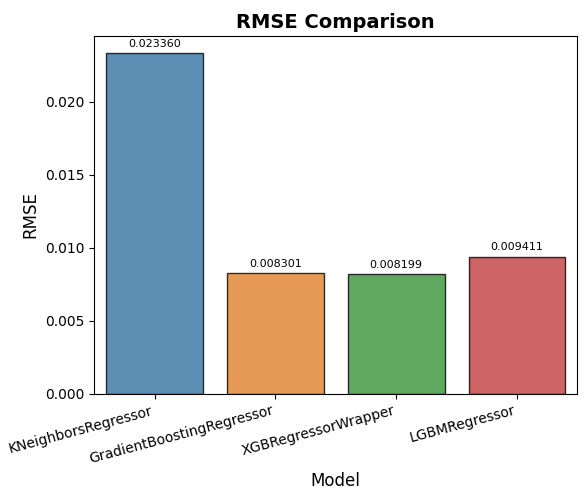
**Figure 4**: R² Score comparison among four regression models.



**Figure 5:** MAE (Mean Absolute Error) across the models.



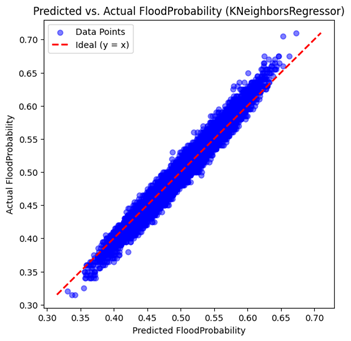
**Figure 6:** MSE (Mean Squared Error) values.



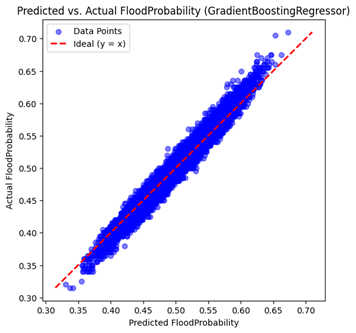
**Figure 7:** RMSE (Root Mean Squared Error) comparison.

**5.3 Predicted vs Actual Flood Probability**

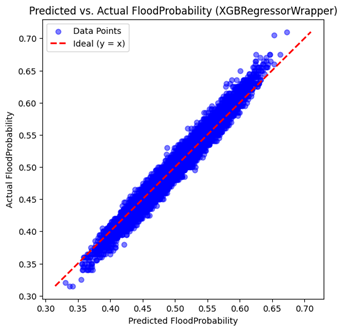
The scatter plot of actual vs. predicted values is an essential visualization technique used to assess the performance of regression models. It offers a visual representation of how closely the model's predictions match the actual values. In an ideal scenario, if the model performs flawlessly, all points will align with the diagonal line (y = x), showing that the predicted values are exactly equal to the actual values. Deviations from the line signify errors in the model's predictions. By analyzing the scatter plot, one can visually detect trends, biases, and the presence of any systematic errors in the model. Clusters of points deviating from the diagonal may indicate underfitting or overfitting, whereas a random spread of points suggests that the model has captured most of the variance in the data.



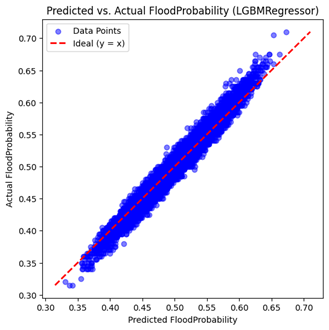
**Figure 8: S**catter plot for KNN Regressor



**Figure 9:** Scatter plot for Gradient Boosting Regressor.



**Figure 10: S**catter plot forXGBoost Regressor.



**Figure 11: S**catter plot for LGBM Regressor.

Figure 8 represents the K-Nearest Neighbors (KNN) Regressor. While the predictions generally align with the red ideal line, a few deviations suggest moderate variance in prediction accuracy, indicating reasonable but less precise performance compared to ensemble models. Figure 9 illustrates the performance of the Gradient Boosting Regressor. Most predicted values closely follow the red diagonal line, showing minimal error and strong agreement with actual values, highlighting the model’s robust predictive capability. Figure 10 shows a tight clustering of points along the ideal line, reflecting high prediction accuracy. The minimal spread indicates that this model performs consistently well in estimating flood probabilities. Figure 11 shows good alignment of predicted and actual values, though with slightly more dispersion than XGBoost and Gradient Boosting. Nonetheless, it still delivers reliable results with acceptable error margins.

