Open Peer Review on Qeios

Flood Prediction Using Artificial Neural Networks: A Case Study in Temerloh, Pahang

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

1 University Malaysia Pahang

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 in Malaysia happen every year, especially in East Coast Peninsular Malaysia, due to the Northeast Monsoon and climate change, which may lead to heavy rainfall throughout the end of the year. Temerloh is one of the districts in Pahang that frequently encounters flood events, especially between November and January every year. The study used a dataset from the National Hydrological Network Management System (SPRHiN), which consists of hydrological data, and weather underground for the meteorological data in the location. The correlation analysis found that stream flow and water level are highly correlated to floods, with correlation coefficients (r values) of 0.83 and 0.76, respectively, while the temperature is inversely related to floods with a -0.28 correlation value. A lower temperature has a higher chance of rain and subsequent flooding. The results show that the model, by using an artificial neural network (ANN), has produced an accuracy of 0.9909 and a good performance of the area under the receiver operating characteristic curve (ROC) curve (AUC) at 0.888. The model also shows a low error with the mean squared error (MSE) of 0.009 and the root-mean-squared error (RMSE) of 0.096. The R2 value of 0.768 and F1 value of 0.875 indicate that the model has high precision and recall. Besides predictive modeling, a flood monitoring dashboard was created to visualize the data interactively. This research is vital in understanding the flood factors in Pahang and would offer academic insight for future research in floods.

A.J.A. Rahman, and **N.A. Ramli**

Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuh Persiaran Tun Khalil Yaakob, 26300 Kuantan, Pahang, Malaysia.

*Corresponding author: azuana@umpsa.edu.my

Keywords: Machine Learning, Artificial Neural Network, Flood in Pahang, Flood Monitoring Dashboard.

1. Introduction

Flooding is one of the natural disasters that has been a problem in various parts of the world. Floods can be defined as dry terrain areas that have been submerged or overflowed by water due to hydrological and meteorological conditions. Malaysia is no exception to 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, which is in the period from November until January. The recent flood in the state of Pahang in 2021 left 63,394 people affected from 17,581 families, with 3500 houses lost and casualties (Department of Statistics Malaysia, 2021). This is the worst flood that has hit the state in history. In addition, businesses have 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 a flood, but recent years have seen more frequent occurrences and more disastrous impacts. 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 surface runoff. Flood possibility increases when there are rises in temperature which elevate the rainfall (Ramayanti et al., 2022). In addition, the density of the population, land use, geographical location, and geological conditions contribute to flooding in an area (Ighile et al., 2022). Other than that, elevation plays a vital role as well, according to Al-Areeq et al. (2022). Despite multiple efforts in flood mitigation, they are not enough to prevent recurring events. Therefore, a new approach needs to be taken in order to reduce the flood severity and enhance the preparedness towards floods. Traditionally, flood prediction is done by using a hydrological rainfall and runoff model. However, this modelling is not very efficient as it requires precise topography, and the data need to be collected from rain precipitation over a certain period. Recent developments in technology have introduced a few techniques which improve flood prediction. One of the developments is a physical-based model that has high effectiveness in simulating possible multiple flood scenarios, but the model requires collecting data over an extended period and its complex prediction technique has led to the method not being preferable by many. Thus, researchers have turned to the technology of machine learning to help in improving the efforts and thus avoid major losses due to floods.

One of the machine learning methods that has been used in modelling floods is artificial neural network (ANN) techniques, as it was applied by Kia et al. (2010) in the Johor River Basin. With the help of a geographic information system (GIS), the research was able to construct a flood map of the area with a satisfactory comparison result between the predicted and the real record. The other machine learning techniques that can be used are logistic regression and support vector machine (SVM). However, the modelling using these two methods does not produce results as good as ANN (Kanwar, 2022). Producing a reliable and accurate flood predictive model is important in preventing the area from flooding, but also in preparing and protecting from the worst outcome. Besides modelling, it is important to investigate the relationship between the variables through correlation analysis in order to know which factors have a significant impact on the flood. Lastly, flood monitoring is easier to be done through data visualization using a dashboard so the decision making can be done faster by government agencies.

