UNIVERSITI TEKNOLOGI MARA

SPATIO-TEMPORAL DISTRIBUTION AND LONG SHORT TERM MEMORY (LSTM) MODELS OF $PM_{2.5}/PM_{10}$ RATIOS AT SEVERAL LOCATIONS IN WEST COAST OF PENINSULAR MALAYSIA

NUR DAMIA BINTI ZAHARI

MSc

FEBRUARY 2022

UNIVERSITI TEKNOLOGI MARA

SPATIO-TEMPORAL DISTRIBUTION AND LONG SHORT TERM MEMORY (LSTM) MODELS OF $PM_{2.5}/PM_{10}$ RATIOS AT SEVERAL LOCATIONS IN WEST COAST OF PENINSULAR MALAYSIA

NUR DAMIA BINTI ZAHARI

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Applied Statistics

Faculty of Computer and Mathematical Sciences

February 2022

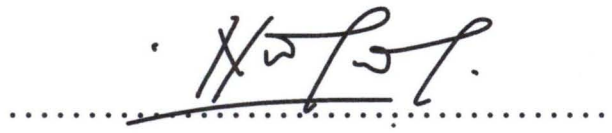
AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Nur Damia binti Zahari
Student I.D. No.	:	2020480796
Programme	:	Master of Science in Applied Statistics – CS702
Faculty	:	Computer and Mathematical Sciences
Thesis Title	:	Spatio-Temporal Distribution and Long Short Term Memory (LSTM) Models of PM2.5/PM10 Ratios at Several Locations in West Coast of Peninsular Malaysia
Signature of Student	:	.....
Date	:	February 2022

APPROVED BY:



DR. NORSHAHIDA SHAADAN

Supervisor

Faculty of Computer and Mathematical Sciences

Universiti Teknologi MARA

ABSTRACT

Air pollution has become an important issue of concern worldwide. The important characteristic of particulate matter (PM) air pollutant is the size of the PM itself. Different size of PM is caused by different sources and different duration of deposition. The size also shows different degree of negative impact towards human health with different depth of penetration to the human respiratory tract when inhaled. The ratio of $PM_{2.5}/PM_{10}$ able to describe the component of the small size particulate matter from PM_{10} that can be an indicator for the level of severity when expose to the pollutant in the environment. The knowledge is important because $PM_{2.5}$ is more dangerous than PM_{10}. However, many researchers have neglected the importance of the ratio of $PM_{2.5}$ /PM_{10} where studies on the $PM_{2.5}$ and PM_{10} ratio were only focusing on the exploratory analysis especially in Malaysia insinuating the needed of modelling the ratio $PM_{2.5}/PM_{10}$. Thus, this study aims to explore the spatio-temporal distribution of $PM_{2.5}/PM_{10}$ ratios and to propose a time series prediction model using Long Short Term Memory (LSTM) at several locations in the west coast of Peninsular Malaysia. This study utilizes daily and hourly data for the period of 2.5 years (July 2017 to December 2019) period from five air quality monitoring stations including Bukit Rambai, Klang, Manjung, Nilai and Seberang Jaya. Exploratory analysis, Robust Mann-Kendall, Pettitt Test and ADF test were conducted to describes the the spatio-temporal distribution of $PM_{2.5}/PM_{10}$ ratios. The study discovered that the average $PM_{2.5}/PM_{10}$ ratios for all five stations were around 0.7 which is more than half of total PM_{10} daily. Based on Robust Mann- Kendall test, it is found that there is a significant increasing trend in daily maximum $PM_{2.5}/PM_{10}$ ratios at Klang, Nilai and Seberang Jaya. In this study, a prediction model for daily maximum $PM_{2.5}/PM_{10}$ ratios is also obtained using LSTM. In determining the best LSTM model, several experimentations were employed to assess the influence of data splitting ratio for training and testing in the model development process as well as the epoch size and number of LSTM layers on the model performance. The results of the analysis have provided the evidence that splitting ratio and hyperparameter setting such as epoch and layer size affect the performance of LSTM models. Thus, the size of splitting ratio, model epoch and layer cannot be generalized for all data sets in different locations as their values give different effect on LSTM model's performance. The results of the experimentation suggested a suitable LSTM model for each study location with minimum errors. In conclusion, LSTM model is shown to be a promising time series prediction model for predicting daily maximum $PM_{2.5}/PM_{10}$ ratio.

ACKNOWLEDGEMENT

My deepest gratitude to Allah S.W.T for his blessings throughout my master's journey. It is indeed a humbling journey. I have made this with my own effort, but it would not have been possible if not for kind support and help from many individuals who have helped me achieving my goal of completing my Masters. Through this acknowledgement, I would like to take the opportunity to thank each one of these individuals.

First and foremost, I would like to express my sincerest gratitude to my supportive supervisor, Dr. Norshahida Shaadan who had guided me along the way of completing this research. It is thanks to her compassion and understanding that led me successful. She is very patient in answering all of my question whenever, even at night which seemed unfavourable to many. Without her invaluable input and continuous motivation, I would have not completed my research. I would like to acknowledge the support from my co-supervisor, Prof. Dr. Mohd. Talib Latif, for his involvement in this work.

I would also like to express my utmost gratitude to both of my parents who are also my inspiration and motivation to work harder, Zahari bin Rafik and Rosmazura binti Yusof for their endless support and encouragement which helped me in the completion of this research. Special thanks to my siblings, Muhammad Nafiz and Nur Dinie for the emotional support.

My appreciation also goes to my classmates especially Nazirah Alhadi, Harith Farhan, Aisyah Rashif, Fasihah, Syuhada Hamdan and Farhana who willingly helped me out with their capabilities. I would like to express my gratitude to my friends, Syarisa Elisa, Nurul Syahadah, Aina Izzaty and Muhammad Syahmi for their encouragement and always telling me that I can do it whenever I doubted myself. Special thanks to all educators from SK Kelana Jaya, SMK Kelana Jaya, UiTM Machang, UiTM Raub, UiTM Kota Bharu and UiTM Shah Alam, without them I could not be where I am now. I am highly indebted to all the individuals who have given me their support. A special dedication to my late-aunty, Aisyah binti Rafik, who had always believed in me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviations

ANN	Artificial Neural Network
API	Air Pollution Index
CAQM	Continuous Air Quality Monitoring Station
CO	Carbon Monoxide
DoE	Department of Environment
EDC	Environmental Data Centre
EQMP	Air Quality Monitoring Network
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAQM	Manual Air Quality Monitoring Stations
MCAQM	Mobile Continuous Air Quality Monitoring Stations
MSE	Mean Square Error
NO_2	Nitrogen Dioxide
O_3	Ozone
PM	Particulate Matter
PM_{10}	Particulate matter of less than 10 mm
$PM_{2.5}$	Particulate matter less than 2.5 mm
R^2	Correlation Coefficient
RNN	Recurrent Neural Network

CHAPTER ONE
INTRODUCTION

1.1 Background of Study

Air quality becomes a major concern in Malaysia due to rapid population growth and industrialization especially in urban area (Alias et al., 2020). It is found that urbanization aggravates the particulate matter (PM) pollutants which become unfavourable to Malaysia as one of the developing countries.

PM is a general term for extremely small particles and liquid droplets in the atmosphere and it can be classified into two which are $PM_{2.5}$ and PM_{10}. $PM_{2.5}$ is a fine particle with a diameter of 2.5 millimetres or less whereas PM_{10} is coarse particles that has a diameter of less or equal to 10 millimetres. Below is the illustration of $PM_{2.5}$ and PM_{10}.

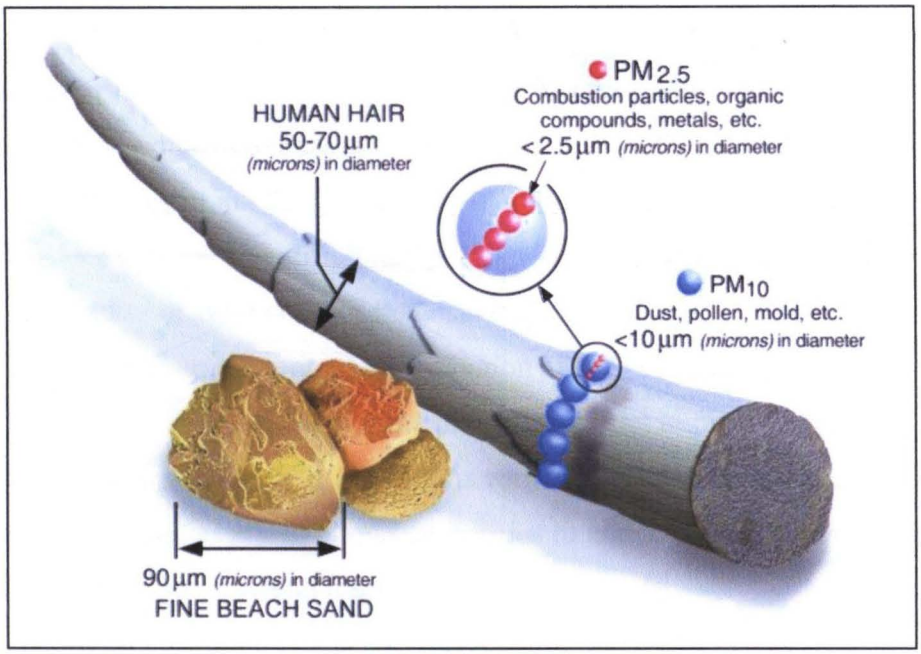


Figure 1.1 Comparison of $PM_{2.5}$ and PM_{10} Sizes
(Source: United States Environmental Protection Agency, 2021)

Figure 1.1 demonstrates that $PM_{2.5}$ and PM_{10} is much smaller than a strain of human hair and fine beach sand. This indicates how these particles can easily entered human body resulting in adverse effects to human health. Study found that $PM_{2.5}$ can

cause severe health problem compared to PM_{10} as it can penetrate through the upper respiratory tract of a human which led to serious respiratory health issues. Besides, PM_{10} are easily removed from the atmosphere with the help of gravitational settling and other processes, whereas $PM_{2.5}$ stayed in the atmosphere in longer duration and can transfer to another region (Harrison, 2001).

PM originates from anthropogenic as well as natural sources. In Malaysia, there are three main sources contribute to worsen air quality including mobile sources, stationary sources, and open burning. $PM_{2.5}$ and PM_{10} originate from different sources, for instance, combustion and mobile sources causes emission of anthropogenic PM resulting in formation of $PM_{2.5}$ whereas mechanical grinding and crushing activities lead to emission of PM_{10} (Munir, 2017). This showed the important characteristics of particulate matter (PM) is the size of the PM itself. Different size of PM originates from different sources, different duration of deposition and different depth of penetration to the human respiratory tract and ratio of $PM_{2.5}/PM_{10}$ able to describe the condition and sources of the PM.

Besides, pollution issues are unavoidable in Malaysia due to its ongoing growth in both its population size as well as urban infrastructure and building projects. West coast of Peninsular Malaysia is known as the location with concentrated industrial activities, highly populated areas and most likely to be affected by the anthropogenic sea-based (Thia-Eng et al., 2000) and land-based (Law et al., 2001) which make it an interesting location for scientific studies (Yap et al., 2002). Hence, it is crucial to monitor air quality in West coast of Peninsular Malaysia.

In this study, spatio-temporal distribution of $PM_{2.5}/PM_{10}$ ratios is analysed. This study also aims to obtain an appropriate predictive model to predict $PM_{2.5}/PM_{10}$ ratios using long short term memory (LSTM). It is very important to build a predictive model and assess the model as it will allows authorities to take a control measures to ensure the health of the population and improve the air quality.

1.2 Problem Statement

Rapid industrialization, urbanization and population growth causes developing countries including Malaysia to have a poor air quality especially in urban and industrial cities. Particulate matter (PM) has been identified to negatively affect human health. Children, senior citizens, people with existing heart or lung diseases and people who exercise or work outdoors. Therefore, it is important to deal with this problem before it is getting worse.

$PM_{2.5}$ and PM_{10} usually being investigated independently using various models such as multiple linear regression (MLR) model, support vector machine (SVM) model and long short term memory (LSTM) model. Several studies focus more on the health effect, spatial distribution, chemical composition, and influential factor analysis. However, these studies have been neglected the ratio of $PM_{2.5}/PM_{10}$. $PM_{2.5}$ is part of PM_{10} and the relationship between them is highly correlated. Hence, it is difficult to identify the contribution of $PM_{2.5}$ and which sources of PM worsen the condition. Origin of the particle, formation process and its impact on human health can be obtained from the ratio of $PM_{2.5}/PM_{10}$ as $PM_{2.5}$ and PM_{10} particles originate from diversity of sources and different chemical properties. High ratio of $PM_{2.5}/PM_{10}$ indicate large contribution from $PM_{2.5}$ which originate from anthropogenic sources. Whereas lower ratio suggests contribution of natural sources such as fugitive dust or sand dust from transferred from another location. Besides, major pollutants in PM_{10} can be identified using $PM_{2.5}/PM_{10}$ ratios and will be helpful to the authorize agency to act on specific emission locally rather than just controlling $PM_{2.5}$.

Small number of studies has been done to predict the $PM_{2.5}/PM_{10}$ ratios in other countries using several advanced method including LSTM, Geographically Weighted Regression (GWR) and Geographically and Temporally Weighted Regression (GTWR) and regression analysis, but very few in Malaysia. Pollutant concentrations are continuous in the behaviour and the series are dynamic in nature. To date, the prediction of the dynamic series of the ratio of $PM_{2.5}/PM_{10}$ is neglected.

Therefore, to fill the gap, a model that consider the long short term memory of the pollutant namely the LSTM, will be employed for this study. The hyperparameter setting and splitting ratio is experimented to obtain for a high predictive performance of LSTM model. These findings will be useful to the future research on modelling the $PM_{2.5}/PM_{10}$ ratios using LSTM model. There is still a paucity of research that need to

be capture on the ratios of $PM_{2.5}/PM_{10}$. Hence, prompting the need for this research.

1.3 Research Questions

The investigation sought answers to the following research questions:

i. What is the spatial-temporal distribution of $PM_{2.5}/PM_{10}$ ratios at several locations in the West Coast of Peninsular Malaysia?

ii. What is the influence of splitting ratio, epochs size and number of LSTM layer on the performance of LSTM model of daily maximum $PM_{2.5}/PM_{10}$ ratios at the study locations?

iii. What is the most appropriate splitting ratio and hyperparameters of LSTM model in predicting daily maximum $PM_{2.5}/PM_{10}$ ratios at study locations?

1.4 Research Objectives

This research was carried out with three main objectives which are:

i. To investigate the spatial-temporal distribution of $PM_{2.5}/PM_{10}$ ratios at several locations in the West Coast region of Peninsular Malaysia.

ii. To investigate the influence of splitting ratio, epochs size and number of LSTM layer on the performance of LSTM model of daily maximum $PM_{2.5}/PM_{10}$ ratios at the study locations.

iii. To determine the most appropriate splitting ratio and hyperparameters of LSTM model in predicting daily maximum $PM_{2.5}/PM_{10}$ ratios at study locations.

1.5 Scope of Studies

This study focused on $PM_{2.5}/PM_{10}$ ratios at several location in West Coast of Peninsular Malaysia. The areas of study; Klang, Bukit Rambai, Seberang Jaya, Nilai and Manjung were selected as it is reported as the area having relatively high value of PM. Besides, these areas are concentrated with industrial activities and highly populated city especially Klang area as it is located close to Kuala Lumpur. Hence, these locations are in risk of high contamination of particulate matter (PM) in the atmosphere. LSTM

prediction model for daily maximum $PM_{2.5}/PM_{10}$ ratios is considered to be developed at the study locations. Statistical experimentation was conducted within the scope of the effect of data splitting ratio of four categories; 60:40, 70:30, 80:20 and 90:10. The effect of hyperparameter including epochs size between 50, 100, 200 and 500 and number of layers between one to four were also investigated in order to determine the best suitable LSTM model at the study locations.

1.6 Significance of Study

This study is beneficial to the society due to the topic itself which is strictly related to the human health and environment. Moreover, the provided data can be use as the guideline to enhance the air quality level in the selected areas. There exist several related studies on $PM_{2.5}/PM_{10}$, however there was lack of studies being carried out on the modelling of $PM_{2.5}/PM_{10}$ ratios prediction. Therefore, this study also discusses on the predictive modelling of the ratios using a new and trending deep learning method namely LSTM. Thus, future research can use the results of the LSTM model obtained as the basis for further enhancement and exploration. The findings of this study can also be helpful to the related government agencies to take precautionary and protective measures if needed. Lastly, this study can increase the knowledge enhancement in the area of ratio of $PM_{2.5}/PM_{10}$. Crucial information can be obtained from the ratio itself for risk analysis and exposure assessment.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

This chapter discussed on the scholarly published article on the various aspects of research related to air pollution. It is constructed to provide insights on several issues such as current scenario in Malaysia and the method used in the past studies to predict the air pollutants. This chapter is organized into four themes covered in this study as presented in Figure 2.1. The first theme introduces the scenario of air pollution in Malaysia. Second theme discuss on the $PM_{2.5}$ and PM_{10} pollutants. The third theme review on the air pollution prediction and the model. Fourth theme assess the related studies on $PM_{2.5}/PM_{10}$ ratio in Malaysia and worldwide. Therefore, this chapter gives a better understanding on current situation of the research conducted on the topic of $PM_{2.5}$ and PM_{10} pollutants and research gap of the study is identified to be filled during the study.

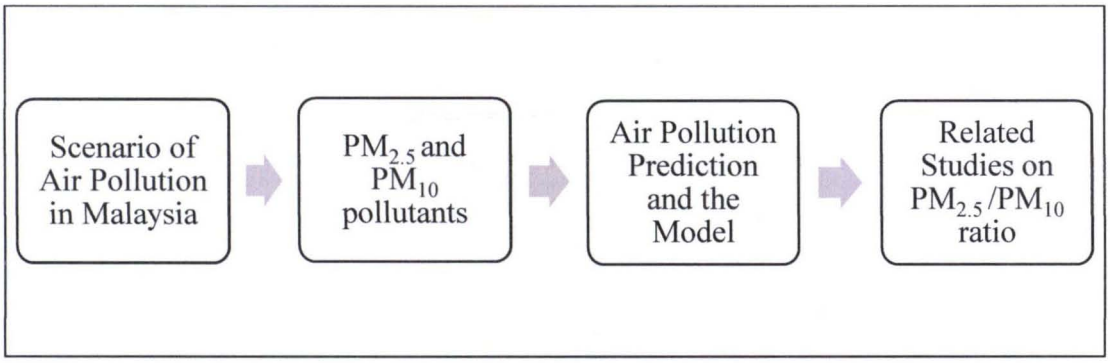


Figure 2.1 Framework for Literature Review

2.2 Scenario of Air Pollution in Malaysia

Air pollution in Malaysia is monitored by the Malaysian Department of Environment (DoE) or originally known as Environment Division under the Ministry of local Government and Environment. It is established on April 15, 1975 with the purpose of preventing, eliminating, and controlling the pollution and enhance the environment, consistent with the principles of the Environmental Quality Act 1974. DoE monitor the air quality through the EQMP Air Quality Monitoring Network to

detect any possible changes in air quality that will negatively affect the environment and also human health. EQMP comprises of three types of monitoring stations including 65 Continuous Air Quality Monitoring Stations (CAQM), 14 Manual Air Quality Monitoring Stations (MAQM) and three Mobile Continuous Air Quality Monitoring Stations (MCAQM). CAQM stations which are strategically located at urban, sub-urban, industrial, and rural areas will transmit the real-time air quality data to the Environmental Data Centre (EDC) on a scheduled basis. Figure 2.2 shows the monitoring stations throughout Malaysia.

Figure 2.2 Monitoring Stations throughout Peninsular Malaysia
(Source: Alyousifi et al., 2021)

Air Pollutant Index (API) was established in 1996 to provide a guideline for public to easily understand the information about air pollution and easily compare the air pollution in Malaysia with the countries in ASEAN. Besides, API is also closely following the Pollutant Standard Index (PSI) developed by the United States Environmental Protection Agency (US-EPA) (Department of Environment Malaysia, 1997). There are five air pollutants included in Malaysia's API to be monitored which are Ozone (O_3), Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2) and particulate matter of less than 10 mm (PM_{10}). In 2017, index for $PM_{2.5}$ pollutant is added into the API system (Department of Environment, 2019). The status indication of the API and its impact on health is listed below.

Table 2.1
API Indication and Its Impact on Health

API	Status	Impact on Health
0 to 50	Good	No bad effect on health
51 to 100	Moderate	Does not badly affect the health
101 to 200	Unhealthy (For sensitive group)	Worsen health condition to sensitive group including elderly, pregnant woman, children and people with heart and lung complications.
201 to 300	Very Unhealthy	Worsen the health condition and low tolerance of physical exercises to people with heart and lung complications and will affect public health.
>300	Hazardous	Hazardous to high-risk people and public health
>500	Emergency	Hazardous to high-risk people and public health

Source: Department of Environmental (2019)

Malaysia Ambient Air Quality Guideline that being used since 1989 has been replaced with new guideline named as New Ambient Air Quality Standard which adopt six types of air pollutants consists of ground level Ozone (O_3), Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), PM_{10} and newly added pollutant, $PM_{2.5}$. The new guideline is divided into three interim target which are target 1 (IT-1) in 2015, interim target 2 (IT-2) in 2018 and the full implementation of the standard in 2020 as shown in Table 2.2.

Table 2.2
Malaysia Ambient Air Quality Guideline

Parameter	Averaging Time	Unit	Existing Guidelines	Malaysia Ambient Air Quality		
				IT-1 (2015)	IT-2 (2018)	Standard (2020)
PM_{10}	1 Year	$\mu g/m^3$	50	50	45	40
	24 Hours		150	150	120	100
$PM_{2.5}$	1 Year	$\mu g/m^3$	-	35	25	15
	24 Hours		-	75	50	35
SO_2	1 Hour	$\mu g/m^3$	350	350	300	250
		Ppm	0.135	0.135	0.115	0.095
	24 Hours	$\mu g/m^3$	105	105	90	80
		Ppm	0.040	0.040	0.035	0.030
CO	1 Hour	$\mu g/m^3$	35	35	35	30
		Ppm	30.6	30.6	30.6	26.2
	8 Hours	$\mu g/m^3$	10	10	10	10
		Ppm	8.75	8.75	8.75	8.75
NO_2	1 Hour	$\mu g/m^3$	320	320	300	280
		Ppm	0.170	0.170	0.160	0.150
	24 Hour	$\mu g/m^3$	75	75	75	70
		Ppm	0.040	0.040	0.040	0.037
O_3	1 Hour	$\mu g/m^3$	200	200	200	180
		Ppm	0.100	0.100	0.100	0.090
	8 Hour	$\mu g/m^3$	120	120	120	100
		Ppm	0.060	0.060	0.060	0.050

Source: Department of Environment (2020)

Malaysia has two monsoon seasons, namely south-west monsoon (May to October) and northwest monsoon (November to March) and transition of wind periods happened in Malaysia which known as inter-monsoon occurred in to May and October to November (Hashim et al., 2018). In early November until ends of March, NE monsoon occurred where heavy rainfall is brought by this monsoon in November to January followed by hot weather in February until March (Hashim et al., 2018). Then, first inter-monsoon happened in April to May followed by SW monsoon in longest period of May to October where hot and dry weather is expected in June to September. However, light and variable winds is expected to happen during two inter-monsoon (Hashim et al., 2018).

A study conducted by Amil et al. (2016) showed that highest monthly mean mass of $PM_{2.5}$ is in the month of June followed by September where south-west monsoon happened during both months. Generally, Peninsular Malaysia has recorded higher $PM_{2.5}$ concentration during south-west monsoon which also known as dry season Amil et al. (2016) and also other Asian cities (Reid et al., 2013). The study further argued that haze is another factor for increasing $PM_{2.5}$ concentration which mostly occurred during south-west monsoon. Among all cities in Peninsular Malaysia, Petaling Jaya has the highest $PM_{2.5}$ concentration (Ee-Ling et al., 2015; Tahir, Koh & Suratman, 2013). Area that affected the most during transboundary haze which released vast

amount of particulate matter from Sumatra in Malaysia is Klang Valley area (Sentian et al., 2019).

