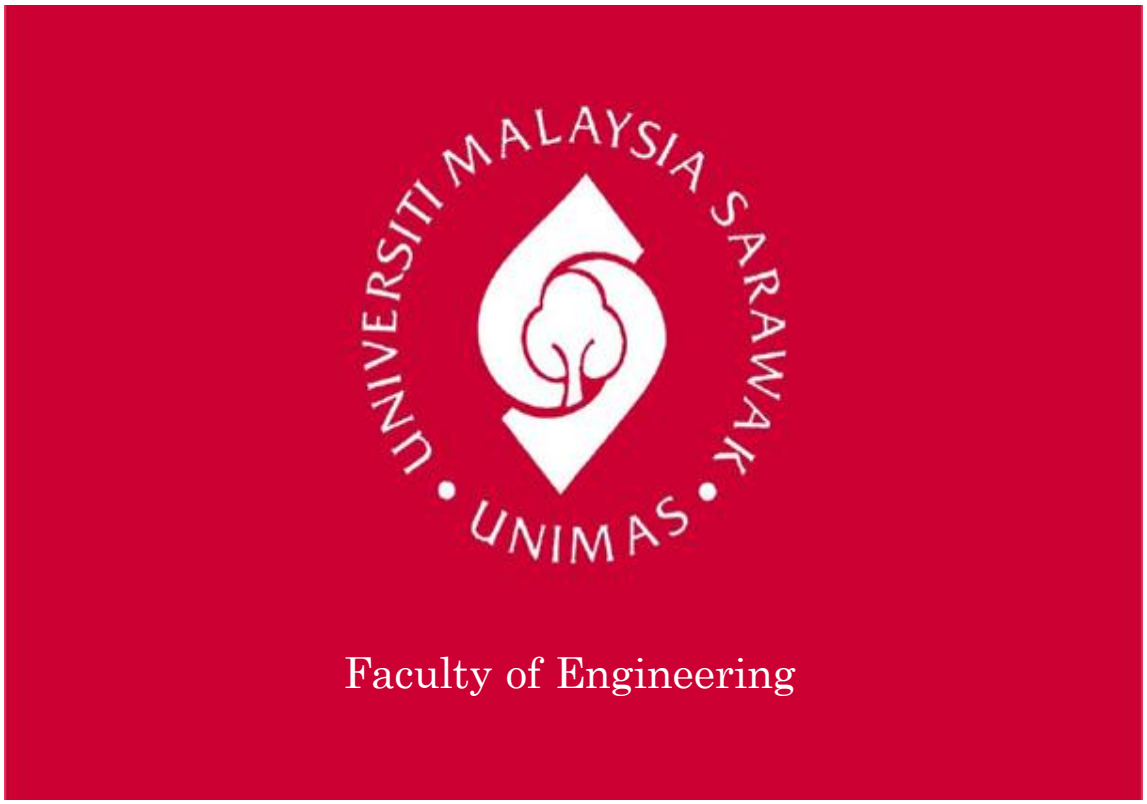

MACHINE LEARNING-BASED WEATHER FORECASTING MODEL TO PREDICT FLASH FLOOD IN MALAYSIA

Juliano Jude Jitas

Bachelor of Engineering

Electrical and Electronics Engineering with Honours

2023

UNIVERSITI MALAYSIA SARAWAK

Grade: _____

Please tick (√)

Final Year Project Report ☑

Masters ☐

PhD ☐

DECLARATION OF ORIGINAL WORK

This declaration is made on the 20 day of July 2023.

Student's Declaration:

I JULIANO JUDE JITAS, 70031, FACULTY OF ENGINEERING,

(PLEASE INDICATE STUDENT'S NAME, MATRIC NO. AND FACULTY) hereby declare that the work entitled MACHINE LEARNING-BASED WEATHER FORECASTING MODEL TO PREDICT FLASH FLOOD IN MALAYSIA is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person.

