Enhanced Adaptive Attention-Based Model for Real-Time Urban Flood Detection Using IoT and Multimodal Data

1. Data Exploration and Preprocessing

- The dataset was loaded and initial exploration was performed using df.head(), df.tail(), df.info(), and df.describe().

- Missing values were handled by removing rows with any missing data using df.dropna().

- A correlation matrix was generated using df.corr() and visualized using a heatmap to understand relationships between features.

- The dataset was split into features (X) and target variable (y) - FloodProbability.

- Features were scaled using StandardScaler to ensure they have zero mean and unit variance.

- The data was split into training and testing sets using train_test_split with a test size of 33%.

IMPORTANT FEATURES OF DATASET:

The dataset used for this analysis comprises numerous features, providing essential insights into environmental and human factors that influence flood occurrence and severity. These features include:

1. MonsoonIntensity: Higher volumes of rain during monsoons increase the probability of floods.

2. TopographyDrainage: The drainage capacity is based on the region's topography. Efficient drainage can help drain rainwater and reduce the risk of floods.

3. RiverManagement: The quality and effectiveness of river management practices. Proper river management, including dredging and bank maintenance, can improve water flow and reduce floods.

4. Deforestation: The extent of deforestation in the area.* Deforestation reduces the soil's ability to absorb water, increasing surface runoff and the risk of floods.

5. Urbanization: The level of urbanization in the region. Urban areas have impermeable surfaces (asphalt, concrete), which reduce water infiltration, raising the risk of floods.

6. ClimateChange: The impact of climate change on the region. Climate change can lead to more extreme precipitation patterns, including torrential rains that can cause floods.

7. DamsQuality: The quality and maintenance status of dams. Well-maintained dams can control floods, and dams with structural problems can break and cause catastrophic floods.

8. **Siltation:** The extent of siltation in rivers and reservoirs. The accumulation of sediments in rivers(siltation) reduces drainage capacity and increases the risk of floods.

9. Agricultural Practices: The types and sustainability of agricultural practices. The intensification of agriculture can lead to deforestation, excessive use of fertilizers and pesticides, and inappropriate irrigation practices, reducing soil biodiversity and increasing the risk of floods.

10. Encroachments: The degree of encroachment on flood plains and natural waterways. Construction in flood-prone areas impedes the natural flow of water and increases the risk of floods.

11. **IneffectiveDisasterPreparedness:** ** The lack of emergency plans, warning systems, and simulations increases the negative impact of floods.

12. DrainageSystems: Well-maintained and adequately sized drainage systems help drain rainwater and reduce the risk of floods.

13. CoastalVulnerability: Low-lying coastal areas are prone to flooding from storm surges and sea level rise.

14. Landslides: Steep slopes and unstable soils are more prone to landslides.

15. Watersheds: Regions with more watersheds may have a higher or lower risk of flooding, depending on various factors.

16. DeterioratingInfrastructure: Clogged culverts, damaged drainage channels, and other deficient infrastructure can increase the risk of floods.

17. PopulationScore: Densely populated areas can suffer more severe losses.

18. WetlandLoss: Wetlands act as natural sponges, absorbing excess water and helping to prevent floods.

19. InadequatePlanning: Urban planning that does not consider the risk of flooding increases the vulnerability of communities.

20. PoliticalFactors: Factors such as corruption and a lack of political will to invest in drainage infrastructure can make it difficult to manage flood risk.

21. FloodProbability: The overall probability of flooding in the region. This is the target variable for predictive analysis. This is the target Variable.