The research aims to acquire a reliable dataset that can give a better understanding of the factors that impact the flood in Temerloh through the National Hydrological Network Management System (SPRHiN). In addition, the research targets to develop an accurate flood prediction model by using artificial neural networks and subsequently producing a user-friendly and interactive dashboard that can visualize and analyse 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 modelling is done by using Artificial Neural Network (ANN). The research is significant because it offers an opportunity to understand the factors that affect 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 part of flood mitigation efforts. In addition, a reliable Power BI dashboard could provide insights for future studies on the flood factors and flood risk in other areas 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

Flooding is a disaster that can significantly affect human beings. There are four categories of floods: flash floods, urban floods, river floods, and coastal floods. In Malaysia, the most common floods are flash floods and monsoon floods. Several factors can contribute to floods, including slope, altitude, and topography. Besides, floods in Malaysia are caused mainly by prolonged heavy rain and poor urbanisation 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 the 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, 8 relevant papers have been selected for the study. From the previous research, the best machine learning technique and the relevant flood factors can be utilized for the research. The summary of selected papers that is 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 factors: rainfall, water level, weather, durations, maximum temperature, and minimum	Logistic Regression (LR) and Artificial Neural Network	Each technique has excellent F-measure, but LR has slightly better result of 0.997 while ANN has a value of 0.989. LR also has a better AUC of 0.985, and ANN has an AUC of 0.975

(Halim et al., 2022)	temperature	(ANN)	AUC of 0.965, and ANN has an AUC of 0.975.
An Artificial Neural Network Model for Flood Simulation using GIS: Johor River Basin, Malaysia (Kia et al., 2012)	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 flooding, with an R^2 value of 0.931, and slope and land use are the subsequent important factors with R^2 values of 0.962 and 0.986, respectively.
Application of GIS and Machine Learning to Predict Flood Areas in Nigeria (Ighile et al., 2022)	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 results 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 outcomes also able to determine curve number, land use, SPI, and aspect as the most important factors.
Deep Neural Network Classifier for Flash Flood Susceptibility (Kanwar et al., 2022)	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 values of accuracy, precision, and AUC of 0.8523, 0.9459, and 0.8727, respectively. LR is the second-best model with an accuracy of 0.75, precision of 0.9, and AUC of 0.7893. SVM has the lowest performance, but after improvement with Grid Search, it is able to achieve an 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 a 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 (Al-Areeq et al., 2022)	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 an 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, 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 areas 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, 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

To achieve all three objectives of this research, 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 make sure the data obtained is relatable and appropriate for the research and has a high relation as well as integrity in order to achieve the research objectives.