20/7/2023
Date submitted

Name of the student (Matric No.)
JULIANO JUDE JITAS (70031)

Supervisor's Declaration:

I SHIRLEY ANAK RUFUS (SUPERVISOR'S NAME) hereby certifies that the work entitled MACHINE LEARNING-BASED WEATHER FORECASTING MODEL TO PREDICT FLASH FLOOD IN MALAYSIA (TITLE) was prepared by the above named student, and was submitted to the "FACULTY" as a * partial/full fulfillment for the conferment of Bachelor of Engineering (Hons) in Electrical and Electronics (PLEASE INDICATE THE DEGREE), and the aforementioned work, to the best of my knowledge, is the said student's work.

Received for examination by: SHIRLEY ANAK RUFUS Date: 20/7/2023
(Name of the supervisor)

I declare that Project/Thesis is classified as (Please tick (√)):

☐ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*
☐ **RESTRICTED** (Contains restricted information as specified by the organisation where research was done)*
☑ **OPEN ACCESS**

Validation of Project/Thesis

I therefore duly affirm with free consent and willingly declare that this said Project/Thesis shall be placed officially in the Centre for Academic Information Services with the abiding interest and rights as follows:

- This Project/Thesis is the sole legal property of Universiti Malaysia Sarawak (UNIMAS).
- The Centre for Academic Information Services has the lawful right to make copies for the purpose of academic and research only and not for other purpose.
- The Centre for Academic Information Services has the lawful right to digitalise the content for the Local Content Database.
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis for academic exchange between Higher Learning Institute.
- No dispute or any claim shall arise from the student itself neither third party on this Project/Thesis once it becomes the sole property of UNIMAS.
- This Project/Thesis or any material, data and information related to it shall not be distributed, published or disclosed to any party by the student except with UNIMAS permission.

Student signature _____ (Date) 20/7/2023.

Supervisor signature: _____ (Date) 20/7/2023

Current Address:
Faculty of Engineering, University of Malaysia, Sarawak, Jalan Datuk Mohammad Musa, 94300 Kota Samarahan, Sarawak

Notes: * If the Project/Thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach together as annexure a letter from the organisation with the period and reasons of confidentiality and restriction.

[The instrument is duly prepared by The Centre for Academic Information Services]

MACHINE LEARNING-BASED WEATHER FORECASTING MODEL TO PREDICT FLASH FLOOD IN MALAYSIA

JULIANO JUDE JITAS

A dissertation submitted in partial fulfilment
of the requirement for the degree of
Bachelor of Engineering
Electrical and Electronics Engineering with Honours

Faculty of Engineering

Universiti Malaysia Sarawak

2023

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to the following individuals who have provided invaluable support, guidance, and assistance throughout the completion of my final year project:

First and foremost, I would like to extend my sincere appreciation to my supervisor, Madam Shirley Anak Rufus. Their unwavering commitment, expert knowledge, and insightful feedback have been instrumental in shaping the direction and success of this project. Their guidance and encouragement have been truly invaluable, and I am immensely grateful for their mentorship.

I would also like to thank the faculty, especially to Electrical and Electronic Department for their continuous support and encouragement throughout my academic journey. Their dedication to teaching and their willingness to share their expertise have been crucial in my growth as a student and researcher.

Additionally, I would like to acknowledge the assistance and contributions of my classmates and friends who have provided valuable insights, shared resources, and engaged in fruitful discussions. Their encouragement and camaraderie have made this project a collaborative and enriching experience.

Finally, I would like to express my heartfelt gratitude to my family for their unwavering support, love, and understanding throughout my academic journey. Their encouragement and belief in my abilities have been a constant source of motivation.

I am deeply indebted to all the individuals mentioned above, as well as anyone else who has contributed in any way to the successful completion of this project. Their collective efforts have greatly enriched my learning experience and shaped the outcome of this final year project.

