Open Peer Review on Qeios

RESEARCH ARTICLE

Flood Prediction by using Artificial Neural Network: A Case Study in Temerloh, Pahang

Ahmad Jazli Abdul Rahman[1], Nor Azuana Ramli[2]

1 Centre for Mathematical Sciences, University Malaysia Pahang, Malaysia
2 University Malaysia Pahang, Malaysia

Funding: The authors would like to thank Sistem Pengurusan Rangkaian Hidrologi Nasional (SPRHiN) Malaysia for providing related data for this study.

Potential competing interests: No potential competing interests to declare.

Abstract

Floods are natural disasters that can cause significant property damage and sometimes result in loss of life. In Malaysia, floods occur every year, particularly on the East Coast of Peninsular Malaysia, due to the Northeast Monsoon and the impacts of climate change, which can lead to heavy rainfall at the end of the year. Temerloh, a district in Pahang, frequently experiences flooding events, especially between November and January. Despite various efforts in flood mitigation and preparation, the damage to both citizens and property each year results in costs amounting to thousands of Ringgits and the time needed to clean up the aftermath of floods. To address this issue, this research examined the hydrological and meteorological factors contributing to the floods in Temerloh and developed a machine-learning model capable of predicting future flood occurrences. The study utilized a dataset from the National Hydrological Network Management System (SPRHiN), which includes hydrological data and meteorological information for the specific location. The correlation analysis revealed a strong relationship between stream flow and water level to floods, with correlation coefficients (r values) of 0.83 and 0.76, respectively. In contrast, temperature exhibited an inverse relationship with floods, showing a correlation value of -0.28; this suggests that lower temperatures are associated with a higher likelihood of rain and subsequent flooding. The results indicated that the model, developed using an artificial neural network (ANN), achieved an impressive accuracy of 0.9909 and demonstrated strong performance, as evidenced by an area under the receiver operating characteristic (ROC) curve (AUC) value of 0.888. The model also exhibited low error rates, with a mean squared error (MSE) of 0.009 and a root mean squared error (RMSE) of 0.096. Additionally, the R^2 value of 0.768 and the F1 score of 0.875 indicate that the model possesses high precision and recall. Furthermore, a flood monitoring dashboard was created to provide interactive data visualization. This research is essential for understanding the factors contributing to flooding in Pahang and will offer valuable insights for future studies on floods.

Corresponding author: Nor Azuana Ramli, azuana@umpsa.edu.my

1. Introduction

Floods are a natural disaster that has been a problem in various parts of the world. It can be defined as a dry terrain area submerged or overflowed by water due to a hydrological and meteorological state. Malaysia is no exception from this problem, immensely because Malaysia has high precipitation throughout the year, receiving 3297.34 mm of rain in 2021 (Trading Economics, n.d.). Across the East Coast of Peninsular Malaysia, the heaviest rainfall is during the Northeast Monsoon Season, from November until January. The recent flood in Pahang in 2021 left 63,394 people affected from 17,581 families, with 3500 houses losing causalities[1]. This is the worst flood that has hit the state in history. In addition, the business has been left crippled and cost millions of ringgits. Recovery to a normal state costs another thousand hours of manpower and money. This is not the first time Pahang has faced flood, but it has seen more frequent occurrences and disastrous impacts in recent years. Thus, there is a need to assess the factors that highly contribute to this misfortunate event.

Numerous reasons can contribute to flooding in an area. One of the factors is the terrain, which affects the direction and rate of the surface runoff. Flood possibility increases when there is rising in temperature, which elevates the rainfall (Ramayanti et al., 2022). In addition, the population density, land use, geographical location, and geological conditions contribute to flooding in an area[2]. Elevation also plays a vital role, according to Al-Areeq et al.[3]. Despite multiple efforts in flood mitigation, more is needed to prevent recurring events. Therefore, a new approach must be taken to reduce the severity of floods and enhance preparedness for them. Traditionally, flood prediction is done by using hydrological rainfall and runoff models. However, this modeling could be more efficient as it requires precise topography, and the data must be collected from the rain precipitation over a certain period. Recent developments in technology have introduced a few techniques that improve flood prediction. One of the developments is a physical-based model that is highly effective in simulating possible multiple flood scenarios. However, the model requires collecting data over an extended period, and its complex prediction technique has led to the method not being preferred by many. Thus, researchers have turned to machine learning technology to help improve efforts and avoid significant losses due to floods.

One of the machine learning methods used in modeling the flood is artificial neural network (ANN) techniques, as has been applied by Kia et al.[4] in the Johor River Basin. With the help of a geographic information system (GIS), the research constructed a flood map of the area with a satisfactory comparison result between the predicted and the actual record. The other machine learning techniques that can be used are logistic regression and support vector machine (SVM). However, the modeling using these two methods does not produce results as good as ANN[5]. Producing reliable and accurate flood predictive modeling is vital in preventing the area from flooding and preparing and protecting from the worst outcome. Besides modeling, it is important to investigate the relationship between the variables through correlation analysis to know which factors significantly impact the flood. Lastly, flood monitoring is more accessible through data visualization using a dashboard so the government agency can make decisions faster.