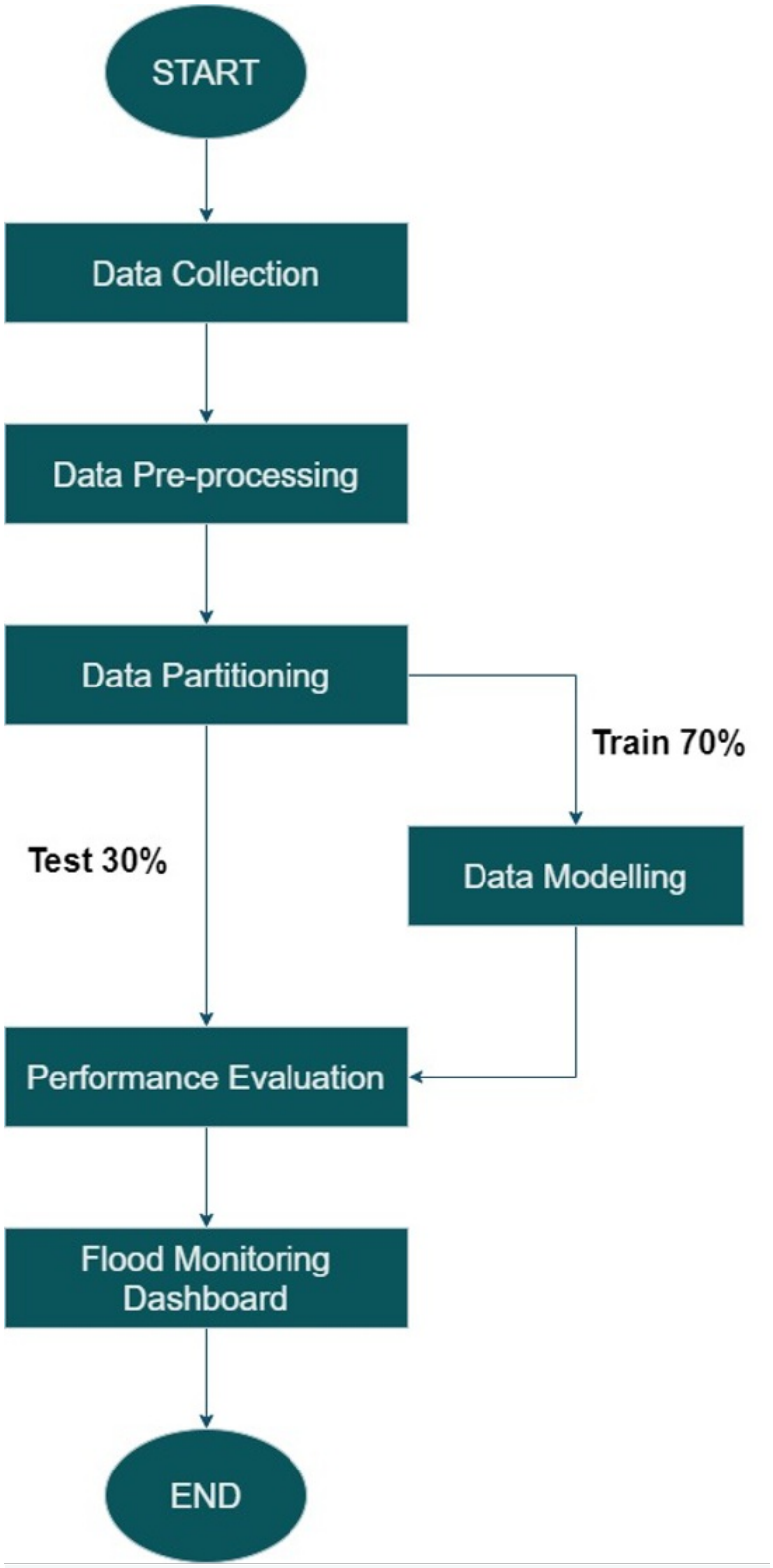

Figure 1. Research Framework Flowchart

Before the dataset can be applied to the machine learning model, it needs to go through a data pre-processing stage to ensure the quality of the data. Next, the data partitioning will divide the data into two components: a 70% train set and a 30% test set to be fed into the Artificial Neural Network (ANN) model, where the machine learning and prediction are done. The performance of the model will be evaluated through four evaluations which are 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 types of data—rainfall, streamflow, water level, and temperature data—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. There are three data pre-processing methods that were applied in this research. Firstly, data transformation was done in order to change the value, format, or structure of data into more meaningful and useful data, which includes data encoding for string data and data formatting from Fahrenheit into Celsius. Data integration was done manually to combine all 10 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 data. This includes a linear interpolation technique that dictates the value of a function at any intermediate points.

3.2. Model Development

Neural networks are one of the machine learning models, and they are a subset of deep learning that mimics how a human brain works. An artificial neural network (ANN) is a concept that simulates how input data is transferred and processed to reach a conclusion at the output. The neural network works by determining the underlying pattern of the data and subsequently learns to make the model better. For this research, there are activation functions at the hidden layer and output layer to segregate the important data, suppress irrelevant information, and help pass through only relevant information to the next layer. Another method to be applied to the model in order to minimize the differences between the predicted and actual output is the learning rate. The machine learning model performance developed in this study was evaluated through four evaluations, which are confusion matrix, area under the Receiver Operating Characteristics (ROC) curve (AUC), mean squared error (MSE), and root-mean-squared error (RMSE). The performance score needs to have good outcomes on four criteria, which are accuracy, recall, precision, and F1-score. Accuracy is to evaluate the number of correct predictions compared to the total predictions, while 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 points. The F1-score is basically 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 was taken between the date range of 1 January 2021 and 31 December 2022. There are 10 separate datasets, and each dataset has 13 columns, and 32 rows embody the data collected by days and clustered by month. These data are pre-processed through data transformation, data integration, and data cleaning. From the output, it is noticed that there are 15 missing data points in the “Rainfall” column, 34 data points in the “Water Level” column, and 14 data points in each of the “Stream

Flow," "Weather," and "Flood" columns. Irrelevant rows are removed, and individual missing data points are replaced with new values through linear interpolation and null value replacement. To find the strongest factors that contribute to floods, correlation analysis is done to evaluate the relationship between the factors and flood occurrences, as observed in Figure 2. Stream flow and water level have a very strong relationship with floods, with values of 0.83 and 0.76, respectively. However, weather has an inversely proportional relationship with floods. This is understandable as lower temperatures are highly prone to rainy days.