ABSTRACT

Monsoon season especially Northeast monsoon often bring high rainfall which can leads to flood disaster in rapid city such as Klang Valley in Selangor, Malaysia. The flood disaster will damaged infrastructure, premises, agriculture and vehicles in the zone. To overcome this problem, a weather forecasting model shall be build to predict rainfall with flood level to provide a flood warning system in future. This study aims to investigate the capability of machine learning and analyze the performance of the machine learning in predicting floods. 3 years of hourly historical weather dataset is obtained from Open-Meteo website consisting of 26328 data points applied into three different algorithms, linear regression, decision tree, and gradient boosting to predict the rainfall, which can help in distinguish the possibility of flooding in the area. The performance of each models is evaluated by two metrics, $R^2\ score$ and root-mean-square-error (RMSE) to analyze how well the model can explain the variance between weather features and flood, along with how far the predicted value deviated from the actual value. The results demonstate that gradient boosting model has the best and consistent performance through the stages, achieving 0.92 of $R^2\ score$ and 0.33 of RMSE among the three algorithms. The gradient boosting model also able to predict the rainfall and flood possibility by releasing flood warning system according to the rainfall predicted.

ABSTRAK

Musim tengkujuh terutamanya monsun Timur Laut sering membawa jumlah hujan yang tinggi, di mana ia menyebabkan bencana banjir di bandar yang pesat seperti Lembah Klang di Selangor, Malaysia. Bencana banjir akan merosakkan infrastruktur, premis, pertanian dan kenderaan di zon tersebut. Untuk mengatasi masalah ini, sebuah model ramalan cuaca hendaklah dibina untuk meramalkan hujan dengan paras banjir bagi menyediakan sistem amaran banjir pada masa akan datang. Kajian ini bertujuan untuk menyiasat keupayaan pembelajaran mesin atau dikenali sebagai "Machine Learning" dan menganalisis prestasi pembelajaran mesin dalam meramal banjir. Set data cuaca sejarah untuk setiap jam selama 3 tahun diperoleh daripada laman web "Open-Meteo" yang terdiri daripada 26328 data yang digunakan dalam tiga jenis algoritma yang berbeza, iaitu "Linear Regression", "Decision Tree", dan "Gradient Boosting" untuk meramalkan hujan, yang boleh membantu dalam menentukan kebarangkalian banjir di kawasan tersebut. Prestasi setiap model dinilai dengan dua metrik, iaitu skor R^2 dan ralat akar-min-kuadrat untuk menganalisis sejauh mana model boleh menerangkan varians antara ciri-ciri cuaca dan banjir, bersama-sama dengan sejauh mana nilai ramalan menyimpang daripada nilai sebenar. Keputusan menunjukkan bahawa model "Gradient Boosting" mempunyai prestasi yang terbaik dan konsisten melalui peringkat-peringkat, dimana ia mencapai 0.92 skor R^2 dan 0.33 RMSE ketiga-tiga algoritma. Model "Gradient Boosting" juga mampu meramalkan hujan dan kebarangkalian banjir dan sekali gus mengeluarkan sistem amaran banjir mengikut ramalan hujan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANFIS	Adaptive Neuro-Fuzzy Inference Systems
ANNs	Artificial Neural Networks
DT	Decision Tree
EPSs	Ensemble Prediction Systems
LR	Logistic Eegression
NADMA	Malaysian National Disaster Management Agency
MLP	Multilayer Perceptron
SVM	Support Vector Machines
WNN	Wavelet Neural Networks

CHAPTER 1

INTRODUCTION

1.1 Background

Floods is the most common sort of natural catastrophe, happen when an excess of water submerges a normally dry ground [1]. Floods can happen due to the water overflowing at the floodplain or caused by extreme rainfall in an area [2]. In Malaysia, flooding is a regular occurrence and happens occasionally throughout the year, primarily can be categorized into two groups, flash flood and monsoon flood.

Flash floods commonly occurred due to sudden heavy rainfall in a short period of time in the area. Conventionally, flash floods usually happened in rapid cities such as Selangor, Pahang, Melaka, Negeri Sembilan, Johor, Kuala Lumpur and Kelantan [3]. Those cities often have a high level of urbanization with growing built-up regions causing the difficulty to upgrade the streams, and sewage. The accelerated growth of the cities that come with undeveloped old, existed drainage system is inefficient and no longer able to handle and direct the water from heavy rains [4]. One of the latest flash floods that occurred in rapid city of Malaysia is Petaling Jaya, Selangor on 8 November 2022 [5].