2.3 $PM_{2.5}$ and PM_{10} pollutants and their impact

Air pollution is a mixture of gaseous and particulate component in large quantity for a long period of time that could affect the human health and other living things. According to Hamanaka and Mutlu (2018), clinical studies has highlighted that particulate matter (PM) bring bigger impact to the health of human compared to other pollutants. Level of severity of PM towards human health depending on the sizes of PM (Khan et al., 2016; Ross et al., 2013). PM can be classified according to size of particles which are coarse (PM_{10}), fine ($PM_{2.5}$), and ultrafine ($PM_{0.1}$) (Hamanaka & Mutlu, 2018). Particles with diameter of 2.5 micrometers are called $PM_{2.5}$ while diameter of particles sizes of 10 micrometers known as PM_{10}.

According to Sentian et al. (2019), scientific studies have been conducted on the impact of short-term and long-term exposure of living things towards the air pollutants differ in term of its severity. Mohd Zahid et al. (2018) reported that particulate matters with extremely small sizes have higher chances to access the body of human compared to particles in large size as illustrate in the figure 2.3.

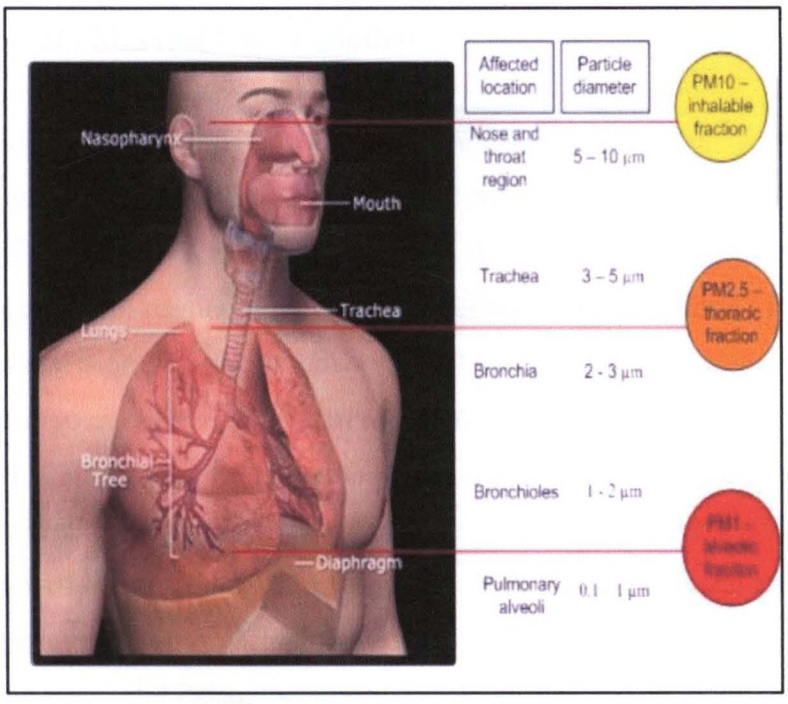


Figure 2.3 PM in Human Lung (Source: Department of Environment, 2018)

Figure 2.3 demonstrate that PM_{10} can easily passes across the nose and throat of a human (Mohd Zahid et al., 2018) while that $PM_{2.5}$ can penetrate through the upper respiratory tract of a human into the bronchioles and alveoli which can be deposited in the lung and access major organ systems of human body (Yunesian et al., 2019) that can lead to serious health issue such as mutagenic and respiratory problem (Ali-Mohamed & Jaffar, 2000; Amoatey, Omidvarborna, & Baawain, 2018). Hamanaka and Mutlu (2018) also stated that particulate matter can also increase the possibility of death from cardiovascular disease such as heart failure, ischemic or thrombotic stroke and even ischemic heart disease. Besides, Samoli et al. (2008) reported that unemployed and elderly are more vulnerable towards short-term exposure of PM while Pope et al. (2004) discovered that among never, former, and current smokers that are exposed to chronic $PM_{2.5}$ were positively associated with mortality from ischemic heart disease. However, former and current smokers have higher risk to die from cardiac arrest, arrhythmia and heart failure but not for never smokers' group. Short and long-term studies have discovered that higher risk of fatal and non-fatal ischemic heart (Cesaroni et al., 2014; Pope et al., 2004; Xie et al., 2015) disease for an individual that are exposed to $PM_{2.5}$ and also myocardial infarction (Madrigano et al., 2013; Nawrot et al., 2011). This shows that effect of $PM_{2.5}$ are more serious than PM_{10} and should not be taken lightly.

2.3.1 Sources of $PM_{2.5}$ and PM_{10} Pollution

Generally, issue of air pollution is contributed by various sources which can be categorized into two, namely natural sources and anthropogenic sources (Mohd Zahid et al., 2018). Eruption of volcanoes, unintended fires in forestry, storm of ashes and road are the natural sources contributed to air pollution while industry factories and power stations emission along with industrial fuel burning are the common production of anthropogenic sources (Mohd Zahid et al., 2018). Due to the numerous sources of air pollution, particulate matter may contain a wide range of chemical elements.

Department of Environment (2020) reported that power plants (39%), industries (29%), emission of motor vehicles (12%) and others (12%) are the main sources of PM emission in Malaysia. States located in East Coast of Peninsular Malaysia namely, Kelantan, Terengganu and Pahang were heavily affected by the particles originate from the sea such as sea-spray aerosol coming from combination of organic matter and

inorganic sea salt as it is close to the coastal region (Ismail et al., 2015).

According to Lin et al. (2014), emission from a country can be transmitted to neighbouring country. Southern part of Thailand, Singapore, Brunei and Malaysia are the example of a countries that affected by fire emission from neighbouring country which is Indonesia (Quah, 2002). It is reported that Taiwan also affected by the fire emission in Laos, Thailand and Myanmar (Chang & Song, 2010). It can be seen that sources of PM_{10} concentration of most of country in Asia are the fire emission from neighboring country. This happened when combination of forest and agriculture burning created massive amount of aerosols emission during dry season within a high-pressure region causes building of layer of smoke called haze incident.

Khan et al. (2015) revealed that combustion of gasoline, diesel, and heavy oil; natural gas and coal burning are the main sources of $PM_{2.5}$ concentration in semi-urban area in Bangi. Meanwhile research conducted by Rahman et al. (2011) indicate that emission of two stroke engine, motor vehicles, soil, industry and biomass burning are the sources of $PM_{2.5}$ concentration in Klang Valley.

Besides, Lestiani et al. (2008) successfully identified the sources of $PM_{2.5}$ emission in Bandung and Lembang, Indonesia as soil, biomass burning, emission of two stroke engine and motor vehicle, road dust, secondary sulphate and sea salt. Besides, Reid et al., (2013) discussed that the major factors of $PM_{2.5}$ concentration were two-stroke engine vehicle emission, high emission of particles and incomplete combustion products whereas $PM_{2.5}$ concentration in Bangkok, Thailand is affected by the mobile sources and Manila, Philippines heavily influenced by the diesel truck and emission of bus. Singaporean $PM_{2.5}$ sources are dust of soil, biomass burning and automobiles, combustion of fuel oil, sea salt and metallurgical industry (Balasubramanian et al., 2003). In Beijing, Zheng et al. (2005) reported that the sources are coming from dust, combustion of coal, secondary sulphate and ammonia, exhaust of gasoline and diesel, biomass aerosol, vegetative detritus and cigarette smoke whereas $PM_{2.5}$ sources of India are observed to be biomass burning, marine aerosol, motor vehicles, brake and tyre wear, secondary PM and soil. It can be said that the sources of $PM_{2.5}$ varied for every region in a country. However, mostly the affected areas were affected by the biomass burning and emission of vehicle.

2.4 Prediction Model for Air Pollution

Air pollution is in alarming situation, not only Malaysia, but almost all the country in the world. Numerous studies have been conducted to predict the air pollution around the world to improve the air quality in specific locations. According to Saeed et al. (2017) various algorithms for developing predictive models from single time series data have been proposed including predictive modelling and time series forecasting method.

2.4.1 Linear Models

Generalized Additive Models (GAM) is the extension of Generalized Linear Model (GLM) which allows linear model to capture the non-linear relationship. Li et al. (2017) implemented GAM to robustly forecast the $PM_{2.5}$ concentrations in Shandong Province in China and capture the relationships between $PM_{2.5}$ concentrations and predictor variables while considering the non-linear effect and variability of spatial and spatiotemporal of the predictors. The stability of the model is improved by using the ensemble learning and the variogram is implemented to capture the daily residuals. The result obtained indicate that higher R^2 value of 0.89 is achieved when PM_{10} included as the predictor compared to model with absence of PM_{10} as the predictor input ($R^2 = 0.86$).

Several studies have been conducted using Multiple Linear Regression (MLR) to predict $PM_{2.5}$ and PM_{10} in Malaysia with different variables. A study done by Abdullah et al. (2017) used the MLR model to predict long term PM_{10} concentration in different monsoon using meteorological factors while Alifa et al. (2020) used the same model to examine the impact of urban, wildfire, and meteorological variables on observed PM_{10} patterns at various temporal scales. Both studies concluded that meteorological factors that affecting the variability of PM_{10} concentration varies by monsoon. Alifa et al. (2020) further added that high monthly median of PM_{10} concentration in urban areas compared to less populated area is due to the increase of anthropogenic emissions that have been detected in populated cities proved that urbanization can lead to serious air quality problem in future and should be observe closely.

2.4.2 Non-linear Machine Learning Models

Several studies have been conducted using non-linear model to cater the problem of non-linear phenomena since nature of air pollution data to be nonlinear. Pan (2018) implemented Extreme Gradient Boosting (XGBoost) to predict the hourly $PM_{2.5}$ concentrations in China. The difference between XGBoost and boosting is that it utilized advanced regularization to improve model generalization capabilities which resulting in better training performance (Zhang et al., 2018). The study concluded that XGBoost is indeed powerful with high level of accuracy and low level of fitting probability as it outperforms other machine learning algorithm, namely, RF, Support Vector Machine for Regression (SVMreg), MLR and Decision Tree Regression (DTR) in predicting the hourly $PM_{2.5}$ concentrations.

On the other hand, Abdullah et al. (2019) employed both linear; MLR model and two different non-linear model; multi-layer perceptron and radial basis function (RBF) and compare the performance of those models using several pollutants and meteorological variables. The result showed that non-linear models outperform linear model with RBF as the best model in predicting the next day of PM_{10} concentration.

Study conducted by Shafii et al. (2020) employed SVM model to compare the prediction accuracy of four different types of kernel function which include linear kernel, polynomial kernel, Radial Basis Function (RBF) kernel and sigmoid kernel. The result exhibit that model that employed RBF kernel with parameter of 100 is the best at accurately predicting API of $PM_{2.5}$. The study further justified that SVM can reduce the problem of overfitting in training and works well in high dimensionality data than other conventional method. However, result obtained by a study conducted by Masood and Ahmad (2020) with the objective, to forecast the $PM_{2.5}$ concentration with meteorological factors and pollutants concentrations as an input variable, discovered that ANN has even better prediction performance than SVM model as it can deal with the nature of air pollution data that has complex multidimensionality problems.

Study conducted by Alpan and Sekeroglu (2020) utilizing machine learning (ML) algorithm including decision tree (DT), Support Vector Regression (SVR) and random forest (RF) model to predict six different pollutants by taking meteorological data into consideration. Result showed that RF has the capability in predicting pollutants concentrations with high R^2 of 0.7411 to 0.8654.

A study by Lin (2021) employed complex model such as boosting, random forest and neural network and concluded that these models have smaller error in predicting both $PM_{2.5}$ and PM_{10} concentration in China compared to simpler model like MLR. It is discovered that boosting has better prediction performance than other models with R^2 of 84.2% and 75.7% for $PM_{2.5}$ and PM_{10}, respectively. This showed that non-linear model is more suitable to use in predicting pollutant data due to the nature of the data itself.

2.4.3 Time Series Models

A comparison of time series model and interactive multiple model (IMM) have been studied by Li, Li and Wang (2019) where autoregressive (AR) model is formed, and the model is then transformed into state of equation to proceed with Kalman filtering. Next, the AR model and the AR-Kalman model are compared with the IMM algorithm to identify which model accurately predict the $PM_{2.5}$ concentration at different levels of air quality. The result found that IMM models is better at predicting $PM_{2.5}$ concentration than single AR model and single AR-Kalman model as the model has lower MAPE at different levels of air quality.

Study by Alyousifi et al. (2020) to forecast the API in Klang, Malaysia using Markov method time series model is conducted to address the limitation of fuzzy time series which is implementation of random partition of discourse's universe. This study is performed using partition method Markov weighted fuzzy time-series model based on the optimum partition method. The API index in this study is determined by choosing the highest value of five main pollutants which include PM_{10}, O_3, CO_2, SO_2, and NO_2. The performance of proposed model is then compared with eight fuzzy time series models and three conventional time series models (ARIMA, GARCH, and ARIMA–GARCH). The result indicates that the proposed model outperformed others fuzzy time series models and conventional time series models.

A comparison between Nonlinear Autoregressive Exogenous (NARX) Neural Network and Support Vector Machine (SVM) regression is performed to accurately predict the API in Malaysia (Mustakim & Mamat, 2021). The NARX model has utilized series-parallel feedforward network while six different kernels are used in this study which include Coarse Gaussian, Cubic, Fine Gaussian, Linear, Medium Gaussian and Quadratic kernel to be compared with the NARX model. The performance of the

mentioned models is evaluated using RMSE and R^2 values. The result of the study showed that NARX model performed better than SVM models.

2.4.4 Deep Learning Models

A study by Suleiman, Tight and Quinn (2020) conducted with the aim to compare three methods which are include Deep Learning (DL) algorithms, Extreme Learning Machine (ELM) and RF in modelling the roadside of $PM_{2.5}$ and PM_{10} concentrations. The study concluded that all the proposed models including RF have high performance with average performance of 0.94 and 0.90 (R); 98% and 99% (FAC2); 4.5 and 9.2 (RMSE) and 0.84 and 0.83 (IOA) for predicting $PM_{2.5}$ and PM_{10} respectively. The study further discuss that these models have faster training speed and scalability compared to traditional ANN when using high performance computation.

Another powerful deep learning method is employed to predict air quality which is LSTM RNN method as the model has a good proven record of success with time series data (Belavadi et al., 2020). This study examines the performance of the LSTM RNN model in predicting the air quality in India and design a low-cost sensor node for future monitoring air quality. The study concluded that model that fits every city under any condition does not exist since different city has different temporal variance, for instance Bengaluru metropolitan city has high temporal variance which might bring an adverse effect to the performance of the model.

2.4.5 Spatial Models

Another study also implemented LSTM model but with geo layering called Geo-LSTM. The study conducted by Ma et al. (2019) in Washington with the purpose of predicting the air pollution at non-monitored sites using known data and observed points is implemented to address the neglection of other studies in considering the long-short temporal trend and spatial associations or air pollution. The model is then compared with several other models including SVR, LASSO, Ridge Regression, Gradient Boosting Decision Tree (GBDT), RF, ANN, RNN, ordinary LSTM, Inverse Distance Weighting (IDW), Original Kriging (OK) and Cube Spline (Cs). The result demonstrate

that Geo-LSTM outperform other models as the model has the ability to learn the non-linearity of temporal sequence from the long-term dependencies and also consider the spatial-temporal of air pollutants.

Another method is utilized to overcome the limitation of previous study in modelling the spatial of air pollutants which is Spatial interpolation in GIS. Recent study by Tella, Balogun and Faye (2021) in Malaysia has employed Geographic Information System (GIS), Multivariate Regression Model (MVR) to explore the relationship between PM_{10} and main climate factors including temperature, wind speed and humidity. Performance of the model is evaluated and concluded with regression model of R^2, RMSE and MAE of 0.298, 12.737 and 10.343 respectively. The study also found that temperature, humidity, and wind speed were the important parameter associated with the PM_{10} concentration in the study area.

2.4.6 Hybrid Models

Hybrid model is also being used to model the pollutant. Hybrid model named as Comprehensive Forecasting Model (CFM), a combination result of ARIMA, ANN and Exponential Smoothing Method (ESM) using entropy weighting method to predict $PM_{2.5}$ concentration in China (Liu and Li, 2015). The proposed model was found accurately predict the $PM_{2.5}$ concentration with higher accuracy compared to single forecasting method. Furthermore, the model able to overcome the limitation of single model by balancing the deviation of each single model and predict more accurately the despite heavy workload compared to single method.

Du et al. (2020) proposed a novel hybrid model to model the six different air pollutant in three cities in China. The study first developed a multi-objective algorithm named MOHHO which then being introduced to the ELM models for parameters' tuning. Finally, the time series prediction is performed using the optimized ELM model. The study used several metrics to compare with five models including ARIMA, LSSVM, EMD-MOHHO-ELM, CEEMD-MOHHO-ELM and ICEEMDAN-MOHHO-ELM. It was concluded that hybrid model has higher prediction accuracy and more stable compared to other five models.

Hybrid model of ML model combined with time series model has been conducted by Shahriar et al. (2021) in India to compare the performance of Integrated Moving Average (ARIMA), ARIMA-ANN, ARIMA-SVM and Principle Component

Regression (PCR), DT and CatBoost deep learning in predicting $PM_{2.5}$ concentration. The result indicate that CatBoost has the best performance followed by hybrid model of ARIMA and ANN. Combination of ARIMA and ANN and also DT have an acceptable result while PCR model does not perform well in predicting $PM_{2.5}$ concentration. As previous study by Liu and Li (2015) concluded that hybrid model is better in predicting $PM_{2.5}$ concentration compared to single model, this study extended the research by comparing hybrid model with deep learning model and concluded that deep learning method is better than hybrid model in predicting $PM_{2.5}$ concentration.

To summarize the literature on the models including both predictive and time series forecasting method applied in the past studies, Table 2.3 is presented for a brief explanation.

2.5 Related Studies on $PM_{2.5}/PM_{10}$ Ratio

Tahir, Koh and Suratman (2013) reported that researchers have been studying on the air particulate pollution in Malaysia since late 1990s but most of the studies only being reported in the grey literature. Researchers started publishing studies on the PM_{10} in 1995 while research on $PM_{2.5}$ only started in 2004 focusing on the nitrogen dioxide and particulates ($PM_{2.5}$ and PM_{10}) levels in Kota Kinabalu Malaysia. Until recently, numerous studies have been conducted to predict $PM_{2.5}$ and PM_{10} in Malaysia and worldwide; however, less consideration is taken on the ratio– prediction of $PM_{2.5}$ and PM_{10} (Zhao et al, 2019). To the knowledge of the researcher, limited number of studies have been reported focusing solely on the ratio of $PM_{2.5}/PM_{10}$ in Malaysia. The following review supports the statement.

Table 2.3
Summary of Past Studies using Different Models

Authors	Study Location	Objectives	Air Pollution Parameter	Models	Findings
Tella, Balogun and Faye (2021)	Malaysia	To examine the relationship between PM_{10} and climate parameters.	PM_{10}	GIS, MVR and Pearson Correlation	High predictive performance of Regression model.
Shahriar et al. (2021)	India	To predict the ambient $PM_{2.5}$ and access the model performance.	$PM_{2.5}$	ARIMA, ANN, ARIMA-SVM, PCR, DT and CatBoost	CatBoost performed best at predicting $PM_{2.5}$ for all the stations.
Mustakim and Mamat (2021)	Malaysia	To produce one step ahead prediction model to predict API in Malaysia	PM_{10}, O_3, CO_2, SO_2, and NO_2	NARX and SVM	NARX outperformed SVM regression models.
Lin (2021)	China	To predict $PM_{2.5}$ and PM_{10}	$PM_{2.5}$ and PM_{10}	MLR, DT, Boosting, Neural Network and RF	Boosting outperform others model.
Alyousifi et al. (2020)	Malaysia	To forecast the API using Markov method time series model	PM_{10}, O_3, CO_2, SO_2, and NO_2	Markov weighted fuzzy time-series model	Markov model outperformed others fuzzy and conventional time series models.
Suleiman, Tight and Quinn (2020)	United Kingdom	To model air quality.	PM_{10} and $PM_{2.5}$ concentrations	ML methods, RF, ELM and DL algorithms.	Models performed slightly better in predicting $PM_{2.5}$ compared to PM_{10}.
Masood and Ahmad (2020)	India	To develop $PM_{2.5}$ prediction model with meteorological and pollutant input	$PM_{2.5}$ concentration	SVM, ANN	ANN shows better prediction accuracy than SVM for $PM_{2.5}$ prediction.
Shafii et al. (2020)	Malaysia	To access API of $PM_{2.5}$ using SVM with different types of the kernel.	API of $PM_{2.5}$	SVM	RBF model with its parameters of cost and gamma of 100 is best at forecasting API.
Alpan and Sekeroglu (2020)	China	To predict the values of six primary pollutant concentrations.	PM_{10} and $PM_{2.5}$	RF, DT and SVM	RF has a high prediction capacity with R^2 of 0.74 to 0.86.

Table 2.3
(cont'd)

Authors	Study Location	Objectives	Air Pollution Parameter	Models	Findings
Alifa et al. (2020)	Malaysia	To investigate the influence of urban, wildfire emissions and meteorological variables on the PM_{10} patterns	PM_{10}	MLR	The influence of meteorology varies by region and season
Belavadi et al. (2020)	India	To forecast air quality using LSTM RNN.	PM_{10} and $PM_{2.5}$	LSTM RNN	Performance of the model decreases as the variation increases.
Du et al. (2020)	China	To predict daily $PM_{2.5}$ and PM_{10}.	$PM_{2.5}$ and PM_{10}	ELM	The proposed hybrid model provides stable and better performance compared to others
Abdullah et al. (2019)	Malaysia	To compare the forecasting ability of linear non-linear models	PM_{10}	MLR and Multi-layer perceptron	Non-linear model (Radial Basis Function) outperforms the linear model
Ma et al. (2019)	Washington	To estimate the air pollution at non-monitored locations	$PM_{2.5}$	LSTM neural network and Geo-LSTM	Geo-LSTM has better performance than traditional methods.
Li, Li and Wang (2019)	Malaysia	To predict $PM_{2.5}$ concentration according to different air quality levels.	$PM_{2.5}$	IMM, AR, MA, and ARMA models.	IMM algorithm provide more accurate and effective prediction of $PM_{2.5}$.
Pan (2018)	China	To predict $PM_{2.5}$ concentration	$PM_{2.5}$ and PM_{10}	XGBoost algorithm	XGBoost algorithm outperforms other data mining methods.
Li et al. (2017)	China	To predict $PM_{2.5}$ concentrations	$PM_{2.5}$	GAM	The proposed method can be used to predict $PM_{2.5}$ exposure with good accuracy.
Abdullah, Ismail and Fong (2017)	Malaysia	To predict PM_{10} concentration level using MLR models	PM_{10}	MLR	The MLR models each monsoon has R^2 in range 0.57 to 0.68
Liu and Li (2015)	China	To predict the time series data of $PM_{2.5}$	$PM_{2.5}$	ARIMA, ANN and ESM	Forecasting model had better applicability.

2.5.1 Exploratory Analysis

A study in late 1990s conducted by Parkhurst et al. (1999) to describe the comprehensive study on the TVA dichotomous sampler of $PM_{2.5}/PM_{10}$ ratio in United State found a strong correlation between $PM_{2.5}$ and PM_{10} with $PM_{2.5}$ contributing on the average of 67% to the PM_{10}. It is also discovered that the ratio of $PM_{2.5}/PM_{10}$ is in range of 0.5 to 0.6 in urban area which is lower than other urban location in EPA sites. This implies that smaller fraction of $PM_{2.5}$ found in PM_{10} in western US compared to the East area where $PM_{2.5}$ is found to be more dominant. Study also discovered that $PM_{2.5}/PM_{10}$ ratios exhibit seasonal trend where the ratio found to be higher during summer season (June to August) due to emission of biogenic, cultural activities and higher speed of wind.

Besides, Blanco-Becerra, Gáfaro-Rojasv and Rojas-Roa (2015) explored whether $PM_{2.5}/PM_{10}$ ratio in Columbia, is reduced by the scavenging effect in rainy seasons with the comparison to dry sessions in Colombia using the exploratory analysis. The study supported the hypothesis that rain's scavenging action reduces the PM_{10}, to a smaller extent of $PM_{2.5}$ concentrations. However, it is not plausible to claim that rainy season reduces the risk affiliated with PM as the analysis of the study was inconclusive regarding the effect of rain on $PM_{2.5}$ concentrations which is also the weakness of this study.

