A Comparative Study Using Machine Learning and Flood Data to Predict Floods

Ariel Dominic A. Alfonso
Ateneo de Naga University
Naga City, Camarines Sur, Philippines
arialfonso@gbox.adnu.edu.ph

Xier Gabriel M. Mangunay
Ateneo de Naga University
Naga City, Camarines Sur, Philippines
xmangunay@gbox.adnu.edu.ph

James Edward Q. Vidola
Ateneo de Naga University
Naga City, Camarines Sur, Philippines
jvidola@gbox.adnu.edu.ph

Raphael Henry Garay
Ateneo de Naga University
Naga City, Camarines Sur, Philippines
rgaray@gbox.adnu.edu.ph

ABSTRACT

Typhoons are considered one of the most devastating natural calamity in the Philippines. This natural calamity resulted in major casualties, taking properties and a hundred lives. In such cases, people get left in their homes and await rescue due to flood. Because of this, various research were conducted to solve the problem, one of which is machine learning. However, efficiency and accuracy come into question. What is the best machine learning algorithm to use for flood prediction? The researchers have decided to do a comparative study of machine learning algorithms to predict river and urban flood height using weather data and flood reports from various local government agencies in the Philippines.

The dataset used to train the machine learning algorithms came from weather data from the Philippine Atmospheric, Geophysical, and Astronomical Services Administration (PAGASA), urban flood reports by the Metropolitan Manila Development Authority (MMDA), River floods reports and water level by the Effective Flood Control Operational System (EFCOS), and digital elevation maps by the National Mapping and Resource Information Authority (NAMRIA). These datasets were consolidated to accommodate the training and testing for urban and river flood prediction separately. The 2021 dataset columns were separated to validate the predictions through a Streamlit application that generates results for each area.

This research provides an overview of three different machine learning algorithms that the researchers will be comparing throughout this study. The Random Forest (RF), Artificial Neural Network (ANN), and Support Vector Machines (SVM). The algorithms were used to predict the flood height and water level for urban and river floods respectively. The evaluation methodologies used for the models are Cross-Validation and Linear Regression showing the standard deviation, average score accuracy, and mean squared error. The result show that Random Forest is the most effective machine learning model using the Philippine data. In urban flood prediction, RF displayed the highest accuracy at 61% while having a low mean squared error of 31 making it the most balanced model than the others. Similarly, RF demonstrated the highest average accuracy at 99.9511% in five rivers while maintaining the lowest mean squared error across seven rivers out of nine.

CCS CONCEPTS

• **Computing methodologies → Feature selection**; **Classification and regression trees**; **Cross-validation**.

KEYWORDS

machine learning, neural networks, gaze detection, text tagging

ACM Reference Format:
Ariel Dominic A. Alfonso, Xier Gabriel M. Mangunay, James Edward Q. Vidola, and Raphael Henry Garay. 2018. A Comparative Study Using Machine Learning and Flood Data to Predict Floods. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Typhoons and rainstorms have ravaged multiple areas in the Philippines. In alone, fifteen (15) tropical cyclone have been recorded in the Philippines, five under the category of typhoons, four were tropical depressions, three were tropical storms, and four were severe tropical storms [2]. That alone caused floods in multiple areas leaving people unprepared and unable to stock basic necessities such as food and water. This was more devastating as COVID-19 continues to plague the country.In addition, given that we are still in the midst of the pandemic, evacuation centers were overcrowded and social distancing were not observed allowing the quick spread of the virus and disease. This caused severe impacts on the mental health and well-being of those affected [7].

From 1980 to 2020, floods became the second leading cause of annual hazard occurrence in the Philippines, making up 23.13 percent while storms and typhoons were at the top with 46.94 percent, also the leading cause of flash floods[4]. MMoreover, in the near future, the risk of floods in the Philippines will continue to rise due to climate change. In Davao Oriental alone the risk of flooding increases in 27 barangays with high-risk floods out of 183 barangays, and in the short term run, might increase to 39 barangays if there is without disaster risk management involved [1]. Buba et al. also

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

state that the findings of their investigation showed that among other characteristics of flood susceptibility, elevation is the most crucial element, followed by land use. 4, 445 Ha of the 6, 258 Ha total area i classified as very susceptible, and 1, 815 Ha as moderately vulnerable [**?**].

Rossi et.al mentioned that news pages would always report natural events such as weather forecasts and disasters currently happening such as typhoons, earthquakes, and floods according to Rossi et al. [8]. On the other hand, Social Media today such as Facebook, TikTok, Instagram, YouTube, and X (formerly Twitter) help provide daily and crucial information to the public on what is happening in their country [9] . Some users of the said social media websites would stream the events live for the users to see what is happening in real-time. Allowing the public to have access to a plethora of constant updates.