Monsoon flood in another way is cause by the heavy rain due to the monsoon season. Monsoon flood is particularly the main type of flood in Malaysia. The Southwest Monsoon, which begins in late May and lasts through September, and the Northeast Monsoon, which lasts from November to March, are the two monsoon regimes that define the weather of Malaysia [6]. While the Southwest Monsoon often displays drier weather, the Northeast Monsoon brings significant rain, particularly to the states on the east coast of Peninsular Malaysia, west of Sarawak, and east of Sabah, that causing the monsoon flood in Malaysia. In more details, the strong cold of air current (monsoon burst) caused by northwesterly wind from Siberia before switching to northeasterly wind is formed due to the temperature differential between the land and the water that heated by the sunrays,

interact with the low-pressure system and cyclonic eddies at the Equator, produced a strong wind and heavy rain.

Python is a general-purpose, high-level programming language that is widely used for a variety of purposes, including web development, data analysis, machine learning, and scientific computing. Python also known for the simplicity, readability, and flexibility, which makes it a popular choice among programmers and data scientist.

In terms of weather forecasting, Python is a useful tool because it provides a number of libraries and frameworks that can be used to process and analyse weather data. For example, Python can be used to scrape data from weather websites, visualize weather patterns using data visualization libraries like Matplotlib and Seaborn, and even build machine learning models such as linear regression model, decision trees model, and neural network model to make weather predictions through different features of the weather data. Versatility of Python and powerful data analysis capabilities become an ideal choice for weather forecasting and other related tasks.

Logistic regression (LR) is a statistical approach for assessing a dataset in which there are one or more independent variables that predict an outcome. The method can be applied in situations where the number of determining factors ranges from one to many [7]. While decision tree (DT) is a machine learning algorithm to make a prediction based on a series of decisions made based on the values of the input features [8].

Gradient Boosting is a potent machine learning technique that combines many weak models, such decision trees, to produce a robust prediction model. The method corrects faults caused by earlier models repeatedly, with each new model focused on flaws generated by the ensemble [9].

1.2 Problem Statement

In Malaysia, flash floods or monsoon floods often occurred at some part of east coast of Peninsular Malaysia, Sabah, and Sarawak [10]. According to the Malaysian National Disaster Management Agency (NADMA), the states of Johor and Selangor have the highest number of flood incidents in Malaysia. In 2021, these two states accounted for more than half of all flood incidents losses in the country [11]. Flooding is the most frequently occurring natural disaster in Malaysia, affecting an average of 300,000 people each year.

According to the document released by Department of Irrigation and Drainage, the primary causes of flooding in Malaysia are increased runoff rates brought on by urbanization, loss of flood storage brought on by development encroaching upon and occupying flood plains and drainage corridors, insufficient drainage systems or failure of localized drainage improvement works, locally persistent heavy rainfall, and tidal backwater effect [10]. Consequently, during the monsoon season that brings a high number of heavy rainfalls, monsoon floods will occur more frequently compared to non-monsoon season. The impact from the monsoon floods caused a big number of losses to Malaysia with overall losses of RM 6.1 billion, consisting of RM 2.0 billion of public assets and infrastructure damage, RM 1.6 billion worth of living quarters damage, RM 1.0 billion of vehicles damage, RM 0.5 billion and RM 90.6 million of damage cost to business premises and agriculture [11].

Back to the people, the locals were uninformed and unprepared for the floods, let alone the intensity and scope of the excessive precipitation and flooding disaster that occurred during this specific season. This is frequently caused by the circumstance where individuals are unfamiliar with extreme flooding situations [12]. The low flood risk awareness can have serious consequences. Lack of knowledge about flood risk can result in sloppy planning and preparedness, such as neglecting to have an evacuation strategy in place. Lack of knowledge about flood risk can also result in complacency, which prevents people and communities from taking the essential safeguards to save their lives and property in the case of a flood. People could be caught off guard by the quick start of floods and unable to flee or seek refuge in a safe area, which can cause catastrophic injuries or fatalities.