A study focusing on the estimation of spatial parameter is conducted to predict the ratio of $PM_{2.5}/PM_{10}$ is conducted by Chu, Huang and Lin (2015) in Taiwan. The study implemented fuzzy c-means to cluster the spatial heterogeneity first, then the PM_{10}-$PM_{2.5}$ relationship is modeled using Global Ordinary Least Squares Regression, Geographically Weighted Regression (GWR) and Geographically and Temporally Weighted Regression (GTWR). R^2 value is used as the metric to measure the performance of the models. The study concluded that GTWR has better performance than GWR model in terms of R^2. It is found that the relationship of PM_{10}-$PM_{2.5}$ exhibit spatio-temporal variation where R^2 (0.85) value of OLS model for winter season is higher due to accumulation of aerosol, lower air temperature and presence of pollutants and dust in winter season.

There is very limited study on $PM_{2.5}/PM_{10}$ ratio in Malaysia. A study conducted by Mat Shukri et al. (2017) in identifying relationship between $PM_{2.5}$ and PM_{10} and meteorological factors at Roadside of Penang Bridge indicate that the ratio of

$PM_{2.5}/PM_{10}$ is at the lowest in July and September while highest in month of august. However, ratio of $PM_{2.5}/PM_{10}$ is deliberated with very minimal discussion in these studies and only included as a small part in descriptive analysis of the study.

Another study employed Theil-sen approach to examine the long-term temporal including the annual, seasonal, and monthly trends of $PM_{2.5}/PM_{10}$ ratio in Bahrain (Coskuner, Jassim and Munir, 2018). The study concluded that PM pollution is dominated by the coarse particles and the ratio suggested a decreasing pattern in all seasons. A positive correlation found between $PM_{2.5}/PM_{10}$ ratio while the ratio demonstrated negative relationship with temperature and wind speed parameter which can be assumed that ambient temperature and windy days increases the PM_{10} concentrations.

A study conducted in China, by Zhao et al (2019) using statistical and correlation analysis with the purpose of examining the relationship of ratios and AQI, rate of change of the ratios and the meteorological parameters' impact on the ratio. The study concluded that diverse economic development and industrial types have resulted in rising trend from northwest to southwest. Besides, higher ratio indicates that greater possibility of high AQI which imply that air pollution will be more critical. The study discovered that relative humidity, precipitation and were positively impact the ratio while sunshine duration, wind speed and temperature contribute negatively to the ratios.

2.5.2 Predictive Analysis

A laboratory study is performed to determine the relative breakage and $PM_{2.5}/PM_{10}$ ratios created in this study by stimulating the impacts of erosion of saltation-size aggregate for a range of soils in United State (Hagen, 2004). The ratios and relative breakage are then modelled using regression analysis with saltation-size of sand and precipitation, and fraction of clay and annual precipitation as the independent variables, respectively. The result of the study showed that average $PM_{2.5}/PM_{10}$ ratios is 0.154 with increasing saltation-size and decreasing precipitation. The predicted values of the study range between 0.1 to 0.3 with R^2 of 0.53.

A recent study conducted in Wuhan, China with the aim of predicting $PM_{2.5}/PM_{10}$ ratios using LSTM neural network based on space, time and random pattern (Wu et al., 2020) as LSTM is known to be a dynamic model that has good capability in remembering historical information. The model is then compared with three other

different model which include SVM, Back Propagation Neural Network and chi-squared automatic interaction detection (CHAID) using measured value and error rate. It is found that LSTM model has the smallest average error of 15.76 but none of the model can accurately predict ratio larger than 0.9. Nevertheless, LSTM is the only model found to be able to predict ratio larger than 0.8 on the last day. Table 2.5 summarized the past studies related to $PM_{2.5}/PM_{10}$ ratio.

2.5.3 Characteristics and the Importance of $PM_{2.5}/PM_{10}$ Ratios

According to Sugimoto et al. (2016), ratio can be interpreted as higher ratio of $PM_{2.5}/PM_{10}$ which is greater than 0.5 indicating presence of anthropogenic sources while smaller ratio which is smaller than 0.5 suggest that coarse particle exist in considerable involvement which possibly might related to natural sources such as dust storm. For instance, higher ratio is discovered in eastern of United State with ratio of approximately 0.7 (USEPA, 2004), Oman with median ratio of 0.69 (Alattar et al., 2019), United Kingdom with median ratio of 0.65 (Munir, 2017), while ratio in Saudi Arabia is found to be relatively low with ratio of 0.33 (Khodeir et al., 2012).

As can be seen in the Table 2.5 that studies on the $PM_{2.5}/PM_{10}$ ratio were only focusing on the exploratory analysis both in Malaysia and worldwide insinuating the needed of modelling the ratio of $PM_{2.5}/PM_{10}$. Many researchers have neglected the importance of the ratio of $PM_{2.5}/PM_{10}$. The knowledge of $PM_{2.5}/PM_{10}$ ratio can exhibit information regarding prediction of $PM_{2.5}$ without actually directly measure the $PM_{2.5}$. In fact, several successful proven in a study in California and Taipei (Blanchard et al., 2011; Chu, Huang & Lin, 2015), but these study lack of evidence of spatio-temporal variability.

Several study have successfully proven the indirect measurement of $PM_{2.5}$ using ratio of $PM_{2.5}/PM_{10}$ (Blanchard et al., 2011; Chu et al., 2015; Hwa-Lung & Chih-Hsin, 2010), a few other have research on the spatio-temporal variability (Zhao et al., 2019) , which proven that $PM_{2.5}/PM_{10}$ ratio varies in time and space; and numerous study conducted to demonstrate the variability of PM concentrations (Akinlade et al., 2015; Ghim et al., 2015; Hu et al., 2019; Zhang & Cao, 2015), but a little study has been conducted to predict the ratio of $PM_{2.5}/PM_{10}$ especially in Malaysia. To the knowledge of researcher, this study is the first research conducted to predict the ratio of $PM_{2.5}/PM_{10}$ in Malaysia.

Table 2.5
Summary of Past Studies related to $PM_{2.5}/PM_{10}$ Ratio

Author	Study Location	Objectives	Method	Findings
Wu et al. (2020)	China	To predict $PM_{2.5}/PM_{10}$ ratios	LSTM	LSTM able to predict higher ratio of greater than 0.8 and had better stability.
Zhao et al. (2019)	China	To investigate the relationship between the ratios and AQI.	Statistical method, Correlation analysis and Temporal Variations.	Higher ratio will result in larger possibility of higher AQI.
Coskuner, Jassim and Munir (2018)	Bahrain	To analyse long-term temporal trends of ratios	Theil-Sen approach	$PM_{2.5}/PM_{10}$ ratio showed a decreasing pattern in all seasons.
Mat Shukri et al. (2017)	Malaysia	To identify relationship between $PM_{2.5}$ and PM_{10} and meteorological factors	Descriptive Statistics	Ratio of $PM_{2.5}/PM_{10}$ is lowest in July and September while highest in month of August.
Blanco-Becerra, Gáfaro-Rojasv and Rojas-Roa (2015)	Colombia	To determine whether scavenging effect reduces ratio	Exploratory Analysis	$PM2_{.5}/PM_{10}$ ratio is reduced to 0.36.
Chu, Huang and Lin (2015)	Taiwan	To discover the spatio-temporal pattern of PM and predicting the ratios.	fuzzy c-means clustering, OLS regression, GWR and GTWR	GTWR provides a relatively high goodness of fit and sufficient space-time explanatory power.
(Hagen, 2004)	United State	To determine the relative breakage and $PM_{2.5}/PM_{10}$ ratios.	t-test and regression analysis	Predicted values of the study range between 0.1 to 0.3 with R^2 of 0.53
Parkhurst et al. (1999)	United State	To describe study of TVA dichotomous sampler $PM_{2.5}/PM_{10}$.	Summary Statistics	A strong association between $PM_{2.5}$ and PM_{10} is found with on the average of 67% contribution of $PM_{2.5}$ to the PM_{10}.

2.6 Conclusion

In conclusion, several gaps are identified from the review of past studies done on the issue of $PM_{2.5}/PM_{10}$ pollution. Firstly, the most obvious gap is limited study on the ratio of $PM_{2.5}/PM_{10}$ in Malaysia and even worldwide. Most of the studies focus on the $PM_{2.5}$ and PM_{10} pollutants individually and neglected the importance of examining the ratio of $PM_{2.5}/PM_{10}$. Zhao et al. (2019) has emphasized the importance of ratio that is crucial information can be obtained from ratio including the source of particles, formation process and impact of particulate matter derive from diverse sources and chemical properties towards human health. Hence, this study was conducted to fill the research gap.

Next, several studies have been conducted using numerous model which include the machine learning model and forecasting model to predict the $PM_{2.5}$ and PM_{10} concentrations. However, it has neglected the long-short temporal trend which is the nature of pollution data. LSTM models has proven to be superior and have number of success with time series data while considering long-short temporal trend (Belavadi et al., 2020). Besides, LSTM has been implemented in other country to predict the air pollution is found to be excellent in predicting, hence this study employed LSTM model to predict $PM_{2.5}/PM_{10}$ ratio in Malaysia.

CHAPTER THREE
METHODOLOGY

3.1 Introduction

This chapter discusses the methodology to achieve the objectives of the study mentioned in chapter 1. This chapter is partitioned into several parts. The first part defines the research design of the study and the justification on the chosen research design. Second part elaborate on the study location and the data used in this study. Third part illustrates the research framework followed by fourth part explaining on the data pre-processing. fifth part explained in detail each method used to achieve first objective and sixth part explained the theory of LSTM model and seventh part discussed in detail each process in model development of LSTM model. Lastly, the summary of the methodologies is reviewed to conclude this chapter.

3.2 Research Design and Method

This study concerned with the analysis on the ratio of $PM_{2.5}/PM_{10}$. The research design for this study is descriptive, exploratory, and longitudinal as it is involved in investigating the spatial and temporal pattern and distribution of $PM_{2.5}/PM_{10}$ ratios and modelling the time series of the ratios.

3.3 Data and Study Location

Secondary data is acquired for several study location. Data acquisition and study location are discussed in this section.

3.3.1 Data

Data of hourly $PM_{2.5}$ and PM_{10} concentrations are obtained from Department of Environment (DOE). The $PM_{2.5}$ and PM_{10} concentrations is recorded hourly in a span of two and half years (July 2017-Dec 2019) for several states in Western Coastal Region of Peninsular Malaysia including Klang, Selangor; Bukit Rambai, Melaka; Seberang Jaya, Pulau Pinang; Nilai, Negeri Sembilan; and Manjung, Perak. These are the

locations that often reported having high particulate matter concentration (Ya'acob & Mar Iman, 2020). A recent study conducted by Ya'acob and Mar Iman (2020) reported that these location have exceeded the PM_{10} concentration value guideline provided by New Malaysia Ambient Air Quality Standard (MAAQS) which is 50 μg/m3 especially Klang and Bukit Rambai with 1.36 and 1.32 times, respectively, more polluted than the national average. Besides, PM_{10} is contributed by the $PM_{2.5}$ which suggest that $PM_{2.5}$ and PM_{10} are highly related (Chu et al., 2015). Hence, an increase in $PM_{2.5}$ concentration will increase the PM_{10} concentration as well. Table below displayed the details of the dataset that will be implemented in this study.

Table 3.1
Details of Dataset

Location	Klang, Bukit Rambai, Seberang Jaya, Nilai and Manjung
Unit	mg/m^3
Dataset Span	2.5 years (hourly data)
Type of Data	Continuous
Air Pollution Parameter	$PM_{2.5}/PM_{10}$ ratio

In addition, ratio of $PM_{2.5}/PM_{10}$ are computed by dividing $PM_{2.5}$ concentration with PM_{10} concentration. The formula is as follows:

$$\text{ratio of } PM_{2.5}/PM_{10} = \frac{PM_{2.5} \text{ concentrations}}{PM_{10} \text{ concentrations}} \quad (3.1)$$

3.3.2 Location of Data

Malaysia has tropical weather all year, but due to its proximity to water, the climate is frequently humid. The mean relative humidity varies in range of 72% to 87%. Normally, the minimum relative humidity is discovered in January and February while maximum in the month of November. The chosen study area which located in west coast of peninsular Malaysia usually having temperature between 27°C to 32°C in daytime and 21°C to 24°C in nighttime with rainfall ranges in range of 200 cm to 400 cm annually. Climate in Malaysia is divided by two monsoon seasons which included south-west monsoon and north-east monsoon (MET, 2019).

South-west monsoon happened in late of May or early of June until end of October which resulting study area to be drier than any other months with high temperature and less rainfall. Meanwhile north-east monsoon occurs in November until March following year. Klang is an urban area situated in Selangor which comprised of housing area, commercial area and industrial activities. Meanwhile Seberang Jaya is a sub-urban site located in Pulau Pinang where the main economic activities are commercial services and business area, and this location is surrounded by residential areas and educational facilities. Bukit Rambai, Manjung and Nilai are located in the industrial hub in Peninsular Malaysia. Nilai is located in Negeri Sembilan with population approximately 38,612 in 2010 and has recorded the highest Average Daily Traffic (ADT) which indicate that this location has highest traffic-related air pollutants (Ya'acob & Mar Iman, 2020). Table 3.2 showed the coordinates and background of the study locations and selected study locations are depicted in the following Figure 3.1.

Table 3.2
Coordinates and Background of Study Locations

No	Location	Background	Latitude and Longitude
1	Klang	Urban	3.0449° N, 101.4456° E
2	Bukit Rambai	Industrial	2.2530° N, 102.1890° E
3	Seberang Jaya	Sub-urban	5.3819° N, 100.3989° E
4	Nilai	Industrial	2.8025° N, 101.7989° E
5	Manjung	Industrial	4.4022° N, 100.7098° E

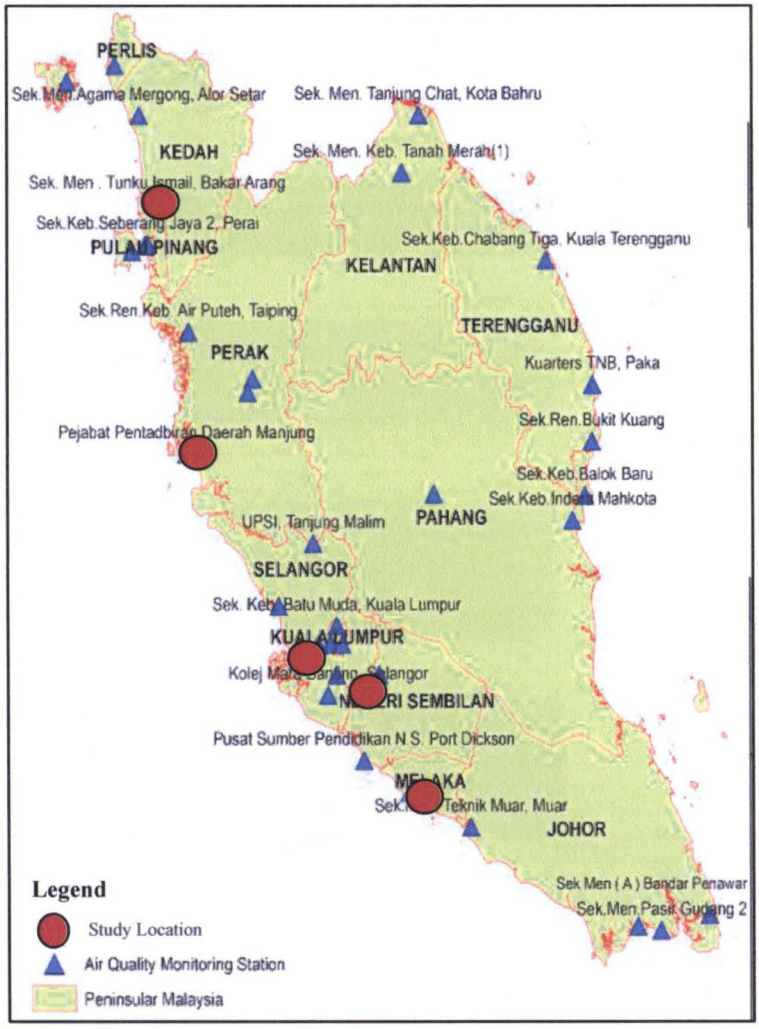


Figure 3.1 Location of Study Area (Source: Alyousifi et al., 2021)

3.4 Research Framework

There are several steps implemented to achieve the objectives of the study. First step is acquiring data from DoE. Second step is data examination in pre-processing process to ensure the quality of data and able to build a good model. Descriptive methods for spatial and temporal analysis are utilized to achieve the first objective in third step. Fourth step is developing the LSTM model. An experimentation to assess the effect of data splitting ratio into training and testing on the model performance as well as the effect of different hyperparameter, epoch size and number of LSTM layer were conducted. Training dataset is used to establish the model while model evaluation is performed using test dataset. Lastly, the third objective is achieved by comparing the performance of the best LSTM models obtained in objective 2 using several performance indicators. The research framework of this study is illustrated below.

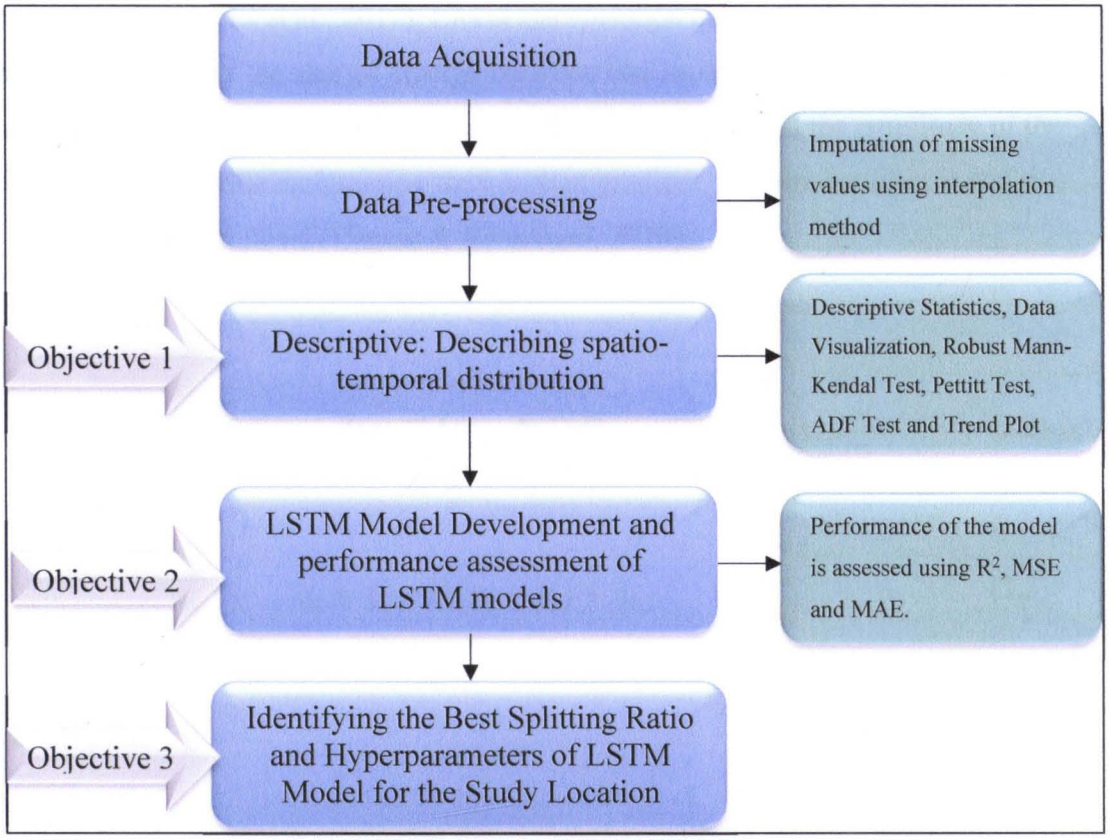


Figure 3.2 Flow of Research Methodology

3.5 Data Pre-Processing

Data pre-processing played an important role in ensuring good quality of data. This stage concern with diagnosing the data and one of important steps in this stage is data imputation.

3.5.1 Missing Values

Missing value in time-series data is one of main issue that needs to be cater. Missing values for some timestamps might be discovered during pre-processing stage. Hence, it is crucial to deal with those missing data in time-series data even if it is in small amount. Linear interpolation is one of a popular method in estimating hourly monitoring data of PM (Bekkar et al., 2021). To handle the issue of missing data, popular method known as linear interpolation will be implemented to fill the missing values in the dataset (Che et al., 2018). In this method, before and after of similar timestamps will be identified and each missing value will be filled with average value

of those timestamps (Li et al., 2017). Linear interpolation method indicate that constant gradient is involve in the rate of change between one sample point to the next point. This assumption implies that the amplitude of the i^{th} point is x_i and the amplitude of the $i + 1^{th}$ point is x_{i+1} while maintaining the constant gradient, j^{th} point between x_i and x_{i+1} can be computed as follows (Usman & Ramdhani, 2019):

$$\frac{x_{i+1} - x_i}{(i+1) - i} = \frac{x_j - x_i}{j - i} \tag{3.2}$$

Or

$$x_j = (j - i)(x_j - x_i) + x_i \tag{3.3}$$

3.5.2 Outlier

Outlier is the observations that found to be inconsistent with the rest of the data. In the context of air pollution, measurements obtained from automatic continuous air quality monitoring station might contain outlier. The outlier occurs due to the several reasons such as measurement error, presence of external factors affecting the observed variable and natural variability of pollutants in the atmosphere (Veselík et al., 2020). In the context of ratio of $PM_{2.5}/PM_{10}$, the outlier exist when $PM_{2.5}$ is larger than PM_{10} and this can be described as unreasonable phenomenon since $PM_{2.5}$ is a subset of PM_{10} (Spandana et al., 2021). Hence, existence of outlier due to the measurement error is assigned as missing value and the value is replaced using interpolation method. Besides, deviation of measurement from the rest of air pollution data signifies as extreme phenomenon. Therefore, outlier is not removed from dataset as this study employed LSTM model which known as a flexible method that able to predict volatile movements in the data (Shah et al., 2019) and presence of outliers do not influence the performance of the model. All the analysis conducted in this study also does not influence by the outlier.

3.6 Describing Trend and Homogeneity Pattern

Descriptive analysis is performed in third phase to achieve the second objective. Descriptive analysis is conducted to investigate the spatial variability and the temporal trend of $PM_{2.5}/PM_{10}$ ratios at the study locations. This phase involved descriptive statistics such as mean and median, minimum value, maximum value, skewness as well as kurtosis. Several data visualization is used including histogram to visualize the distribution of the data and spider chart to display the frequency of the extreme cases. Besides, stationarity, homogeneity and trend analysis which include ADF test, Pettitt test, and Robust Mann-Kendall test are also conducted.

3.6.1 Robust Mann-Kendall

Mann-Kendall is a non-parametric test which implies that it does not required to follow certain distribution. Another reason this method is preferred because it is less affected by missing values in the dataset. Given in equation (3.4) is the Mann-Kendall Statistic (S) (Gilbert, 1987).

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sign(x_j - x_k) \tag{3.4}$$

Where x_j and x_k represent the ratio of $PM_{2.5}/PM_{10}$ values at the j^{th} and k^{th} point, respectively, n signify the number of observations in this study and $sign(x_j - x_k)$ is the sign function in equation (3.5).

$$sign(x_j - x_k) = \begin{cases} +1, if\ x_j - x_k > 0 \\ 0, if\ x_j - x_k = 0 \\ -1, if\ x_j - x_k < 0 \end{cases} \tag{3.5}$$

Robust Mann-Kendall test is performed in this study to test whether the time series have a monotonic upward or downward trend. Monotonic indicate a consistent gradual change in direction over time. Robust Mann-Kendall is analysed rather than original Mann-Kendall due to its robustness towards presence of autocorrelation which

make it easier to detect trend (Mallick et al., 2021). Robust Mann-Kendall test with p-value < α (0.05) suggests presence of monotonic trend (Hamed & Rao, 1998) and tau value ranges from 1 to -1 suggest a decrease or increase of monotonic trend (Brossart et al., 2018).

3.6.2 Augmented Dickey-Fuller (ADF) Test

Data that has not too large variance and approaching to the average value is said to be stationary (Bierens & Song, 2006). There are several ways to test the stationarity of the time series data including Augmented Dickey-Fuller (ADF) test. Null hypothesis of ADF test stated that time series data can represent the unit root that is non-stationary. When ADF values have large negative values, the null hypothesis will be rejected which indicate that the model is stationary. Stationary data is indeed easier to model as it have more defined pattern (Hamami & Dahlan, 2020). However, it is discovered that LSTM is known to be a solution to deal and model non-linear and non-stationary data (Press, 2018).

