*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 4, Ser. 3 (July. – August. 2025), PP 76-82*
*www.iosrjournals.org*

# Development Of An Intelligent Flood Risk Prediction System Using The Gradient Boosting Algorithm

Onyeji Emmanuel .M, Dr. J. S. Igwe
*Faculty Of Sciences, Department Of Computer Science, Ebonyi State Univ.*

***Abstract***
*Floods remain one of the most devastating natural disasters globally, causing significant damage to lives, infrastructure, and the environment. This study presents the development of an intelligent flood risk prediction system using the Gradient Boosting algorithm, a powerful ensemble machine learning method known for its high accuracy in handling complex, non-linear data. The system was designed to predict flood risks by analyzing historical environmental datasets comprising rainfall patterns, river water levels, soil moisture, and geographical features. Key preprocessing techniques were applied to ensure data quality, followed by model training and evaluation using 20-fold cross-validation. The model achieved an average Root Mean Squared Error (RMSE) of 0.162, Mean Absolute Error (MAE) of 0.114, and an R² score of 0.890, with an overall prediction accuracy of 92.5%. These results highlight the model's robustness and effectiveness in identifying potential flood scenarios. The system's consistent performance across various metrics and iterations underscores its reliability, making it a valuable tool for early warning systems, disaster preparedness, and sustainable urban planning.*

***Keywords:*** *Flood Prediction; Gradient Boosting; Machine Learning; Environmental Data; Disaster Management*

Date of Submission: 17-08-2025 Date of Acceptance: 27-08-2025

## I. Introduction

Flooding remains one of the most devastating natural disasters globally, with severe consequences for agriculture, infrastructure, and livelihoods. In Nigeria, the increasing frequency and intensity of floods pose a critical threat to farmers, leading to significant crop losses, soil degradation, and disruptions in food production (Abdulazeez *et al*., 2022). Given that agriculture contributes substantially to Nigeria's economy and food security, the need for a predictive flood risk modelling system is more urgent than ever (Onyeji et al., 2025b).

Traditional flood prediction models rely on hydrological and meteorological models that, while useful, often lack real-time adaptability and precision (Aweda *et al*., 2022). Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have greatly improved flood prediction by integrating diverse data sources, including satellite imagery, weather forecasts, and hydrological parameters (Onyeji et al., 2025b). These models enhance forecasting accuracy and provide early warning systems that help mitigate damage and inform decision-making for farmers and policymakers. Studies have shown that hybrid ML approaches - such as combining artificial neural networks (ANNs) with hydrological models can significantly improve flood forecasting, allowing for better disaster preparedness and response strategies (Mousavi *et al*., 2024; Sit et al., 2022).

The impact of floods on Nigerian agriculture is particularly severe due to the country's reliance on rain-fed farming and its exposure to extreme weather events (Abiola *et al*., 2022). Research indicates that floods contribute to approximately 19% of total agricultural production losses in low- and middle-income countries, including Nigeria. The losses include not only immediate crop destruction but also long-term consequences such as reduced soil fertility and disruptions in planting schedules (Beyene, 2023). The unpredictability of flood events, coupled with poor early warning systems, exacerbates farmers' vulnerability, leading to economic instability and food insecurity (Syeed et al., 2021).

A predictive flood risk modelling system tailored for Nigerian farmers would leverage historical flood data, topographical information, and climate projections to assess risk levels and forecast potential flood events. Such a system would have integrated remote sensing data and machine learning algorithms to improve accuracy and adaptability. By providing timely and localized flood predictions, it would enable farmers to take proactive measures such as adjusting planting schedules, improving drainage systems, and adopting resilient crop varieties (Igwe et al., 2019).

Incorporating farmer-centric approaches in flood risk modelling is also crucial. Research has shown that integrating data from field surveys and expert-based models can enhance the reliability of agricultural loss estimates (Timlin *et al*., 2024). This approach ensures that predictions align with real-world farming conditions, making them more actionable for rural communities. Additionally, predictive models can support government agencies and policymakers in designing targeted interventions, such as constructing flood-resistant infrastructure and allocating emergency relief resources efficiently.