2 OBJECTIVES

The main objective of the study is to determine which of the Machine Learning algorithms are the most accurate in predicting floods using Philippine data. In order to achieve the main objective, the following specific objectives must be accomplished:

- Gather weather data, urban and river flood data, and digital elevation models from the local government such as PAGASA, NAMRIA, MMDA, and EFCOS
- Extract and convert geological and weather data features from gathered data sets using the Geological Information System (GIS).
- Consolidate the MMDA and EFCOS datasets to train and test machine learning algorithms.
- Implement the Random Forest Algorithm, Support Vector Machine Algorithm, and the Artificial Neural Network Algorithm in the given dataset to predict floods
- Analyze data from the results of the Random Forest models, Artificial Neural Network models, and Support Vector Machine models
- Determine the accuracy, standard deviation, and mean squared error of the selected machine learning models using Linear Regression and Cross-validation
- Compare the results of the created machine learning models
- Develop a simple Streamlit web app to test the models as a proof of concept
- Devise a conclusion based on the acquired data

3 REVIEW OF RELATED LITERATURE

3.1 Flood prediction using Machine Learning Models

3.1.1 Random Forest Algorithm. Nowadays, Machine Learning and Artificial Intelligence (AI) are used to create flood detection systems. In the study of Hashi et al., they used multiple Machine Learning algorithms with an emphasis on Deep Learning. These are Naive Bayes, Random Forest, J48, and Convolutional Neural Network. Naive Bayes and the J48 algorithm were used to classify the floods such as low medium and a high chance of flood. Random Forest was used for the high amount of data sets, and the Convolutional Neural Network deep learning approach was used to predict the chance of having floods in their area. The results show that the Random Forest Algorithm was the most accurate with 98 percent followed by the other algorithms at 84 to 88 percent It was noted that this study could be improved by using crowdsourced and real-time data to identify if an area is flooding completely [5]. The research will be utilized to determine better machine methods for predicting floods and its effectiveness.

Further research in the Random Forest algorithm was done by Chen et al. wherein Random Forest and Radial Basis Function Neural Network was used to evaluate the risk of flooded areas in China particularly the Yangtze River. Geographic information system (GIS), seasonal rainfall data, gross domestic product (GDP), and urban impervious area ratio were all analyzed for the model. In their results, it has been proven that the risk assessments the researchers created were in line with the real-life situation making the random forest model effective in predicting the flooded areas [**?**].

3.1.2 Artificial Neural Network. In the research of Chang et al., A support Vector Machine forecasting model for typhoon floods was developed due to the need for early emergencies regarding floods.they have concluded that ANN models would improve the model performance making it more essential in flood forecasting and prediction [11]. In a research conducted by Tsakiri et al., their research states that it introduces a hybrid model for forecasting river flood events with an example of the Mohawk River in New York. Time series analysis and artificial neural networks are combined for the explanation and forecasting of the daily water discharge using hydrogeological and climatic variables. For the prediction of the water discharge time series, each component has been described by applying the multiple linear regression models (MLR), and the artificial neural network (ANN) model. The MLR retains the advantage of the physical interpretation of the water discharge time series. Furthermore, the researchers were able to establish the necessity of time series decomposition before using any model. A comparison of the models shows that using ANN to forecast flood events on decomposed time series increases forecast accuracy. The hybrid model, which combines artificial neural networks and time series decomposition, can forecast up to 96 percent of the explanation for the time series of water outflow [**?**].

In the research of Falah et al., Artificial Neural Networks were used for flood susceptibility mapping in data-Scarce Urban Areas. The factors for the flood prediction include elevation, slope, drainage use, and the flood inventory, the results stated that the approach the researchers made provided a clear presentation of the areas susceptible to flooding. This makes the Artificial Neural Network good for flood susceptibility mapping [3].

3.1.3 Support Vector Machine. Support vector machine (SVM) is a highly favored technique in flood modeling, operating as a supervised learning machine that uses statistical learning theory and the structural risk minimization rule. Its training algorithm constructs models that allocate new non-probabilistic binary linear classifiers, which reduce the empirical classification error while increasing the geometric margin through inverse problem-solving. By training on past data, SVM can predict a quantity ahead in time. Furthermore, the SVM has been expanded as a regression tool, called support vector regression (SVR), over the last twenty years.