3.6.3 Pettitt Test

Pettitt test (Pettitt, 1979) is a nonparametric approach adapted from rank-based Mann-Whitney test in detecting any abrupt shift point in the series. The reason Pettitt test is widely applied in time series because it is highly sensitive to breaks in time series (Baruah et al., 2022). In this study, Pettitt test is employed to determine possible abrupt changes in ratio of $PM_{2.5}/PM_{10}$ rather than a stable period of upward and downward trend. The null hypothesis established as $H_0: k^* = n$, which imply there is no change of point in the series, against alternative hypothesis $H_1: 1 \leq k^* < n$ indicating there is a change of point in the series. Equation (3.6) and ((3.7) are used in this test to detect shift point:

$$U_{k,n} = \sum_{i=1}^{k} \sum_{j=k+1}^{n} D_{ij} \tag{3.6}$$

$$D_{ij} = sgn(X_i - X_j) = \begin{cases} 1, if\ X_i - X_j > 0 \\ 0, if\ X_i - X_j = 0 \\ -1, \quad Otherwise \end{cases} \tag{3.7}$$

Statistic of $U_{k,n}$ signifies whether distribution of time series $\{X_1, X_2, \ldots, X_{k^*}\}$ and $\{X_{k*+1}, X_{k*+2}, \ldots, X_n\}$ are equal. As for t and $U_{k,n}$, where $1 \leq k < n$, the Pettitt test is expressed in equation (3.8).

$$K_n = \max_{1 \leq k < n} |U_{k,n}| \tag{3.8}$$

The distribution is limited to $2\exp\{-6k^2/(n^2 + n^3)\}$.

3.7 LSTM Model: Theory and Concept

LSTM stands for Short-Term Long-Term Memory. LSTM is one of the best methods in handling time-series data or spatial temporal reasoning (Awan, Minerva & Crespi, 2020). In contrast to traditional neural network, LSTM employ memory unit rather than neurons. It is a type of artificial neural network designed to recognize patterns in sequences of data which takes time and sequence into account, or in other words, it can handle data with a temporal dimension. Besides, this study implemented LSTM with RNN instead of simple RNN since the simple RNN cannot handle high volume of samples over several thousand timesteps which would be resulting in major computational bottleneck and poor accuracy of the model. It is known that air pollution temporally varied and long-term exposure to PM effect the human health, hence the best predictor to predict future PM ratios is the previous PM ratios over long periods of time. The illustration of the difference between simple RNN and LSTM with RNN models are presented in Figure 3.3 and Figure 3.4, respectively.

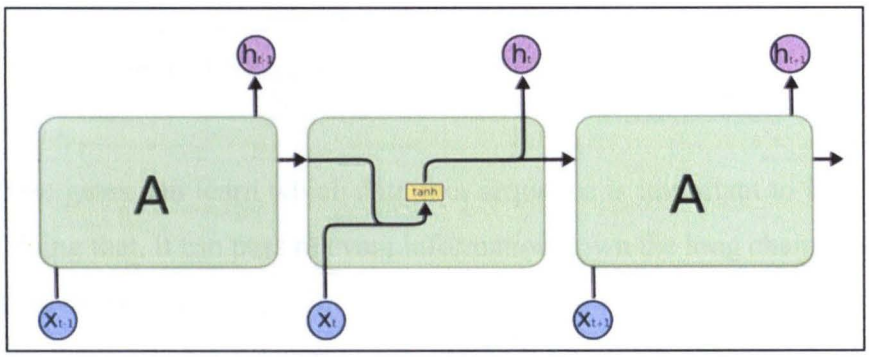


Figure 3.3 Simple RNN (Source: Olah, 2018)

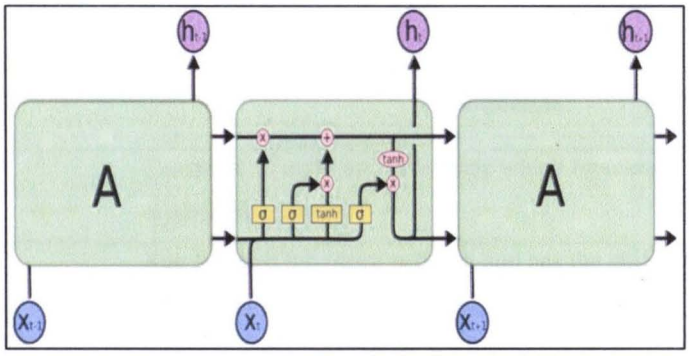


Figure 3.4 LSTM RNN (Source: Olah, 2018)

Figure 3.3 displayed simple RNN model with one layer and no gated memory cells while Figure 3.4 illustrate the four layers LSTM model with gated memory and sigmoid activation function. LSTM is a type of RNN introduced by Hochreiter and Schmidhuber in 1997 (Hochreiter & Schmidhuber, 1997) with the ability to recognize the sequential data and predict the future scenario while considering the temporal dimension. The usual hidden layer is substituted with LSTM cells that comprised with several gates functioning as controller of the input flow as illustrated in the Figure 3.5.

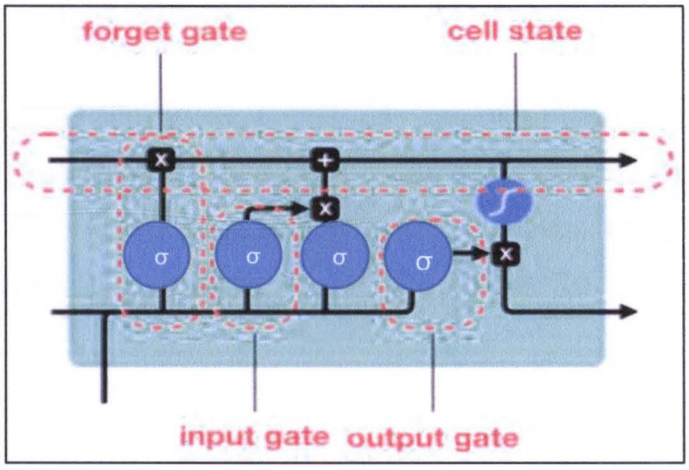


Figure 3.5 General Framework (Source: Phi, 2018)

These gates can learn which data in a sequence is important to keep or throw away. By doing that, it can pass relevant information down the long chain of sequences to make predictions. These operations are used to allow the LSTM to keep or forget information. Table 3.3 shows several gates and its function (Partheeban, 2021).

Table 3.3
LSTM Gates and its Function

Gate	Function
Input Gate	Comprise of input and determine which input value should be used to modify the memory
Cell State	Run through the entire network and has the ability to add or drop the information with the help of gates.
Forget Gate Layer	Decides which fraction of information that should be discarded or kept which decided by the sigmoid function.
Output Gate	Consist of output generated by the LSTM and decide the output using the input and block memory.
Sigmoid (σ) layer	Generates number between 0 to 1 to describe how much of each component should be allowed to pass through. The network also decides which values is not important and can be forgotten and which values to let through the passing point.
Tanh Layer	Generates new vector which will be added to the state and gives weightage to the values that are passed deciding their level of importance ranging from -1 to 1. As the vectors going through the neural network and undergoes various mathematical operation, the value continues to be multiplied. Hence, Tanh functions ensure that values will always be between -1 to 1.

As displayed in table above, LSTM cells consists of various gate that act as the controller of the input flow. In addition, it is also comprising of sigmoid layer, tanh layer and point wise multiplication operation **(x)** and addition **(+)**. The mathematical symbols and notation for each gate is represented in Figure 3.6.

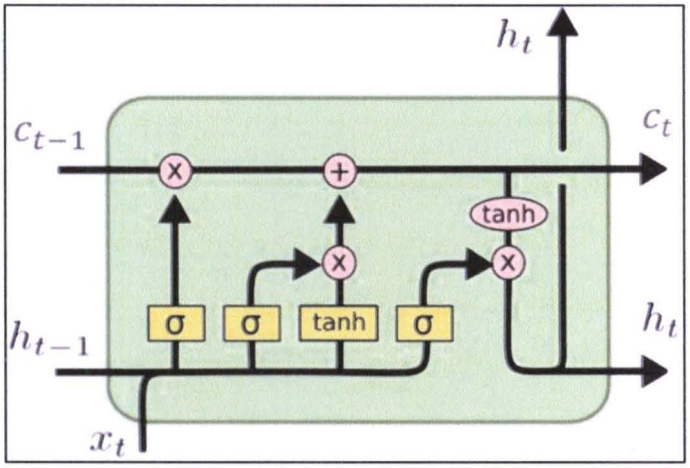


Figure 3.6 LSTM the framework of gates (Source: Olah, 2018)

The cell state is updated based in the gate output and mathematically, it can be represented in the following equations (Selvin et al., 2017):

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f \quad (3.9)$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i \quad (3.10)$$

$$c_t = tanh(W_c.[h_{t-1}, x_t] + b_c \quad (3.11)$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o \quad (3.12)$$

$$h_t = o_t * tanh(c_t) \quad (3.13)$$

Where x_t is the input vector which is the historical ratio of $PM_{2.5}/PM_{10}$; h_t is the output vector which is the predicted ratio of $PM_{2.5}/PM_{10}$; $c_t$is the cell state vector; f_t is the forget gate vector; $i_t$is the input gate vector; $o_t$is the output gate vector and W,b are the parameter weight matrix and bias vector. σ is the sigmoid layer act as decision maker in deciding which input to pass through while tanh squish the value to be within -1 to 1. LSTM learns the two weights during propagation through time stamp *t*.

3.8 LSTM Model Development Process and Validation

To achieve the objective of building predictive model involving the dynamic series of $PM_{2.5}/PM_{10}$ ratios, a recent deep learning named LSTM model is explored. To identify the best appropriate model to be built, several experimentations on the splitting ratio, epoch size and number of LSTM layer are conducted. Development of LSTM model consists of several phases to obtain the model. The methodology to build the model is illustrated in Figure 3.7.

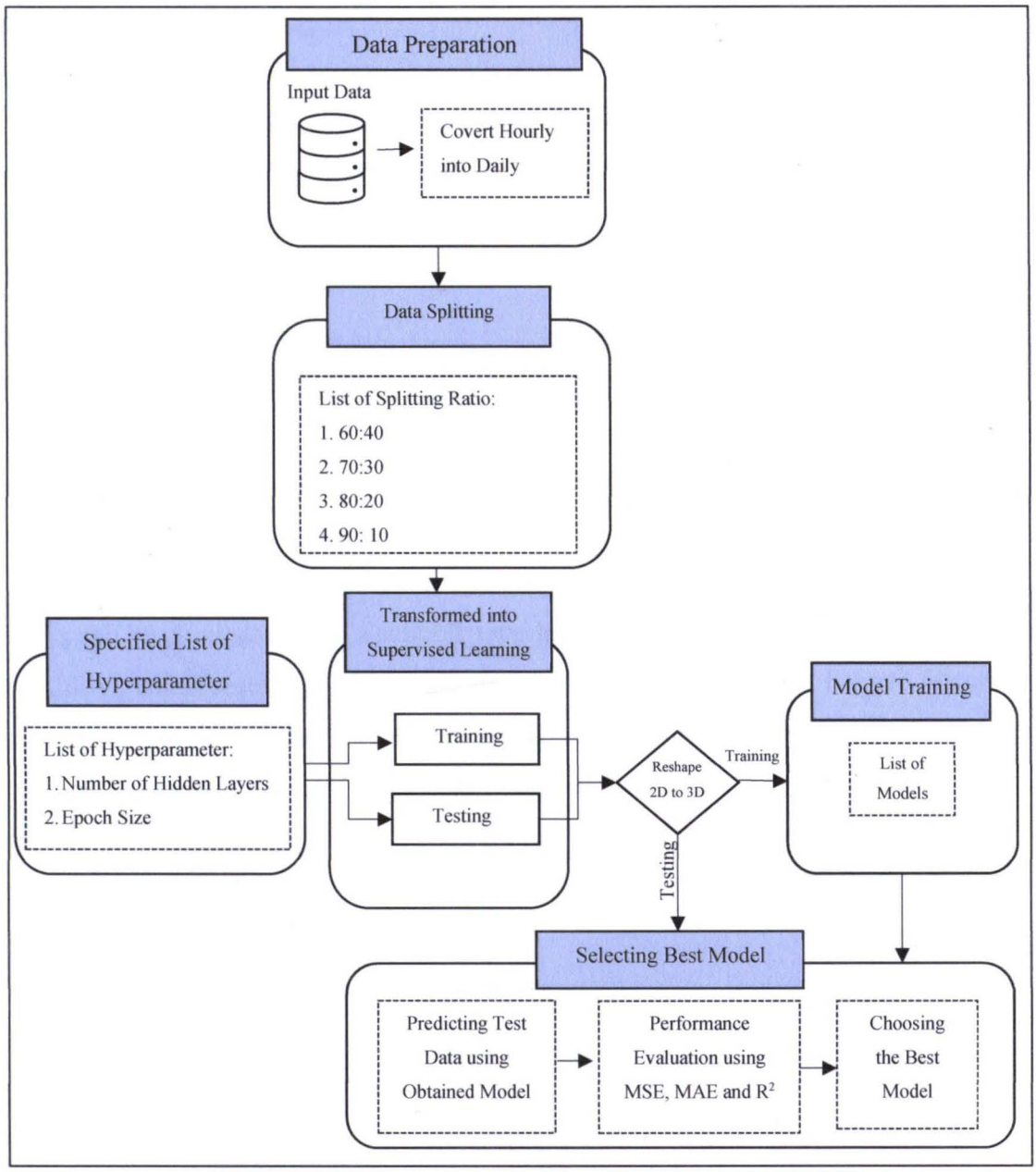


Figure 3.7 Process of LSTM Model Development

Each phase in the process of LSTM model development is discussed in the following subsection.

3.8.1 Data Preparation

In this study, daily maximum ratio of $PM_{2.5}/PM_{10}$ is used. The daily maximum ratios of $PM_{2.5}/PM_{10}$ is obtained by acquiring the maximum ratios in each day of timestamp using python.

3.8.2 Data Splitting

The data is split into specified splitting ratio. Four splitting ratios are examined to determine which splitting ratio produce the best performance of LSTM model. The splitting ratio that are being assessed are 60:40, 70:30, 80:20, and 90:10 for the proportion of train and test, respectively.

3.8.3 Transformed into Supervised Learning

Time series data need to be transformed into supervised learning mode to employed LSTM model and lag size is needed for the transformation to succeed. Lags is crucial in time series as prediction is obtained based on past values. Redundant features can be eliminated with a suitable time lag in forecasting a time series (Ribeiro et al., 2011), which resulting in better accuracy of the prediction model. Lagged feature acts as an input to feed the model with enough past values to predict future value. However, too many lags can cause overfitting to the model. Therefore, this study restricted the lagged size to 3. The reshaping process to supervised learning are illustrated in equation (3.14) and (3.15) (Abbasimehr et al., 2020).

$$S_{input} = [S_1, S_2, \dots, S_{N-L}]^T = \begin{bmatrix} S_n(t_1) & S_n(t_2) & \cdots & S_n(t_L) \\ S_n(t_2) & S_n(t_3) & \cdots & S_n(t_{L+1}) \\ \vdots & \vdots & \cdots & \vdots \\ S_n(t_{N-L}) & S_n(t_{N-L+1}) & \cdots & S_n(t_{N-1}) \end{bmatrix} \quad (3.14)$$

Where N is the total observation and L is size of lag. Equation (3.15) is the corresponding values of output:

$$S_{output} = [O_1, O_2, \dots, O_{N-L}] = \begin{bmatrix} S_n(t_{L+1}) \\ S_n(t_{L+2}) \\ \vdots \\ S_n(t_N) \end{bmatrix} \quad (3.15)$$

3.8.4 Specified List of Hyperparameter

Both datasets are tested using different hyperparameter to obtain the best parameter to predict daily $PM_{2.5}/PM_{10}$ ratios. Some parameters such as lag size, number of neurons, optimizer and activation function is predetermined in preliminary stage and employed in the model to assess the performance of LSTM model with different epochs size and number of layers. The hyperparameter are as follows:

3.8.4.1 Optimizer

The role of optimizer in a network is to minimize the objective function of the network. One of the most commonly employed optimizers is stochastic gradient descent (SDG) as it is proven to be efficient in optimizing large volume of published machine learning systems. However, the drawback of SGD is its sensitivity towards selection of learning rate resulting it difficult to tune in. Situation where when large rate is chosen, it might affect the divergence of the system in terms of objective, but small learning rate will slow down the learning process. Hence, several others optimization algorithm have been introduced to overcome such situation, including Adam (Kingma and Ba, 2014), Adagrad (Duchi et al., 2011), Adadelta (Zeiler, 2012), Nadam (Dozat, 2015), and RMSProp (Hinton et al., 2012).

This study implemented Adaptive moment estimation (Adam) optimizer as it does not needed data to be in stationary state and works well with sparse gradient. Besides, Adam optimizer only involve first-order gradients with small requirement of memory (Kingma & Ba, 2014). This optimizer has an ability to compute its own adaptive learning rate from first and second moment estimation of the gradient for different parameters.

3.8.4.2 Activation Function

Activation function act as facilitator in introducing the non-linearity to the learning process which causing the network to learn the complex patterns. Activation function is important in network as it can affect the complexity and performance of the network as well as the convergence of the algorithm (Gecynalda et al., 2011). There are several activation function, sigmoid (σ), hyperbolic tangent (tanh), Softmax and Rectified Linear Unit (ReLU) (Agarap, 2018). Most widely used activation function in recent studies is ReLu as it has excellent consistencies. Despite many other activation functions which include devised alternatives to ReLu have been introduced, but none of those function succeed in achieving the level of consistency of ReLu. This activation function is defined in equation (3.16).

$$f(x) = \max(x, 0) \tag{3.16}$$

ReLU is an identity function for the non-negative inputs and zero function for the negative inputs. Another well-known activation function is sigmoid (log-sigmoid) which has demonstrated to have a good performance upon being applied in LSTM network (Farzad et al., 2019). Therefore, these two activations function is implemented in the network of this study. Sigmoid activation function is defined in equation (3.17).

$$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.17}$$

3.8.4.3 Number of Neurons

Large number of neurons could overfit the training dataset, but small number of neurons causes LSTM model not to be able to memorize the important information for prediction (Reimers & Gurevych, 2017). Since this study is a univariate approach and only utilize one parameter which is the $PM_{2.5}/PM_{10}$ ratios, hence number of neurons in LSTM layer is controlled to small number which is 8 number of neurons in LSTM layer. Whereas single neuron with default linear activation function in the output layer of LSTM model is used to make prediction.

3.8.4.4 Epochs

One training epoch is described as one iteration over all training input (Reimers & Gurevych, 2017). Small number of epochs causes the model unable to capture the training patterns, but large number of epochs will overfit the model. It is not an easy task to obtain optimal size of epochs. Therefore, four different size of epochs are evaluated in this study which are 50, 100, 200 and 500 to achieve model with high performance.

3.8.4.5 Number of LSTM Layer

Selection of appropriate number of LSTM layers to be included is domain specific. Different number of optimal LSTM model can be obtained for each input feature and characteristics of data. Therefore, four different hidden layers are evaluated including one layer, two layers, three layers and four layers to determine which number of LSTM layers is appropriate for ratios of $PM_{2.5}/PM_{10}$ data. Each layer has the same number of neurons which is 8.

3.8.5 Selection of the Best Model

The performance of both training and testing models are evaluated using mean absolute error (MAE), mean squared error (MSE) and coefficient of Determination (R^2). The final model is obtained with the smallest MAE and MSE, and highest R^2. The mathematical representations of the evaluation metrics are as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - O_i| \tag{3.18}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2 \tag{3.19}$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}(P_i - \bar{P})(O_i - \bar{O})}{n.S_{pred}.S_{obs}}\right)^2 \tag{3.20}$$

Where,

N = Total number of measurements at specific site

P_i = Predicted values of $PM_{2.5}/PM_{10}$ ratios

O_i = Observed values of $PM_{2.5}/PM_{10}$ ratios

$\bar{P}$ = Mean of predicted values of ratio of $PM_{2.5}/PM_{10}$

$\bar{O}$ = Mean of the observed values of $PM_{2.5}/PM_{10}$ ratios

S_{pred} = Standard deviation of predicted values

S_{obs} = Standard deviation of observed values

MAE is very valuable for continuous data as it is insensitive to outliers and does not deal with big errors while MSE is valuable for dataset containing outliers (Awan, Minerva & Crespi, 2020). Besides, R^2 will be used to measure the ability of the model to explain and predicts the future outcomes. Smaller MAE and MSE and higher R^2 indicate better prediction. Another method employed in this study to measure the predictive performance of LSTM model is loss per epoch graph where the loss functions is assigned to mean square error. Displayed in Figure 3.8 is the illustration of graph with good fit. As shown in Figure 3.8, a good fit of learning curve is when the training loss and validation loss reach the point of stability but should expect some gap between training and validation loss. In this study, testing loss is used instead of validation loss. Continuation of training a good fit will causes overfitting issue.

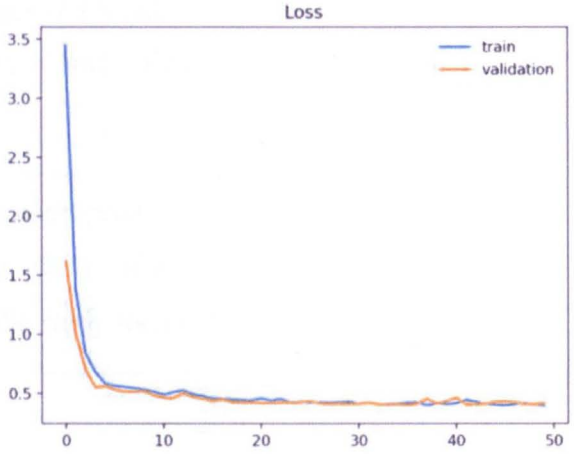


Figure 3.8 Figure of Good Fit Case

3.9 Summary

In summary, there are three objectives that have been achieved in this study. First objective is to investigate the spatial and temporal distribution of $PM_{2.5}/PM_{10}$ ratios at several location in the West Coast region of Peninsular Malaysia which is achieved by performing descriptive analysis, modified Mann-Kendall, Pettitt Test and trend plot. Second objective is to investigate the influence of splitting ratio, epochs size and number of LSTM layer on LSTM model for $PM_{2.5}/PM_{10}$ ratios at the study locations. The third objective is to determine the most appropriate LSTM model in predicting $PM_{2.5}/PM_{10}$ ratios at the study locations based on the results obtained in objective 2. Below is the summary of the objectives and the method to achieve the stated objectives.

Table 3.4
Summary of Objectives and Method Implemented

Objectives	Method
To investigate the spatial-temporal distribution of $PM_{2.5}/PM_{10}$ ratios at several locations in the West Coast region of Peninsular Malaysia	Descriptive Analysis, Modified Mann-Kendall, Pettitt Test and Trend Plot
To investigate the influence of splitting ratio, epochs size and number of LSTM layer on the performance of LSTM model of daily maximum $PM_{2.5}/PM_{10}$ ratios at study locations.	LSTM modelling
To determine the most appropriate LSTM model in predicting daily maximum $PM_{2.5}/PM_{10}$ ratios at the study locations	MSE, MAE, and R^2

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 Introduction

This chapter examined the results obtained to achieve the objectives of the study using the method discussed in Chapter 3. The data have been pre-processed where missing values were replaced using interpolation method. Descriptive analysis, modified Mann-Kendall, Pettitt Test, and trend plot was analysed using R programming to discern the spatial and temporal distribution of $PM_{2.5}/PM_{10}$ ratios while LSTM model was built and assessed using Python. In this chapter, section is divided into three where first section discussed the results to achieve first objective; investigating spatial and temporal distribution of $PM_{2.5}/PM_{10}$ ratios at selected study locations, second section examined the results to satisfied second objective; investigating the influence of splitting ratio, epochs size and number of LSTM layer on LSTM model for $PM_{2.5}/PM_{10}$ ratios at study locations, followed by third section discussed the results to achieve third objective; determining the most appropriate splitting ratio and hyperparameters of LSTM model in predicting $PM_{2.5}/PM_{10}$ ratios at the study locations and lastly, the conclusion for this chapter.