The research aims to acquire a reliable dataset that can better understand the factors impacting the flood in Temerloh through the National Hydrological Network Management System (SPRHiN). In addition, the research aims to develop an accurate flood prediction model by using artificial neural networks and producing a user-friendly and interactive dashboard that can visualize and analyze the available dataset using Microsoft Power BI. The study is focused on physical factors that highly impact floods, including rainfall, water level, streamflow, and temperature, and the modeling is done using an Artificial Neural Network (ANN). The significant research offers an opportunity to understand the factors that affected the flood in Temerloh, Pahang. Besides, the research will benefit the state government and locals in the area so they can take precautions before the flood occurs. The study also can be used as a guide to other parties in planning the development of an area and as flood mitigation efforts. In addition, a reliable Power BI dashboard could provide insights into future studies on the flood factors and flood risk in another area in Pahang. Lastly, the study will benefit academicians as it will be one of the references that can be added to the list of the latest technology in predicting floods.

2. Literature Review

Floods are disasters that can significantly cost human beings. There are four categories of floods: flash, urban, river, and coastal. In Malaysia, the most common floods are flash floods and monsoon floods. Several factors can contribute to flood, including slope, altitude, and topography. Besides, floods in Malaysia are caused mainly by prolonged heavy rain and poor urbanization planning. To reduce the impact of floods on society and property, a functioning flood management system needs to be established. There are four stages of flood management: flood prevention, preparedness, response, and recovery. In assisting flood management, a few technologies are beneficial and efficient, such as mobile phone short message system (SMS), information and communication technology (ICT), and geographic information systems (GIS).

Machine learning techniques and data mining have been used to prepare for the flood for accurate and reliable flood prediction. Only significant factors must be selected to produce accurate flood predictions using machine learning. From the literature review, it is observed that there are gaps in flood analysis and Flood Susceptibility Map (FSM) in Temerloh, Pahang. Therefore, this paper will address the gap by conducting a detailed area analysis. After reviewing multiple research projects, eight relevant papers have been selected for the study. Based on previous research, the best machine learning technique and relevant flood factors can be utilized. The summary of selected papers that are significant to this study is presented in Table 1.

Table 1. Significant papers on flood factors and machine learning techniques used.

Research Title/ Author/ Year	Flood factors	Machine Learning technique(s)	Results
Flash Flood Prediction in Selangor Using Data Mining Techniques [6]	6 factors: rainfall, water level, weather, durations, maximum temperature, and minimum temperature	Logistic Regression (LR) and Artificial Neural Network (ANN)	Each technique has excellent F-measure, but LR has slightly better result of 0.997 while ANN has value of 0.989. LR also has better AUC of 0.985, and ANN has AUC of 0.975.
An Artificial Neural Network Model for Flood Simulation using GIS: Johor River Basin, Malaysia	7 factors: land use, rainfall, slope, elevation, flow accumulation, soil, and geology	Artificial Neural Network (ANN)	Sensitivity analysis showed elevation has the highest weight in flood with R^2 value of 0.931, with slope and land use is the subsequent important factors with R^2 value of 0.962 and 0.986 respectively