In the research of Yan et al., a flash flood forecast was developed in urban areas in China specifically the Jinlong River Basin in Hangzhou. Support vector machine and numerical simulations were used in forecasting. 77 of the rainfall events they collected were used for training the models while 33 of the events were used for testing. In their results, the computation of the support vectormodels only took 3.07 milliseconds while the numerical simulations took 53 hours to complete. In addition, both of the models accuracy was over 90 percent. This means that the support vector machine was more efficient and effective in the forecasting of floods [**?**].

3.1.4 Other Algorithms. Similarly, in the research conducted by Sella Nevo et al., which was "Flood Forecasting with Machine Learning Models in an Operational Framework" they have utilized Google's flood forecasting system which is used to deliver accurate flood signals to selected agencies. The forecasting system consists of a particularly alert system to warn citizens of potential floods, data validation, and machine learning which was a thresholding model and manifold model. In their results, the system delivered one hundred million alarm systems to the citizens who were in areas commonly affected by flood[40]. The researchers will make use of this study to determine how machine learning models are used in flood forecasting.

Furthermore, in the research made by Xiangfu Kong et al., they have predicted the post flood mapping for road networks. The datasets features they have used are Taxi GPS Data, the flood riskareas and the precipitation data of their study area which is Shenzhen, China, The historical taxiGPS data provided the areas wherein the taxi would not pass depending on how strong the rainfallis. In their results, they have recommended the taxi GPS data for determining the roads or placesthat are flooded. This is essential in our research as it provides an opportunity for the researchersto devise a study that does not rely on hardware sensors. However, in their data, The researchershave realized that GPS data alone cannot completely determine the flooded areas as some would still pass by flooded areas even if it is flooded hence the 60 percent accuracy [6].

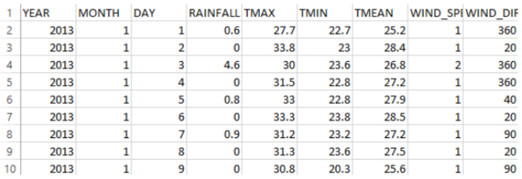
4 METHODOLOGY

4.1 Project Planning

The end goal will be to determine which is the best machine learning algorithm for predicting floods and that will provide an alternative and better flood prediction. This will be achieved by first creating a system design that will serve as the basis for setting up the data analysis, training of algorithms and report of data sets.

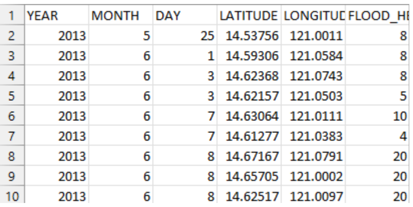
4.2 Datasets

The Philippine Atmospheric, Geophysical, and Astronomical Services Administration (PAGASA) will be utilized by gathering all the rainfall data from a specific time period. The features of the dataset from PAGASA consist of the year, month, day, rainfall, maximum and minimum temperature, wind speed, wind direction, elevation, flood height, and coordinates of the locations that will be utilized in the prediction of floods. Through the help of MMDA, urban flood reports from 2013 until 2021 will be utilized in this study. Meanwhile for river flood prediction, water level data from EFCOS on years 2000-2021 will be used instead.



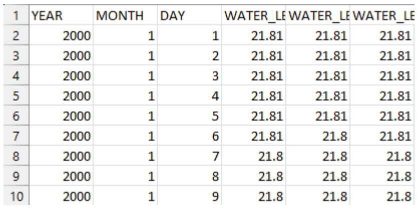
1	YEAR	MONTH	DAY	RAINFALL	TMAX	TMIN	TMEAN	WIND_SPI	WIND_DIF
2	2013	1	1	0.6	27.7	22.7	25.2	1	360
3	2013	1	2	0	33.8	23	28.4	1	20
4	2013	1	3	4.6	30	23.6	26.8	2	360
5	2013	1	4	0	31.5	22.8	27.2	1	360
6	2013	1	5	0.8	33	22.8	27.9	1	40
7	2013	1	6	0	33.3	23.8	28.5	1	20
8	2013	1	7	0.9	31.2	23.2	27.2	1	90
9	2013	1	8	0	31.3	23.6	27.5	1	20
10	2013	1	9	0	30.8	20.3	25.6	1	90

Figure 1: Dataset from PAGASA



1	YEAR	MONTH	DAY	LATITUDE	LONGITUD	FLOOD_HI
2	2013	5	25	14.53756	121.0011	8
3	2013	6	1	14.59306	121.0584	8
4	2013	6	3	14.62368	121.0743	8
5	2013	6	3	14.62157	121.0503	5
6	2013	6	7	14.63064	121.0111	10
7	2013	6	7	14.61277	121.0383	4
8	2013	6	8	14.67167	121.0791	20
9	2013	6	8	14.65705	121.0002	20
10	2013	6	8	14.62517	121.0097	20