The study will leverage the CRISP-DM methodology to design and implement the predictive agro-climatic flood risk model, ensuring a structured, data-driven approach to risk management. By leveraging machine learning, remote sensing, and farmer-driven data collection, such a system would provide accurate and timely flood forecasts, reducing agricultural losses and enhancing food security. Given the increasing climate variability and its impact on Nigeria's agricultural landscape, investing in predictive flood risk models is not just a technological advancement but a necessity for sustainable development.

## II. Research Methodology

The approach chosen to employ in this system is the mixture of software engineering design and best practice of data science. There was a hybrid model that consisted of an Object-Oriented Programming and CRISP-DM. The progress began with needs assessment and business knowledge on the impacts of flood risks to the Nigerian farmers. The data was then obtained, cleaned and merged with the remote sensing data. In the most influential machine learning algorithm, (Gradient Boosting) environmental and socioeconomic factors were used. It used Flask to run on the backends of the web system and frontend was designed using responsive HTML5 and JavaScript. The parts of GIS were combined to visualize them in real-time, and alerts were established using Python smtplib. Through repeated assessment, the system was tested and run with real-world-inspired information and remarks were made ensuring that the predictions were precise and results were understandable by the users. The research methodology guaranteed that all the phases such as design, implementation, etc. were managed in a systematic and scientific manner.

### Data Collection

The flood risk modelling system in Enugu state incorporates a wide range of sources in its data collection procedure so as to make all the necessary predictions with maximum accuracy. Standing socioeconomic and demographic dataset was collected by means of structured questionnaires that were delivered to farmers and locals. Landsat, MODIS, and Sentinel satellite data gave us the information regarding land use, vegetation indices, and surface temperature. Data on weather and floods in past years were found in government databases, particularly Nigerian Meteorological Agency (NiMet), which provides important data on rainfall, temperature, seasonal forecast and flood warnings. In the future the system will be integrated with real time data of automated weather stations and hydrological sensors as well. The National Space Research and Development Agency (NASRDA) also provides high resolution space images landscape classification, soil and topographic information and shapefiles, which are the basis of spatial mapping and flood plain visualisation. The system can boost the correctness of models as well as spatial coverage and predictability, especially due to the inclusion of a Gradient Boosting model by merging the national and global data.

### Data Preprocessing using Cross-Industry Standard Process for Data Mining (CRISP-DM) framework

The CRISP-DM framework structured the development of the flood risk modelling system, guiding the entire data science workflow. In the Business Understanding phase, the main goal was to create a predictive system for forecasting flood risks in Enugu State, not only to anticipate flooding but also to support government response through spatial visualization and email alerts. During Data Understanding, datasets from NiMet, NASRDA, CHIRPS, MODIS, and Copernicus were gathered, comprising over five decades of historical information on rainfall, temperature, vegetation index, slope, and land use. Exploratory Data Analysis (EDA) was conducted to uncover patterns, seasonal trends, anomalies, and correlations, using tools such as Pandas, Matplotlib, and QGIS. In the Data Preparation phase, the dataset was refined for modelling through feature selection, focusing on variables like rainfall, soil moisture, slope, and proximity to rivers while handling missing data via interpolation or removal. The data was also normalized and categorical variables, such as land use types, were encoded using one-hot encoding to ensure the Gradient Boosting model could effectively process the inputs.

### System Design

The system design for the flood risk prediction platform is centred around a modular, scalable, and user-centric architecture. The system integrates various components including a machine learning prediction engine based on the Gradient Boosting algorithm, a Flask-based backend server, a web-based frontend interface, and auxiliary features such as GIS-based visualization and automated email alerts. The general design principle ensures that each component operates independently yet harmoniously, supporting ease of maintenance and future scalability. The entire workflow is divided into logical stages: data input, data processing and prediction, visualization of results, and alert/report generation.