Features Correlation Heatmap

Correlation Graph (Heatmap)
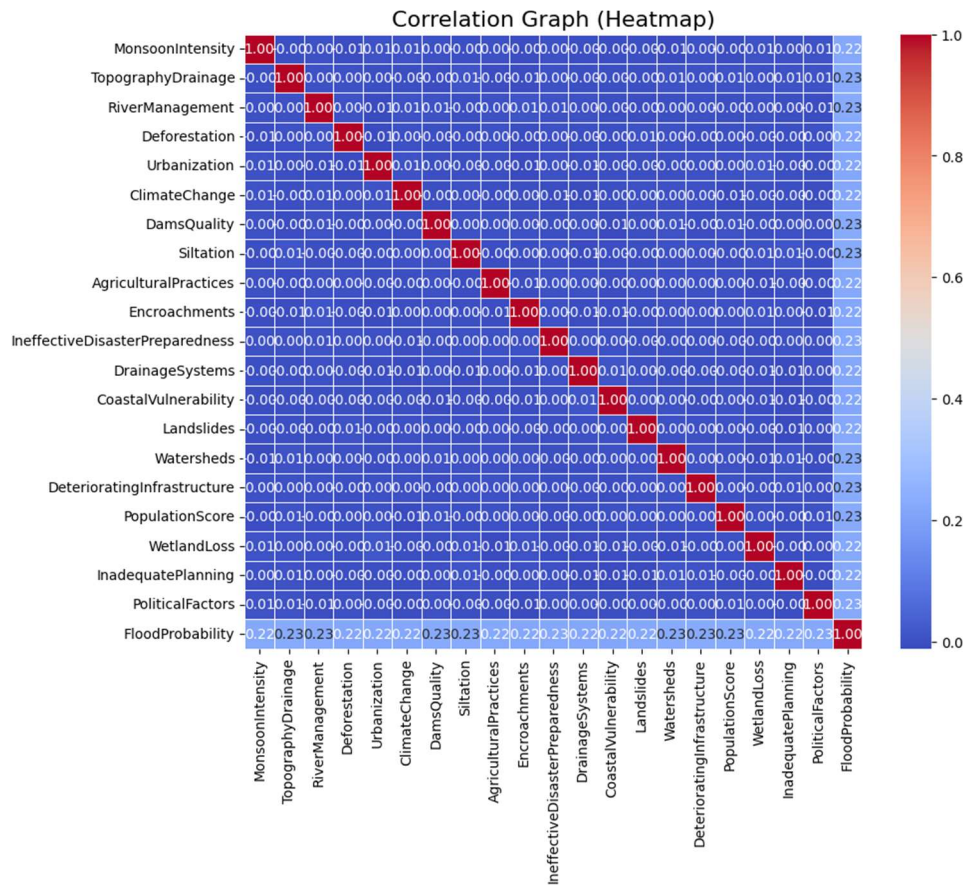
2. Model Selection and Hyperparameter Tuning

- A variety of regression models were considered, including:

    o KNeighborsRegressor

    o RandomForestRegressor

    o DecisionTreeRegressor

    o GradientBoostingRegressor

    o XGBRegressor

    o LGBMRegressor

    o SVR

- RandomizedSearchCV was used to tune the hyperparameters of each model using a predefined parameter grid.

- The best hyperparameters for each model were determined based on the R² score.

3. Model Evaluation

- The performance of each model was evaluated using metrics such as:

- **RMSE:** Average size of prediction errors, penalizing larger errors more.
- **R² score:** How well the model fits the data, 1 is perfect.
- **MAE:** Average size of prediction errors, ignoring direction.
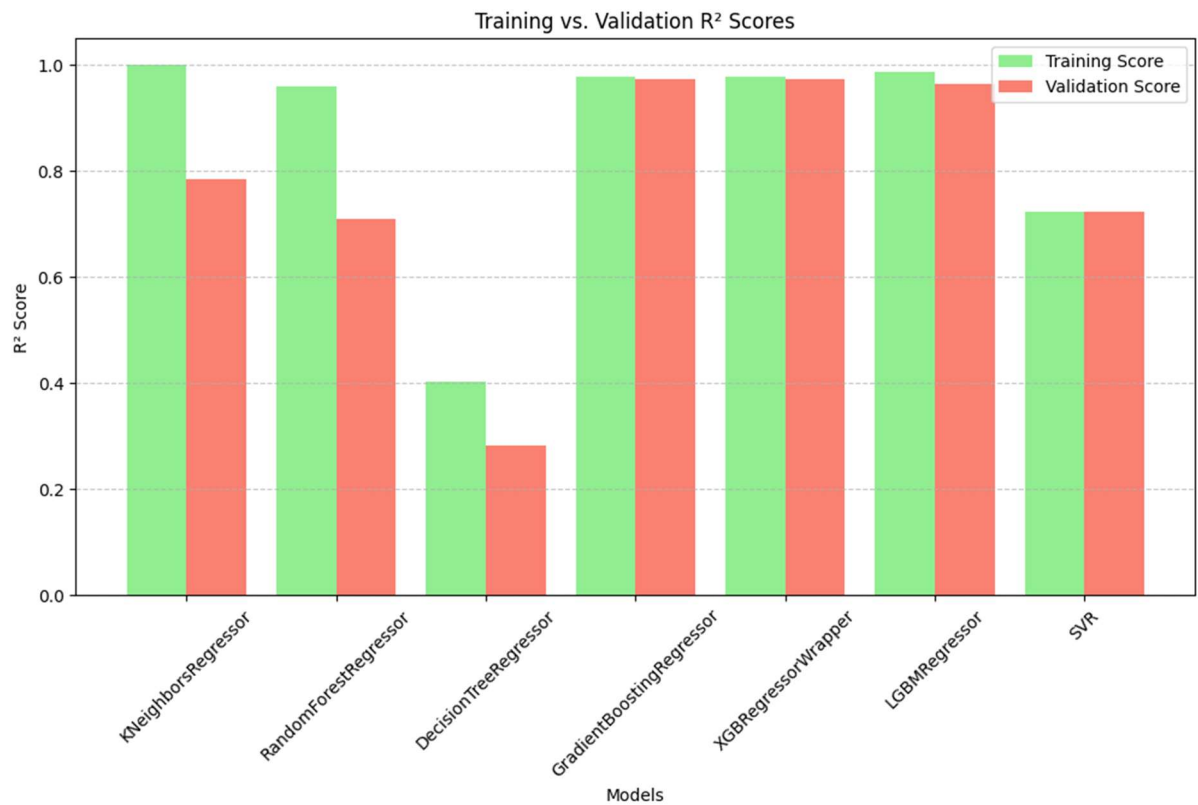- **EVS:** How much of the target's variance the model explains, 1 is perfect.