**6. Conclusion and Future Work**

This research uncovered most of the promising role of machine-learning regression in forecasting flood probabilities using a variety of environmental and social-economic determinants. All models displayed significant performance levels through the key metrics: score, mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). Gradient Boosting, XGBoost, and LGBM outperformed traditional approaches. RandomizedSearchCV hyperparameter optimization significantly improved model prediction accuracy without loss of generality. The adaptability of the model on Kaggle's flood analysis data set shows that it can be relied upon to forecast disaster-prone areas, helping in early warning systems and disaster preparedness.

The future study would be on incorporating the real-time satellite and IoT sensor data to increase the extent of responsivity. In addition, the new deep learning architectures, namely, Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs), can be attempted to capture temporal and spatial dependencies of flood-related data for more dynamic and accurate flood forecasting solutions.

**References**

[1] Riccardo, M., et al. (2022). Deep learning methods for flood mapping: A review of existing applications and future research directions. *Hydrology and Earth System Sciences, 26*, 4345–4378.

[2] H., et al. (2015). World’s 15 countries with the most people exposed to river floods. *World Resources Institute*.

[3] FitzGerald, G., Du, W., Jamal, A., Clark, M., & Hou, X.-Y. (2010). Flood fatalities in contemporary Australia (1997–2008). *Emergency Medicine Australasia, 22*(2), 180–186.

[4] National Geographic Society. (2011, November 7). Flood. *National Geographic Society*.

[5] Greatrex, H., et al. (2023, July). Causes, impacts, and mitigation strategies of urban pluvial floods in India: A systematic review. *International Journal of Disaster Risk Reduction, 93*, 103751.

[6] Holden, E., et al. (2020, April 23). Flooding will affect double the number of people worldwide by 2030. *The Guardian*.

[7] Mir, H. (2023, August 8). Issue brief on ‘Urban Flooding in Pakistan’. *Institute of Strategic Studies Islamabad*.

[8] Yang, L., et al. (2020). Assessment of community vulnerability to different types of urban floods: A case for Lishui City, China. *Sustainability (Switzerland), 12*(19), 7865.

[9] Pathak, R., et al. (2020). Social, economic and environmental assessment of urban sub-catchment flood risks using a multi-criteria approach: A case study in Mumbai City, India. *Journal of Hydrology, 591*, 125216.

[10] Tarun, G., et al. (2023, May). Flood prediction using machine learning. *International Journal for Research in Applied Science and Engineering Technology, 11*(5), 363–367.

[11] Mosavi, A., et al. (2018, October). Flood prediction using machine learning models: Literature review. *Water, 10*(11), 1536.

[12] Syeed, M. M. A., et al. (2022, June). Flood prediction using machine learning models. *International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*.

[13] Ke, Q., et al. (2020, November). Urban pluvial flooding prediction by machine learning approaches – A case study of Shenzhen city, China. *Advances in Water Resources, 145*, 103719.

[14] Balamurugan, S., & Manojkumar, P. (2019, October). Study of short-term rain forecasting using machine learning based approach. *Wireless Networks*.

[15] Kadiyala, A., & Woo, H. (2022). Flood prediction and analysis on the relevance of features using explainable artificial intelligence. *arXiv*.

[16] Vamshi, K., et al. (2021). A review on rainfall prediction using machine learning and neural network. In *Proceedings*, 2763–2769.

[17] Hashi, A. O., et al. (2021). Real-time flood detection with DL. *IJETTC, 69*(5), 249–256.

[18] Anonymous. (2020). Predicting flood using artificial neural networks. *International Journal of Engineering and Technology, 15*(1), 53–57.

[19] Chang, F.-J., et al. (n.d.). *Flood forecasting using machine learning methods*. *Wireless Networks*.

[20] Dazzi, S., et al. (2021). Flood stage forecasting using machine-learning methods: A case study on the Parma River (Italy).