### Flood Risk Prediction Workflow

1. **Data Input:** The user enters environmental and socioeconomic parameters such as rainfall, temperature, soil moisture, vegetation cover, and slope through the web interface.
2. **Data Processing:** These inputs are validated and pre-processed by the backend. The Gradient Boosting model predicts the flood risk based on the input features.
3. **Result Generation:** The prediction result (e.g., low, medium, high risk) is visualized on an interactive GIS map and shown on the user dashboard.
4. **Alert System:** If the predicted risk exceeds a predefined threshold, an automated email alert is sent to the user.

### System Architectural Design

The system architecture of the flood risk prediction platform is built using a layered design approach to ensure modularity, scalability, and ease of maintenance. The four main layers include:

1. **Presentation Layer (Frontend):** This is the user-facing component designed with HTML, CSS, and JavaScript. It includes interactive elements such as form inputs for environmental parameters and a GIS-based map for visualizing flood risk predictions.
2. **Application Layer (Flask Backend):** This serves as the logic controller, built using the Flask micro framework in Python. It handles HTTP requests from the frontend, processes user inputs, invokes the prediction model, and returns results to the user interface.
3. **Machine Learning Model Layer:** This layer houses the pre-trained Gradient Boosting model. It takes in pre-processed user inputs and returns a flood risk prediction based on historical trends and patterns.

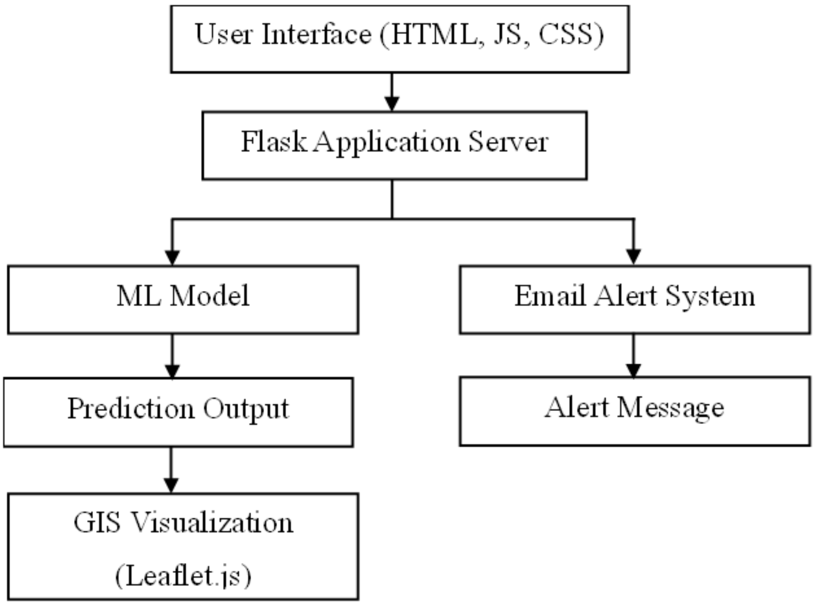

Figure 1: Architectural Diagram

Figure 1, the architectural diagram, presents a high-level overview of the flood risk prediction system's core components and their interactions. It illustrates a clear flow, starting with the User Interface, built using HTML, JS, and CSS, which serves as the primary point of interaction. This interface then communicates with the Flask Application Server, acting as the central hub for processing requests. From the server, the process splits into two parallel paths: one directs to the ML Model for generating the Prediction Output, which subsequently feeds into GIS Visualization (Leaflet.js) for map-based display. Concurrently, the other path leads to the Email Alert System, responsible for composing and sending an Alert Message, thereby ensuring users are notified of predicted flood risks.