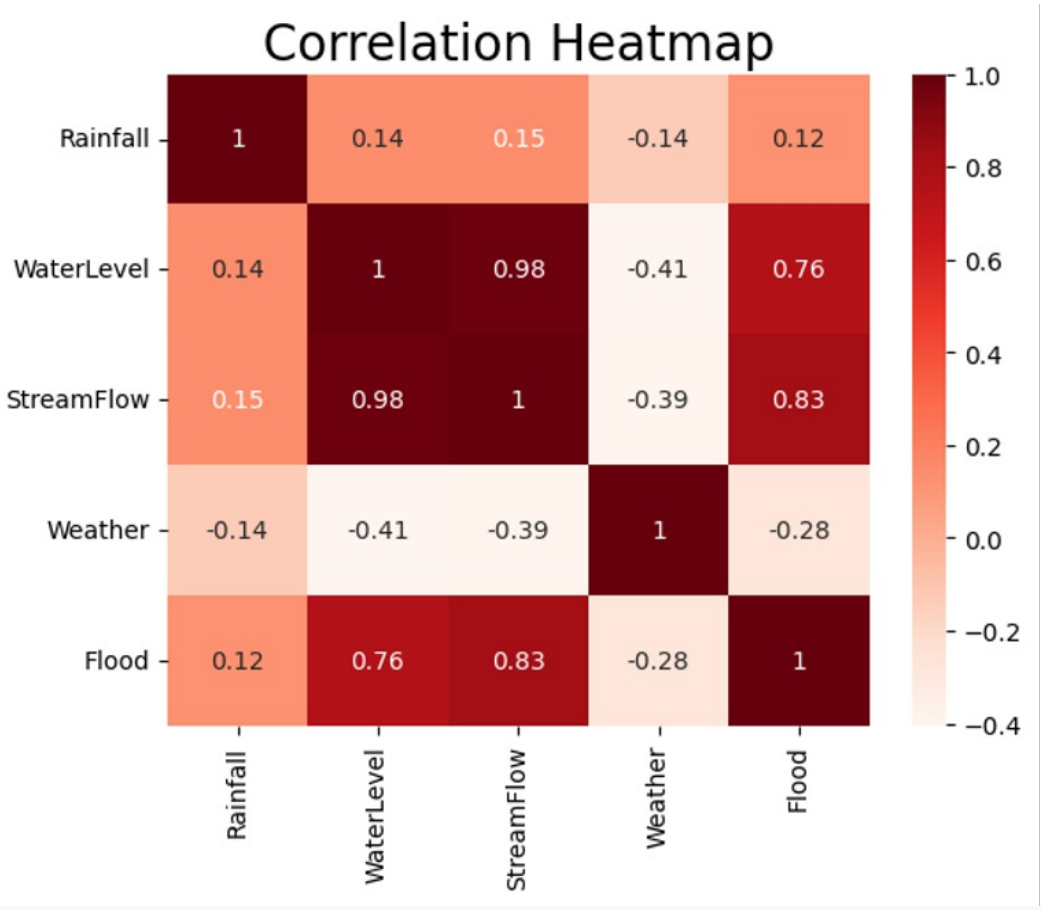

Figure 2. Correlation Analysis

Afterwards, the model development is started by assigning the input and output variables. The rainfall, water level, streamflow, and weather attributes are the inputs, while flood is the target variable for the experiment. Since the dataset is relatively small, the Holdout method is used to avoid overfitting, where the data is randomly split into training and testing sets numerous times. Next, data scaling is done to ensure that data with a big magnitude does not dominate the calculation and that the result would not dismiss data with small magnitude.

The next step is to build the ANN structure in the Python programming. The neural network is constructed to have one input layer, two hidden layers with six neurons in each layer, and one output layer to balance between the model complexity and the process time and required machine capacity. A learning curve is plotted to observe how well the training and validation performance is evaluated by examining the loss in both sets. The output of the training and

validation loss calculation can be observed in Figure 3, which indicates that the model is fitting well to the training data. The final step for the machine learning modelling is the prediction.