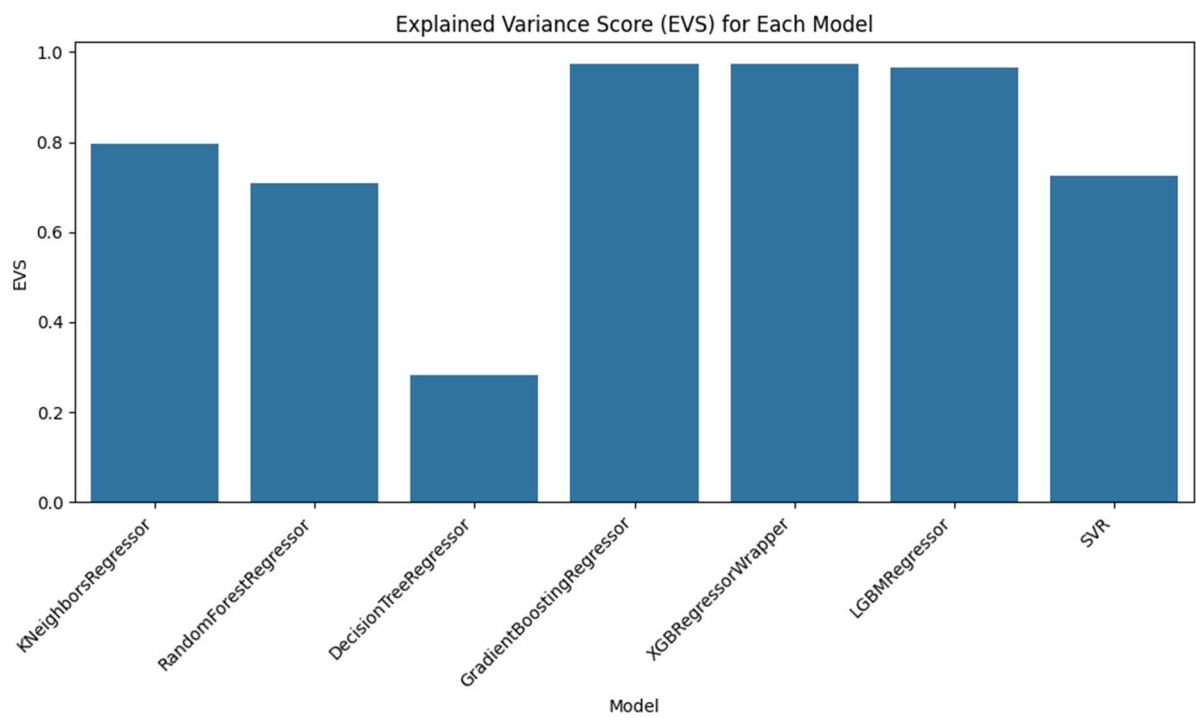## Performance Comparison of Regression Models: RMSE vs. MAE
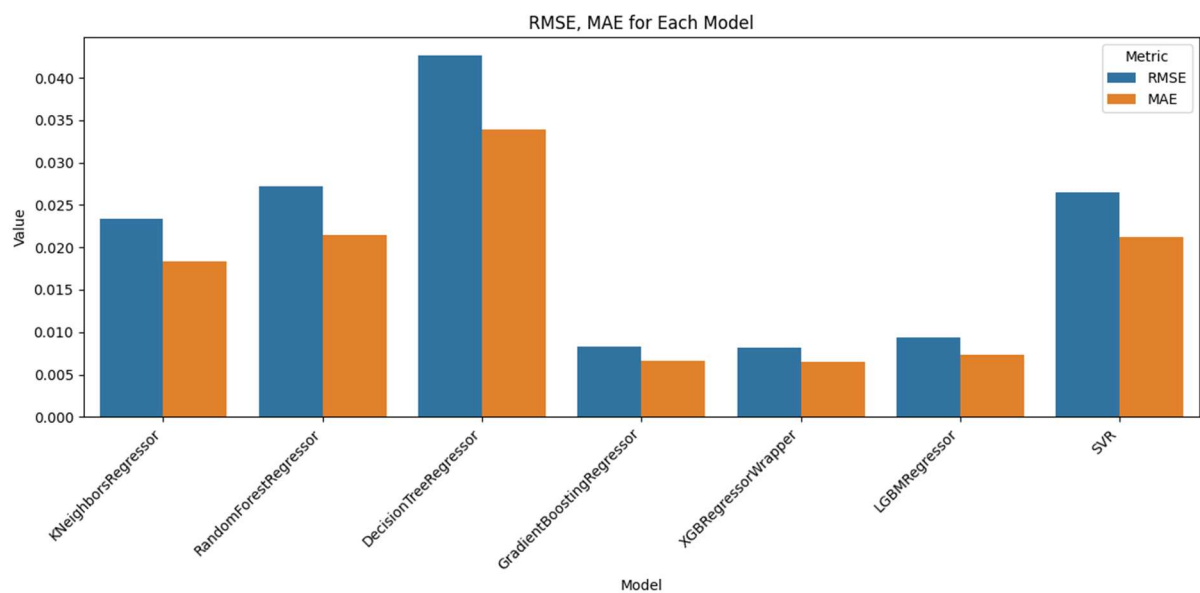