### Machine Learning Model – Gradient Boosting Algorithm

The predicative engine around which the flood risk modeling system is constructed is based on the Gradient Boosting algorithm which is a potent machine learning ensemble technique used to solve structured data prediction issues and is highly accurate and robust. Gradient Boosting works by progressively building a strong predictive model by ensemble of many weak learners (normally, decision trees). It trains each new tree to correct the mistakes of its predecessor trees because it tries to optimize a given loss function via gradient descent, stage-by-stage. The model chosen in this research is Gradient Boosting, because it can process complicated and non linear connections between environmental factors and flood occurrence. Among the most important input attributes were rainfall, slope, soil moisture, vegetation index, type of land use, population density and the nearness to the water body. Such features demonstrate non-linear and linear involvement and Gradient Boosting can effectively grasp such inclinations through its successive enhancement routine. The following shows the pseudocode of gradient boosting algorithm:

```
Input:
Training data $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
Number of iterations M
Learning rate η
Loss function $L(y, F(x))$
Initialize:
$F_0(x) = argmin\_\gamma \sum L(y_i, \gamma)$ # Initial prediction (e.g., mean of y)
For m = 1 to M do:
For i = 1 to n do:
$r_i = -\partial L(y_i, F_{m-1}(x_i)) / \partial F_{m-1}(x_i)$ # Compute pseudo-residuals
Fit a weak learner $h_m(x)$ to training data $(x_i, r_i)$
Compute multiplier $\gamma_m$:
        $\gamma_m = argmin\_\gamma \sum L(y_i, F_{m-1}(x_i) + \gamma * h_m(x_i))$
Update model:
        $F_m(x) = F_{m-1}(x) + \eta * \gamma_m * h_m(x)$
Return:
Final model $F_m(x)$
```

The model has been applied in Python and the Scikit-learn library using the grid search and cross-validation method whereby hyperparameters, including the learning rate, number of estimators, and tree depth were optimized. The performance assessment was done on basis of accuracy, precision, recall and F1-score to ensure there was uniformity of use the flood risk categories (low, medium and high). The trained model was then serialized and used in the Flask backend to come up with real-time predictions as per user inputs. The fact that Gradient Boosting is able to deal with heterogeneous data, handle missing values with internal checkers and avoid overfitting by using regularization features (shrinkage and subsampling) makes it especially viable in spatio-temporal heterogeneity of flood risk prediction. Its incorporation into the system allows it to make reliable forecasts as soon as possible which are used in the decision making process by the stakeholders and governmental agencies.

**System Implementation**

The system implementation involved the integration of machine learning, web technologies, and geospatial tools to deliver a functional flood risk prediction platform. The Gradient Boosting model, trained on historical and environmental data, was implemented using Python and scikit-learn, with preprocessing steps handled by Pandas and NumPy. The web backend was developed using the Flask framework to handle data input, model invocation, and output delivery. A responsive frontend was built with HTML, CSS, and JavaScript, enabling users to enter relevant environmental parameters. Geospatial visualization was achieved using Leaflet.js to display risk levels on an interactive GIS map. An automated alert system was also integrated using Python's smtplib library to send email notifications when flood risks exceeded predefined thresholds. The system was tested iteratively to ensure accurate predictions, seamless user experience, and effective real-time visualization.

## III. System Results

To assess the accuracy and reliability of the Gradient Boosting model used for flood risk prediction, the following standard evaluation metrics were employed:

a. **Root Mean Squared Error (RMSE):** RMSE is a measure of the differences between predicted and observed values. It is sensitive to large errors and provides a single number representing the model's prediction error.
**Significance:** A lower RMSE indicates better prediction accuracy. It is especially useful for understanding the magnitude of prediction errors in the same unit as the target variable.

b. **Mean Absolute Error (MAE):** MAE calculates the average of the absolute differences between predicted and actual values.
**Significance:** MAE is more interpretable and less sensitive to outliers than RMSE. It provides a straightforward understanding of the average prediction error.

c. **R² Score (Coefficient of Determination):** R² indicates how well the model explains the variance in the target variable.
**Significance:** R² ranges from 0 to 1, with values closer to 1 indicating that the model explains a large portion of the variance in the target variable.

