DLSU Research Congress 2025
De La Salle University, Manila, Philippines
June 26 to 28, 2025

Predicting Flood Occurrences in Caloocan City Using Select Machine Learning Models

Alyana Joy De Leon[1], Samira Allyson Fortuno[2], Andrea Nicole Madolora[3], Michaela Lynn Santos[4], and Shirley Chu[1,*]

De La Salle University - Integrated School (Manila Campus)
[1]alyana_deleon@dlsu.edu.ph
[2] samira_fortuno@dlsu.edu.ph
[3]andrea_madolora@dlsu.edu.ph
[4]michaela_lynn_santos@dlsu.edu.ph
**Corresponding Author:* shirley.chu@dlsu.edu.ph

Abstract: Flooding has become a significant global economic problem and public health concern due to global warming and increased rainfall. In Metro Manila, the City of Caloocan is particularly susceptible to flooding because of its flat terrain and proximity to various bodies of water. This study investigates the application of selected machine learning models, namely, Decision Trees, Logistic Regression, Support Vector Machine, and Artificial Neural Network, for predicting flood occurrences in Caloocan. The models use climatological data from 2020 to 2025 to enhance prediction accuracy and inform local flood mitigation strategies. To address the class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed, and the dataset was partitioned into training and testing sets using a 75%-25% split. Principal Component Analysis (PCA) was applied to reduce data dimensionality before model training. Model performance was evaluated using Accuracy, Precision, and Recall, with results visualized through Confusion Matrices. The findings demonstrated each model's effectiveness in predicting flood occurrences, with Decision Trees and Artificial Neural Networks generally performing better with higher Accuracy, Precision, and Recall scores than Logistic Regression and Support Vector Machine. This research contributes to data-driven disaster risk reduction initiatives and provides valuable insights for local government units and communities in enhancing flood resilience within the City of Caloocan.

Key Words: Prediction; Forecasting; Floods; Machine Learning; Decision Tree; Logistic Regression; Support Vector Machine; Artificial Neural Network

1. INTRODUCTION

1.1 Background of the Study

Flooding has become a widespread and destructive natural disaster, severely impacting cities worldwide. With their dense populations and extensive infrastructure, urban areas are particularly vulnerable, as highlighted by Mignot et al. (2019) and Abebe et al. (2018). In the Philippines, cities like Caloocan in Metro Manila face recurring flooding issues, which disrupt transportation and commerce. Flooded streets cause traffic congestion and pose health risks, as they can become breeding grounds for disease-carrying mosquitoes and waterborne contaminants (Plyushteva & Schwanen, 2022). To mitigate these risks, the Department of Public Works and Highways (DPWH) has proposed flood control wall projects from 2018 to 2024 to protect areas along creek banks. While these walls provide immediate protection, they offer a localized solution and do not address the root causes of flooding (Wang et al., 2022). This study aims to address the limitations of conventional flood forecasting models by employing machine learning techniques specifically tailored to the unique flooding conditions in Caloocan City. It seeks to evaluate the effectiveness of selected models in accurately predicting flood events and supporting data-driven disaster preparedness efforts.

Machine learning models have become crucial in flood prediction due to their ability to capture complex, nonlinear relationships among environmental variables, offering a more effective alternative to traditional hydrological models that often rely on simplified assumptions and involve uncertainties (Moges et al., 2020). These models have shown strong performance in flood forecasting when trained on historical weather data. Support Vector Machines (SVM) and Logistic Regression are effective for binary classification tasks, while Decision Trees offer interpretability and handle nonlinearity well. Furthermore, Artificial Neural Networks (ANNs) can learn and generalize from complex data, providing robust predictions even with limited prior knowledge of underlying processes. Recent studies highlight that integrating machine learning with real-time environmental data improves the reliability and accuracy of flood forecasts by capturing key parameters that are difficult to observe directly (Razali et al., 2020). Therefore, these models serve as valuable tools for predicting flood risks in Caloocan City.

1.2 Scope and Limitations

This study focuses on flood forecasting in Caloocan and its immediate surroundings by analyzing daily climatological and flood advisory data sourced from PIMOH and local Facebook pages, which are assumed to be accurate and reliable. The predictive model was developed using machine learning techniques, specifically Support Vector Machines (SVM), Decision Trees, Logistic Regression, and Artificial Neural Networks (ANN), to identify the most effective approach to forecasting flood events based on historical weather data. However, a key limitation of the study lies in the potential impact of human interventions, such as flood wall construction, which can alter natural water flow, redistribute flood risks, and introduce inconsistencies in data, thereby affecting the accuracy and reliability of the machine learning models.