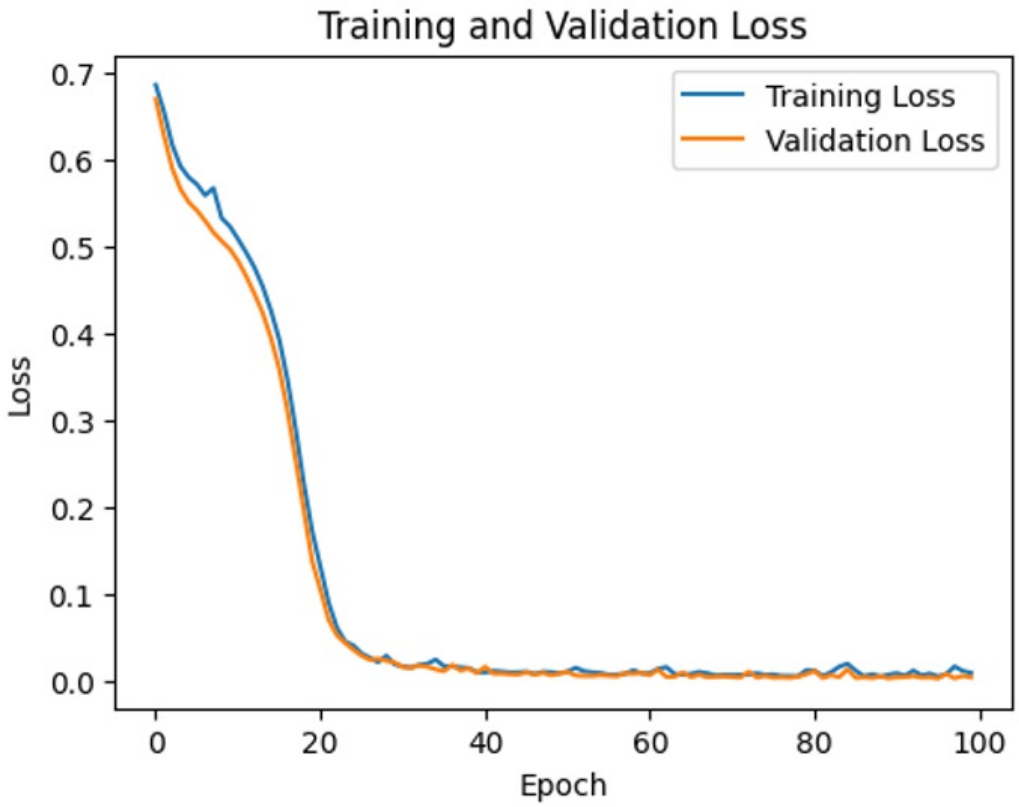

Figure 3. Learning curve of the ANN model fitted to the training data

To evaluate the performance of the data, the result is validated through a few evaluations. The first evaluation is a confusion matrix which evaluates the accuracy of the data prediction. From Figure 4, most of the predictions are done accurately, where 210 “No Flood” and 4 “Flood” data points are correctly predicted. Only 5 instances exist where the “Flood” data are predicted as “No Flood,” and no event exists where the model predicted “No Flood” as “Flood.” The accuracy of the model is calculated to be 0.9909, which is very high.