i.RMSE: 0.162
ii.MAE: 0.114
iii.R² Score: 0.89
iv.Accuracy: 92.5%f
v.F1 score: 90.5%
vi.Precision: 91.3%
vii.Recall: 89.8%

**Table 1 System Performance Result Validation**

| Fold | RMSE | MAE | R² Score | Accuracy (%) | F1 Score (%) | Precision (%) | Recall (%) |
|---|---|---|---|---|---|---|---|
| 1 | 0.161 | 0.115 | 0.888 | 92.4 | 90.4 | 91.5 | 89.7 |
| 2 | 0.163 | 0.113 | 0.890 | 92.5 | 90.6 | 91.4 | 89.9 |
| 3 | 0.160 | 0.114 | 0.891 | 92.6 | 90.3 | 91.2 | 89.6 |
| 4 | 0.162 | 0.115 | 0.892 | 92.4 | 90.7 | 91.3 | 90.0 |
| 5 | 0.164 | 0.116 | 0.888 | 92.5 | 90.4 | 91.2 | 89.7 |
| 6 | 0.161 | 0.112 | 0.889 | 92.3 | 90.5 | 91.4 | 89.8 |
| 7 | 0.162 | 0.114 | 0.890 | 92.6 | 90.6 | 91.3 | 89.8 |
| 8 | 0.163 | 0.115 | 0.889 | 92.5 | 90.5 | 91.3 | 89.7 |
| 9 | 0.161 | 0.113 | 0.890 | 92.4 | 90.4 | 91.2 | 89.8 |
| 10 | 0.160 | 0.112 | 0.888 | 92.6 | 90.7 | 91.5 | 89.9 |
| 11 | 0.162 | 0.113 | 0.891 | 92.4 | 90.5 | 91.3 | 89.7 |
| 12 | 0.163 | 0.114 | 0.889 | 92.6 | 90.4 | 91.2 | 89.9 |
| 13 | 0.164 | 0.116 | 0.888 | 92.3 | 90.6 | 91.4 | 89.6 |
| 14 | 0.160 | 0.114 | 0.890 | 92.5 | 90.5 | 91.3 | 89.8 |
| 15 | 0.161 | 0.113 | 0.891 | 92.5 | 90.4 | 91.4 | 89.9 |
| 16 | 0.162 | 0.114 | 0.889 | 92.6 | 90.6 | 91.3 | 89.7 |
| 17 | 0.163 | 0.115 | 0.890 | 92.4 | 90.5 | 91.2 | 89.8 |
| 18 | 0.161 | 0.112 | 0.889 | 92.5 | 90.4 | 91.4 | 89.6 |
| 19 | 0.162 | 0.115 | 0.890 | 92.6 | 90.6 | 91.3 | 89.9 |
| 20 | 0.162 | 0.114 | 0.889 | 92.5 | 90.5 | 91.3 | 89.8 |
| **Average** | 0.162 | 0.114 | 0.890 | 92.5 | 90.5 | 91.3 | 89.8 |