Comparison of R² Scores for Different Regression Models
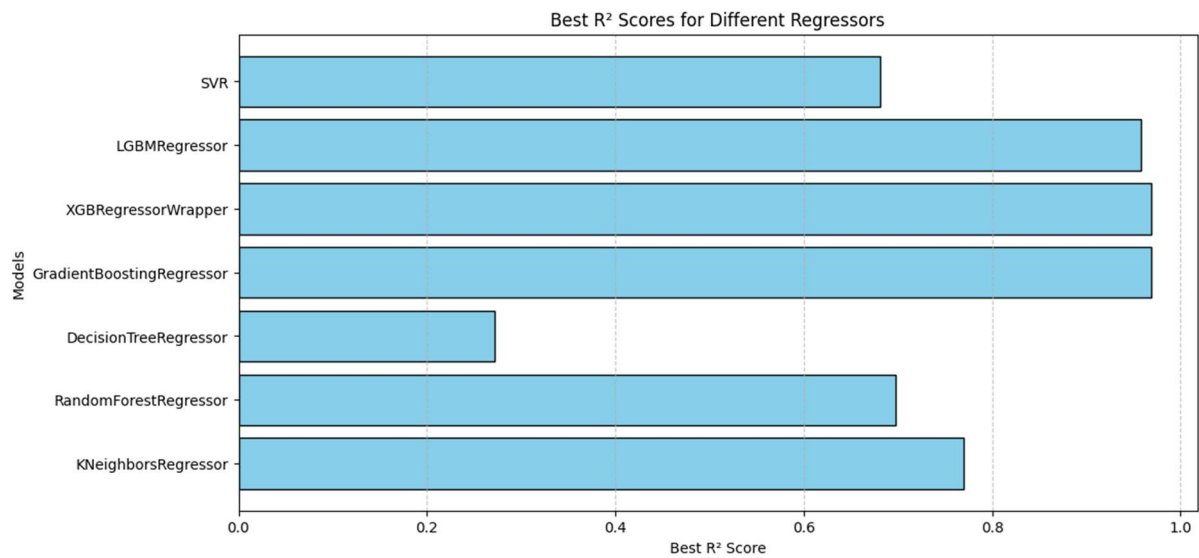
Training vs. Validation R² Scores for Different Regression Models