[4]			respectively.
Application of GIS and Machine Learning to Predict Flood Areas in Nigeria [2]	15 factors: soil type, aspect ratio, elevation, roughness, distance to the road, water and rail, curvature, curve number, slope, stream power index (SPI), topographic wetness index (TWI), land cover, and temperature	Artificial Neural Network (ANN) and Logistic Regression (LR)	The validation result revealed that ANN has better AUC accuracy (0.764) compared to LR (0.625). In addition, ANN has better performance success of 0.964 compared to LR (0.677). The outcome also able to determine curve number, land use, SPI and aspect as the most important factors.
Deep Neural Network Classifier for Flash Flood Susceptibility [5]	Four factors: slope, aspect, elevation, and curvature	Logistic Regression (LR), Support Vector Machine (SVM) and Deep Learning Artificial Neural Network (DL-ANN)	The best model observed is DL-ANN with value of accuracy, precision, and AUC of 0.8523,0.9459 and 0.8727 respectively. LR is the second-best model with accuracy of 0.75, precision of 0.9 and AUC of 0.7893. SVM has the lowest performance but after improved with Grid Search, it able to achieve accuracy score of 0.8068, with precision (0.853) and AUC (0.8090).
Performance Comparison of Two Deep Learning Models for Flood Susceptibility Map in Beira Area, Mozambique (Ramayanti et al., 2022)	10 factors: slope, plan curvature, profile curvature, depth of the valley, distance of the river, aspect, altitude, topographic wetness index (TWI), slope length, and land use	Group Method of Data Handling (GMDH) and Convolutional Neural Network (CNN)	The highest flood prone area is the area with lower slope, lower altitude and near the river. Better AUC (0.90) and RMSE (0.022) showed that CNN is the better model compared to GMDH with AUC of 0.87 and RMSE of 0.089.
Computational Machine Learning Approach for Flood Susceptibility Assessment Integrated with Remote Sensing and GIS Techniques from Jeddah, Saudi Arabia [3]	14 factors: slope, elevation, topography, lithology, aspect, land cover, land use, stream power index (SPI), plan curvature (PC), distance river, convergence index (CI), flow accumulation (FA), soils, and precipitation.	Bagging Ensemble (BE), Logistic Model Tree (LT), Kernel Vector Support Machine (k-SVM), and K-Nearest Neighbour (KNN)	The study divides the factors into two combinations, C1 and C2. The most impactful factors are topographic Wetness Index (TWI) and distance river (DR). BE is the best model with AUC of 0.97 for C1 and 0.83 for C2, followed by LT (0.97 for C1, 0.8 for C2), k-SVM (0.93 for C1 and 0.75 for C2) and lastly is KNN (0.89 for C1 and 0.65 for C2).
Flash Flood Susceptibility Mapping of Sungai Pinang Catchment using Frequency Ratio (Saleh et al., 2022)	10 factors: aspect, curvature, slope, Stream Power Index (SPI), Normalized Difference Vegetation Index (NDVI), Topographic Wetness Index (TWI), rainfall, land use/ land cover (LU/LC), distance from river, and elevation.	Frequency Ratio (FR) and Ensemble Frequency Ratio and Analytical Hierarchy Process (FR-AHP)	The outcome shows that flood is highly susceptible to occur in the convex and urban area with lower elevation and low slope angle. In addition, FR (0.8833) has better accuracy than FR-AHP (0.8562).
Spatial Prediction of Flood in Kuala Lumpur City of Malaysia using Logistic Regression (Tella et al., 2022)	10 factors: slope, altitude, drainage density, distance to river, rainfall, NDVI, TWI, NDWI, MNDWI and LULC.	Logistic Regression (LR)	The results highlighted that the most critical factors are distance to river, MNDWI, TWI and LULC. The LR model produced 0.84 accuracy, 0.91 precision, 0.72 recall, and 0.80 F1-score.

3. Material and Methods

A good research procedure needs to be established to produce excellent results. Figure 1 shows the operational process flow for the research. Collecting data is the first step in the research procedure. It is extremely important to ensure that the data obtained is relatable, appropriate, and of high relational integrity to achieve the research objectives.