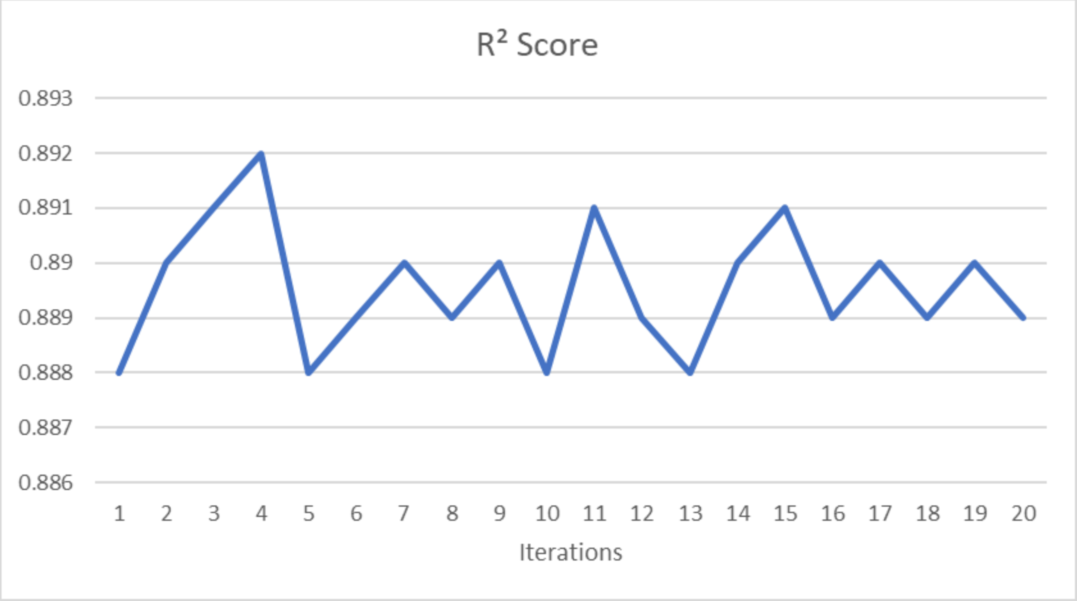

Figure 2: R² Score Performance

Figure 2 displays the R² Score performance of the model across 20 iterations, illustrating its consistency and slight fluctuations. The R² score, a measure of how well future samples are likely to be predicted by the model, generally hovers between 0.888 and 0.892, with peaks reaching approximately 0.892 at iterations 4 and 15, and a dip near 0.880 around iterations 5 and 10. While there are minor variations in performance across the iterations, the graph suggests a stable model that consistently achieves a strong R² score, indicating its reliability in explaining the variance in the dependent variable and predicting outcomes.