1.3 Significance of the study

Machine learning-based flood forecasting models designed explicitly for Caloocan City address the city's persistent urban flooding concerns. By enhancing predictive accuracy and facilitating the timely issuance of early warnings, these models have the potential to mitigate economic disruptions and support more effective governmental response strategies. Furthermore, the study may increase operational confidence among local stakeholders, including businesses and community planners. The insights derived from the models can also inform infrastructure development and urban planning efforts, particularly in the design of flood-resilient structures and drainage systems. Ultimately, the research aims to support a more resilient and adaptive urban environment, improving disaster preparedness and protecting lives and property in Caloocan City.

2. METHODOLOGY

2.1 System Architecture

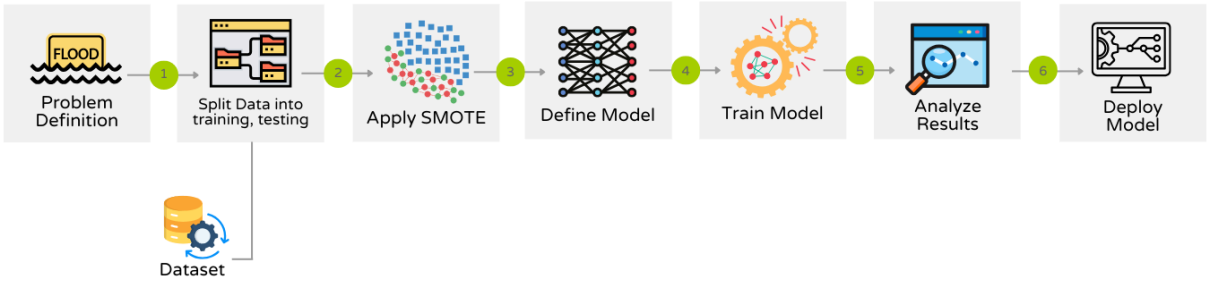

Figure 1. System Architecture of Decision Tree and ANN

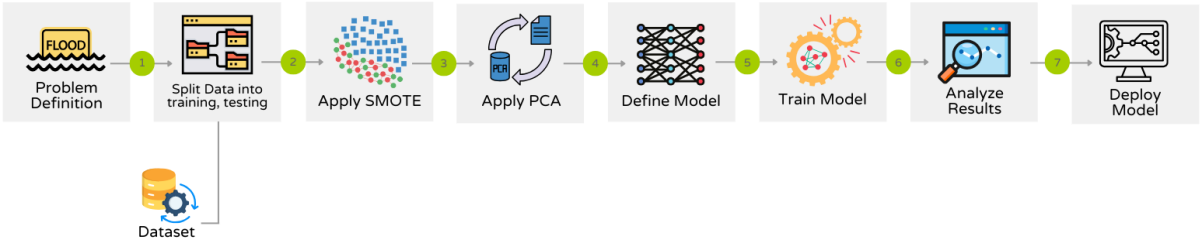

Figure 2. System Architecture of Logistic Regression and SVM

2.2 Data Collection & Preparation

Climatological data and flood condition updates from January 2020 to December 2024 were collected to support flood forecasting, with data from January 1, 2025, onward continuously being gathered for real-time validation and performance assessment of the model. Weather-related variables such as temperature, rainfall, and wind speed were sourced from the PIMOH Weather website, while flood occurrences were tracked using updates from the Caloocan City DRRMO, primarily through their Facebook page. Supplementary data from the Caloocan Public Information Office, online newspapers, and social media platforms were also used to enhance the model's accuracy and reliability. The dataset underwent thorough cleaning to ensure quality—missing or inconsistent entries were addressed through interpolation or verified using secondary sources. Standardizing formats for dates and units was also applied, ensuring a consistent and high-quality dataset for input into the machine learning models.

2.3 Data Analysis

The flood forecasting model's performance was evaluated using the Confusion Matrix, Precision, Recall, and Accuracy metrics. The Confusion Matrix presented a 2x2 breakdown of the model's predictions, highlighting true positives (correctly predicted floods), true negatives (correctly predicted no floods), false positives (predicted floods that did not occur), and false negatives (missed flood events). Precision and Recall were calculated based on the data from the Confusion Matrix.