Figure 2: Dataset from MMDA



1	YEAR	MONTH	DAY	WATER_LE	WATER_LE	WATER_LE
2	2000	1	1	21.81	21.81	21.81
3	2000	1	2	21.81	21.81	21.81
4	2000	1	3	21.81	21.81	21.81
5	2000	1	4	21.81	21.81	21.81
6	2000	1	5	21.81	21.81	21.81
7	2000	1	6	21.81	21.8	21.81
8	2000	1	7	21.8	21.8	21.8
9	2000	1	8	21.8	21.8	21.8
10	2000	1	9	21.8	21.8	21.8

Figure 3: Dataset from EFCOS

Data acquired from namria will be the Digital Elevation Map covering the entirety of Metro Manila. This map will be used to extract elevation and slope values for each flood points from the given coordinates provided by the MMDA dataset.

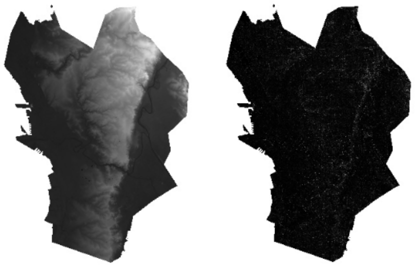


Figure 4: Data extracted from Digital Elevation Map provided by NAMRIA. Left is the elevation and right is the slope.

4.3 Data Consolidation

In order to create a consistent and cohesive format, data must be gathered from diverse systems, databases, or files. For this research, the data that needs to be consolidated are gathered from different government agencies such as EFCOS, MMDA, PAGASA, and NAMRIA. The data gathered from these government agencies will be combined into one Excel sheet in order for easier access and navigation for training and testing the model. Furthermore, the researchers kept the various features separate records in the data set. Each flood report in this situation and its corresponding coordinates and attributes would be represented as a separate entry. For this research, the data would be separated for each station in Metro Manila.

4.4 Prediction Process

In this study, we employ machine learning algorithms to forecast floods, utilizing essential datasets that encompass latitude, longitude, flood reports, and rainfall information. Latitude and longitude serve to pinpoint precise flood locations by measuring north-south and east-west distances from the equator and prime meridian, respectively. Flood reports detail the depth of floods in specific areas, while precipitation data encompasses various atmospheric conditions like rain, snow, sleet, or hail. These datasets are pivotal for predicting future flood events by analyzing rainfall patterns and flood severity. For enhanced precision, Digital Elevation Models from NAMRIA are utilized to extract elevation and slope data for the study locations, employing QGIS for this purpose. The extracted data, presented in CSV format, aligns with existing precipitation data obtained from PAGASA.

4.5 Proof of Concept

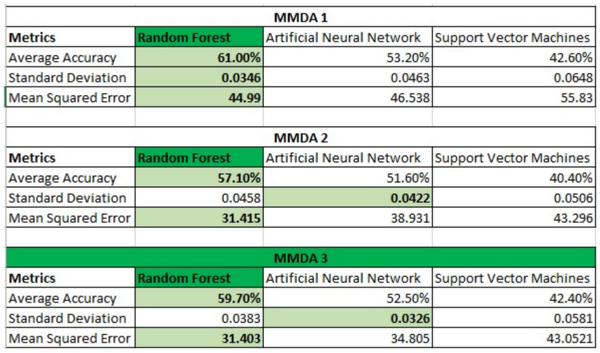
The researchers will create a simple web app that will incorporate an interface that users can easily recognize and be able to see the results of flood prediction from years 2000 until 2021. The most recent year of the data will be left out in order to test the accuracy and effectiveness of the proof of concept. For the remaining years of data, 80 percent will be used in training while 20 percent for testing the model. For the development of the web application, the basis would be the created Jupyter notebook models. the content will be implemented into a Python code such as the application of the regression models, the train test split and the standardscaler. All of these will be made into a .pkl file wherein it can be used by the streamlit application code.

5 RESULTS AND DISCUSSION

5.1 Urban Flood Predictions

5.1.1 Model Comparison for Jupyter Notebook. Three datasets have been used to train each of the three algorithms: MMDA 1: Flood Reports + 3200 (400/year) randomly selected points with no rain. MMDA 2: Flood Reports + 3200 (400/year) randomly generated points with no rain. MMDA 3: Flood Reports + 3000 randomly selected and generated points with no rain (1500 each).