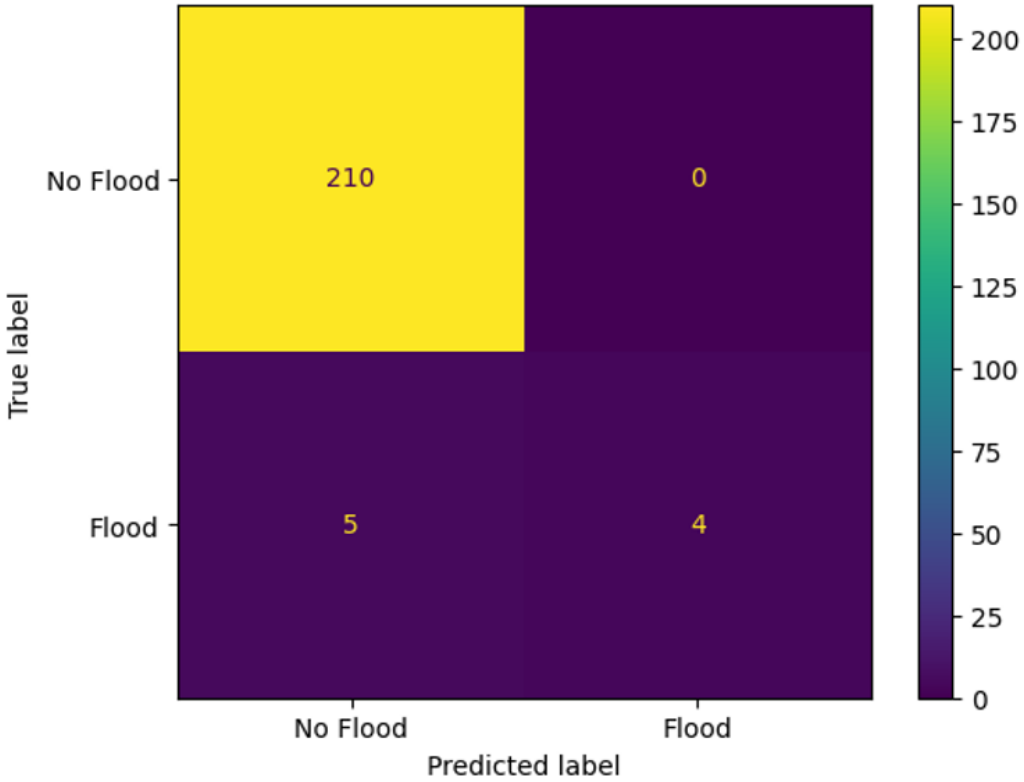

Figure 4. Confusion Matrix

Next, the AUC evaluation is executed, and from Figure 5, the performance of the classifier is considered good and nearly excellent, with a value of 0.888.

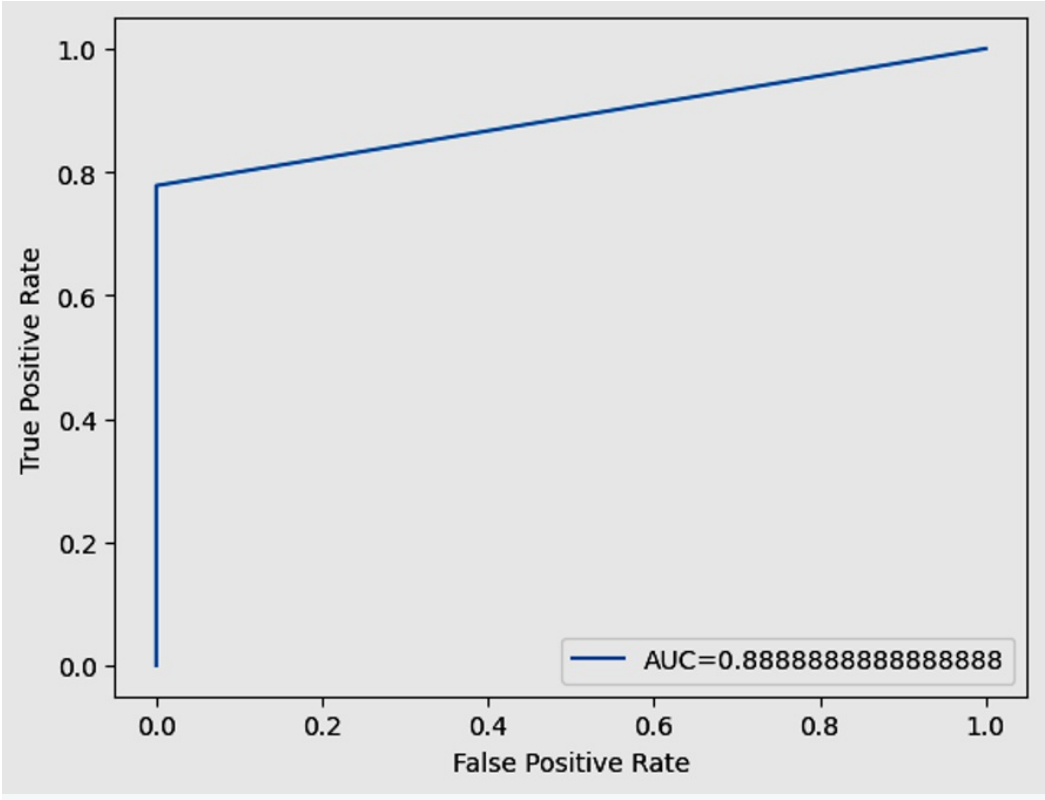

Figure 5. Area under the Receiver Operating Characteristics curve (AUC)