4.2 Data Pre-processing

Descriptive analysis of before and after imputation was obtained to highlight the difference of mean, median, minimum, and maximum value before and after the imputation was performed on the dataset. Descriptive analysis before the data being imputed is obtained in Table 4.1. Total number of hourly data for each location is 21864 observations. Each timesteps in time series data is unique and dropping the missing value would create a holes in the series (Setiawan, 2020). Therefore, this study implemented linear interpolation to interpolate the missing value. Interpolation method required before and after timesteps value to be able to interpolate, however, first timesteps in Bukit Rambai dataset was missing. Thus, the missing value was replaced with mean value of particular day in Bukit Rambai. Bukit Rambai station has the highest percentage of missing value of 5.68% (1241 out of 21864) while Seberang Jaya station

has the smallest percentage of missing value of 0.75% (165 out of 21864). The mean, median, minimum, and maximum value before imputation was compared to after the dataset being imputed.

Table 4.1
Table of Missing Value Before Imputation

Station	Number of Observations	Missing Value (%)	Mean	Median	Minimum	Maximum
Bukit Rambai	21864	5.68% (1241)	0.8050	0.6736	0.0012	19.8450
Klang	21864	1.68% (368)	0.7006	0.7154	0.0076	0.9991
Manjung	21864	1.40% (307)	0.6961	0.7188	0.0088	0.9992
Nilai	21864	2.27% (496)	0.6880	0.7085	0.0035	0.9997
Seberang Jaya	21864	0.75% (165)	0.7104	0.7159	0.0115	0.9999

Next, descriptive statistics after the imputation performed on the dataset was obtained and displayed in Table 4.2. The mean, median, minimum, and maximum value of Bukit Rambai which had the highest percentage of missing value were 0.8050, 0.6736, 0.0012, and 19.8450, respectively. After the imputation was performed on the dataset, the mean, median, minimum, and maximum value were 0.6822, 0.7053, 0.0081, and 0.9985, respectively. High value of maximum value was due to instrument malfunction which resulted in higher concentration in $PM_{2.5}$ compared to PM_{10}. This was an unreasonable phenomenon since $PM_{2.5}$ is a subset of PM_{10} (Spandana et al., 2021). Hence, the data was assigned as missing value and interpolated to replace those missing value. Meanwhile, Seberang Jaya had the smallest percentage of missing value with mean, median, minimum, and maximum value before imputation of 0.7104, 0.7159, 0.0115, and 0.9999, respectively, and after imputation of 0.7105, 0.7160, 0.0109, and 0.9999, respectively. It can be concluded that the mean, median, minimum, and maximum value of all station had small changes after the imputation was performed on the data except for Bukit Rambai station.

Table 4.2
Table of Missing Value After Imputation

Station	Mean	Median	Minimum	Maximum
Bukit Rambai	0.6822	0.7053	0.0081	0.9985
Klang	0.6996	0.7142	0.0076	0.9991
Manjung	0.6961	0.7189	0.0088	0.9992
Nilai	0.6859	0.7076	0.0035	0.9997
Seberang Jaya	0.7105	0.7160	0.0109	0.9999

4.3 Investigating Spatial-temporal Distribution of $PM_{2.5}/PM_{10}$ Ratios at Study Locations

Terminologically, spatial-temporal refers to location and time, respectively. This study utilized time series data of $PM_{2.5}/PM_{10}$ ratio that can be described by coherent patterns across both space and time where ratios of different locations were assessed at several temporal period such as daily and monthly period. Ratios of $PM_{2.5}/PM_{10}$ from July 2017 to December 2019 had been analysed and the descriptive analysis, histogram, spider chart, statistical analysis, and trend plot were obtained to investigate the spatial-temporal distribution of $PM_{2.5}/PM_{10}$ Ratios.

4.3.1 Descriptive Analysis of $PM_{2.5}/PM_{10}$ ratios

Descriptive statistics of $PM_{2.5}/PM_{10}$ ratios for all station for the recorded data from July 2017 to December 2019 was obtained to gain better understanding on the behaviour of ratios data. The descriptive statistics was tabulated and displayed in Table 4.3. It can be seen from Table 4.3 that there was fairly large difference between the median ratio and maximum ratio for all the stations indicating a presence of outliers in the dataset. This is due to the fact that nature of PM concentration tends to increase above average and suddenly decrease during peak hour of the day showed the volatility of the data itself. Since LSTM model able to predict volatile movements in the data, outliers were not omitted from the dataset.

Table 4.3
Table of Descriptive Analysis of Ratios July 2017-Dec 2019

Station	Bukit Rambai	Klang	Manjung	Nilai	Seberang Jaya
Median	0.71	0.71	0.72	0.71	0.72
Standard Deviation	0.15	0.15	0.15	0.14	0.14
Minimum	0.008	0.008	0.009	0.003	0.011
Maximum	0.998	0.999	0.999	0.999	0.999
Skewness	-0.89	-0.69	-0.97	-1.07	-0.62
Kurtosis	1.16	0.88	1.38	2.02	1.42

To have better understanding on the distribution of the ratios, histogram of $PM_{2.5}/PM_{10}$ ratios was presented Figure 4.1. Figure 4.1 depicts the distribution of $PM_{2.5}/PM_{10}$ ratios in each location where all locations were found to be skewed to the left. This correspond to the median value of all the stations that were larger than 0.5 suggesting presence of anthropogenic sources (Sugimoto et al., 2016). The highest frequency of hourly ratios was discovered to be ratio of 0.75 which was relatively higher than a cut-off point given by Sugimoto et al. (2016). This showed that most of the time, contribution of $PM_{2.5}$ was high in the atmosphere at all locations when compared to PM_{10}.

Claimed made on the distribution of the histogram was supported by the skewness value in Table 4.3. Nilai station recorded highest skewness and kurtosis value of -1.07 and 2.02, respectively. All the stations had approximately similar median of $PM_{2.5}/PM_{10}$ ratios of 0.7. Overall, it was discovered that the ratios depicted high value which was larger than 0.5 and negatively skewed for all the station. The median ratios of study location was comparatively higher than reported median ratios of $PM_{2.5}/PM_{10}$ in United Kingdom which was 0.65 (Munir, 2017) and Oman with median ratio of 0.69 (Alattar et al., 2019).

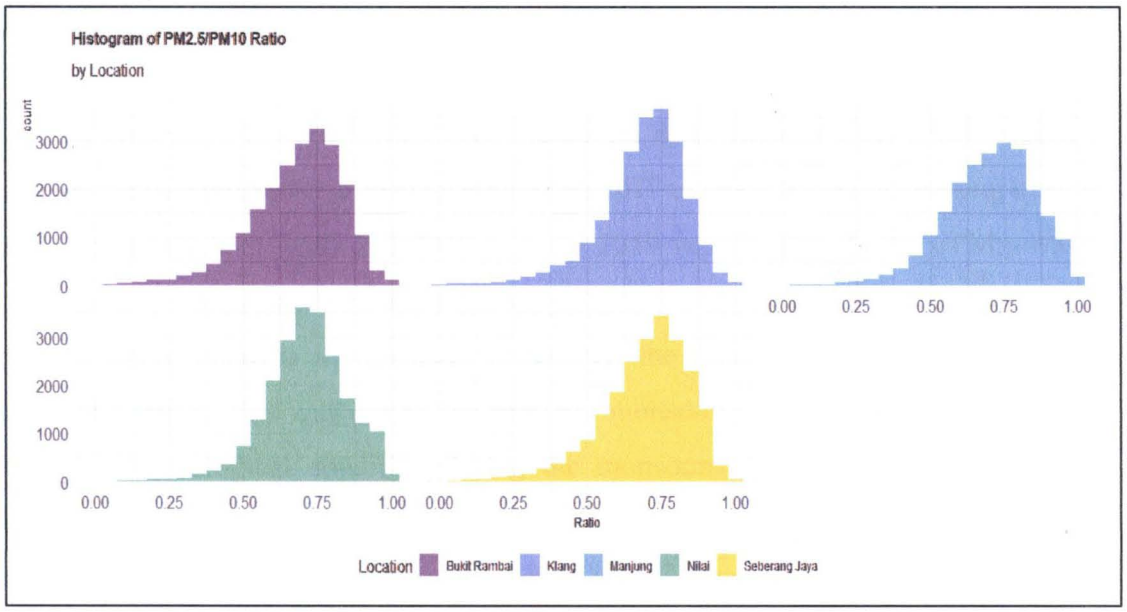


Figure 4.1 Histogram of Hourly $PM_{2.5}/PM_{10}$ ratios by Location

Several trend plots were visualized to describe the trend of $PM_{2.5}/PM_{10}$ ratios at the study location. Trend plots were visualized with two different temporal period where Figure 4.2 displayed daily median of $PM_{2.5}/PM_{10}$ ratios in 2017-2019 whereas Figure 4.3 exhibited daily maximum $PM_{2.5}/PM_{10}$ ratios in 2017 to 2019 for each study location. The daily median of $PM_{2.5}/PM_{10}$ ratios visualized to describe the trend of day-to-day ratio in 2017 to 2019. The trend plot revealed the characteristics of time series data is dynamic with sudden changes in the trend make forecasting task difficult. The NE, Inter I, SW and Inter II in Figure 4.2 and Figure 4.3 represent northeast monsoon, inter-monsoon I, southwest monsoon, and inter-monsoon II, respectively. Malaysia experienced four monsoon seasons which are Northeast Monsoon, first inter-monsoon, Southwest Monsoon and second inter-monsoon.

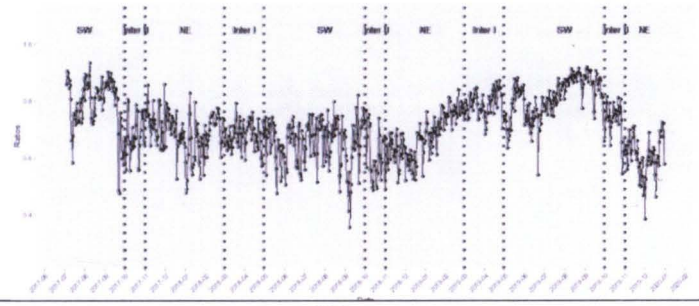
It can be seen that all industrial areas which were Bukit Rambai, Manjung and Nilai had similar pattern. However, Nilai showed different pattern during SW monsoon period in 2018 due to slight haze in mid-August coming from the local and transboundary haze from neighbouring countries. Fortunately, it lasted for short period of time due to humid weather condition (Department of Environment, 2018). Higher ratio was seen in 2019 compared to 2018 in industrial area due to the increase in emission of industrial sources in 2019 impacted the industrial area (Department of Environment, 2019). Besides, an increasing pattern in 2019 compared to 2018 was seen in Bukit Rambai, Klang, Manjung and Seberang Jaya. This was due to the

transboundary haze coming from neighbouring country in certain periods (Department of Environment, 2019).

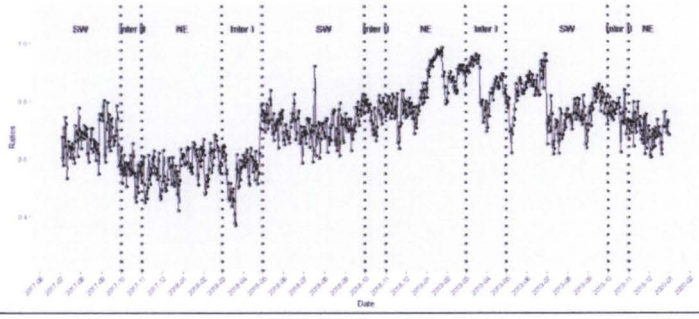
In contrast, Figure 4.3 illustrates maximum daily ratio of $PM_{2.5}/PM_{10}$ which described the extreme cases in each location from 2017 to 2019. A clear increasing trend was detected in Klang from mid-2018 to early 2019 where ratio of $PM_{2.5}/PM_{10}$ was approaching to 1 in December 2018 to January 2019. A noticeable trend was seen in all locations where changes in monsoon influenced the increasing and decreasing of $PM_{2.5}/PM_{10}$ ratios. This was due to the seasonal monsoon variation which associated with the meteorological condition during the monsoons apart from anthropogenic emission. During SW monsoon, the trend was more likely to increase due to the dry weather in June to September and similar pattern was discovered in NE monsoon season due to the hot weather. This showed that seasonal monsoon influenced the changes in ratio of $PM_{2.5}/PM_{10}$ due to the meteorological conditions in particular monsoon season. Similar deduction was made in China indicate that seasonal changes in China affected the ratio of $PM_{2.5}/PM_{10}$ within the country where ironically, higher ratio was found in winter season due to the stagnant weather condition and high emission from heating appliances (Fan et al., 2021).

In summary, the background of the location and changes in seasonal monsoon highly influence the ratio of $PM_{2.5}/PM_{10}$ due to the emission of anthropogenic sources and meteorological condition within each monsoon season.

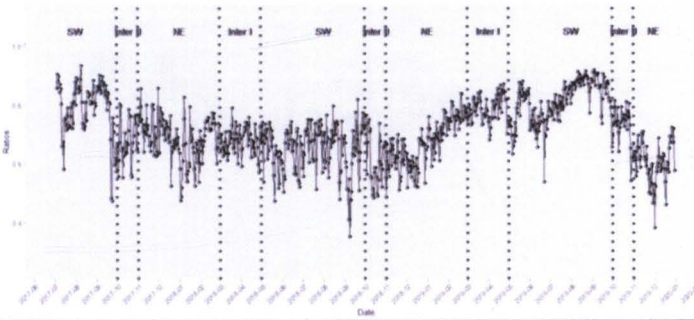
Bukit Rambai



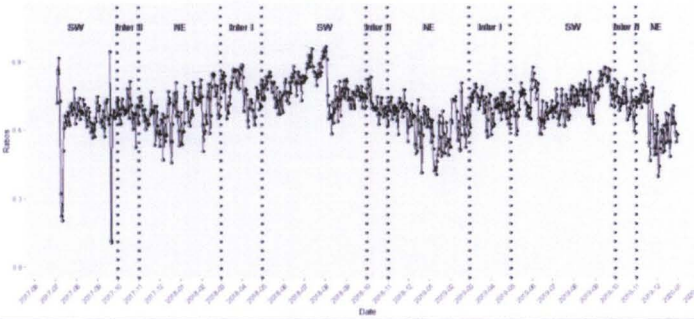
Klang



Manjung



Nilai



Seberang Jaya

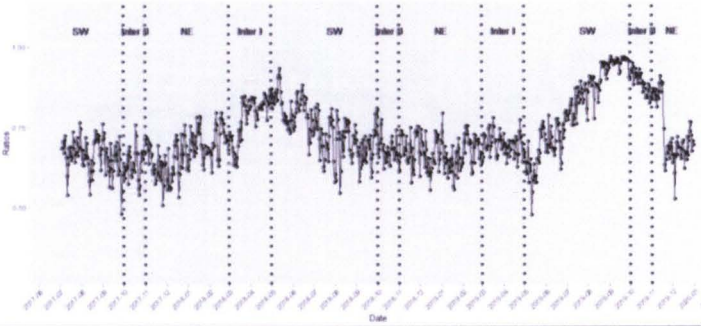


Figure 4.2 Trend Plot of Daily Median $PM_{2.5}/PM_{10}$ Ratios by Location

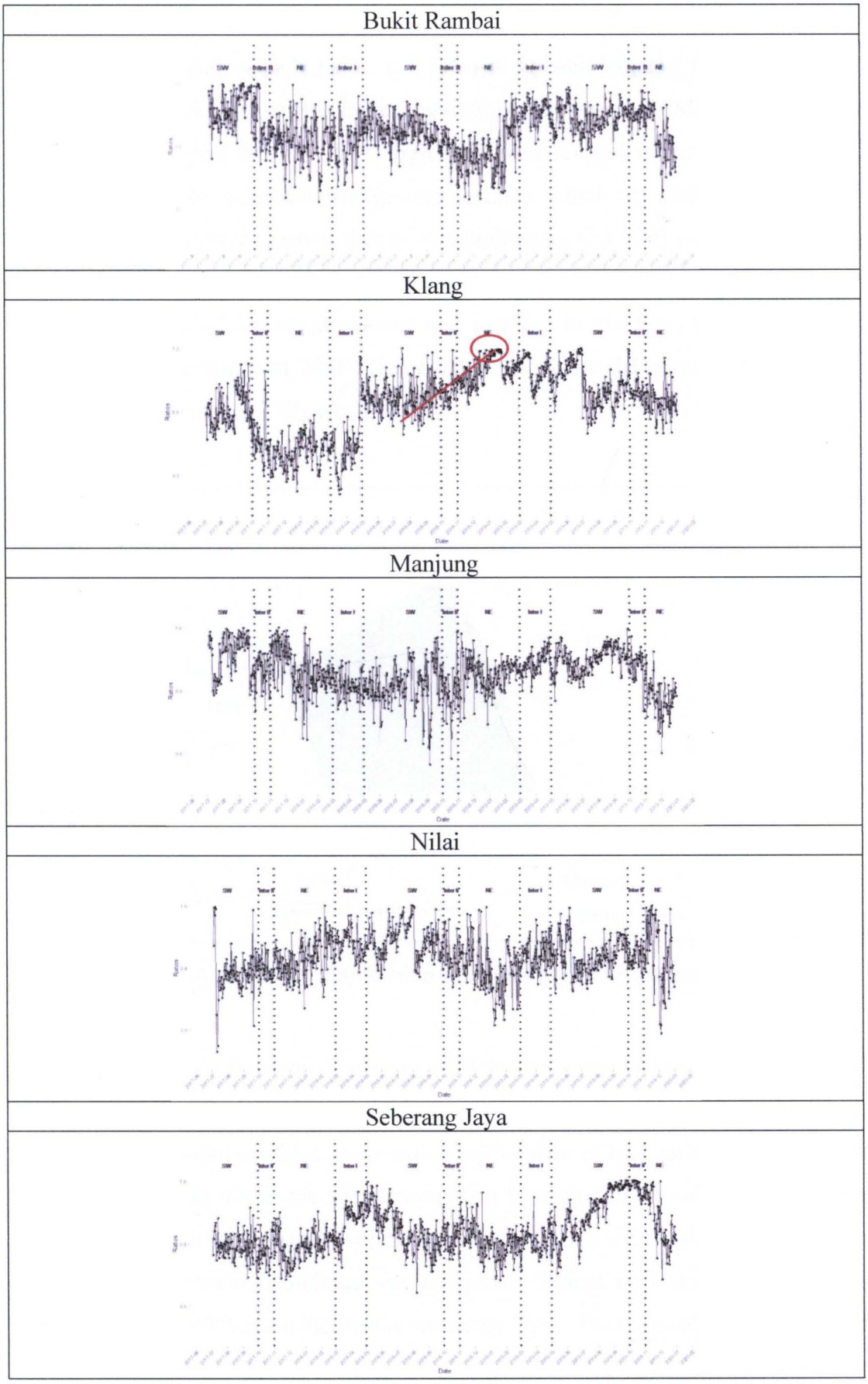


Figure 4.3 Trend Plot of Daily Maximum $PM_{2.5}/PM_{10}$ Ratios by Location

Detail number of high ratios, (>0.7) cases in each location was displayed in Figure 4.4. Figure 4.4 depicts the frequency of high ratio cases of $PM_{2.5}/PM_{10}$ ratios in study location for year 2017 to 2019. High cases were referred to maximum ratio that exceeded 0.7. Both industrial background location which are Seberang Jaya and Manjung had the highest number of days with high cases. Out of 911 days in 2.5 years, 889 days had the ratio of larger than 0.70 and this is concerning. This might be due to the large emission of industrial sources and increase in number of motor vehicles (Department of Environment, 2018). Surprisingly, Klang had the lowest number of high ratio cases with a total of 770 cases.

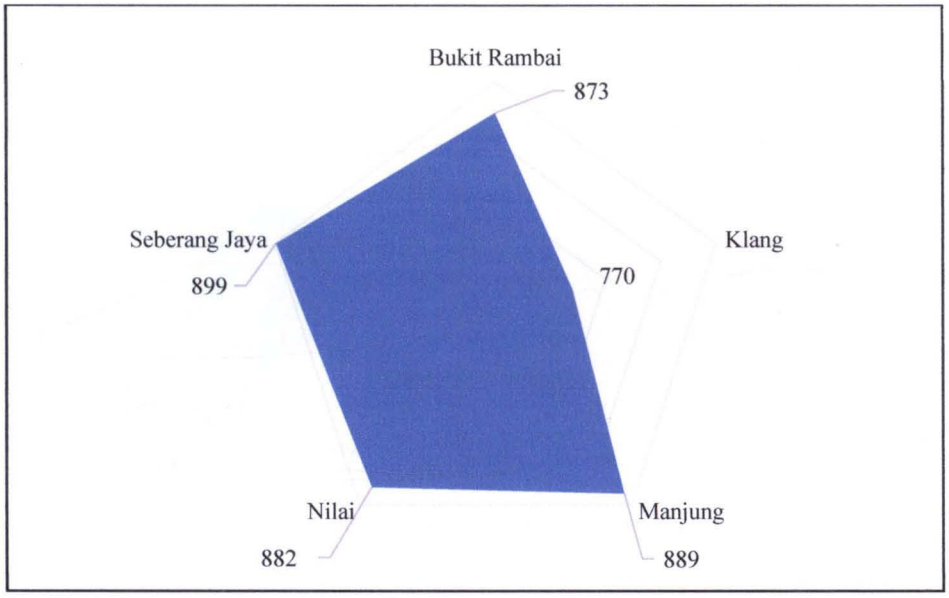


Figure 4.4 Frequency distribution of High $PM_{2.5}/PM_{10}$ ratios (>0.7)

Spider Chart of monthly frequency of high ratios cases by location was obtained in Figure 4.5 to explore further on the high ratio cases in monthly period. The figure depicts similar pattern in all locations with small number of high ratios cases in month of January to March due to the NE monsoon that brought a heavy rain. It showed that $PM_{2.5}$ was easier to be deposited within wet weather in line with findings by Wu et al. (2018) indicating that wet condition was 92% more efficient in eliminating $PM_{2.5}$ while dry condition was 63% more inclined in removing PM_{10}. The ratios of $PM_{2.5}/PM_{10}$ tend to drop when $PM_{2.5}$ decreases proved that ratio of $PM_{2.5}/PM_{10}$ can described the contribution of the $PM_{2.5}$ itself.

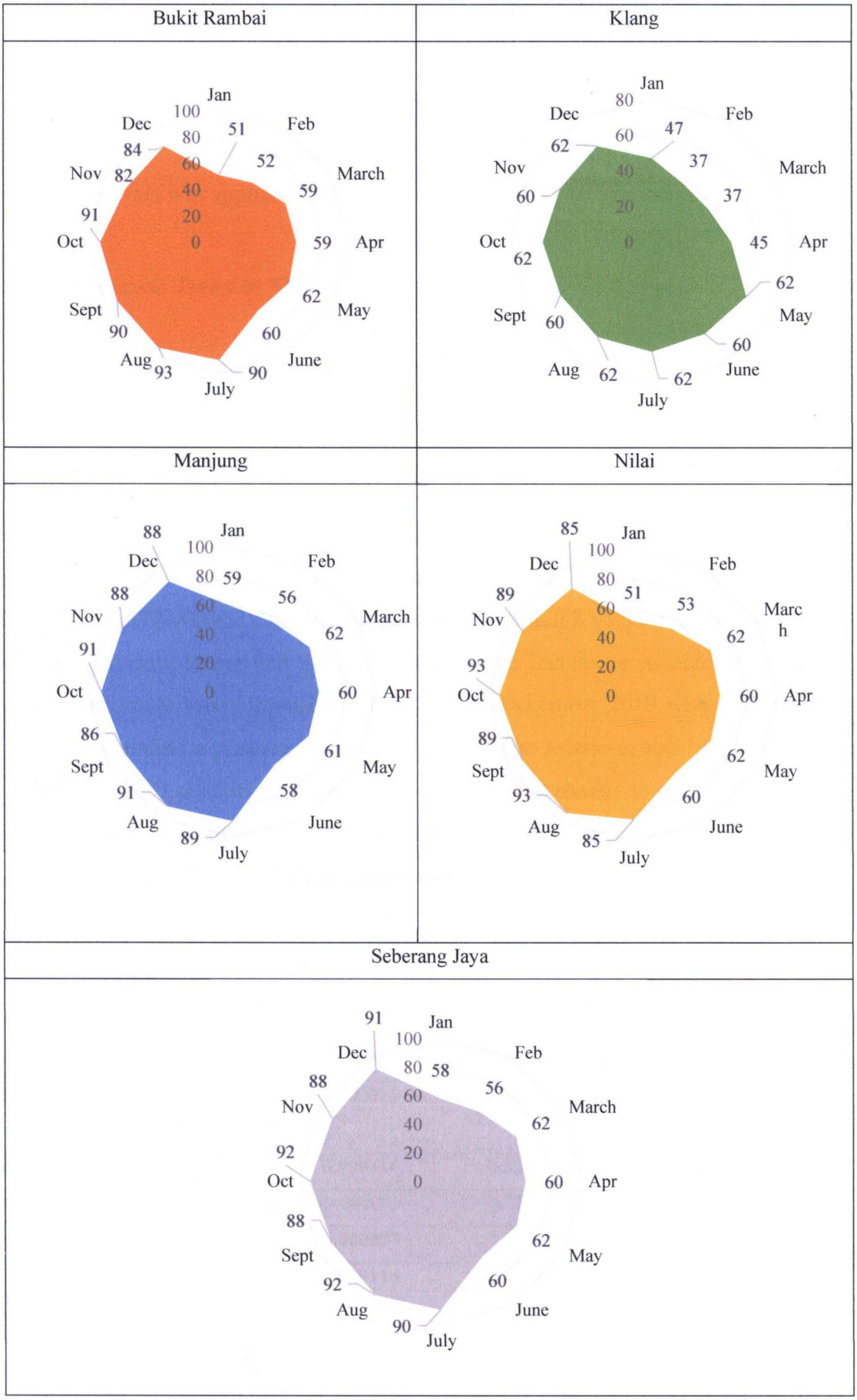


Figure 4.5 Frequency Distribution of High $PM_{2.5}/PM_{10}$ Ratio (>0.7) by Month

4.3.2 Trend, Homogeneity and Stationarity of $PM_{2.5}/PM_{10}$ Ratios

Robust statistical analysis was used in this study to statistically test the monotonic trend, abrupt changes and stationarity of $PM_{2.5}/PM_{10}$ ratios. Robust statistical analysis was applied due to the nature of the $PM_{2.5}/PM_{10}$ ratios data.