Explained Variance Score (EVS) Comparison Across Regression Models

Explained Variance Score (EVS) for Each Model

| Model	RMSE	R2	MAE	EVS
KNeighborsRegressor	0.023360	0.784627	0.018292	0.797302
RandomForestRegressor	0.027147	0.709129	0.021435	0.709456
DecisionTreeRegressor	0.042628	0.282786	0.033952	0.282817
GradientBoostingRegressor	0.008301	0.972805	0.006566	0.972807
XGBRegressorWrapper	0.008199	0.973470	0.006471	0.973479
LGBMRegressor	0.009411	0.965043	0.007340	0.965061
SVR	0.026502	0.722799	0.021214	0.724340

## Final Recommendation: XGBoost (XGBRegressorWrapper)

Why XGBoost?

1. **Lowest RMSE (0.008199)** → Minimizes error in predictions

2. **Highest R² (0.973470)** → Explains 97.34% of variance in the data

3. **Lowest MAE (0.006471)** → Less deviation from actual values

4. **Highest EVS (0.973479)** → Explains maximum variation in the target variable

5. **Efficient on structured/tabular data** → Handles missing values, outliers, and feature importance well.

6. **Scalability** → Works efficiently on large datasets due to parallelization.
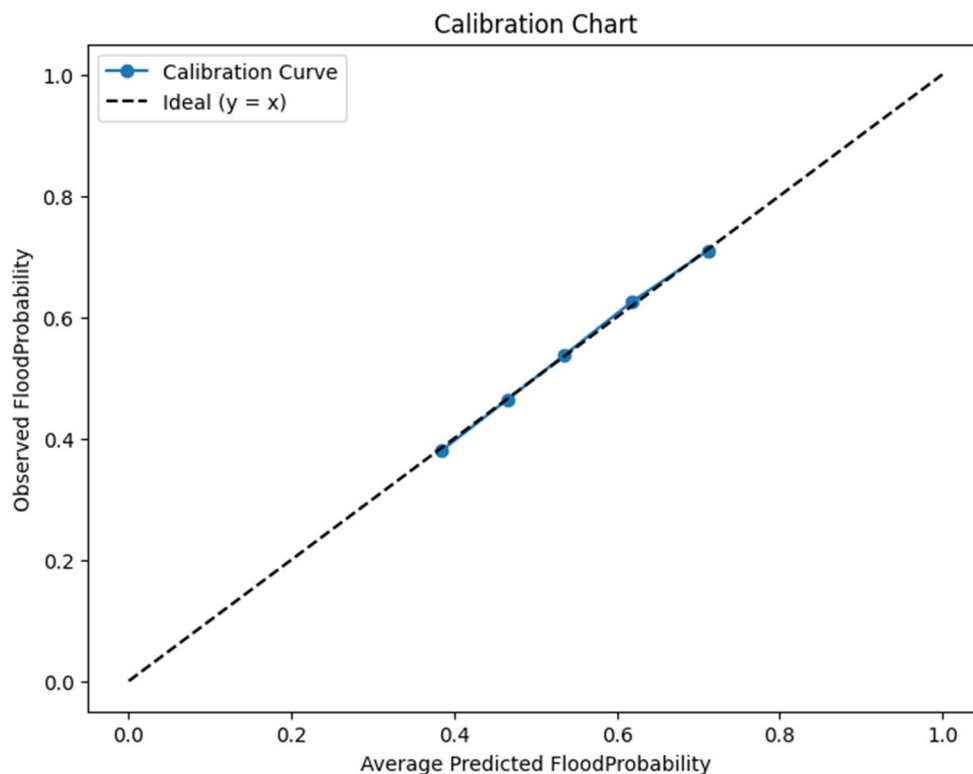
Result Analysis