Figure 5: Tabulated metrics for Urban Flood Prediction



MMDA 1

Metrics	Random Forest	Artificial Neural Network	Support Vector Machines
Average Accuracy	**61.00%**	53.20%	42.60%
Standard Deviation	**0.0346**	0.0463	0.0648
Mean Squared Error	**44.99**	46.538	55.83

MMDA 2

Metrics	Random Forest	Artificial Neural Network	Support Vector Machines
Average Accuracy	**57.10%**	51.60%	40.40%
Standard Deviation	0.0458	**0.0422**	0.0506
Mean Squared Error	**31.415**	38.931	43.296

MMDA 3

Metrics	Random Forest	Artificial Neural Network	Support Vector Machines
Average Accuracy	**59.70%**	52.50%	42.40%
Standard Deviation	0.0383	**0.0326**	0.0581
Mean Squared Error	**31.403**	34.805	43.0521

(1) **Random Forest**
Despite MMDA 1 having the highest accuracy among the three at 61%, it has a higher mean squared error compared to MMDA 2 and MMDA 3 at 44.99. Among the three, MMDA 3 is the most balanced model, with an accuracy of 59.70% and the lowest Mean Squared Error at 31.

(2) **Artificial Neural Network**
The highest accuracy among the three is from MMDA 1 however like the previous algorithm, the mean squared error is also significantly higher than the other two at 46.54 again making the model from MMDA 3 the more balanced among the three by having a 52.50% accuracy while also having a significantly lower mean squared error at 34.80

(3) **Support Vector Machines**
Models developed from this algorithm have the lowest accuracy in predicting urban floods at 40.40% to 42.60% while having a significantly higher mean squared error compared to the two algorithms at least 43.3. Despite this, the MMDA 3 dataset remains the most balanced model having a decent accuracy and maintaining a low mean squared error.

Based on the results using gathered and generated data to train three models for each dataset, Random Forest proves to be the most effective algorithm in predicting urban floods using Philippine data combined with a dataset with user flood reports from MMDA, enriched with generated rows of existing points on days with zero rainfall, and randomly generated points within Metro Manila on days with zero rainfall.

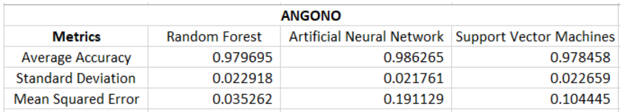
5.2 River Flood Predictions

In this section, the researchers analyze and compare the results obtained from three machine-learning models for each of nine (9) chosen rivers in Metro Manila.

5.2.1 Model Comparison for Jupyter Notebook.

(1) **Angono**

Figure 6: Model Comparison of prediction results for Angono dataset

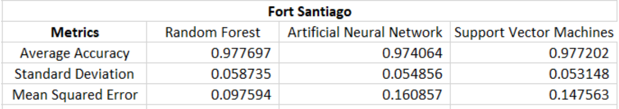


ANGONO			
Metrics	Random Forest	Artificial Neural Network	Support Vector Machines
Average Accuracy	0.979695	0.986265	0.978458
Standard Deviation	0.022918	0.021761	0.022659
Mean Squared Error	0.035262	0.191129	0.104445

The figure above compares the three models wherein it shows the Artificial Neural Network having the highest average accuracy and lowest standard deviation. In contrast, Random Forest has the lowest mean squared error.

(2) **Fort Santiago**

Figure 7: Model Comparison of prediction results for Fort Santiago dataset

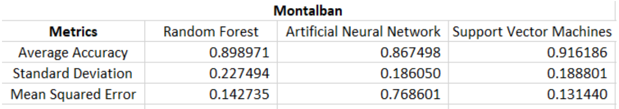


Fort Santiago			
Metrics	Random Forest	Artificial Neural Network	Support Vector Machines
Average Accuracy	0.977697	0.974064	0.977202
Standard Deviation	0.058735	0.054856	0.053148
Mean Squared Error	0.097594	0.160857	0.147563

The figure above compares the three models wherein it shows Random Forest having the highest average accuracy and the lowest mean squared error. In contrast, Support Vector Machines have the lowest standard deviation.

(3) **Montalban**

Figure 8: Model Comparison of prediction results for Montalban dataset

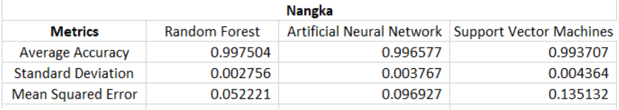


Montalban			
Metrics	Random Forest	Artificial Neural Network	Support Vector Machines
Average Accuracy	0.898971	0.867498	0.916186
Standard Deviation	0.227494	0.186050	0.188801
Mean Squared Error	0.142735	0.768601	0.131440

The figure above compares the three models wherein it shows Support Vector Machines having the highest average accuracy and the lowest mean squared error. In contrast, Artificial Neural Network has the lowest standard deviation.