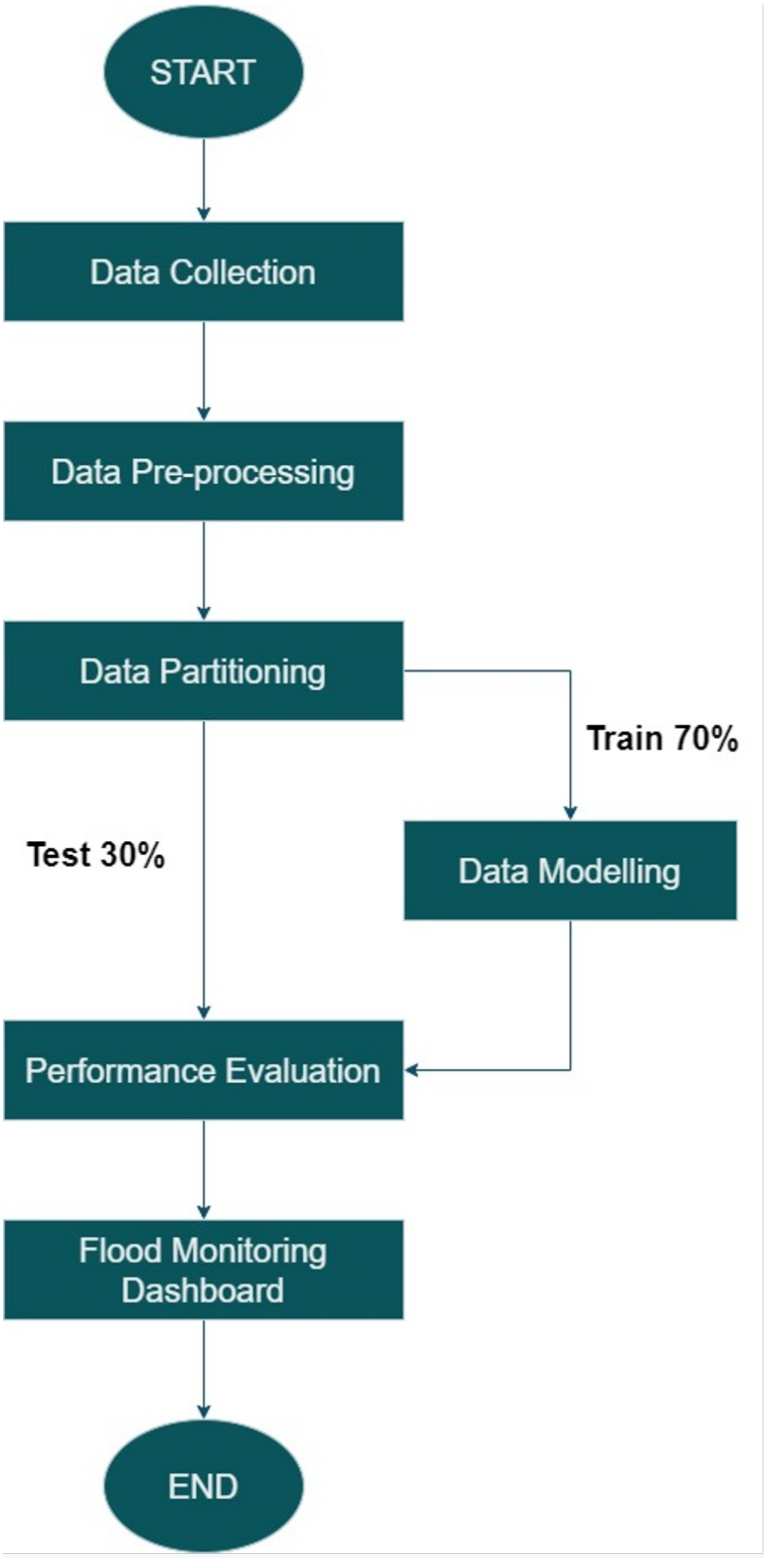


Figure 1. Research Framework Flowchart

Before the dataset can be applied to the machine learning model, it must go through the pre-processing stage to ensure the data quality. Next, the data is divided into a 70% train set and a 30% test set to be fed into the Artificial Neural

Network (ANN) model, where machine learning and prediction are developed. The performance of the model is evaluated through four evaluations, which are the confusion matrix, area under the Receiver Operating Characteristics (ROC) curve (AUC), mean squared error (MSE), and root-mean-squared error (RMSE). Finally, an interactive Flood Monitoring Dashboard is generated for data exploration and visualization.

3.1. Data Preprocessing

For this research, four data types, rainfall, streamflow, water level, and temperature, were acquired from two different sources. Rainfall, streamflow, and water level data were requested from the National Hydrological Network Management System (SPRHiN), and the temperature data was extracted from the Weather Underground (wunderground.com) website. Three data pre-processing methods were applied in this research. Firstly, data transformation was done to change the data's value, format, or structure into more meaningful and valuable data, including a data encoder for string data and data formatting from Fahrenheit to Celsius. Data integration was done manually to combine all ten datasets into one, which eases the process of understanding and evaluating the data. In addition, data cleaning was done to deal with missing, incorrect, duplicated, irrelevant, and improperly formatted. This includes a linear interpolation technique that dictates the value of a function of any intermediate points functions.

3.2. Model Development

A neural network is one of the machine learning models, and it is a subset of deep learning that mimics how a human brain works. An artificial neural network (ANN) is a concept that simulates how the input data is transferred and processed to conclude the output. The neural network works by determining the underlying pattern of the data and subsequently learning to improve the model. For this research, activation functions at the hidden and output layers segregate the important data, suppress irrelevant information, and help pass only relevant information to the next layer. Learning rate is another method to be applied in the model to minimize the differences between the predicted and actual output. The machine learning model performance developed in this study was evaluated through four evaluations, which are the confusion matrix, the area under the Receiver Operating Characteristics (ROC) curve (AUC), mean squared error (MSE), and root-mean-squared error (RMSE). The performance score needs a good outcome based on four criteria: accuracy, recall, precision, and F1 score. Accuracy evaluates the number of correct predictions compared to the total prediction. Recall or sensitivity is the capability to find the relevant information within the dataset, and precision is how the model can identify only the relevant data point. The F1 score is the best combination of precision and recall.

4. Results and Discussion

Since the research focuses on the recent big flood that hit Pahang in 2021-2022, the data for all four attributes were taken from 1 January 2021 to 31 December 2022. There are ten separate datasets, and each dataset has 13 columns, and 32 rows embody the data collected by day and clustered by month. These data were pre-processed through data transformation, integration, and cleaning. From the output, it is noticed that there are 15 missing data in the “Rainfall” columns, 34 data in the “Water Level,” and 14 data in each of the “Stream Flow,” “Weather,” and “Flood” columns. An irrelevant row was removed, and individual missing data was replaced with new values through linear interpolation and null values. To find the strongest factor that contributes to flood, correlation analysis was done to evaluate the relationship between the factors and the flood occurrences, as observed in Figure 2. Stream flow and water level have a very strong relationship with flood, with values of 0.83 and 0.76, respectively. However, weather has an inverse relationship with floods. This is understandable as the lower temperature is highly prone to rainy days.