4.3.2.1 Monotonic Trend in Time Series

Presence of monotonic trend in the $PM_{2.5}/PM_{10}$ ratios can be verified using Mann-Kendall test. The monotonic trend of $PM_{2.5}/PM_{10}$ ratio was tested, and the result of Robust Mann-Kendall test is tabulated in Table 4.4.

Statistically significant monotonic trend for daily maximum ratios was discovered across three station which were Klang, Nilai and Seberang Jaya with p-value less than 0.05. These stations had positive Tau value indicating that there was a positive trend with a small average change in each day due to small S value. Meanwhile, result indicated that only Klang and Seberang Jaya had an increasing monotonic trend of monthly maximum ratios throughout July 2017 to December 2019 with positive tau value demonstrating a positive monotonic trend. These locations also had S value of 0.0032 and 0.0001 describing the small average positive change in each month. Nilai showed a monotonic trend in daily maximum data, however, when analysed using monthly maximum ratio, the trend was no longer significant.

Table 4.4
Table of Modified Mann-Kendall Test Result

Station	Daily Maximum Ratios			Monthly Maximum Ratios		
	Tau	P-value	Sen's Slope (S)	Tau	P-value	Sen's Slope (S)
Bukit Rambai	0.0328	0.1382	0.000015	-0.0851	0.5207	-0.0004
Klang	**0.3466**	**0.0000**	**0.000214**	**0.3241**	**0.0125**	**0.0032**
Manjung	-0.0112	0.6099	-0.000005	-0.0851	0.5207	-0.0004
Nilai	**0.0652**	**0.0031**	**0.000027**	0.0115	0.9431	0.0001
Seberang Jaya	**0.2801**	**0.0000**	**0.000118**	**0.4230**	**0.0011**	**0.0035**

4.3.2.2 Abrupt Changes (Homogeneity) in Trend

Pettitt test is tested to determine possible abrupt changes in ratio of $PM_{2.5}/PM_{10}$ rather than a stable period of upward and downward trend. Table 4.5 displays the result for monthly maximum ratios of Pettit test.

Result obtained indicated that two stations were statistically significant with p-value less than 0.05. Abrupt change was detected in Klang in June 2018. Similarly, Seberang Jaya showed a sudden change in May 2019. The sudden change in Klang was due to the dry and hot weather in the period of June to September during SW monsoon (Hashim et al., 2018). Besides, it was also discovered that abrupt changes in May 2019 in Seberang Jaya where the maximum ratios increased from April 2019 to June 2019. This happened due to the transition period of inter-monsoon to SW monsoon and continued to increase from May to June as dry and hot weather occurred during this period.

Table 4.5
Table of Pettitt Test Result

Station	Bukit Rambai	Klang	Manjung	Nilai	Seberang Jaya
P-value	0.2328	**0.0026**	**0.0499**	0.6559	0.0446
Change Point at Time K	6	12	7	18	23
Change in Date	Dec 2017	**June 2018**	**Jan 2018**	Dec 18	May 2019

4.3.2.3 Stationarity of Data

ADF test was performed on the maximum daily of $PM_{2.5}/PM_{10}$ ratios to obtain an idea on how strongly the ratio affected by the trend. Table 4.6 displays the tabulated result of ADF test result. Result indicated that test on two stations which were Klang and Seberang Jaya dataset failed to reject H_0 implying that the time series dataset was non-stationary with p-value of 0.17 and 0.23, respectively. Meanwhile, Bukit Rambai, Manjung and Nilai dataset were found to be stationary with p-value of 0.01. Fortunately, LSTM not only can learn stationary data, but also deal with non-stationary data (Press, 2018) which is the nature of time series data. Hence, it is not necessary to dealt with non-stationary data.

Table 4.6
Table of ADF Test Result

Station	Dickey-Fuller	P-value	Decision Rule
Bukit Rambai	-4.0616	**0.01**	Reject H_0
Klang	-2.9734	0.17	Failed to reject H_0
Manjung	-4.2724	**0.01**	Reject H_0
Nilai	-5.5019	**0.01**	Reject H_0
Seberang Jaya	-2.8215	0.23	Failed to Reject H_0

4.4 Investigating the Influence of Splitting Ratio, Epochs Size and Number of LSTM Layer on LSTM model for $PM_{2.5}/PM_{10}$ Ratios

Tuning of hyperparameter as well as the splitting ratio played an important role in the performance of LSTM model in each dataset. Each dataset has different optimal tuning configuration and splitting ratio. Hence, several hyperparameters and splitting ratio were tuned to each dataset of study location in order to obtain model with best performance. Several hyperparameters that were tuned were number of epochs and number of LSTM layer. Before the model was tuned, several splitting ratios were examined to obtain the optimal splitting ratio that produced the best performance. Each dataset was assigned with the following hyperparameter as a default. Activation function was pre-determined in the pre-liminary stage where sigmoid was assigned as activation function for Bukit Rambai, Klang and Nilai dataset and ReLu was assigned to Manjung and Seberang Jaya dataset. The default lag size, number of neurons, number of layer and epochs size were assigned to 3, 8, 1 and 50, respectively in this study while utilizing Adam optimizer.

Table 4.7
Table of Default Hyperparameter Settings

Hyperparameter Settings		
Lag Size	:	3
Number of Neuron	:	8
Optimizer	:	Adam
Activation Function	:	Sigmoid (Bukit Rambai, Klang and Nilai), ReLu (Manjung and Seberang Jaya)
Number of Layer	:	1
Epochs Size	:	50

4.4.1 Splitting Ratio

The splitting ratio was tested with proportion of train to test of 60:40, 70:30, 80:20 and 90:10. Each splitting proportion was tested at number of epochs of 50. The performance of the model for each splitting ratio for Bukit Rambai is displays in Table 4.8. The result obtained showed that there was a decrease in MSE of training dataset when training size increased from 60% to 90% of overall dataset. Also, increase of R^2 was discovered as the size of training dataset increased. Small difference detected from all performance measured indicated that there was no overfitting issue in this dataset. It can be said that the best splitting ratio for Bukit Rambai was 90:10. The reason was due to its smallest MSE in train dataset as well as highest R^2 compared to performance of others splitting ratio.

Table 4.8
Table of Splitting Ratio of Bukit Rambai

Splitting Ratio	Dataset	Model Performances		
		MSE	MAE	R^2 (%)
60:40	Train	0.0062	0.0618	39.51
	Test	0.0060	0.0635	43.00
70:30	Train	0.0063	0.0626	40.52
	Test	0.0049	0.0585	40.73
80:20	Train	0.0062	0.0629	42.66
	Test	0.0050	0.0586	41.46
90:10	Train	**0.0057**	**0.0600**	**43.79**
	Test	0.0061	0.0660	48.84

Meanwhile, performance of Klang dataset showed an opposite result from Bukit Rambai. As the training size increased, the model was more likely to overfit as large difference between R^2 of train and test dataset was discovered from Table 4.9. This was due to variation in the training dataset itself. 60% of training dataset holds better variability compared to bigger training size. Ratio of 60 to 40 also produce small MSE and MAE with small gap between both of it. Thus, it can be concluded that 60:40 was the best splitting ratio for Klang dataset.

Table 4.9
Table of Splitting Ratio of Klang

Splitting Ratio	Dataset	Model Performances		
		MSE	MAE	R^2 (%)
60:40	Train	**0.0107**	**0.0878**	**73.32**
	Test	0.0132	0.0983	**62.34**
70:30	Train	0.0127	0.0952	80.83
	Test	0.0074	0.0698	50.91
80:20	Train	0.0136	0.0978	82.20
	Test	0.0026	0.0402	13.14
90:10	Train	0.0125	0.0927	81.00
	Test	0.0025	0.0385	2.00

In contrast to result obtained for Klang, Manjung dataset showed a better performance when larger amount of data in training dataset was fed to the network. The gap of all performance measured between train and test dataset was also small for splitting ratio of 90:10. Besides, this splitting ratio for Manjung dataset also had the highest R^2 for train dataset. Hence, the optimal splitting ratio for Manjung dataset was 90:10.

Table 4.10
Table of Splitting Ratio of Manjung

Splitting Ratio	Dataset	Model Performances		
		MSE	MAE	R^2 (%)
60:40	Train	0.0081	0.0712	33.26
	Test	0.0059	0.0597	56.91
70:30	Train	**0.0063**	**0.0627**	31.34
	Test	0.0063	0.0629	61.53
80:20	Train	0.0068	0.0650	33.79
	Test	0.0090	0.0752	64.85
90:10	Train	0.0075	0.0685	**37.86**
	Test	0.0086	0.0747	41.54

Nilai dataset also required larger train dataset of 80% of overall dataset. Table 4.11 revealed that 80% of train dataset produced better R^2 value with small gap between train and test dataset. MSE and MAE value for splitting ratio of 80 to 20 also appeared to have small value with small difference between error in train dataset and test dataset. Hence, it can be concluded that 80:20 ratio was the best splitting ratio for Nilai dataset.

Table 4.11
Table of Splitting Ratio of Nilai

Splitting Ratio	Dataset	Model Performance		
		MSE	MAE	R^2
60:40	Train	**0.0054**	**0.0579**	28.13
	Test	0.0056	0.0575	20.11
70:30	Train	0.0060	0.0603	44.05
	Test	0.0050	0.0550	27.54
80:20	Train	0.0056	0.0584	**41.70**
	Test	0.0053	0.0552	**30.07**
90:10	Train	0.0057	0.0589	40.80
	Test	0.0094	0.0768	30.81

Seberang Jaya dataset also needed more data in training dataset for the model to be able to capture the training pattern. Even though splitting ratio of 60:40 had the smallest error, however, 90:10 splitting ratio produced highest R^2 with small gap between train and test dataset. Hence, it can be deduced that the ratio of 90:10 was the optimal splitting ratio for the Seberang Jaya dataset.

Table 4.12
Table of Splitting Ratio of Seberang Jaya

Splitting Ratio	Dataset	Model Performance		
		MSE	MAE	R^2 (%)
60:40	Train	**0.0047**	**0.0546**	46.05
	Test	0.0095	0.0800	76.92
70:30	Train	0.0060	0.0610	45.92
	Test	0.0121	0.0893	78.33
80:20	Train	0.0050	0.0559	43.06
	Test	0.0083	0.0728	73.37
90:10	Train	0.0079	0.0703	**63.89**
	Test	0.0121	0.0882	72.29

Overall, four datasets required large train dataset to produce better results. The best splitting ratio of Bukit Rambai, Manjung, and Seberang Jaya were ratio of 90:10 and Nilai with 80:20 splitting ratio. In contrast, only Klang dataset had the best splitting ratio of 60:40, indicating it only needed small training size to capture the training pattern and more training data fed to the network will cause overfitting. This proved that not all dataset requires large amount of data in training dataset which contradict with the findings by Wu et al. (2021) that stated LSTM model with larger ratio of 80:20 or 90:10 produce better prediction accuracy. Hence, it can be said that splitting ratio is dependent on the characteristics of the data itself.

4.4.2 Number of Epochs

The model with the best splitting proportion for each study location was assessed with different number of epochs to determine whether an increase in the number of epochs affect the performance of the model. The number of epochs that were examined included 50, 100, 200 and 500 epochs. Figure 4.6 exhibits the performance of the model with different size of epochs for Bukit Rambai dataset. The result indicated that splitting ratio of 90:10 and 100 epochs presented better performance compared to other epochs size. Based on Figure 4.6, epochs of 100 showed a possibility to converge compared to other epochs size that demonstrated no indication to converge. Therefore, 100 epochs were taken as the best epoch size for Bukit Rambai dataset.

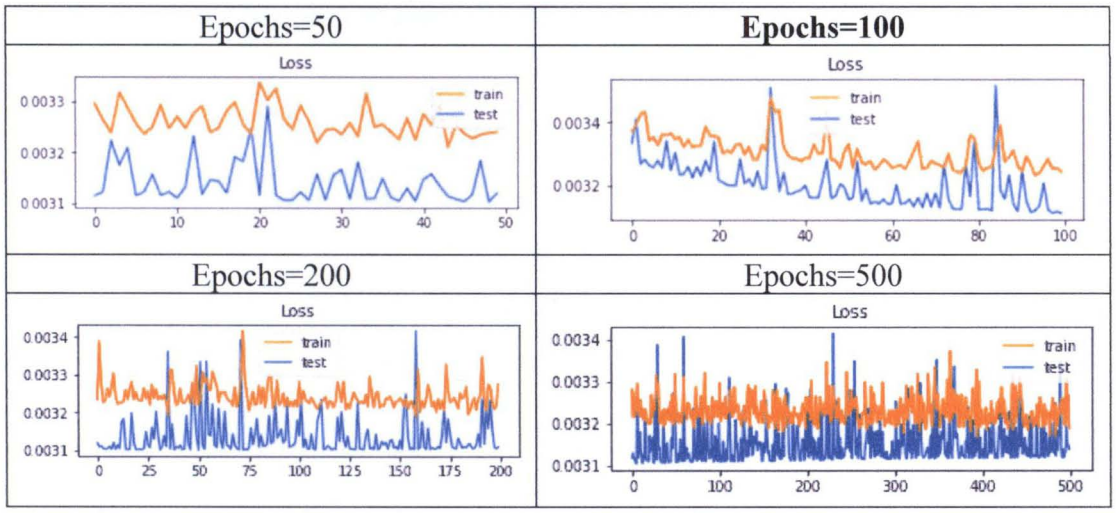


Figure 4.6 Loss per Epochs of Bukit Rambai Dataset

Next, Figure 4.7 reveals the performance of the models for different epochs size for Klang dataset based on the loss per epochs graph. It can be seen that when the network was trained for 100 times, the test set was more likely to converge. Hence, epochs of 100 were chosen as the best size of epochs for Klang dataset in achieving better performance of LSTM model.

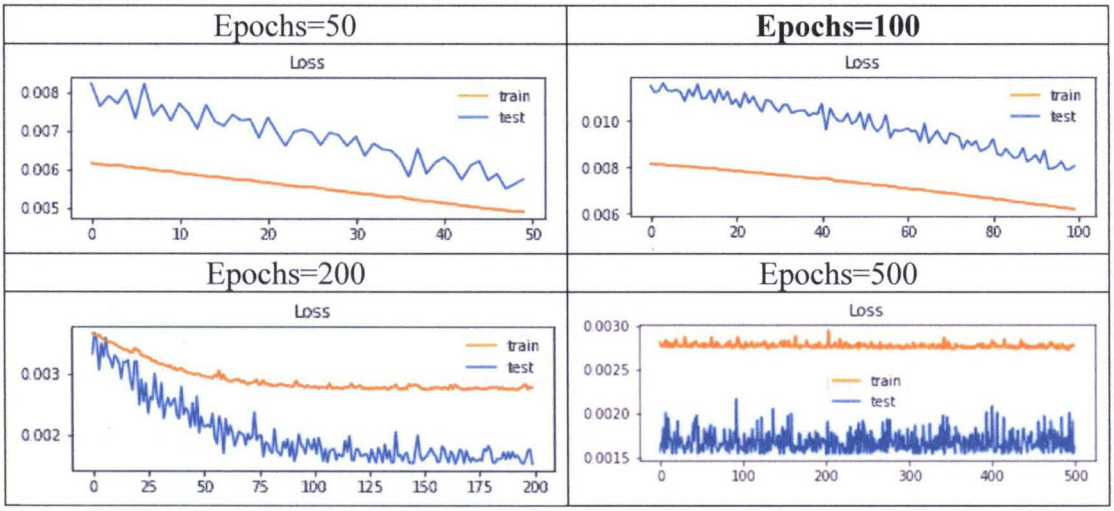


Figure 4.7 Loss per Epochs of Klang Dataset

It was apparent from Figure 4.8 that both epochs size of 50 and 200 were able to make the test set to converge for Manjung dataset. However, considering the running time to be executed, epoch size of 50 took shorter running time when compared to epoch size of 200. This claim was supported by Hastomo et al. (2021) that stated larger number of epochs can reduce the error but took longer running time. Hence, epoch size of 50 was chosen as the best size of epoch for Manjung dataset.

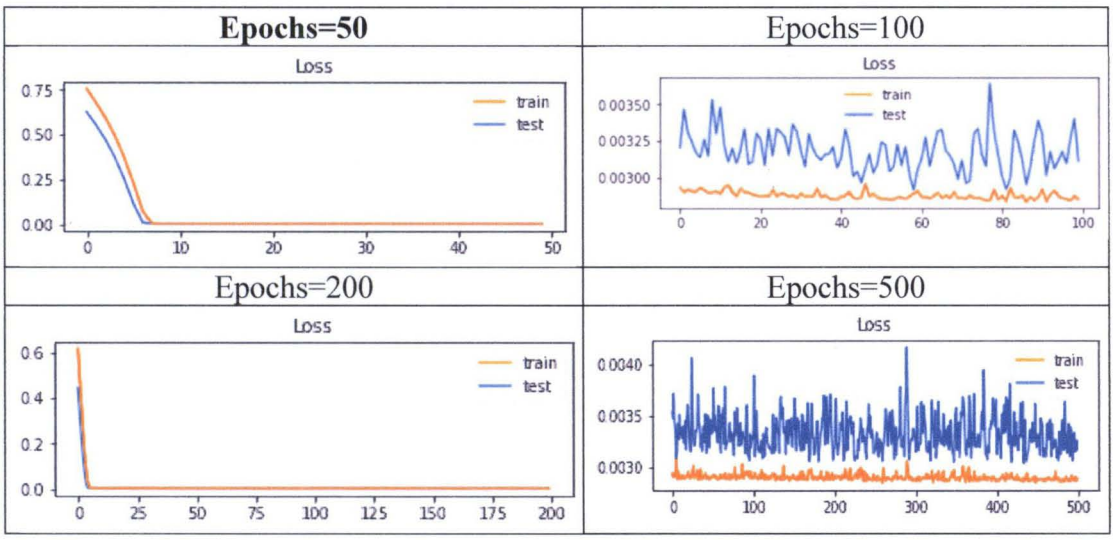


Figure 4.8 Loss per Epochs for Manjung Dataset

Meanwhile, Nilai dataset was observed to converge well when the epochs size was assigned to 50 and 100 based on Figure 4.9. However, 50 epochs took shorter time to be executed compared to larger size of epochs. Therefore, 50 epochs were selected as the best epoch size for Nilai dataset.

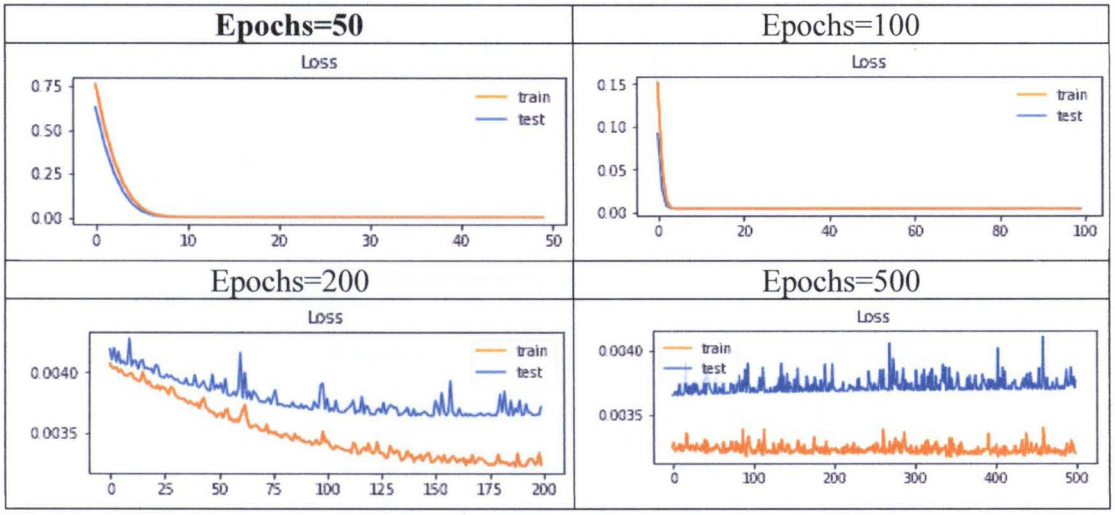


Figure 4.9 Loss per Epochs for Nilai Dataset

Differ to Nilai dataset, 200 of epochs were selected as the best epoch size for Seberang Jaya dataset. This was due to the fact that convergence of test set was detected when epochs size was set to 200 while other epochs size showed no indication to converge.

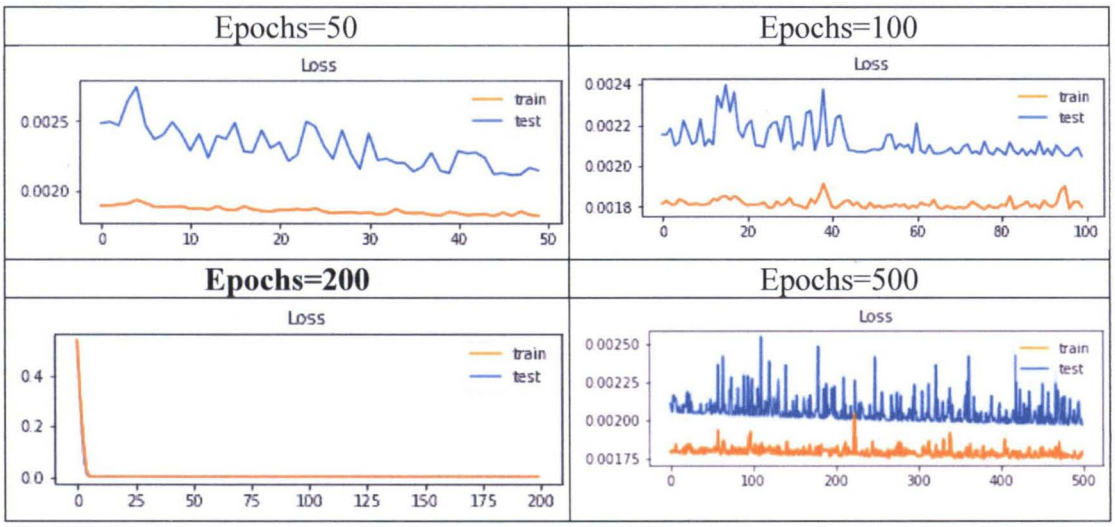


Figure 4.10 Loss per Epochs for Seberang Jaya Dataset

To summarize, Bukit Rambai and Klang needed 100 epochs for the test set to be converged while Manjung and Nilai datasets only required 50 epochs for the convergence to happen. Meanwhile, Seberang Jaya dataset required more epochs size of 200 to be able to converge well. This showed that different dataset requires different size of epochs to train the model in order to achieve better prediction performance. This contradict with the finding by Hastomo et al. (2021) that stated LSTM required smaller epochs to be able to achieve high prediction accuracy.