(4) **Nangka**

Figure 9: Model Comparison of prediction results for Nangka dataset

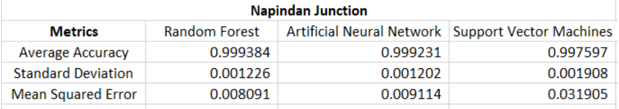


Nangka			
Metrics	Random Forest	Artificial Neural Network	Support Vector Machines
Average Accuracy	0.997504	0.996577	0.993707
Standard Deviation	0.002756	0.003767	0.004364
Mean Squared Error	0.052221	0.096927	0.135132

The figure above compares the three models wherein it shows Random Forest having the highest average accuracy, the lowest mean squared error, and the lowest standard deviation.

(5) **Napindan Junction**

Figure 10: Model Comparison of prediction results for Napindan Junction dataset

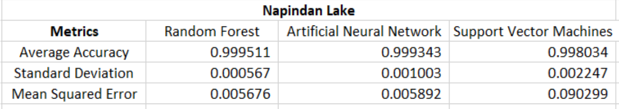


Napindan Junction			
Metrics	Random Forest	Artificial Neural Network	Support Vector Machines
Average Accuracy	0.999384	0.999231	0.997597
Standard Deviation	0.001226	0.001202	0.001908
Mean Squared Error	0.008091	0.009114	0.031905

The figure above shows the comparison of the three models wherein it shows Random having the highest average accuracy and the lowest mean squared error, and Artificial Neural Network having the lowest standard deviation.

(6) **Napindan Lake**

Figure 11: Model Comparison of prediction results for Napindan Lake dataset

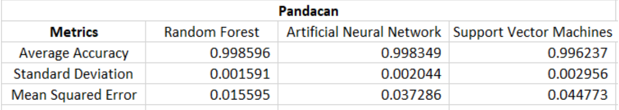


Napindan Lake			
Metrics	Random Forest	Artificial Neural Network	Support Vector Machines
Average Accuracy	0.999511	0.999343	0.998034
Standard Deviation	0.000567	0.001003	0.002247
Mean Squared Error	0.005676	0.005892	0.090299

The figure above shows the comparison of the three models wherein it shows Random Forest having the highest average accuracy, the lowest standard deviation, and the lowest mean squared error.

(7) **Pandacan**

Figure 12: Model Comparison of prediction results for Pandacan dataset

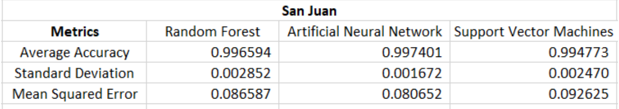


Pandacan			
Metrics	Random Forest	Artificial Neural Network	Support Vector Machines
Average Accuracy	0.998596	0.998349	0.996237
Standard Deviation	0.001591	0.002044	0.002956
Mean Squared Error	0.015595	0.037286	0.044773

The figure above compares the three models wherein it shows Random Forest having the highest average accuracy, the lowest standard deviation, and the lowest mean squared error.

(8) **San Juan**

Figure 13: Model Comparison of prediction results for San Juan dataset

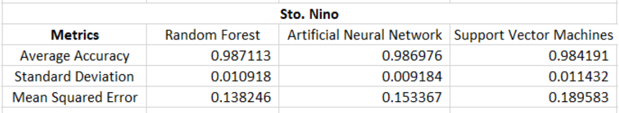


San Juan			
Metrics	Random Forest	Artificial Neural Network	Support Vector Machines
Average Accuracy	0.996594	0.997401	0.994773
Standard Deviation	0.002852	0.001672	0.002470
Mean Squared Error	0.086587	0.080652	0.092625

The figure above compares the three models wherein it shows the Artificial Neural Network having the highest average accuracy, the lowest standard deviation, and the lowest mean squared error.

(9) **Santo Niño**

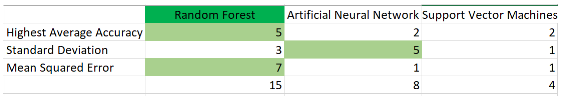
Figure 14: Model Comparison of prediction results for Santo Niño dataset



	Sto. Nino		
Metrics	Random Forest	Artificial Neural Network	Support Vector Machines
Average Accuracy	0.987113	0.986976	0.984191
Standard Deviation	0.010918	0.009184	0.011432
Mean Squared Error	0.138246	0.153367	0.189583

The figure above compares the three models wherein it shows Random Forest having the highest average accuracy and lowest mean squared error. In contrast, the Artificial Neural Network has the lowest standard deviation.