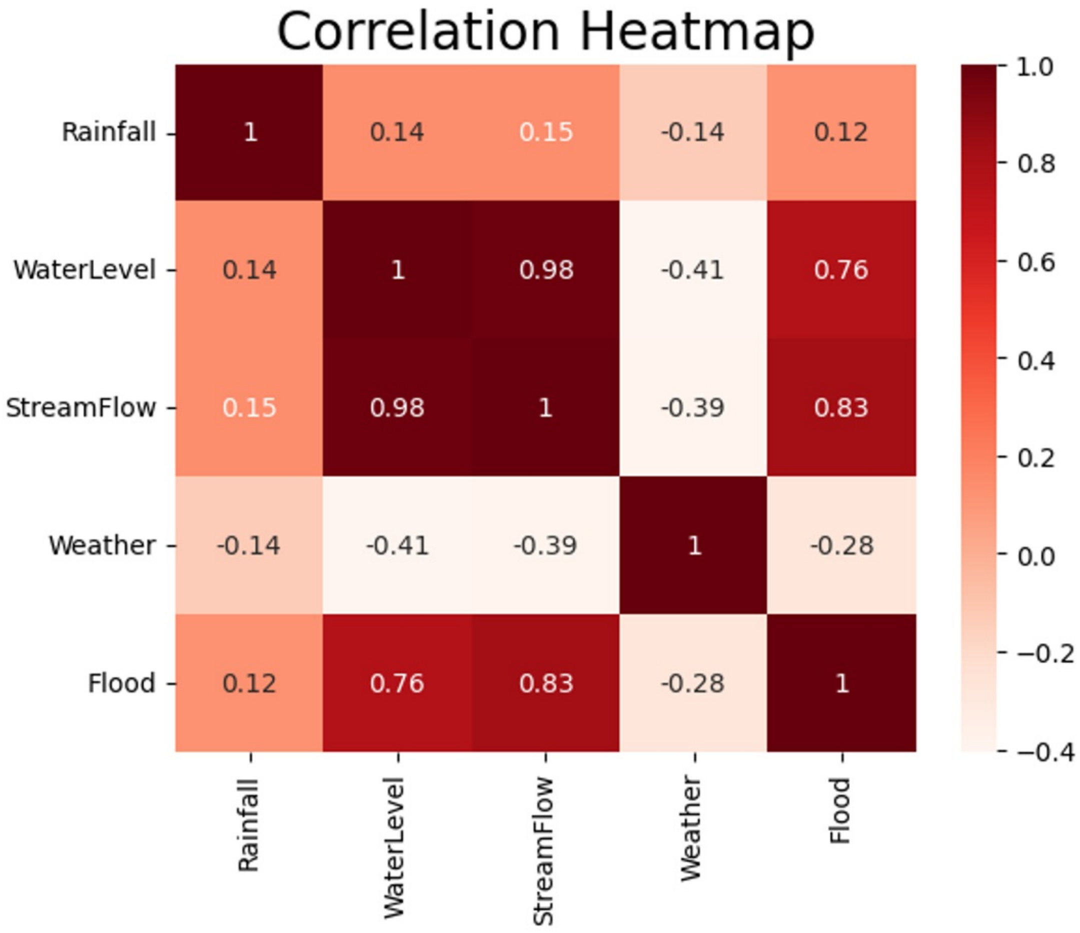


Figure 2. Correlation Analysis

Afterward, the model development started by assigning the input and output variables. The input of rainfall, water level, streamflow, and weather attributes are the input, while flood is the target variable for the experiment. Since the dataset is relatively small, the Holdout method was used to avoid overfitting, where the data was randomly split into training and testing sets numerous times. Next, data scaling was done to avoid data with a big magnitude dominating the calculation, and eventually, the result would dismiss the data with a small magnitude.

The next step is to develop an ANN structure in Python programming. The neural network was constructed to have one input layer, two hidden layers with six neurons in each layer, and one output layer to balance the model complexity and the processing time and required machine capacity. A learning curve was plotted to observe the performance of the training and validation by evaluating the loss in both sets. The output of the training and validation loss calculation can be observed in Figure 3, indicating that the model fits the training data well. The final step for machine learning modeling is prediction.