Lastly, the performance and error evaluation is done through MAE, MSE, RMSE, which have values of 0.009, 0.009, and 0.096, respectively, indicating the error in the prediction is very low. R^2 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. Other than that, the F1 value of 0.875 shows the prediction has strong precision and recall.

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

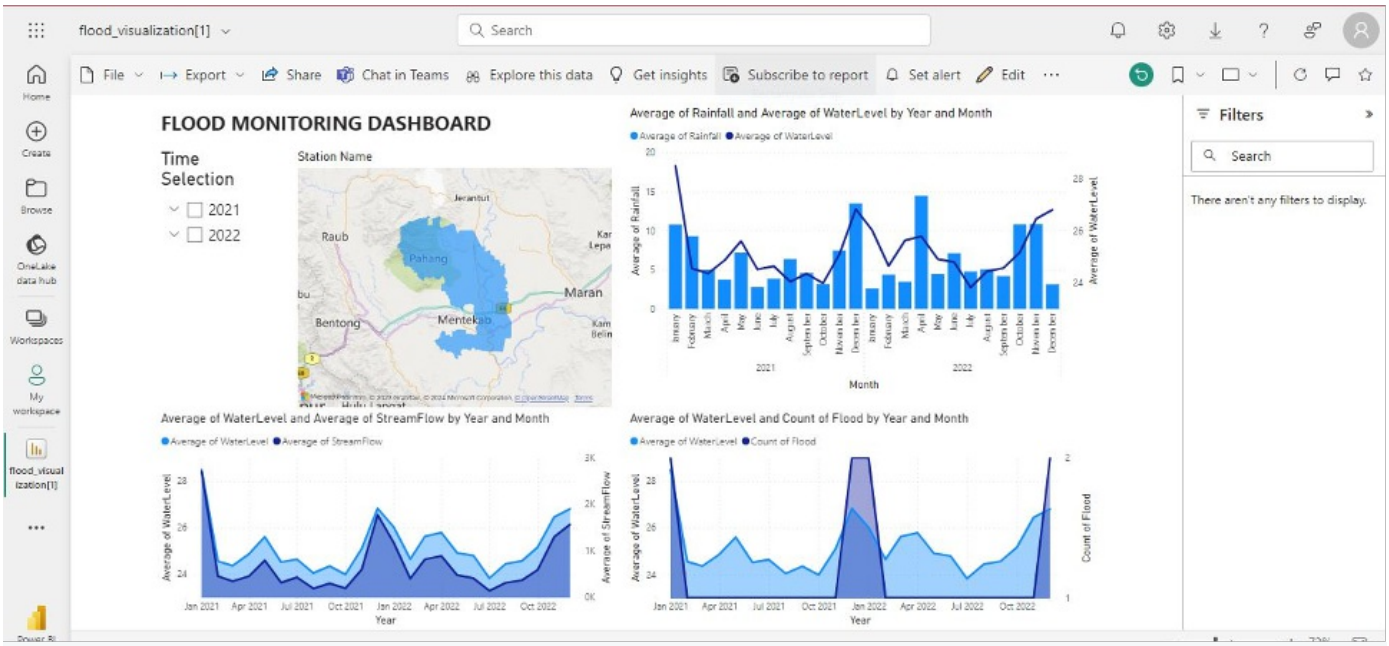

Figure 6. Flood Monitoring Dashboard

5. Conclusion

The research has taken initiative to develop a machine learning model by using an artificial neural network (ANN) approach that has 0.9909 accuracy. A confusion matrix and area under the Receiver Operating Characteristics curve (AUC) are produced to validate the accuracy result. From the result evaluation as well, it is found that the prediction has a very low error with MSE of 0.009 and RMSE of 0.096 but has a high sensitivity with R2 value of 0.768, and the F1 value of 0.875 indicates that the prediction has strong precision and recall. The study also was able to determine the factors that highly contribute to flood through correlation analysis, which shows that flood is highly impacted by stream flow (0.83) and water level (0.76). Rainfall has a weak relation to flood (0.12), while temperature has an inversely relationship with flood, which indicates the lower the temperature, the higher the chance of flood. A Flood Monitoring Dashboard has been

developed by using Microsoft Power BI, which provides interesting and interactive data visualization that helps to relate the factors to each other.