Precision measured how many of the predicted floods were floods and was calculated as:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad \text{(Eq. 1)}$$

The recall assessed how many of the actual floods were correctly predicted and was calculated as:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negative} \quad \text{(Eq. 2)}$$

Finally, Accuracy was calculated as the proportion of correct predictions (both floods and no floods) to the total predictions, given by:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Predictions} \quad \text{(Eq. 3)}$$

3. RESULTS AND DISCUSSION

3.1 Overview of the Dataset

The dataset used for flood prediction exhibited a significant class imbalance. Initially, it contained 28 floods and 1,877 non-flood events. These flood and non-flood events represent the daily weather data collected from January 2020 until March 2025. After removing 244 rows with missing values in key features (e.g., rainfall, wind data), only 11 flood events remained, intensifying the imbalance. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, generating synthetic flood instances to balance the classes.

Table 1. Dataset Summary

	Flood Occurrences	Non-Flood Occurrences
Original	28	1,877
After Data Cleaning	11	1,650
After the Application of SMOTE	1,650	1,650

The data was divided into training and test sets in a ratio of 75%:25% to ensure the models learn from climatological patterns while minimizing overfitting. This resulted in 2475 training samples (1244 from the original data) and 825 testing samples (417 from the original data). SMOTE was applied to the 4 cases of flood occurrences to balance the model and decrease the risk of the model being biased to the majority class.

Table 2. Class Distribution of Original and Synthetic Samples in the Test Set

	No Flood	Flood
Original	413	4
Synthetic	0	408

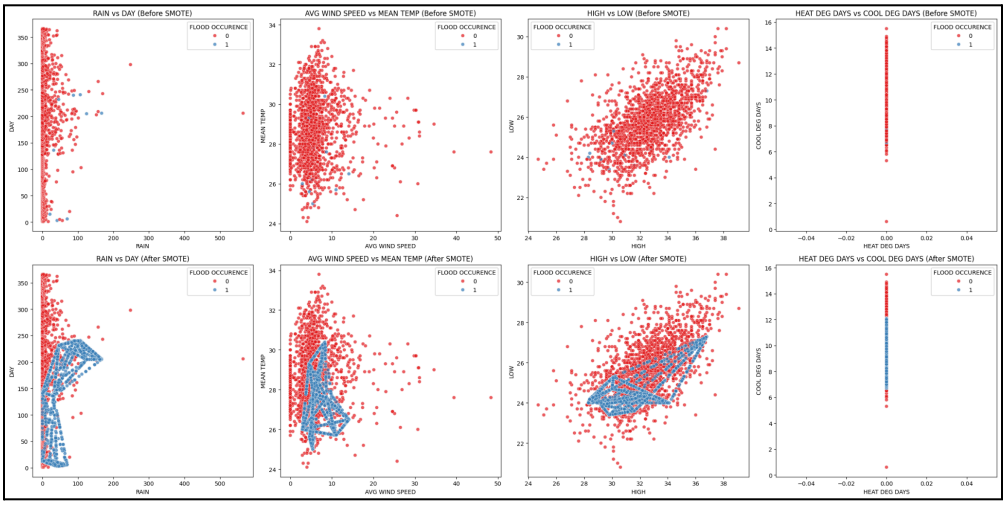

Figure 3. Scatterplots of the dataset before and after the application of SMOTE

The final balanced dataset was used to select machine learning models using features such as mean, low, and high temperatures, heat and cool degree days, average and highest wind speeds, rainfall, and dominant wind direction. The output was binary: 1 for flood, 0 for no flood.

3.2 Model Performance and Evaluation

3.2.1 Decision Tree

Table 3. Confusion Matrix with SMOTE applied to the dataset

	Predicted No Flood	Predicted Flood
Actual No Flood	403	10
Actual Flood	4	408

Table 4. Confusion Matrix with the original dataset

	Predicted No Flood	Predicted Flood
Actual No Flood	403	10
Actual Flood	0	4

The model showed strong performance on both the SMOTE-augmented and original datasets. In the SMOTE confusion matrix, only 14 instances were misclassified out of 825 test samples, 4 of which were synthetic flood cases mistakenly labeled as non-flood. Importantly, all four actual flood events from the original data were correctly classified, suggesting that including synthetic data did not hinder the model's ability to detect real flood occurrences. It is also worth noting that the original dataset exhibited very low precision due to the limited number of actual flood cases and a relatively high number of false positives. These results indicate that SMOTE effectively enhanced flood class representation and improved generalization.