Following the serious flood happened in Kuala Lumpur back to 1971, the government responded by taking several proactive measures to address the issue [13]. The Permanent Flood Control Commission was established, flood disaster relief equipment was established, river basin studies were conducted, drainage master plans for major towns were prepared, structural and non-structural measures were implemented, flood forecasting and warning systems were established, and a nationwide network of hydrological and flood data collection stations were established.

The majority of the previously listed procedures are expensive and time-consuming to set up. The Malaysian government has invested more than RM 3 billion on structural flood mitigation measures since the 1970s [13]. These include both structural and non-structural techniques. However, given the rising project costs and the flooded areas, this sum is still insufficient and will continue to grow. Therefore, in order to lessen the consequences of future floods, non-structural measures like flood forecasting and warning systems are more preferred since they may assist a responsible authority in organizing an efficient emergency response to floods.

1.3 Objectives

The primary goal of this study is to develop a machine learning model that can accurately predict the flash flood during monsoon season occurring in a specific location based on various factors such as weather conditions and the previous flood data. The model will be used to achieve the following objectives:

1. To identify the relationship between monsoon season and flood.
2. To investigate the capability of machine learning in predicting floods.
3. To analyze the performance of machine learning model in predicting floods.

1.4 Project Scope

1. Develop three different machine learning-based weather forecasting model by using Python (Jupyter Lab) to predict flood in Klang, Selangor.
2. Decide on the location of study area, which is Klang, Selangor.
3. Obtain the data of weather and flood from Open-Meteo website with the time period between 2020 to 2023.
4. Apply three machine learning algorithms, Linear Regression, Decision Tree, and Gradient Boosting method.
5. Analyze how the features of the weather can cause flood conditions in Klang, Selangor.
6. Evaluate and fine-tune the machine learning-based model for precise results.
7. Comparison between three different machine learning-based weather forecasting models to examine the accuracy and validity of the model.

Chapter 2

LITERATURE REVIEW

2.1 Overview

This chapter covers the study of literature review on the development of machine learning-based weather forecasting model to predict flood in Malaysia. The study included how severe is the flood towards the community by showing the statistics of flood events with the estimated losses and deaths. To develop the model, two main study had been done, the features required to predict the flood event, and which algorithm should be applied for the model.

2.2 Flood Statistics

Floods occur everywhere across the world. In order to assess the severity of flood events, some research on flood data was conducted in both the global and Malaysian contexts.

2.2.1 Global

Floods may happen everywhere in the world, and they can vary greatly in frequency, intensity, and impact. Numerous things can result in flooding, such as a lot of rain, snowmelt, broken dams, and storm surges. The effects of floods can vary depending on a number of variables, such as the topography of the area, its level of development, and the preparedness of local community and response activities. Table 2.1 shows the statistics of major flood events that occurred globally since 2010.

Table 2.1: Global major flood statistics [13]

Year	Country	Severity	Estimated Losses	Causes
2022	Pakistan [14]	• 7.9M people affected • 1600 deaths	-	• Monsoon rainfall
2011	South-East Asia [15] • Thailand • Cambodia • Myanmar • Vietnam	-	-	• Typhoon • Monsoon rainfall
2011	Laos [15]	• 3000 deaths	-	• Typhoon
2010	Pakistan [15]	• 20 Million people affected • 2000 deaths	US$ 43 billion	• Monsoon rainfall

2.2.2 Malaysia

In Malaysia, monsoon flood is a regular occurrence, especially during the monsoon season, which normally lasts from November to March. These floods have the potential to seriously harm homes, agriculture, and infrastructure while also interfering with vital services like transportation. Numerous states and districts received the most flood reports in December 2022 during monsoon season. The affected states and districts, as well as the total number of afflicted persons, are included in Table 2.2.