4.4.3 Number of LSTM Layer

Model with optimal splitting ratio and epoch size was assessed with several number of LSTM layer which were 1, 2, 3 and 4 layers. Each layer had the similar number of neurons which was 8 neurons. The performance of Bukit Rambai dataset for different number of LSTM layer was displayed in Table 4.13. The result suggested that as the number of layers increased, MSE value does not increase nor decrease. In contrast, there was a slight decrease in MAE and R^2 when the number of layers being added. Therefore, two layers of LSTM model was selected as the most appropriate number of LSTM layers for Bukit Rambai Dataset as it had highest R^2 of train dataset.

Table 4.13
Table of Number of Layer for Bukit Rambai

Number of Layer(s)	Dataset	Model Performance		
		MSE	MAE	R^2 (%)
1	Train	0.0063	0.0630	43.03
	Test	0.0067	0.0674	49.16
2	Train	**0.0057**	0.0600	**43.79**
	Test	0.0061	0.0657	48.83
3	Train	**0.0057**	0.0602	42.69
	Test	0.0061	0.0663	45.60
4	Train	**0.0057**	**0.0598**	41.22
	Test	0.0061	0.0656	43.41

The performance of the LSTM model for Klang dataset was shown in Table 4.14 where it was discovered that there was no sign of overfitting when increased in number of layers being made. Besides, a slight increment was seen in the value of MSE and MAE as number of layers being added. Since the difference in MSE and MAE between each layer was quite small, hence model with highest R^2 was chosen as the best model. In this case, two layers were the best number of layers for Klang dataset as it had the highest value of R^2 for both train and test dataset.

Table 4.14
Table of Number of Layers for Klang

Number of Layer(s)	Dataset	Model Performance		
		MSE	MAE	R^2 (%)
1	Train	0.0107	0.0878	73.32
	Test	0.0132	0.0983	62.34
2	Train	**0.0106**	**0.0876**	**74.04**
	Test	0.0150	0.1065	64.57
3	Train	0.0106	0.0875	70.51
	Test	0.0157	0.1096	60.68
4	Train	0.0106	0.0875	73.42
	Test	0.0159	0.1101	62.79

Next, the performance of Manjung dataset in Table 4.15 exhibits a slight decreased in MSE and MAE value as the number of layers being added. However, R^2 value decreased when more layer was added in the network indicating the capability of the model to predict decreases as more numbers of layer added. Therefore, the best number of layers for Manjung dataset was a single LSTM layer.

Table 4.15
Table of Number of Layers for Manjung

Number of Layer(s)	Dataset	Model Performance		
		MSE	MAE	R^2 (%)
1	Train	0.0063	0.0630	**37.62**
	Test	0.0075	0.0699	41.63
2	Train	0.0066	0.0642	35.94
	Test	0.0076	0.0703	39.35
3	Train	**0.0059**	**0.0607**	35.10
	Test	0.0074	0.0700	36.73
4	Train	**0.0059**	0.0612	32.76
	Test	0.0070	0.0678	36.10

Table 4.16 displayed the performance for Nilai dataset where sign of overfitting was detected when more than two layers added to the network. This is because the difference between R^2 of train and test dataset was fairly large. A small decrease in the error (MSE and MAE) values were identified when the number of layers increased. Hence, two layers of LSTM layer were the best number of layers for Nilai dataset as it had highest R^2 of 42.11%.

Table 4.16
Table of Number of Layers for Nilai

Number of Layer(s)	Dataset	Model Performance		
		MSE	MAE	R^2 (%)
1	Train	0.0056	0.0584	41.70
	Test	0.0053	0.0552	30.07
2	Train	**0.0055**	**0.0582**	**42.11**
	Test	0.0053	0.0551	30.04
3	Train	**0.0055**	**0.0582**	37.02
	Test	0.0053	0.0550	20.77
4	Train	**0.0055**	**0.0582**	37.44
	Test	0.0053	0.0549	20.42

The result for Seberang Jaya dataset in Table 4.17 indicated that not much difference can be seen from MSE and MAE values when number of layers increased, but a slight decreased in R^2 value was detected as more numbers of LSTM layer being added. Therefore, the best number of layers for Seberang Jaya dataset was single LSTM layer as it had highest R^2 value and small error.

Table 4.17
Table of Number of Layers for Seberang Jaya

Number of Layer(s)	Dataset	Model Performance		
		MSE	MAE	R^2 (%)
1	Train	0.0078	0.0700	**62.99**
	Test	0.0119	0.0879	70.86
2	Train	0.0078	0.0699	61.24
	Test	0.0120	0.0880	68.58
3	Train	0.0078	0.0698	60.55
	Test	0.0122	0.0885	67.66
4	Train	**0.0076**	**0.0687**	59.88
	Test	0.0119	0.0884	67.01

In summary, the number of LSTM layers had an impact on the performance of LSTM model. Dataset of two study locations which were Manjung and Seberang Jaya only needed single LSTM layer to produce good performance, but Bukit Rambai, Nilai, and Klang dataset required double LSTM layers to have better prediction performance.

This implied that small number of layers was sufficient enough in producing a good predictive LSTM model. This finding coincide with study by McNally et al. (2018) that suggested two layers were sufficient enough in modelling non-linear relationship of time series data.

4.5 Determining the Most Appropriate Splitting Ratio and Hyperparameter of LSTM Model in Predicting Daily Maximum $PM_{2.5}/PM_{10}$ Ratios

This study covered three aspect of hyperparameter setting in modelling LSTM model. Based on result obtained for objective 2, the best splitting ratio and experimented hyperparameter for each study location was displayed in Table 4.18.

Table 4.18
Table of Best Hyperparameter Setting for Each Station

Location	Splitting Ratio	Epochs Size	Number of Layer
Bukit Rambai	90:10	100	2
Klang	60:40	100	2
Manjung	90:10	50	1
Nilai	80:20	50	2
Seberang Jaya	90:10	200	1

With the corresponding performance of MSE, MAE and R^2 tabulated in Table 4.19.

Table 4.19
Table of Performance of each Station with Best Hyperparameters

Performance Measure		Bukit Rambai	Klang	Manjung	Nilai	Seberang Jaya
MSE	Train	0.0057	0.0106	0.0063	0.0055	0.0078
	Test	0.0061	0.0150	0.0075	0.0053	0.0119
MAE	Train	0.0600	0.0876	0.0630	0.0582	0.0700
	Test	0.0657	0.1065	0.0699	0.0551	0.0879
R^2	Train	43.79	74.04	37.62	42.11	62.99
	Test	48.83	64.57	41.63	30.04	70.86

It can be concluded that Bukit Rambai dataset require 90:10 splitting ratio, 100 epochs and double LSTM layer to operate with best prediction performance while Klang dataset needed 60:40 splitting ratio, 100 epochs and double LSTM layer. Meanwhile, Manjung dataset required splitting of 90:10 with 50 epochs and single LSTM layer and Nilai dataset needed 80:20 splitting ratio, 50 epochs and double LSTM layer to perform well. Whereas Seberang Jaya dataset achieved good prediction performance with 90:10 splitting ratio, 200 epochs and single LSTM layer.

4.6 Summary

In conclusion, all the objectives of this study were achieved in this chapter. To summarize the result, objective 1 which is to investigate the spatial and temporal distribution of $PM_{2.5}/PM_{10}$ ratios at several location in west coast of peninsular Malaysia found that the median ratios of all study locations were 0.71 indicating presence of anthropogenic sources. The study also deduced that there was monotonic trend exist in monthly maximum ratios of several station including Klang and Seberang Jaya. Objective 2 which is investigating the influence of splitting ratio, epochs size and number of LSTM layer discovered that each hyperparameter differed in each dataset depending on the characteristics and variation in the data itself. Lastly, the best hyperparameter for each dataset was obtained in Table 4.18 which indicate that the last objective of determining the most appropriate LSTM model for predicting $PM_{2.5}/PM_{10}$ ratios was achieved.

CHAPTER FIVE
CONCLUSION AND RECOMMENDATION

5.1 Introduction

This study had achieved all the objectives stipulated in the beginning of research and discussed in chapter four. In this chapter, it is divided into two sections where first section concluded the findings in chapter 4, followed by second section which discussed the recommendation for any aspect of improvement regarding the subject matter.

5.2 Conclusion

LSTM model is a newly developed model with a powerful prediction ability that are being used in many aspects of research including air quality. Hence, this study implemented LSTM model to examine the impact of three scope of hyperparameter settings which were splitting ratio, epochs size and number of LSTM layer towards the prediction performance of the model to predict daily maximum $PM_{2.5}/PM_{10}$ ratios at several locations in the west coast region of Peninsular Malaysia. The subject of the study was focusing on the ratios of $PM_{2.5}/PM_{10}$ which can reveal the conditions and sources of the particulate matter.

The trend of $PM_{2.5}/PM_{10}$ ratios and their spatial-temporal distribution for 2.5 years (July 2017 to December 2019) in five different study locations were assessed. The median ratio of Bukit Rambai, Klang, Manjung, Nilai and Seberang Jaya were 0.7 relatively higher than ratio discovered in Oman with median ratio of 0.69 (Alattar et al., 2019) and United Kingdom with median ratio of 0.65 (Munir, 2017). Supported by them, these showed a presence of anthropogenic sources in the atmosphere which led to formation of $PM_{2.5}$. The average ratio of $PM_{2.5}/PM_{10}$ given by the statistic median of study locations also exceeded value of 0.5 suggesting higher contribution of $PM_{2.5}$ compared to PM_{10} in the particulate matter component. The results thus concluded that $PM_{2.5}$ was more dominant at the study locations.

Robust statistical analysis, e.g., Robust Mann-Kendall and Pettitt test showed a positive trend in daily maximum of $PM_{2.5}/PM_{10}$ in Klang and Seberang Jaya with abrupt changes in ratios due to the dry weather brought by Southwest monsoon. Besides, the

finding obtained from ADF test indicated that only Klang and Seberang Jaya dataset had non-stationary state of data. However, stationary data is said to be easier modelled especially for time series data (Siami-Namini et al., 2018). Interestingly, these stations produced high R^2 when compared to other three station which found to be stationary. This disclosed that LSTM predict better in non-stationary data compared to stationary data.

Besides, Klang dataset also showed an interesting result that needed only 60% of training data to be able to predict well. Contradict with the finding by Wu et al. (2021) that stated LSTM model with larger ratio of 80:20 or 90:10 produce better prediction accuracy. Selecting the best hyperparameter setting in LSTM network is not a straightforward task and every dataset works best with different hyperparameter. The epochs size required varied by each dataset where Seberang Jaya needed larger epoch of 200 compared to other stations which required less epoch contradict with finding Hastomo et al. (2021) which suggested smaller epochs produced better prediction accuracy of LSTM model. Findings of the study also indicated that one or two layers were sufficient for univariate time series study to produce good prediction performance.

The hyperparameter setting found to be differ for every dataset and is dependent on the characteristics of the data itself. Within the scope of the experimentation criteria, the best hyperparameter for Bukit Rambai, Klang, Manjung, Nilai and Seberang Jaya are 90:10 splitting ratio, 100 epochs and double LSTM layer; 60:40 splitting ratio, 100 epochs and double LSTM layer; 90:10 splitting ratio, 50 epochs and single LSTM layer; splitting of 80:20, 50 epochs and double LSTM layer; and 90:10 splitting ratio, 200 epochs and single LSTM layer, respectively.

With the best hyperparameter setting, the MAE in training set obtained for Bukit Rambai, Klang, Manjung, Nilai and Seberang Jaya are 0.06, 0.09, 0.06, 0.06, and 0.07, respectively and MSE of 0.01 for each study location. The error measure was found to be small. In contrast, the R^2 for Bukit Rambai, Klang, Manjung, Nilai and Seberang Jaya are 43.79%, 74.04%, 37.62%, 42.11% and 62.99%, respectively. Therefore, this study had successfully experimented in three scope of hyperparameter setting and achieved a relatively good result especially in Klang and Seberang Jaya dataset.

5.3 Recommendation

This study has covered three angles of scope of model experimentation assessment which are splitting ratio, epochs size and number of LSTM layer due to limited research time. It is recommended for future research to experiment on different other hyperparameters such as dropout and lag size in affecting the prediction performance of LSTM model.

A study to examine the limitation of LSTM model for $PM_{2.5}/PM_{10}$ ratios is also recommended to be explored further. This study also conducted with limited data obtained from DoE which contain the concentration of $PM_{2.5}$ and PM_{10} recorded hourly from July 2017 to December 2019 for each study location. It is recommended to build a model with more data to obtain a better result.

This study has highlighted on the importance of ratios; Therefore, it is proposed to relevant authorities to provide a guideline on $PM_{2.5}/PM_{10}$ ratios as it is an important measure in providing the source and condition of particulate matter in the atmosphere. This guideline can also be used for future research in exploring on the ratio of $PM_{2.5}/PM_{10}$ to provide more significant findings on the result obtained.

Results of prior analysis at the stage of data preparation and understanding, has detected anomaly readings of the ratio which exceeded 100% ratio in Bukit Rambai. It is expected that this record has come from defaults or malfunction of instruments during the data collection process. Thus, the results suggest for more frequent quality control checking to be conducted by DOE. As for this study this anomaly value was treated as missing values and has been imputed before the data set is used for the analysis.

REFERENCES

Abbasimehr, H., Shabani, M., & Yousefi, M. (2020). An optimized model using LSTM network for demand forecasting. *Computers & Industrial Engineering*, *143*, 106435.

Abdullah, S., Ismail, M., Ahmed, A. N., & Abdullah, A. M. (2019). Forecasting particulate matter concentration using linear and non-linear approaches for air quality decision support. *Atmosphere*, *10*(11). https://doi.org/10.3390/atmos10110667

Abdullah, S., Ismail, M., & Fong, S. Y. (2017). Multiple Linear Regression (MLR) models for long term Pm 10 concentration forecasting during different monsoon seasons. *Journal of Sustainability Science and Management*, *12*(1), 60–69.

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *ArXiv Preprint ArXiv:1803.08375*.

Akinlade, G. O., Olaniyi, H. B., Olise, F. S., Owoade, O. K., Almeida, S. M., Almeida-Silva, M., & Hopke, P. K. (2015). Spatial and temporal variations of the particulate size distribution and chemical composition over Ibadan, Nigeria. *Environmental Monitoring and Assessment*, *187*(8). https://doi.org/10.1007/s10661-015-4755-4

Alattar, N., Yousif, J., Jaffer, M., & Aljunid, S. A. (2019). Neural and Mathematical Predicting Models for Particulate Matter Impact on Human Health in Oman. *WSEAS Trans Env & Dev*, *15*, 578–585.

Ali-Mohamed, A. Y., & Jaffar, A. H. (2000). Estimation of atmospheric inorganic water-soluble aerosols in the western region of Bahrain by ion chromatography. *Chemosphere - Global Change Science*, *2*(1), 85–94. https://doi.org/10.1016/S1465-9972(99)00058-6

Alias, N. F., Khan, M. F., Sairi, N. A., Zain, S. M., Suradi, H., Rahim, H. A., Banerjee, T., Bari, M. A., Othman, M., & Latif, M. T. (2020). Characteristics, Emission Sources, and Risk Factors of Heavy Metals in PM2.5from Southern Malaysia. *ACS Earth and Space Chemistry*, *4*(8), 1309–1323. https://doi.org/10.1021/acsearthspacechem.0c00103

Alifa, M., Bolster, D., Mead, M. I., Latif, M. T., & Crippa, P. (2020). The influence of meteorology and emissions on the spatio-temporal variability of PM10 in

Malaysia. *Atmospheric Research*, *246*(March), 105107. https://doi.org/10.1016/j.atmosres.2020.105107

Alpan, K., & Sekeroglu, B. (2020). Prediction of pollutant concentrations by meteorological data using machine learning algorithms. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *44*(4/W3), 21–27. https://doi.org/10.5194/isprs-archives-XLIV-4-W3-2020-21-2020

Alyousifi, Y, Ibrahim, K., Zin, W. Z. W., & Rathnayake, U. (2021). Trend analysis and change point detection of air pollution index in Malaysia. *International Journal of Environmental Science and Technology*. https://doi.org/10.1007/s13762-021-03672-w

Alyousifi, Yousif, Othman, M., Faye, I., Sokkalingam, R., & Silva, P. C. L. (2020). Markov Weighted Fuzzy Time-Series Model Based on an Optimum Partition Method for Forecasting Air Pollution. *International Journal of Fuzzy Systems*, *22*(5), 1468–1486. https://doi.org/10.1007/s40815-020-00841-w

Amil, N., Latif, M. T., Khan, M. F., & Mohamad, M. (2016). Seasonal variability of PM2.5composition and sources in the Klang Valley urban-industrial environment. *Atmospheric Chemistry and Physics*, *16*(8), 5357–5381. https://doi.org/10.5194/acp-16-5357-2016

Amoatey, P., Omidvarborna, H., & Baawain, M. (2018). The modeling and health risk assessment of PM2.5 from Tema Oil Refinery. *Human and Ecological Risk Assessment*, *24*(5), 1181–1196. https://doi.org/10.1080/10807039.2017.1410427

Awan, F. M., Minerva, R., & Crespi, N. (2020). Improving road traffic forecasting using air pollution and atmospheric data: Experiments based on lstm recurrent neural networks. *Sensors (Switzerland)*, *20*(13), 3. https://doi.org/10.3390/s20133749

Balasubramanian, R., Qian, W. B., Decesari, S., Facchini, M. C., & Fuzzi, S. (2003). Comprehensive characterization of PM2.5 aerosols in Singapore. *Journal of Geophysical Research: Atmospheres*, *108*(16). https://doi.org/10.1029/2002jd002517

Baruah, U. D., Robeson, S. M., Saikia, A., Mili, N., Sung, K., & Chand, P. (2022). Spatio-temporal characterization of tropospheric ozone and its precursor pollutants NO2 and HCHO over South Asia. *Science of The Total Environment*, *809*, 151135.

Bekkar, A., Hssina, B., Douzi, S., & Douzi, K. (2021). Air-pollution prediction in smart city, deep learning approach. *Journal of Big Data*, *8*(1), 1–21. https://doi.org/10.1186/s40537-021-00548-1

Belavadi, S. V., Rajagopal, S., Ranjani, R., & Mohan, R. (2020). Air Quality Forecasting using LSTM RNN and Wireless Sensor Networks. *Procedia Computer Science*, *170*(2019), 241–248. https://doi.org/10.1016/j.procs.2020.03.036

Bierens, H. J., & Song, H. (2006). Semi-Nonparametric Estimation of First-Price Auctions Models with Auction-Speci c Heterogeneity using Simulated Method of Moments. *Mimeo, Pennsylvania State University*.

Blanchard, C. L., Tanenbaum, S., & Motallebi, N. (2011). Spatial and temporal characterization of PM2.5 mass concentrations in California, 1980-2007. *Journal of the Air and Waste Management Association*, *61*(3), 339–351. https://doi.org/10.3155/1047-3289.61.3.339

Brossart, D. F., Laird, V. C., & Armstrong, T. W. (2018). Interpreting Kendall's Tau and Tau-U for single-case experimental designs. *Cogent Psychology*, *5*(1), 1518687.

Cesaroni, G., Forastiere, F., Stafoggia, M., Andersen, Z. J., Badaloni, C., Beelen, R., Caracciolo, B., De Faire, U., Erbel, R., Eriksen, K. T., Fratiglioni, L., Galassi, C., Hampel, R., Heier, M., Hennig, F., Hilding, A., Hoffmann, B., Houthuijs, D., Jöckel, K. H., … Peters, A. (2014). Long term exposure to ambient air pollution and incidence of acute coronary events: Prospective cohort study and meta-analysis in 11 european cohorts from the escape project. *BMJ (Online)*, *348*(January), 1–16. https://doi.org/10.1136/bmj.f7412

Chang, D., & Song, Y. (2010). Estimates of biomass burning emissions in tropical Asia based on satellite-derived data. *Atmospheric Chemistry and Physics*, *10*(5), 2335–2351. https://doi.org/10.5194/acp-10-2335-2010

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-24271-9

Chu, H. J., Huang, B., & Lin, C. Y. (2015). Modeling the spatio-temporal heterogeneity in the PM10-PM2.5 relationship. *Atmospheric Environment*, *102*(1), 176–182. https://doi.org/10.1016/j.atmosenv.2014.11.062

Coskuner, G., Jassim, M. S., & Munir, S. (2018). Characterizing temporal variability of PM 2.5 /PM 10 ratio and its relationship with meteorological parameters in Bahrain. *Environmental Forensics*, *19*(4), 315–326. https://doi.org/10.1080/15275922.2018.1519738

Department of Environment, M. (2018). *Environmental Quality Report 2018*.

Department of Environment, M. (2019). *Environmental Quality Report 2019*. https://enviro2.doe.gov.my/ekmc/digital-content/laporan-kualiti-alam-sekeliling-2019/

Department of Environment, M. (2020). *Environmental Quality Report 2020*.

Department of Environmental. (2019). *API Calculation*.

Dozat, T. (2015). Technical report, Incorporating Nesterov Momentum into Adam. *Proc. ICLR Workshop*.

Du, P., Wang, J., Hao, Y., Niu, T., & Yang, W. (2020). A novel hybrid model based on multi-objective Harris hawks optimization algorithm for daily PM2.5 and PM10 forecasting. *Applied Soft Computing Journal*, *96*, 1–24. https://doi.org/10.1016/j.asoc.2020.106620

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(7).

Ee-Ling, O., Mustaffa, N. I. H., Amil, N., Khan, M. F., & Latif, M. T. (2015). Source contribution of PM2.5 at different locations on the Malaysian peninsula. *Bulletin of Environmental Contamination and Toxicology*, *94*(4), 537–542. https://doi.org/10.1007/s00128-015-1477-9

Eeftens, M., Tsai, M.-Y., Ampe, C., Anwander, B., Beelen, R., Bellander, T., Cesaroni, G., Cirach, M., Cyrys, J., de Hoogh, K., De Nazelle, A., de Vocht, F., Declercq, C., Dedele, A., Eriksen, K., Galassi, C., Gražulevičiene, R., Grivas, G., Heinrich, J., … Hoek, G. (2012). Spatial variation of PM2.5, PM10, PM2.5 absorbance and PMcoarse concentrations between and within 20 European study areas and the relationship with NO2 - Results of the ESCAPE project. *Atmospheric Environment*, *62*, 303–317. https://doi.org/10.1016/j.atmosenv.2012.08.038

Fan, H., Zhao, C., Yang, Y., & Yang, X. (2021). Spatio-Temporal Variations of the PM2.5/PM10 Ratios and Its Application to Air Pollution Type Classification in

China. *Frontiers in Environmental Science*, *9*. https://doi.org/10.3389/fenvs.2021.692440

Farzad, A., Mashayekhi, H., & Hassanpour, H. (2019). A comparative performance analysis of different activation functions in LSTM networks for classification. *Neural Computing and Applications*, *31*(7), 2507–2521.

Gecynalda, S., da Gomes, S., Ludermir, T., & Lima, L. M. M. R. (2011). Comparison of new activation functions in neural network for forecasting financial time series [J]. *Neural Computing and Applications*, *20*(3), 417–439.

Ghim, Y. S., Chang, Y. S., & Jung, K. (2015). Temporal and spatial variations in fine and coarse particles in Seoul, Korea. *Aerosol and Air Quality Research*, *15*(3), 842–852. https://doi.org/10.4209/aaqr.2013.12.0362

Gilbert, R. O. (1987). Statistical Methods for Environmental Pollution Monitoring. *New York: Van Nostrand Reinhold.*

Hagen, L. J. (2004). Fine particulates (PM10 and PM2.5) generated by breakage of mobile aggregates during simulated wind erosion. *Transactions of the American Society of Agricultural Engineers*, *47*(1), 107–112. https://doi.org/10.13031/2013.15876

Hamami, F., & Dahlan, I. A. (2020). Univariate Time Series Data Forecasting of Air Pollution using LSTM Neural Network. *2020 International Conference on Advancement in Data Science, E-Learning and Information Systems, ICADEIS 2020*. https://doi.org/10.1109/ICADEIS49811.2020.9277393

Hamanaka, R. B., & Mutlu, G. M. (2018). Particulate Matter Air Pollution: Effects on the Cardiovascular System. *Frontiers in Endocrinology*, *9*(November), 1–15. https://doi.org/10.3389/fendo.2018.00680

Hamed, K. H., & Rao, A. R. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, *204*(1–4), 182–196.