Flood Probability vs. Monsoon Intensity: Model Predictions vs. Actual Observations
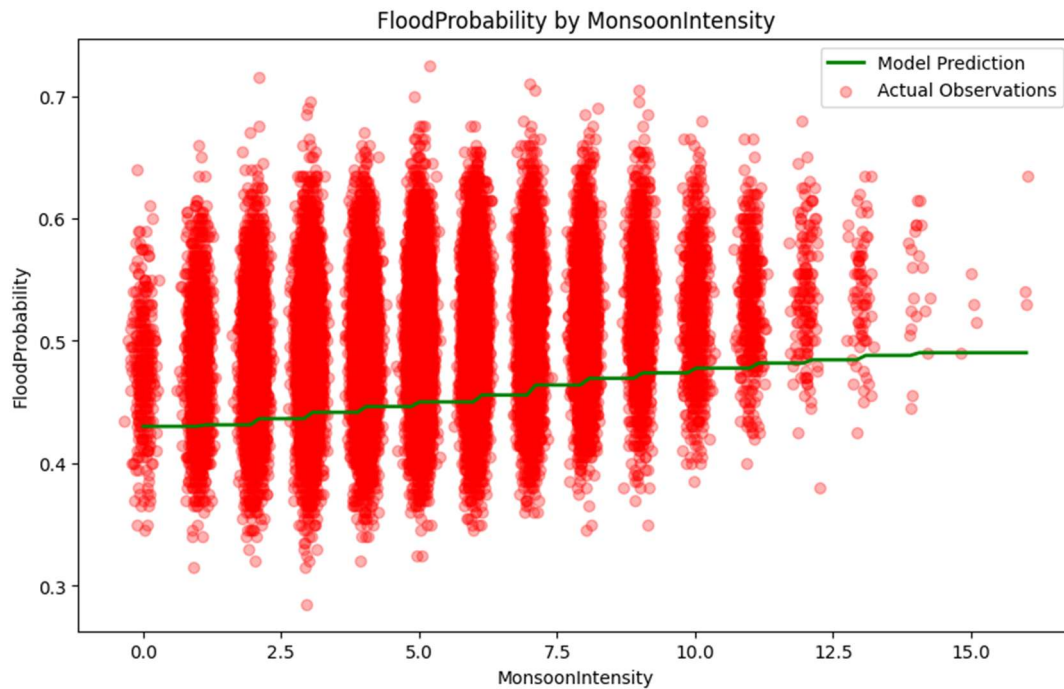
Comparison of Predicted vs. Actual Flood Probability
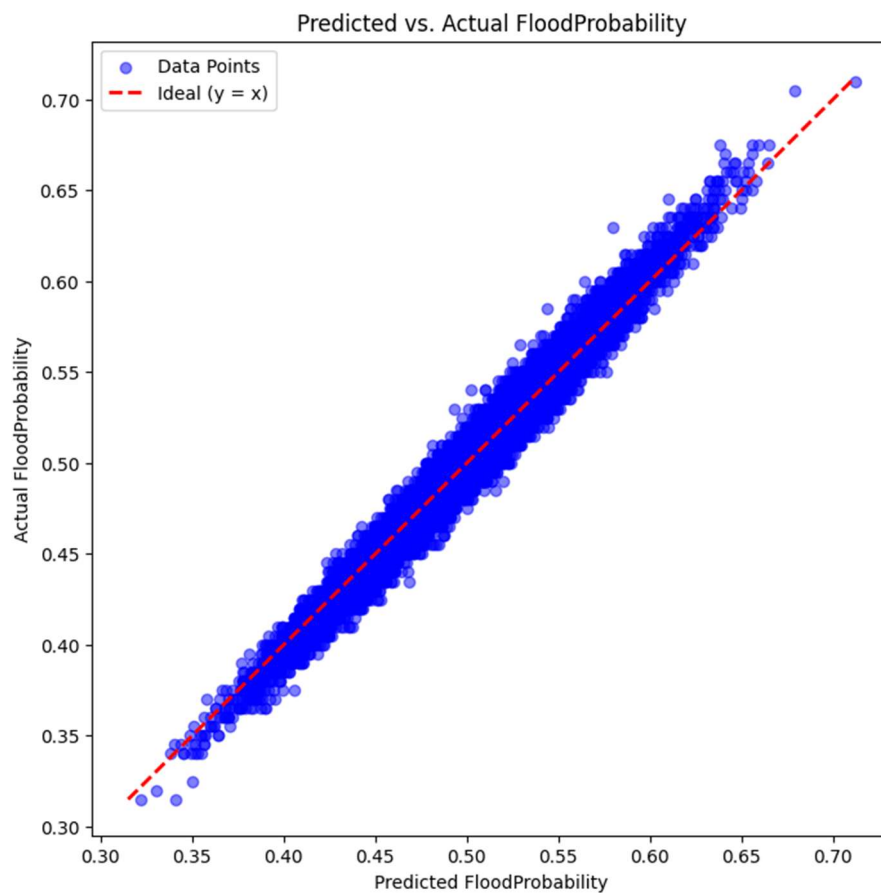
Flood Probability Calibration Curve

Calibration Chart

3. Strengths of the Model

- **High Predictive Accuracy:** Achieves over **97% variance explanation** using XGBoost.
- **Robust Feature Set:** Incorporates **climatic, geographical, and historical flood data**, improving reliability.
- **Scalability:** Efficient for large datasets due to parallelization and gradient boosting optimization.
- **Real-World Applicability:** Can assist **disaster management agencies** in proactive flood risk assessment

4. Conclusion

The **Flood Detection Model successfully predicts flood occurrences with high accuracy** using **XGBoost**, making it a **valuable tool for disaster management and flood risk mitigation**. Further advancements in deep learning and real-time data integration can improve predictive capabilities and responsiveness in flood-prone regions.