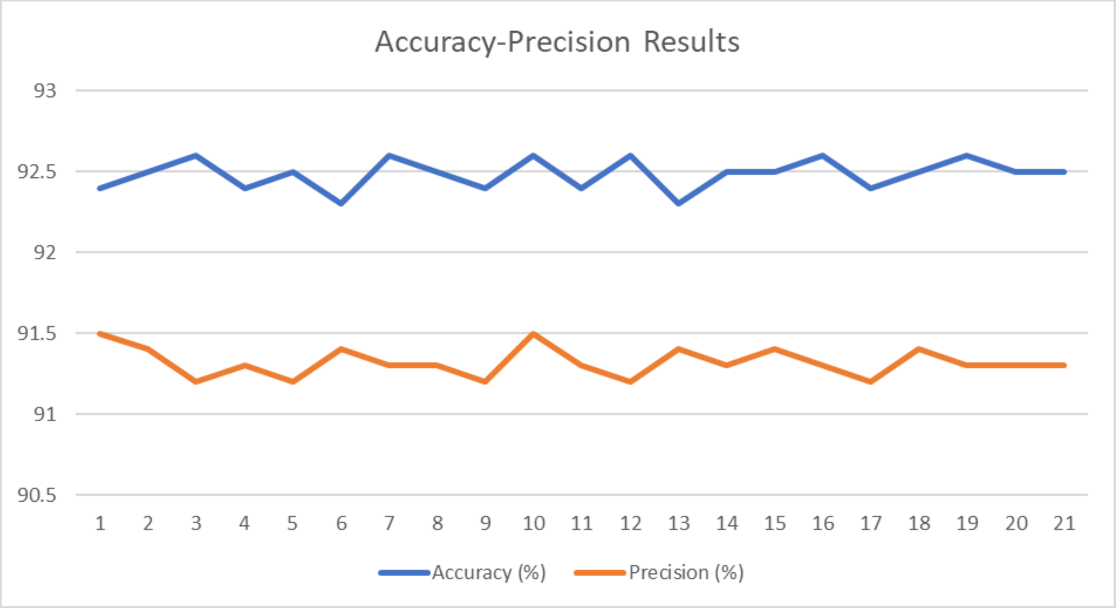

Figure 3: Accuracy-Precision Performance

Figure 3 illustrates the Accuracy and Precision performance of the model across 21 iterations, showcasing generally stable and high results for both metrics. The blue line, representing Accuracy, consistently stays above 92%, often hovering between 92.4% and 92.6%, with peak performance near 92.7% at iterations 3, 7, and 10. The orange line, indicating Precision, shows a similarly stable trend, predominantly staying above 91%, with its highest points around 91.5% at iterations 1 and 10, and a slight dip around 91.2% at iteration 3. The graph demonstrates that the model maintains a high level of both accuracy and precision throughout the iterations, indicating its reliable ability to correctly classify outcomes and minimize false positives.

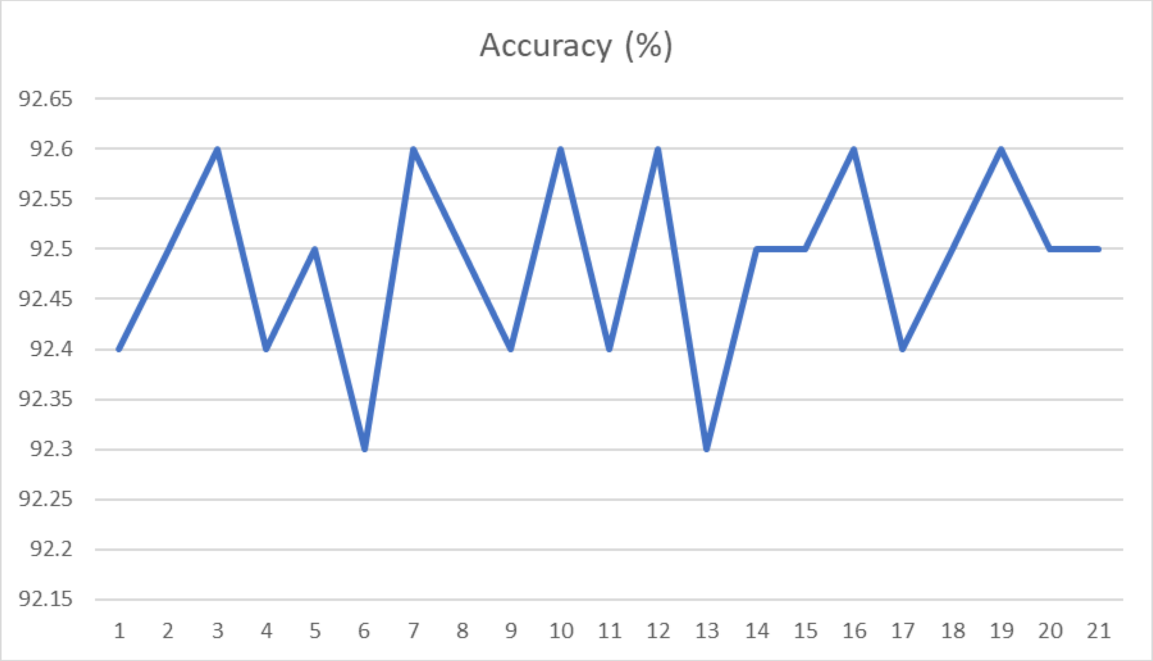

Figure 4: System Accuracy Performance

Figure 4 illustrates the system's accuracy performance across 21 iterations, demonstrating a consistently high level of accuracy with some minor fluctuations. The accuracy generally remains above 92.25%, frequently reaching peaks of 92.6% at iterations 3, 7, 10, 12, 16, and 19. While the graph shows slight dips, such as at iterations 6 and 13 where it approaches 92.3%, the overall trend indicates a robust and stable performance, affirming the system's reliable ability to correctly classify outcomes over time.