Figure 15: Tally of model comparison



	Random Forest	Artificial Neural Network	Support Vector Machines
Highest Average Accuracy	5	2	2
Standard Deviation	3	5	1
Mean Squared Error	7	1	1
	15	8	4

5.2.2 Jupyter Notebook Results. Based on the predictions for each model, results show that Random Forest is more consistent and has the best results in predicting floods when it comes to highest average accuracy and mean squared error. Artificial Neural Network is more consistent when it comes to standard deviation. Support Vector Machines, however, have been the least effective among the three upon comparing the results of the three models.

6 CONCLUSIONS AND RECOMMENDATIONS

6.1 Summary of Findings

This section presents the results and insights gathered from extensive testing of the models of three machine learning algorithms: Random Forest, Support Vector Machines, and Artificial Neural Networks in urban and river flood prediction by using data gathered from the following local government agencies: NAMRIA, EFCOS, PAGASA, and MMDA.

The findings gathered from the study are as follows: In Urban Flood Prediction, all three models displayed an average accuracy between 6140.40

Similarly, in River Flood Prediction, Random Forest also had the highest average accuracy at 99.9511

6.2 Summary of Contributions

6.2.1 Created machine learning models using the data gathered from the Philippine local government.

(1) Using the data gathered from multiple local government organizations such as NAMRIA, MMDA, PAGASA, and EFCOS, the researchers were able to consolidate the data and produce working models for the future machine learning researchers to utilize.
(2) Features include Rainfall, minimum temperature, maximum temperature, mean temperature, wind speed, wind direction, flood height, slope, and elevation.

6.2.2 Compared the effectiveness of machine learning models using Philippine data.

(1) Multiple machine learning models such as Random Forest, Support Vector machines, and Artificial Neural Network were utilized by the researchers for comparison
(2) The results also stated that the Random Forest algorithm would be the most effective machine learning algorithm to utilize when using Philippine Data.
(3) The users can also edit the Jupyter Notebook by changing the CSV file and editing the model to apply their datasets. With this, the users can also view their model's accuracy, standard deviation, and mean squared error.

6.2.3 Created a Streamlit web application for flood prediction.

(1) The process of flood prediction was made easier since just by inputting the values, the predicted values of each model would instantly show.
(2) The targeted users, the machine learning researchers, would have an easier time determining which algorithms are closer to the true value..
(3) The python file can also be edited for the users to change the Metro Manila dataset into their own gathered dataset.

6.3 Conclusions

The research compared three machine learning algorithms in predicting urban and river floods: Random Forest, Artificial Neural Networks, and Support Vector Machines. In terms of both the highest average accuracy and mean squared error, Random Forest emerged as the most effective. On the other hand, Artificial Neural Networks showed superiority in terms of standard deviation met- rics. Ultimately, the study concluded that Random Forest consistently delivered the most effective outputs.

However, it's important to note that the inadequacy of the available data hindered the effective- ness of these models. While a significant amount of EFCOS river flood data exists due to installed water level sensors, the MMDA urban flood reports, which are crucial for comprehensive predic- tions, are limited. The MMDA retained only approximately 2,000 reported flood events, depicting an insufficiency in the locally available data. This inadequacy emphasizes the need for more comprehensive and complete datasets from reliable technologies with continuous serviceability for effective flood prediction. Although the local government does provide accessible flood data for researchers, for future flood prediction and forecasting, advanced technologies and increased sensor deployment will be essential in generating extensive datasets as shown by the river flood results, leading to more accurate predictions

n conclusion, the study highlights the critical role of data quality in leveraging machine learning for flood prediction. While specific datasets proved effective, the inadequacy of data, especially in the context of urban floods, emphasizes the need for more comprehensive, reliable, and complete datasets provided by the local government in predicting floods using machine learning in the Philippines

6.4 Recommendations

The following are recommendations that the researchers have come up with:

- Increase the number of relevant hydrological and geological features
  1. Additional features such as Soil and Land Use Data, and other Hydrological features would allow the models to be more accurate in predicting urban or river floods [?].
  2. For river floods specifically, adding a binary label (0 or 1) for each row would help indicate whether a flood occurred on that day. This is also important as a rise in water levels despite having 0 rainfall could be caused by upstream reservoir/pumping station releases or a rise in tides for coastal rivers such as Fort Santiago.
  3. With the increase of relevant features such as labels for each specific river, developing a consolidated data for training and testing can be viable in creating a unified model which could have the capability to predict metrics for each river with only one code to maintain and modify.
- Add or consider more evaluation metrics
  1. The current number of evaluation metrics used for the evaluation of the models are the Mean Squared Error, Standard Deviation, and the scores on each of the folds in Cross-Validation.
  2. Adding more algorithms such as ROC curve, mean absolute error, root mean squared error and R squared would greatly help evaluate the machine learning models to determine which are the most accurate in predicting urban or river floods.
- Consider other developed regions in the Philippines as a subject for the study
  1. The urban flood data gathered for the research only contains 2000 reports while the river flood data gathered contains nine river areas of Metro Manila. Considering more developed regions of the Philippines with technologies with flood-detecting capabilities that are also flood-prone would greatly help in improving the result and scope of this research.
  2. Increasing the number of Philippine data can increase the effectiveness of each of the machine learning models. In addition, future researchers can also examine how the models work on different types of urban and river flooding in the Philippines.

7 ACKNOWLEDGEMENTS

We thank our dear panelists, Mr. Lowie Vincent Bisana, Mr. Joshua Martinez, and Mr. Adrian Pajarillo for providing us additional insights to improve our paper. To the different local government agencies in Metro Manila - MMDA, PAGASA, NAMRIA, and EFCOS for providing us with the necessary data needed for the machine learning models.

REFERENCES

[1] Jonathan Salar Cabrera and Han Soo Lee. 2018. Impacts of Climate Change on Flood-Prone Areas in Davao Oriental, Philippines. *Water* 10, 7 (2018). https://www.mdpi.com/2073-4441/10/7/893

[2] Statista Research Department. [n. d.]. Number of tropical cyclone events in the Philippines in 2021. https://www.statista.com/statistics/1091502/philippines-tropical-cyclone-events-by-category/

[3] Fatemeh Falah, Omid Rahmati, Mohammad Rostami, Ebrahim Ahmadisharaf, Ioannis N Daliakopoulos, and Hamid Reza Pourghasemi. 2019. Artificial neural networks for flood susceptibility mapping in data-scarce urban areas. In *Spatial modeling in GIS and R for Earth and Environmental Sciences*. Elsevier, 323–336.

[4] World Bank Group. [n. d.]. Philippines Vulnerability. https://climateknowledgeportal.worldbank.org/country/philippines/vulnerability

[5] Abdirahman Hashi, Abdullahi Abdirahman, Mohamed Elmi, Siti Hashi, and Octavio Romo Rodriguez. 2021. A Real-Time Flood Detection System Based on Machine Learning Algorithms with Emphasis on Deep Learning. *International Journal of Engineering Trends and Technology* 69 (05 2021), 249–256. https://doi.org/10.14445/22315381/IJETT-V69I5P232

[6] Xiangfu Kong, Jiawen Yang, Jiandong Qiu, Qin Zhang, Xunlai Chen, Mingjie Wang, and Shan Jiang. 2022. Post-event flood mapping for road networks using taxi GPS data. *Journal of Flood Risk Management* 15, 2 (2022), e12799. https://doi.org/10.1111/jfr3.12799 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jfr3.12799

[7] Ian Christopher Naungayan Rocha, Ana Carla dos Santos Costa, Zarmina Islam, Shubhika Jain, Samarth Goyal, Parvathy Mohanan, Mohammad Yasir Essar, and Shoaib Ahmad. 2021. Typhoons during the COVID-19 pandemic in the Philippines: impact of a double crises on mental health. *Disaster Medicine and Public Health Preparedness* (2021), 1–4.

[8] C. Rossi, F.S. Acerbo, K. Ylinen, I. Juga, P. Nurmi, A. Bosca, F. Tarasconi, M. Cristoforetti, and A. Alikadic. 2018. Early detection and information extraction for weather-induced floods using social media streams. *International Journal of Disaster Risk Reduction* 30 (2018), 145–157. https://doi.org/10.1016/j.ijdrr.2018.03.002 Communicating High Impact Weather: Improving warnings and decision making processes.

[9] Shu-Feng Tsao, Helen Chen, Therese Tisseverasinghe, Yang Yang, Lianghua Li, and Zahid A Butt. 2021. What social media told us in the time of COVID-19: a scoping review. *The Lancet Digital Health* 3, 3 (2021), e175–e194. https://doi.org/10.1016/S2589-7500(20)30315-0

Received 16 November 2023; revised 17 November 2023; accepted 18 November 2023