3.2.2 Logistic Regression

Table 5. Confusion Matrix with SMOTE applied to the dataset

	Predicted No Flood	Predicted Flood
Actual No Flood	262	151
Actual Flood	43	369

Table 6. Confusion Matrix with the original dataset

	Predicted No Flood	Predicted Flood
Actual No Flood	412	1
Actual Flood	4	0

The comparison between the confusion matrices highlights the substantial impact of applying SMOTE to address class imbalance in Logistic Regression. Without SMOTE, the model demonstrated a strong bias toward the majority class, ignoring instances of flooding. Although this configuration yielded a high overall accuracy, it failed to detect the minority class of flood occurrences entirely, undermining the model's utility in flood prediction. In contrast, applying SMOTE resulted in a more balanced classification performance. While it increased both false positives and false negatives, the model became considerably more effective in identifying floods. This compromise highlighted the limitations of relying solely on accuracy as an evaluation metric in imbalanced datasets. It demonstrated the necessity of prioritizing recall when the objective is to detect critical yet infrequent occurrences such as floods.

3.2.3 Support Vector Machine

Table 7. Confusion Matrix with SMOTE applied to the dataset

	Predicted No Flood	Predicted Flood
Actual No Flood	314	99
Actual Flood	80	332

Table 8. Confusion Matrix with the original dataset

	Predicted No Flood	Predicted Flood
Actual No Flood	413	0
Actual Flood	4	0

The confusion matrix without SMOTE reveals that while the model accurately classified 413 no-flood cases, it misclassified all four flood cases as no flood, leading to a recall of 0%. This indicates a strong bias towards the majority class, making the model ineffective for flood prediction despite its high accuracy. However, the model became more balanced when SMOTE was incorporated, accurately identifying 332 flood cases and 316 no-flood cases. On the other hand, 80 Flood cases were incorrectly identified as false negatives, and 97 No Flood cases were incorrectly classified as false positives. As a result, the model's accuracy decreased, but its recall increased, indicating that it is far more effective at identifying flood events. These findings reinforce that SMOTE greatly enhanced the model's ability to identify floods and lessened its bias towards the majority class.

3.2.4 Artificial Neural Network

Table 9. Confusion Matrix with SMOTE applied to the dataset

	Predicted No Flood	Predicted Flood
Actual No Flood	403	10
Actual Flood	1	411

Table 10. Confusion Matrix with the original dataset

	Predicted No Flood	Predicted Flood
Actual No Flood	413	0
Actual Flood	4	0

When trained on the original dataset, the ANN demonstrated a strong ability to identify "No Flood" instances, correctly classifying 413 cases. However, it failed to detect any "Flood" cases, misclassifying all four as "No Flood." This outcome suggests that the model was biased toward the majority class, likely due to class imbalance. In contrast, when SMOTE was applied to balance the dataset, the model significantly improved classification performance across both classes. Specifically, the ANN correctly classified 403 out of 413 "No Flood" instances and 411 "Flood" instances, with only 10 "No Flood" cases misclassified as "Flood." These results indicate that applying SMOTE effectively addressed the class imbalance issue, enabling the model to achieve a more balanced and accurate classification, particularly improving its sensitivity to detecting flood events.

3.2.5. Evaluation Metrics

The tables below present the evaluation metrics of various machine learning models' accuracy, precision, and recall, demonstrating a comparative analysis of their performance with and without the application of SMOTE.

Table 11. Evaluation Metrics of Machine Learning Models with SMOTE

Model	Accuracy	Precision	Recall
Decision Tree	0.98	0.98	0.99
Logistic Regression	0.76	0.71	0.90
Support Vector Machine	0.79	0.77	0.81
Artificial Neural Network	0.99	0.98	1.00

Table 12. Evaluation Metrics of Machine Learning Models with the Original Dataset

Model	Accuracy	Precision	Recall
Decision Tree	0.98	0.29	1.00
Logistic Regression	0.99	0.00	0.00
Support Vector Machine	0.99	1.00	0.00
Artificial Neural Network	0.99	0.00	0.00

4. CONCLUSIONS

In conclusion, upon comparison of the four machine learning models, the Artificial Neural Network model achieved the highest performance, indicating that it could correctly identify almost all actual flood events with very few false positives. The Decision Tree model also performed strongly with scores almost identical to ANN, making it a competitive alternative.