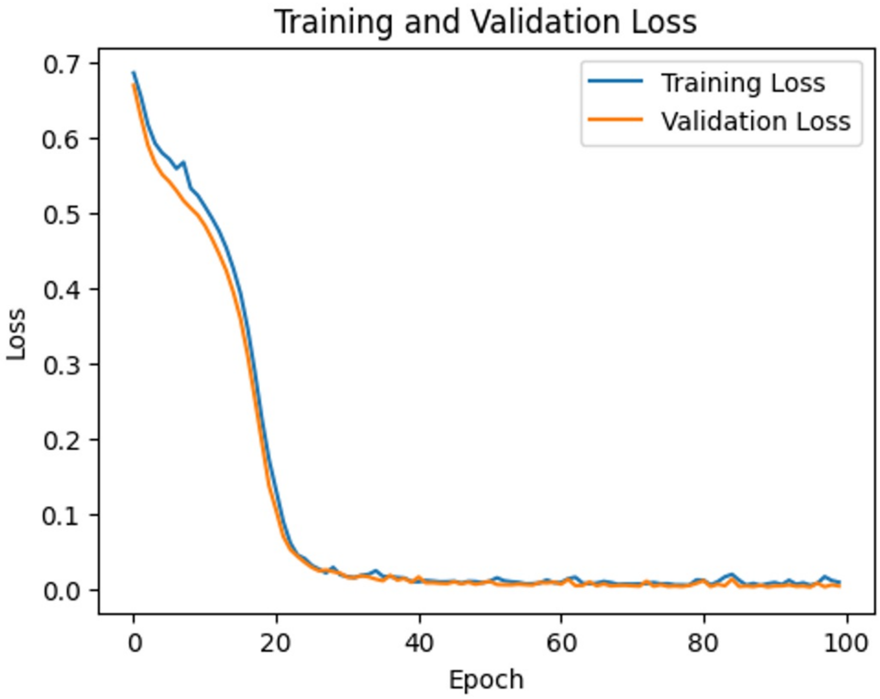


Figure 3. Learning curve of the ANN model fitted into the training data

The result is validated through a few evaluations to evaluate the performance of the data. The first evaluation is the confusion matrix, which evaluates the accuracy of the data prediction. From Figure 4, most predictions are accurate, whereas 210 “No Flood” and 4 “Flood” data are correctly predicted. There are only 5 instances that the “Flood” data predicted as “No Flood” and no event that the model predicted “No Flood” as “Flood.” The model's accuracy is calculated at 0.9909, which is very high.

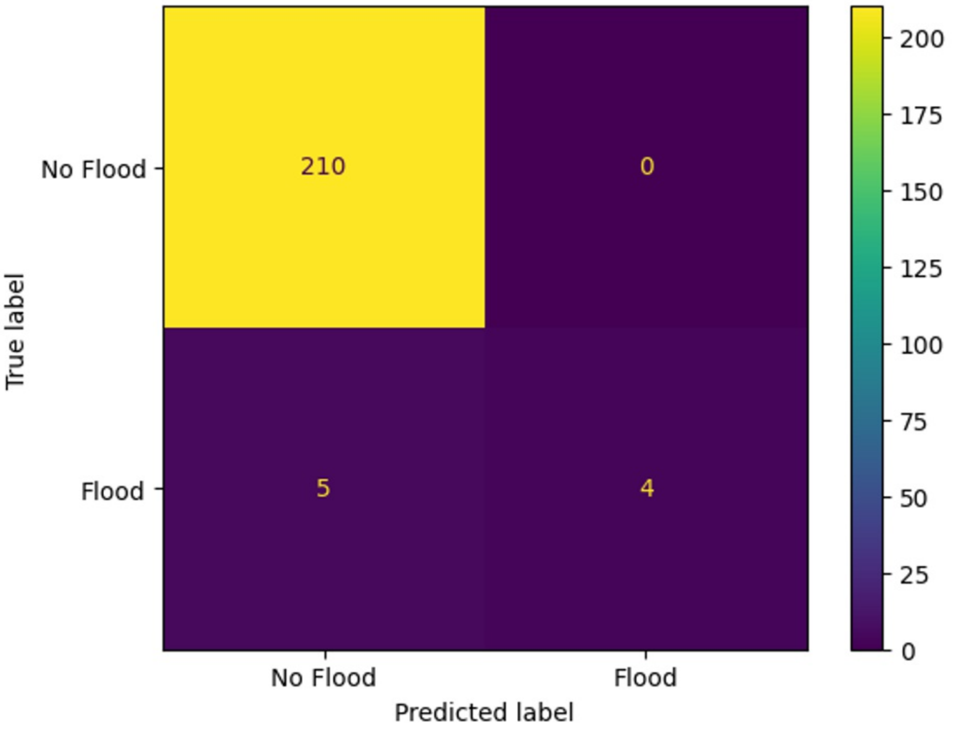


Figure 4. Confusion Matrix

Next, the AUC evaluation is executed. From Figure 5, the classifier's performance is considered good or nearly excellent, with a value of 0.888.