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

References

- Al-Areeq, A. M., Abba, S. I., Yassin, M. A., Benaaf, M., Ghaleb, M., & Aljundi, I. H. (2022).*Computational Machine Learning Approach for Flood Susceptibility Assessment Integrated with Remote Sensing and GIS Techniques from Jeddah, Saudi Arabia.* Remote Sensing, 14(21). https://doi.org/10.3390/rs14215515
- Ang, Kean hua & Ping, Owi. (2018). *Applied GIS in Environmental Sensitivity Development Based Slope Failure.* International Journal of Research. 5. 1286-1289.
- Baharuddin, K.A., Wahab, S.F.A., Rahman, N.H.N.A., Mohamad, N.A.N., Kamaruzaman, T.H.T., Noh, A.Y.M., & Majid, M.R.A. (2015). *The record-setting flood of 2014 in Kelantan: Challenges and recommendations from an emergency medicine perspective and why the medical campus stood dry.* Malaysian Journal of Medical Science, 22(2), 1-7.
- Baheti, P. (2021). *Activation Functions in Neural Networks [12 Types & Use Cases].*Retrieved from https://www.v7labs.com/blog/neural-networks-activation-functions
- Billa, L., Mansor, S., & Mahmud, A. R. (2004). *Spatial information technology in flood early warning systems: An overview of theory, application and latest developments in Malaysia.* Disaster Prevention and Management: An International Journal, 13(5), 356–363. https://doi.org/10.1108/09653560410568471
- Billa, L., Shattri, M., Mahmud, A. R., & Ghazali, A. H. (2006). *Comprehensive planning and the role of SDSS in flood disaster management in Malaysia.* Disaster Prevention and Management: An International Journal, 15(2), 233–240. https://doi.org/10.1108/09653560610659775
- Chee, H.L., Tan, D.T., Chan, N.W. & Zakaria, N.A. (2018). *Applying a system thinking approach to explore root cause of river pollution: A preliminary study of Pinang river in Penang state, Malaysia.* In Proceedings of the 21st IAHR-APD Congress 2018. pp. 1441-1448.
- 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

- Elias, Z., Hamin, Z., & Othman, M. B. (2013). *Sustainable Management of Flood Risks in Malaysia: Some Lessons from the Legislation in England and Wales.* Procedia - Social and Behavioral Sciences, 105, 491–497. https://doi.org/10.1016/j.sbspro.2013.11.052
- Ghapar, A. A., Yussof, S., & Bakar, A. A. (2018). *Internet of Things (IoT) Architecture for Flood Data Management.* International Journal of Future Generation Communication and Networking, 11(1), 55–62. https://doi.org/10.14257/ijfgcn.2018.11.1.06
- Goyal, H. R., & Sharma, S. (2023). *Flood Management System Using Cloud Computing and Internet-of-Things.* 2023 5th Biennial International Conference on Nascent Technologies in Engineering (ICNTE), 1–6. https://doi.org/10.1109/ICNTE56631.2023.10146661
- 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).
- Hamin, Z., Othman, M. B., & Elias, Z. (2013). *Floating on a Legislative Framework in Flood Management in Malaysia: Lessons from the United Kingdom.* Procedia - Social and Behavioral Sciences, 101, 277–283. https://doi.org/10.1016/j.sbspro.2013.07.201
- Ighile, E. H., 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). https://doi.org/10.3390/su14095039
- 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
- Kia, M. B., Pirasteh, S., Pradhan, B., Mahmud, A. R., Sulaiman, W. N. A., & Moradi, A. (2012). *An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia.* Environmental Earth Sciences, 67(1), 251–264. https://doi.org/10.1007/s12665-011-1504-z
- Malaysian Meteorological Department (n.d.). *Malaysia's Climate.* Retrieved from https://www.met.gov.my/en/pendidikan/iklim-malaysia/