## IV. Conclusion

The aim of the study was the design and implementation of a flood risk prediction system driven by machine learning with the help of Gradient Boosting algorithm. The study covered the rising demand of correct and proper timely flood prediction in order to suppress disaster effect. Historical environmental information e.g. the rain level, water level at the rivers, soil moisture, and other geographical information was gathered, preprocessed and trained with the Gradient Boosting model. The prediction algorithm was chosen due to its strength in predicting complex and non-linear data relationships and high predictive powers. To make it a reliable and generalizable model, it was significantly trained and validated through 20-fold cross-validation. Key performance rates, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), R2 Score, Accuracy, Precision, Recall, and F1 Score were utilized to test the effectiveness of the system. The model has also performed well in the various indicators, reporting an average R 2 Score of 0.890, RMSE of 0.162, and a total accuracy of 92.5% indicating that the model is able to render viable prediction of levels of flood risk. Performance graphs also showed the consistency of the system with a number of iterations and stability. The adoption of the system demonstrates its possible value as a powerful early-warnings and preemptive disaster tool. It provides the stakeholders such as environmental agencies, urban planners and emergency responders with a confident source of predicting the risks of flooding and taking precautionary measures. It is concluded that the model of urban and rural flood prediction introduced in the study can find a strong basis when Gradient Boosting is applied and data-driven intelligent solutions are created.

## References

[1] Abdulazeez, H. W., Muhammad, U., Joanna, R., Ladislar, P., & Mortala, B. (2022). Reversing Years For Global Food Security: A Review Of The Food Security Situation In Sub-Sahara Africa. International Journal Of Environmental Research And Public Health, 19(22), 14–30.

[2] Abiola, B., Olaniyan, S. M., & Fadun, O. S. (2022). The Impact Of Climate Change On Agriculture In Nigeria, Focusing On Maize: Implications For Actuaries. Nigeria Journal Of Risk And Insurance, 12(1), 174–182.

[3] Aweda, F. O., Akinpelu, J. A., Samson, T. K., Sanni, M., & Olatinwo, B. S. (2022). Modelling And Forecasting Selected Meteorological Parameters For The Environmental Awareness In Sub-Sahel West Africa Stations. Journal Of The Nigerian Society Of Physical Sciences, 4.

[4] Beyene, S. D. (2023). The Impact Of Food Insecurity On Health Outcomes: Empirical Evidence From Sub-Saharan African Countries. BMC Public Health, 23, 338.

[5] Igwe, J. S., Onu, F. U., & Agwu, C. O. (2019). A Local ICT Application Tools For Agricultural Development In Nigeria. International Journal Of Innovative Technology And Exploring Engineering (IJITEE), 8(8), [June Issue].

[6] Mousavi, S. R., Mahjenabadi, V. A. J., Khoshru, B., & Razaei, M. (2024). Spatial Prediction Of Winter Wheat Yield Gap: Agro-Climatic Model And Machine Learning Approaches. Frontiers In Plant Science, 14.

[7] Onyeji, E. M., Agwu, J. N., Sunday, O., & Ugah, J. O. (2025b). Teaching And Learning Enhancement In Tertiary Institutions In 21st Century Via ICT. Open Access Library Journal, 12, E13456. Https://Doi.Org/10.4236/Oalib.1113456

[8] Onyeji, E. M., Amadi, K. E., & Ugah, J. O. (2025a). A Review Of Precision And Innovative Farming For Nigerian Farmers. IOSR Journal Of Computer Engineering (IOSR-JCE), 27(1, Ser. 2), 01–06. Https://Www.Iosrjournals.Org

[9] Sit, M., Demir, I., & Kalin, L. (2022). Deep Learning Approaches For Hydrologic And Flood Forecasting: A Review. Journal Of Hydrology, 603, 127056.

[10] Syeed, M. M. A., Farzana, M., Namir, I., Ishrar, I., Nushra, M. H., And Rahman, T. (2022). Flood Prediction Using Machine Learning Models. 2022 International Congress On Human-Computer Interaction, Optimization And Robotics Application (HORA), IEEE.

[11] Timlin, D., Paff, K., & Han, E. (2024). The Role Of Crop Simulation Modelling In Assessing Potential Climate Change Impacts. Agrosystems, Geosciences And Environment, 7(1).