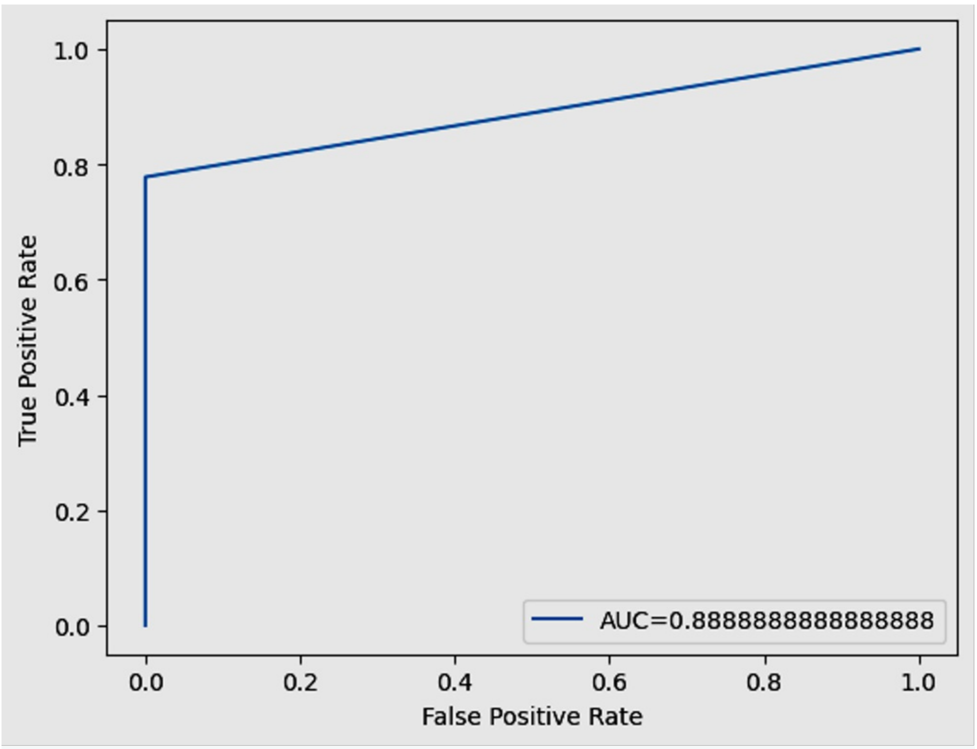
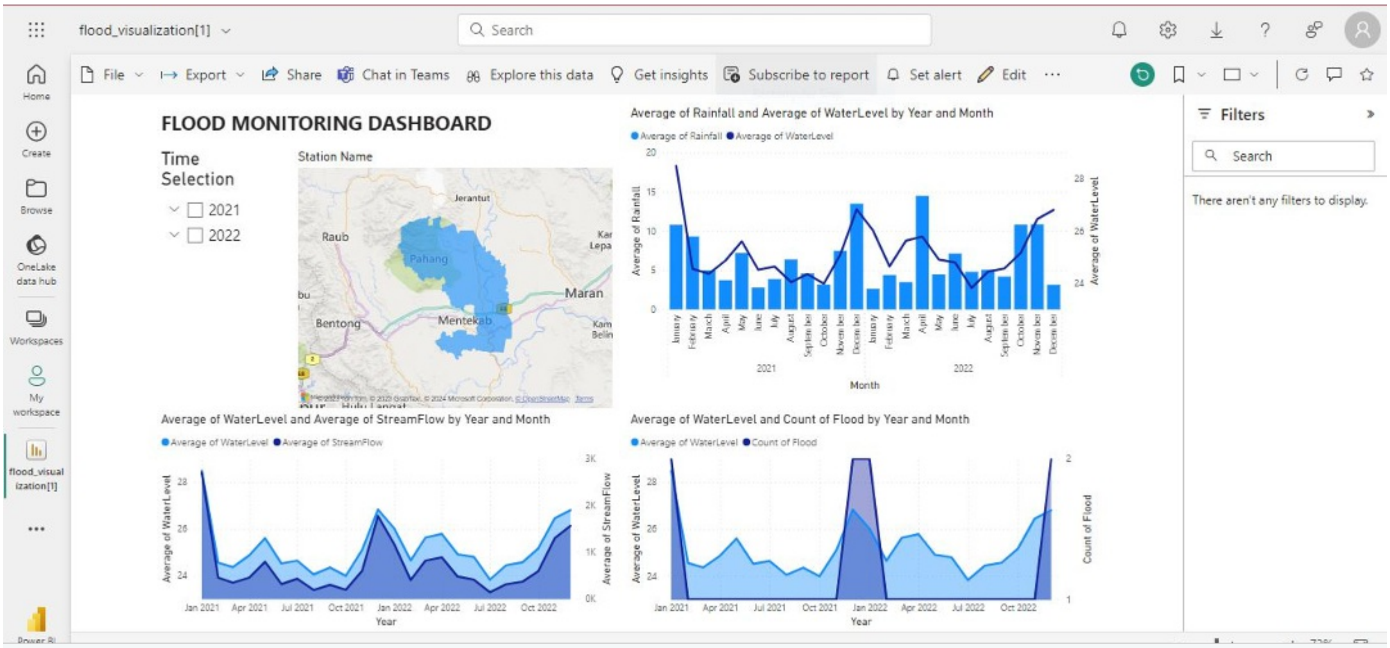