Table 2.2: Area affected by monsoon flood in Malaysia in December 2022 [15]

States	District	Severity
Johor	• Segamat	• 17 families / 53 persons displaced
Kelantan	• Bachok • Jeli • Kota Bharu • Kuala Krai • Machang • Pasir Mas • Pasir Puteh • Tanah Merah • Tumpat	• 6,897 families / 25,353 persons displaced • 4 deaths
Pahang	• Kuantan	• 214 families / 873 persons displaced
Perak	• Hilir Perak	• 17 families / 54 persons displaced
Terengganu	• Besut • Dungun • Hulu Terengganu • Kemaman • Kuala Nerus • Kuala Terengganu • Marang • Setiu	• 10,629 families / 38,806 persons displaced • 1 death

In order to reduce the risks and effects of these disasters, it is crucial for the government and local communities in Malaysia to have efficient flood preparedness and response strategies in place. This may entail taking steps to construct infrastructure that is resistant to flooding, create early warning systems, and evacuate vulnerable populations before a flood.

2.3 Relationship between Flood and Weather Data

Weather information and flood occurrences are closely related. For instance, meteorologists and hydrologists can forecast the possibility of flooding in a certain location by using meteorological information like precipitation, temperature, and wind speed. With the aid of this information, flood warnings and alerts may be sent out to assist people in getting ready for and dealing with future flood occurrences. Through several studies, the most common features that the authors used to predict the flood are temperature, rainfall, humidity, and Table 2.3 shows the specific features utilized by each researcher.

Table 2.3: Data feature and time period that used to predict flood [15] - [16]

Research article	Data features	Period
[12]	• Relative humidity (%) • Temperature (°C) • Pressure (hPA) • Rain (1 or 0) • Rainfall range (0, 1, 2, 3, 4)	2015 – 2019 (5 years)
[13]	• Average rainfall (mm) • Duration of rainfall (hour) • Intensity of rainfall • Drainage system • Likelihood of flood	-
[14]	• Relative humidity (mm) • Daily Rainfall (mm) • Wind speed (meter/second) • Temperature (°C)	1980-2018 (39 years)

Most of the features taken account will be used to predict the extreme rainfall which relates to cause the flood. In [16], the author predicts the occurrence of flood directly through the data set instead of forecasting the extreme rainfall. They also took the drainage system of the study area into account because it might have an impact on the flood factor.

2.4 Machine Learning Algorithm for Flood Prediction

Recent years have seen the application of several machine learning algorithms to the problem of flood prediction, including supervised learning algorithms like decision trees and support vector machines, unsupervised learning algorithms like k-means clustering, and ensemble learning algorithms like random forests. One of the study articles provided a conclusion and comparison of several distinct flood prediction models' algorithms. The study noted that artificial neural networks (ANNs), multilayer perceptron (MLP), adaptive neuro-fuzzy inference systems (ANFIS), wavelet neural networks (WNN), support vector machines (SVM), decision trees (DT), and ensemble prediction systems (EPSs) can all be used to successfully apply flood predictions models [17]. The article demonstrated that ANNs, SVM, and DT are the algorithms that have been investigated the most in 2017 among those previously stated.

The performance of machine learning model is addressed using two metrics: root-mean-square error (RMSE) and accuracy. RMSE is a helpful statistic for assessing the effectiveness of a machine learning model due to the simplicity in interpreting and expressing in the same units as the original data [18]. A lower RMSE value reflects the performance of the model and the improved prediction accuracy. Accuracy in machine learning model is referring to the percentage of accurate predictions it makes out of all possible predictions. It is calculated by dividing the number of accurate guesses by the total number of predictions and is a widely used assessment statistic for classification assignments [18].

Machine learning algorithms come in a wide variety of forms, but they may generally be divided into two categories: supervised learning algorithms and unsupervised learning algorithms. Algorithms for supervised learning are trained using labelled data, which comprises both the input data and the accurate output labels. Using this labelled training data, the algorithm learns to map the input data to the output labels. Support vector machines (SVMs), logistic regression, decision tree and linear regression are some typical examples of supervised learning algorithms.