The Logistic Regression and Support Vector Machine models demonstrated relatively lower performance, with the former scoring the lowest in precision and the SVM having moderate scores across all metrics. These results suggest that while simpler models may still detect floods reasonably well, more complex models, such as Decision Trees and ANN, provide significantly better performance and reliability.

When trained with SMOTE, the ANN and Decision Tree models misclassified only 10 and 14 instances, respectively, out of 825 test samples, reflecting an error rate of merely 1.21% and 1.70%. In contrast, the SVM model misclassified 177 instances (21.45%), while Logistic Regression misclassified 194 (23.52%), indicating much higher error rates and reduced reliability. Furthermore, ANN was the only model that successfully classified all 412 actual flood events in the test set, achieving a 0% false negative rate, critical in real-world flood prediction scenarios where missed alerts can have severe consequences.

However, on the original imbalanced dataset, both Logistic Regression and ANN failed to identify any flood events (recall = 0%), indicating a complete sensitivity breakdown in low-flood conditions. Only the Decision Tree model retained its ability to detect all actual flood events despite the imbalance, albeit with a very low precision (0.29), showing a tendency to overpredict floods to capture all true positives.

The robustness of these models underscores their potential integration into real-world flood forecasting systems, particularly within Caloocan City. Their adoption could foster public trust, promote timely preparedness, and inform urban planning and resource allocation to enhance flood resilience.

Future research may consider exploring strategies for addressing highly imbalanced datasets, such as resampling techniques or algorithm-level adjustments. Additionally, investigating a broader range of machine learning approaches—like ensemble methods and deep learning architectures—could offer insights into improving predictive accuracy. Integrating real-time environmental and sensor data, and extending these models to new flood-prone regions, may also enhance responsiveness and support more effective disaster risk reduction

5. ACKNOWLEDGMENTS

The researchers would like to express their sincere gratitude to their research adviser, Ms. Shirley Chu of the College of Computer Studies, De La Salle University - Manila, for her invaluable guidance, support, and encouragement throughout this study. Her expertise and mentorship greatly contributed to the successful completion of this research. The researchers also extend their appreciation to their panelists, Dr. Thomas James Tiam-Lee and Mr. Arren Matthew Antioquia, for their constructive feedback and insightful recommendations, which helped strengthen the overall quality of the study.

6. REFERENCES

Abebe, Y., Kabir, G., & Tesfamariam, S. (2018). Assessing urban areas' vulnerability to pluvial flooding using GIS applications and Bayesian Belief Network model. Journal of Cleaner Production, 174, 1629–1641. https://doi.org/10.1016/j.jclepro.2017.11.066

Mignot, E., Li, X., & Dewals, B. J. (2019). Experimental modelling of urban flooding: A review. Journal of Hydrology, 568, 334–342. https://doi.org/10.1016/j.jhydrol.2018.11.001

Moges, E., Demissie, Y., Larsen, L., & Yassin, F. (2020). Review: Sources of Hydrological Model Uncertainties and Advances in their analysis. Water, 13(1), 28. https://doi.org/10.3390/w13010028

Plyushteva, A., & Schwanen, T. (2022). " We usually have a bit of flood once a week ": conceptualising the infrastructural rhythms of urban floods in Malate, Manila. Urban Geography, 44(8), 1565–1583. https://doi.org/10.1080/02723638.2022.2105003

Razali, N., Ismail, S., & Mustapha, A. (2020). Machine learning approach for flood risks prediction. IAES International Journal of Artificial Intelligence, 9(1), 73. http://doi.org/10.11591/ijai.v9.i1.pp73-80

Wang, L., Cui, S., Li, Y., Huang, H., Manandhar, B., Nitivattananon, V., Fang, X., & Huang, W. (2022). A review of the flood management: from flood control to flood resilience. Heliyon, 8(11), e11763. https://doi.org/10.1016/j.heliyon.2022.e11763