Figure 5. Area under the Receiver Operating Characteristics curve (AUC)

Lastly, performance and error evaluation are done through MAE, MSE, and RMSE, which have 0.009, 0.009, and 0.096 values, respectively, which indicates the prediction's error is very low. An R2 value of 0.768 proves there is a high variance relationship between the variables, and 76.8% of the observed variation can be explained by the model's inputs. Also, an F1 value of 0.875 shows that the prediction has strong precision and recall.

To gain more insight into the data obtained, the Flood Monitoring Dashboard was created for data exploration and visualization. The dashboard was first configured in Power BI Desktop before it was published to Power BI online. The interactive dashboard consists of 4 visualizations, including 1 map and 3 graphs that can be filtered by year, quarter, month, and day using the slicer.



There aren't any filters to display.

Figure 6. Flood Monitoring Dashboard

5. Conclusion

The research has taken the initiative to develop a machine learning model using an artificial neural network (ANN) approach with 0.9909 accuracy. The confusion matrix and area under the Receiver Operating Characteristics curve (AUC) are produced to validate the accuracy result. The result evaluation also found that the prediction has a deficient error with an MSE of 0.009 and RMSE of 0.096 but has a high sensitivity with an R2 value of 0.768 and F1 value of 0.875, indicating that the prediction has strong precision and recall. The study can also determine the factors contributing to floods through correlation analysis, which shows that floods are highly impacted by stream flow (0.83) and water level (0.76). The rainfall has a weak relation with flood (0.12), while temperature has an inverse relationship with the flood, which indicates that the lower the temperature, the higher the chance of flood. Flood Monitoring Dashboard has been developed using Microsoft Power BI, which provides exciting and interactive data visualization that helps to relate the factors.

The research is important for the authorities to be able to take action in an area that is highly prone to flooding, and it can be used as a guide for other parties in development planning in the future. The machine learning modelling used in this research is expected to assist the academician in future studies on floods in Pahang and worldwide. It is recommended that the research be expanded to other districts and states in Malaysia in order to produce a nationwide Flood Susceptibility Map (FSM). The time and data availability limitation has restricted the research scope, but it acts as the first step towards broader flood mapping. Evaluation using multiple models could help compare the results better. A more comprehensive range of data can be used to train and test the model, which will increase accuracy but will take longer to execute. In addition, it is recommended that the modelling be revisited every year as the condition of the location will change from time to time.

References

1. ^*Department of Statistics Malaysia (2021). Bencana Banjir Temerloh: Fakta dan Angka (DOSM/BPPD/5.2021/Siri 57). Retrieved from https://www.dosm.gov.my/v1/uploads/files/6_Newsletter/Newsletter%202021/DOSM_BPPD_2_2021_Series54.pdf*
2. a, b*Ighile EH, Shirakawa H, Tanikawa H (2022). "A Study on the Application of GIS and Machine Learning to Predict Flood Areas in Nigeria." Sustainability (Switzerland). 14(9). doi:10.3390/su14095039.*
3. a, b*Al-Areeq AM, Abba SI, Yassin MA, Benaaf M, Ghaleb M, Aljundi IH (2022). "Computational Machine Learning Approach for Flood Susceptibility Assessment Integrated with Remote Sensing and GIS Techniques from Jeddah, Saudi Arabia." Remote Sensing. 14(21). doi:10.3390/rs14215515.*
4. a, b*Kia MB, Pirasteh S, Pradhan B, Mahmud AR, Sulaiman WNA, Moradi A (2012). "An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia." Environmental Earth Sciences. 67(1): 251–264. doi:10.1007/s12665-011-1504-z.*
5. a, b*Kanwar, B. (2022). Development of flood prediction models using machine learning techniques (Doctoral dissertation, Missouri University Of Science And Technology). Retrieved from https://scholarsmine.mst.edu/doctoral_dissertations/3171*
6. ^*Halim, M. H., Wook, M., Afiza, N., Razali, M., Hasbullah, A., Erna, H., & Hamid, C. (2022). Flash Flood Prediction In Selangor Using Data Mining Techniques. In Journal of Defence Science, Engineering & Technology (Vol. 5).*