Harrison, R. M. (2001). *Pollution: causes, effects and control*. Royal society of chemistry.

Hashim, N. I. M., Noor, N. M., & Yusof, S. Y. (2018). Temporal characterisation of ground-level ozone concentration in Klang Valley. *E3S Web of Conferences*, *34*, 2047.

Hastomo, W., Karno, A. S. B., Kalbuana, N., & Meiriki, A. (2021). Characteristic parameters of epoch deep learning to predict Covid-19 data in Indonesia. *Journal of Physics: Conference Series*, *1933*(1), 12050.

Hinton, G., Srivastava, N., & Swersky, K. (2012). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited On*, *14*(8), 2.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1007/978-1-4757-5388-2_2

Hu, J., Wang, S., & Mao, J. (2019). Short time PM2.5 prediction model for Beijing-Tianjin-Hebei region based on Generalized Space Time Autoregressive (GSTAR). *IOP Conference Series: Earth and Environmental Science*, *358*(2). https://doi.org/10.1088/1755-1315/358/2/022075

Hwa-Lung, Y., & Chih-Hsin, W. (2010). Retrospective prediction of intraurban spatiotemporal distribution of PM2.5 in Taipei. *Atmospheric Environment*, *44*(25), 3053–3065. https://doi.org/10.1016/j.atmosenv.2010.04.030

Ismail, M., Yuen, F. S., & Abdullah, S. (2015). Trend and status of particulate matter (PM10) concentration at three major cities in east coast of Peninsular Malaysia. *Research Journal of Chemical and Environmental Sciences*, *3*(5), 25–31.

Khan, M. F., Latif, M. T., Saw, W. H., Amil, N., Nadzir, M. S. M., Sahani, M., Tahir, N. M., & Chung, J. X. (2016). Fine particulate matter in the tropical environment: Monsoonal effects, source apportionment, and health risk assessment. *Atmospheric Chemistry and Physics*, *16*(2), 597–617. https://doi.org/10.5194/acp-16-597-2016

Khan, Md Firoz, Latif, M. T., Lim, C. H., Amil, N., Jaafar, S. A., Dominick, D., Mohd Nadzir, M. S., Sahani, M., & Tahir, N. M. (2015). Seasonal effect and source apportionment of polycyclic aromatic hydrocarbons in PM2.5. *Atmospheric Environment*, *106*, 178–190. https://doi.org/10.1016/j.atmosenv.2015.01.077

Khodeir, M., Shamy, M., Alghamdi, M., Zhong, M., Sun, H., Costa, M., Chen, L.-C., & Maciejczyk, P. (2012). Source apportionment and elemental composition of PM2.5 and PM10 in Jeddah City, Saudi Arabia. *Atmospheric Pollution Research*, *3*(3), 331–340. https://doi.org/https://doi.org/10.5094/APR.2012.037

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

Law, A. T., Tan, C. H., Abdul-Rashid, M. K., & Ichikawa, T. (2001). Effect of monsoon seasons on nitrogen distribution in the Straits of Malacca. *Aquatic Resource and Environmental Studies of the Straits of Malacca: Current Research and Reviews*, 39–50.

Lestiani, D. D., Santoso, M., & Hidayat, A. (2008). Karakteristik Black Carbon Partikulat Udara Halus PM2, 5 di Bandung dan Lembang 2004-2005. *Jurnal Sains Dan Teknologi Nuklir Indonesia*, *9*(2), 6.

Li, J., Li, X., & Wang, K. (2019). Atmospheric PM2.5 Concentration Prediction Based on Time Series and Interactive Multiple Model Approach. *Advances in Meteorology*, *2019*. https://doi.org/10.1155/2019/1279565

Li, L., Zhang, J., Qiu, W., Wang, J., & Fang, Y. (2017). An ensemble spatiotemporal model for predicting PM2.5 concentrations. *International Journal of Environmental Research and Public Health*, *14*(5). https://doi.org/10.3390/ijerph14050549

Lin, N.-H., Sayer, A. M., Wang, S.-H., Loftus, A. M., Hsiao, T.-C., Sheu, G.-R., Hsu, N. C., Tsay, S.-C., & Chantara, S. (2014). Interactions between biomass-burning aerosols and clouds over Southeast Asia: Current status, challenges, and perspectives. *Environmental Pollution*, *195*, 292–307. https://doi.org/10.1016/j.envpol.2014.06.036

Lin, X. (2021). The Application of Machine Learning Models in the Prediction of PM2.5/PM10 Concentration. *4th International Conference on Computers in Management and Business, ICCMB 2021*, 94–101. https://doi.org/10.1145/3450588.3450605

Liu, D. J., & Li, L. (2015). Application study of comprehensive forecasting model based on entropy weighting method on trend of PM2.5 concentration in Guangzhou, China. *International Journal of Environmental Research and Public Health*, *12*(6), 7085–7099. https://doi.org/10.3390/ijerph120607085

Ma, J., Ding, Y., Cheng, J. C. P., Jiang, F., & Wan, Z. (2019). A temporal-spatial interpolation and extrapolation method based on geographic Long Short-Term Memory neural network for PM2.5. *Journal of Cleaner Production*, *237*, 117729. https://doi.org/10.1016/j.jclepro.2019.117729

Madrigano, J., Kloog, I., Goldberg, R., Coull, B. A., Mittleman, M. A., & Schwartz, J. (2013). Long-term exposure to PM2.5 and incidence of acute myocardial infarction. *Environmental Health Perspectives*, *121*(2), 192–196. https://doi.org/10.1289/ehp.1205284

Mallick, J., Talukdar, S., Alsubih, M., Salam, R., Ahmed, M., Kahla, N. Ben, & Shamimuzzaman, M. (2021). Analysing the trend of rainfall in Asir region of Saudi Arabia using the family of Mann-Kendall tests, innovative trend analysis,

and detrended fluctuation analysis. *Theoretical and Applied Climatology*, *143*(1), 823–841.

Masood, A., & Ahmad, K. (2020). A model for particulate matter (PM2.5) prediction for Delhi based on machine learning approaches. *Procedia Computer Science*, *167*(2019), 2101–2110. https://doi.org/10.1016/j.procs.2020.03.258

Mat Shukri, M. A., Yusoff, M., Awang, N. R., Jani, M., Ab Kadir, Z., Selvam, B., Sulaiman, M. A., & Abdus Salam, M. (2017). Investigation of relationship between particulate matter (PM2. 5 and PM10) and meteorological parameters at Roadside Area of First Penang Bridge. *Journal of Tropical Resources and Sustainable Sciences*, *5*(1), 33–39.

McNally, S., Roche, J., & Caton, S. (2018). Predicting the price of bitcoin using machine learning. *2018 26th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, 339–343.

MET. (2019). *Weather Phenomena- Monsoon*. Malaysian Meteorological Department. https://www.met.gov.my/pendidikan/cuaca/fenomenacuaca

Mohd Zahid, A. Z., Abdul Malik, N. N. A., & Kassim, J. (2018). Particulate matter study at residential and educational areas in Shah Alam, Malaysia. *MATEC Web of Conferences*, *250*, 1–16. https://doi.org/10.1051/matecconf/201825006010

Munir, S. (2017). Analysing temporal trends in the ratios of PM2. 5/PM10 in the UK. *Aerosol and Air Quality Research*, *17*(1), 34–48.

Mustakim, R., & Mamat, M. (2021). Performance Comparison of Malaysian Air Pollution Index Prediction Using Nonlinear Autoregressive Exogenous Artificial Neural Network and Support Vector Machine. *E3S Web of Conferences*, *287*, 04001. https://doi.org/10.1051/e3sconf/202128704001

Nawrot, T. S., Perez, L., Künzli, N., Munters, E., & Nemery, B. (2011). Public health importance of triggers of myocardial infarction: A comparative risk assessment. *The Lancet*, *377*(9767), 732–740. https://doi.org/10.1016/S0140-6736(10)62296-9

Olah, C. (2018). *Understanding LSTM Networks-Colah's Blog*.

Pan, B. (2018). Application of XGBoost algorithm in hourly PM2.5 concentration prediction. *IOP Conference Series: Earth and Environmental Science*, *113*(1). https://doi.org/10.1088/1755-1315/113/1/012127

Parkhurst, W. J., Tanner, R. L., Weatherford, F. P., Valente, R. J., & Meagher, J. F. (1999). Historic PM2.5/PM10 Concentrations in the Southeastern United

States—Potential Implications of the Revised Particulate Matter Standard. *Journal of the Air and Waste Management Association*, *49*(9), 1060–1067. https://doi.org/10.1080/10473289.1999.10463894

Partheeban, P. (2021). Application of LSTM Models in Predicting Particulate Matter (PM2. 5) Levels for Urban Area. *Journal of Engineering Research*.

Pettitt, A. N. (1979). A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28*(2), 126–135.

Phi, M. (2018). Illustrated Guide to LSTM's and GRU's: A step by step explanation. *Towards Data Science*.

Pope, C. A., Burnett, R. T., Thurston, G. D., Thun, M. J., Calle, E. E., Krewski, D., & Godleski, J. J. (2004). Cardiovascular Mortality and Long-Term Exposure to Particulate Air Pollution: Epidemiological Evidence of General Pathophysiological Pathways of Disease. *Circulation*, *109*(1), 71–77. https://doi.org/10.1161/01.CIR.0000108927.80044.7F

Press, J. (2018). LSTM online training and prediction: Non-stationary real time data stream forecasting. *Wayne State University*.

Quah, E. (2002). Transboundary pollution in Southeast Asia: The Indonesian fires. *World Development*, *30*(3), 429–441. https://doi.org/10.1016/S0305-750X(01)00122-X

Rahman, S. A., Hamzah, M. S., Wood, A. K., Elias, M. S., Salim, N. A. A., & Sanuri, E. (2011). Sources apportionment of fine and coarse aerosol in Klang Valley, Kuala Lumpur using positive matrix factorization. *Atmospheric Pollution Research*, *2*(2), 197–206. https://doi.org/10.5094/APR.2011.025

Reid, J. S., Hyer, E. J., Johnson, R. S., Holben, B. N., Yokelson, R. J., Zhang, J., Campbell, J. R., Christopher, S. A., Di Girolamo, L., Giglio, L., Holz, R. E., Kearney, C., Miettinen, J., Reid, E. A., Turk, F. J., Wang, J., Xian, P., Zhao, G., Balasubramanian, R., … Liew, S. C. (2013). Observing and understanding the Southeast Asian aerosol system by remote sensing: An initial review and analysis for the Seven Southeast Asian Studies (7SEAS) program. *Atmospheric Research*, *122*, 403–468. https://doi.org/10.1016/j.atmosres.2012.06.005

Reimers, N., & Gurevych, I. (2017). Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *ArXiv Preprint ArXiv:1707.06799*.

Ribeiro, G. H. T., Neto, P. S. G. de M., Cavalcanti, G. D. C., & Tsang, R. (2011). Lag selection for time series forecasting using particle swarm optimization. *The 2011 International Joint Conference on Neural Networks*, 2437–2444.

Ross, Z., Ito, K., Johnson, S., Yee, M., Pezeshki, G., Clougherty, J. E., Savitz, D., & Matte, T. (2013). Spatial and temporal estimation of air pollutants in New York City: Exposure assignment for use in a birth outcomes study. *Environmental Health: A Global Access Science Source*, *12*(1), 1–13. https://doi.org/10.1186/1476-069X-12-51

Saeed, S., Hussain, L., Awan, I. A., & Idris, A. (2017). Comparative Analysis of different Statistical Methods for Prediction of PM 2.5 and PM10 Concentrations in Advance for Several Hours. *International Journal of Computer Science and Network Security*, *17*(11), 45–52.

Samoli, E., Peng, R., Ramsay, T., Pipikou, M., Touloumi, G., Dominici, F., Burnett, R., Cohen, A., Krewski, D., Samet, J., & Katsouyanni, K. (2008). Acute effects of ambient particulate matter on mortality in Europe and North America: Results from the APHENA study. *Environmental Health Perspectives*, *116*(11), 1480–1486. https://doi.org/10.1289/ehp.11345

Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, *2017-Janua*(January 2018), 1643–1647. https://doi.org/10.1109/ICACCI.2017.8126078

Sentian, J., Herman, F., Yih, C. Y., & Hian Wui, J. C. (2019). Long-term air pollution trend analysis in Malaysia. *International Journal of Environmental Impacts: Management, Mitigation and Recovery*, *2*(4), 309–324. https://doi.org/10.2495/ei-v2-n4-309-324

Setiawan, I. (2020). Time series air quality forecasting with R Language and R Studio. *2nd International Conference on Applied Science and Technology - Engineering Sciences, ICAST-ES 2019*, *1450*(1). https://doi.org/10.1088/1742-6596/1450/1/012064

Shafii, N. H. B., Alias, R., Zamani, N. F., & Fauzi, N. F. (2020). Forecasting Air Pollution Index (API) PM 2 . 5 Using Support Vector Machine (SVM). *Journal of Computing Research and Innovation*, *5*(3), 43–53.

Rak, A. E., & Salam, M. A. (2021). Potential of arima-ann, arima-svm, dt and catboost for atmospheric pm2.5 forecasting in bangladesh. *Atmosphere*, *12*(1), 1–21. https://doi.org/10.3390/atmos12010100

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018). A comparison of ARIMA and LSTM in forecasting time series. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1394–1401.

Spandana, B., Srinivasa Rao, S., Upadhya, A. R., Kulkarni, P., & Sreekanth, V. (2021). PM2.5/PM10 ratio characteristics over urban sites of India. *Advances in Space Research*, *67*(10), 3134–3146. https://doi.org/10.1016/j.asr.2021.02.008

Sugimoto, N., Shimizu, A., Matsui, I., & Nishikawa, M. (2016). A method for estimating the fraction of mineral dust in particulate matter using PM2.5-to-PM10 ratios. *Particuology*, *28*, 114–120. https://doi.org/10.1016/j.partic.2015.09.005

Suleiman, A., Tight, M. R., & Quinn, A. D. (2020). A comparative study of using Random Forests (RF), Extreme Learning Machine (ELM) and Deep Learning (DL) algorithms in modelling Roadside Particulate Matter (PM10 & PM2.5). *IOP Conference Series: Earth and Environmental Science*, *476*(1). https://doi.org/10.1088/1755-1315/476/1/012126

Tahir, N. M., Koh, M., & Suratman, S. (2013). PM2.5 and associated ionic species in a sub-urban coastal area of Kuala Terengganu, Southern South China Sea (Malaysia). *Sains Malaysiana*, *42*(8), 1065–1072.

Tella, A., Balogun, A. L., & Faye, I. (2021). Spatio-temporal modelling of the influence of climatic variables and seasonal variation on PM10 in Malaysia using multivariate regression (MVR) and GIS. *Geomatics, Natural Hazards and Risk*, *12*(1), 443–468. https://doi.org/10.1080/19475705.2021.1879942

Thia-Eng, C., Gorre, I. R. L., Ross, S. A., Bernad, S. R., Gervacio, B., & Ebarvia, M. C. (2000). The Malacca Straits. *Marine Pollution Bulletin*, *41*(1–6), 160–178.

United States Environmental Protection Agency. (2021). *Particulate Matter (PM) Pollution*. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

USEPA. (2004). Air quality criteria for particulate matter (Final Report, Oct 2004). *Environmental Protection Agency: Washington, DC*. EPA 600/P-99/002aF-bF, 2004

Usman, K., & Ramdhani, M. (2019). Comparison of Classical Interpolation Methods and Compressive Sensing for Missing Data Reconstruction. *2019 IEEE*

International Conference on Signals and Systems, ICSigSys 2019, 29–33. https://doi.org/10.1109/ICSIGSYS.2019.8811057

Veselík, P., Sejkorová, M., Nieoczym, A., & Caban, J. (2020). Outlier identification of concentrations of pollutants in environmental data using modern statistical methods. *Polish Journal of Environmental Studies*, *29*(1), 853–860.

Wu, X., Zhou, J., Yu, H., Liu, D., Xie, K., Chen, Y., Hu, J., Sun, H., & Xing, F. (2021). The Development of a hybrid wavelet-ARIMA-LSTM model for precipitation amounts and drought analysis. *Atmosphere*, *12*(1), 74.

Wu, Y., Liu, J., Zhai, J., Cong, L., Wang, Y., Ma, W., Zhang, Z., & Li, C. (2018). Comparison of dry and wet deposition of particulate matter in near-surface waters during summer. *PloS One*, *13*(6), e0199241.

Wu, Z., Wu, X., Wang, Y., & He, S. (2020). PM2.5/PM10 ratio prediction based on a long short-Term memory neural network in Wuhan, China. *Geoscientific Model Development*, *13*(3), 1499–1511. https://doi.org/10.5194/gmd-13-1499-2020

Xie, W., Li, G., Zhao, D., Xie, X., Wei, Z., Wang, W., Wang, M., Li, G., Liu, W., Sun, J., Jia, Z., Zhang, Q., & Liu, J. (2015). Relationship between fine particulate air pollution and ischaemic heart disease morbidity and mortality. *Heart*, *101*(4), 257–263. https://doi.org/10.1136/heartjnl-2014-306165

Ya'acob, S. H., & Mar Iman, A. H. (2020). The Spatial Influence of Environmental and Anthropogenic Factors on the Pattern of Air Pollution in Malaysia. *2nd International Conference on Tropical Resources and Sustainable Sciences, CTReSS 2020*, *549*(1). https://doi.org/10.1088/1755-1315/549/1/012011

Yap, C. K., Ismail, A., Tan, S. G., & Omar, H. (2002). Concentrations of Cu and Pb in the offshore and intertidal sediments of the west coast of Peninsular Malaysia. *Environment International*, *28*(6), 467–479. https://doi.org/10.1016/S0160-4120(02)00073-9

Yunesian, M., Rostami, R., Zarei, A., Fazlzadeh, M., & Janjani, H. (2019). Exposure to high levels of PM2.5 and PM10 in the metropolis of Tehran and the associated health risks during 2016–2017. *Microchemical Journal*, *150*(June), 104174. https://doi.org/10.1016/j.microc.2019.104174

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *ArXiv Preprint ArXiv:1212.5701*.

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., & Si, Y. (2018). A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and

XGboost. *IEEE Access*, *6*, 21020–21031. https://doi.org/10.1109/ACCESS.2018.2818678

Zhang, Y. L., & Cao, F. (2015). Fine particulate matter (PM 2.5) in China at a city level. *Scientific Reports*, *5*(2014), 1–12. https://doi.org/10.1038/srep14884

Zhao, D., Chen, H., Yu, E., & Luo, T. (2019). PM 2.5 /PM 10 ratios in eight economic regions and their relationship with meteorology in China. *Advances in Meteorology*, *2019*. https://doi.org/10.1155/2019/5295726

Zheng, M., Salmon, L. G., Schauer, J. J., Zeng, L., Kiang, C. S., Zhang, Y., & Cass, G. R. (2005). Seasonal trends in PM2.5 source contributions in Beijing, China. *Atmospheric Environment*, *39*(22), 3967–3976. https://doi.org/10.1016/j.atmosenv.2005.03.036

APPENDICES

APPENDIX 1

Python Coding

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM
from tensorflow.keras.layers import Dense
from tensorflow.keras.layers import Flatten
from keras.layers import Dropout
from sklearn.metrics import r2_score

# import data
br = pd.read_csv('BR',parse_dates=True)

br['DATETIME'] = pd.to_datetime(br['DATETIME'])

BR=br.groupby(pd.Grouper(key='DATETIME', axis=0, freq='1D', sort=True)).max()

# Split data into train and test
train_size= int(len(BR)*0.9)
train, test= BR[0:train_size], BR[train_size:len(br)]
```

```
train=np.array(train)
test=np.array(test)

def create_dataset(dataset, time_step=1):
    dataX, dataY = [], []
    for i in range(len(dataset)):
        # find the end of this pattern
        end_ix = i + time_step
        # check if we are beyond the sequence
        if end_ix > len(dataset)-1:
            break
        # gather input and output parts of the pattern
        seq_x, seq_y = dataset[i:end_ix], dataset[end_ix]
        dataX.append(seq_x)
        dataY.append(seq_y)
    return np.array(dataX), np.array(dataY)
time_step = 3
X_train, y_train = create_dataset(train, time_step)
X_test, ytest = create_dataset(test, time_step)

X_train =X_train.reshape(X_train.shape[0],X_train.shape[1] , 1)
X_test = X_test.reshape(X_test.shape[0],X_test.shape[1] , 1)
```

```
model=Sequential()
model.add(LSTM(8,activation='sigmoid',return_sequences=True,input_shape=(time_step,1)))
model.add(LSTM(8,activation='sigmoid',return_sequences=True))
model.add(LSTM(8,activation='sigmoid',return_sequences=True))
model.add(LSTM(8,activation='sigmoid',return_sequences=False))
model.add(Dense(1))
model.compile(loss='mean_squared_error',optimizer='adam')

model=Sequential()
model.add(LSTM(8,activation='sigmoid',return_sequences=True,input_shape=(time_step,1)))
model.add(LSTM(8,activation='sigmoid',return_sequences=True))
model.add(LSTM(8,activation='sigmoid',return_sequences=False))
model.add(Dense(1))
model.compile(loss='mean_squared_error',optimizer='adam')

model=Sequential()
model.add(LSTM(8,activation='sigmoid',return_sequences=True,input_shape=(time_step,1)))
model.add(LSTM(8,activation='sigmoid',return_sequences=False))
model.add(Dense(1))
model.compile(loss='mean_squared_error',optimizer='adam')
```

```
model=Sequential()
model.add(LSTM(8,activation='sigmoid',return_sequences=True,input_shape=(time_step,1)))
model.add(LSTM(8,activation='sigmoid',return_sequences=False))
model.add(Dense(1))
model.compile(loss='mean_squared_error',optimizer='adam')

model=Sequential()
model.add(LSTM(8,activation='sigmoid',return_sequences=False,input_shape=(time_step,1)))
model.add(Dense(1))
model.compile(loss='mean_squared_error',optimizer='adam')

history=model.fit(X_train,y_train, validation_data=(X_test, ytest),epochs=100)

train_predict=model.predict(X_train)
model.reset_states()
test_predict=model.predict(X_test)

train_predict.shape, y_train.shape
```

```
# Mean Squared Error (MSE)
MSE_train = np.square(np.subtract(y_train,train_predict[:,0])).mean()
MSE_test = np.square(np.subtract(ytest,test_predict[:,0])).mean()
print('Mean Squared Error (MSE) for train: ' + str(np.round(MSE_train, 5)))
print('Mean Squared Error (MSE) for test: ' + str(np.round(MSE_test, 5)))
# # Mean Absolute Error (MAE)
MAE_train = np.mean(abs(train_predict[:,0] - y_train))
MAE_test = np.mean(abs(test_predict[:,0] - ytest))
print('Mean Absolute Error (MAE) for train dataset: ' + str(np.round(MAE_train, 6)))
print('Mean Absolute Error (MAE) for test dataset: ' + str(np.round(MAE_test, 6)))

Ytrain=pd.DataFrame(y_train)
Ytrain.columns=['Actual']
train_pred=pd.DataFrame(train_predict[:,0])
train_pred.columns=['Predicted']
train_df=pd.concat([Ytrain, train_pred], axis=1, join='inner')
train_df.corr()
pow(train_df.corr(),2)
```

```
Ytest=pd.DataFrame(ytest)
Ytest.columns=['Actual']
test_pred=pd.DataFrame(test_predict[:,0])
test_pred.columns=['Predicted']
test_df=pd.concat([Ytest, test_pred], axis=1, join='inner')
test_df.corr()
pow(test_df.corr(),2)

# plot loss
plt.subplot(211)
plt.title('Loss')
plt.plot(history.history['loss'], label='train',color='orange')
plt.plot(history.history['val_loss'], label='test')
plt.legend()
plt.plot(history.history['loss'])
```

APPENDIX 2

Coding of R Programming

```
library(modifiedmk)
mmky(br_py$RATIO)
mmky(k_py$RATIO)
mmky(m_py$RATIO)
mmky(n_py$RATIO)
mmky(sj_py$RATIO)
```

```
library(trend)
trend::pettitt.test(br_py$RATIO)
trend::pettitt.test(k_py$RATIO)
trend::pettitt.test(m_py$RATIO)
trend::pettitt.test(n_py$RATIO)
trend::pettitt.test(sj_py$RATIO)
```

```
#ADF test
library(tseries)
adf.test(br_py$RATIO)
adf.test(k_py$RATIO)
adf.test(m_py$RATIO)
adf.test(n_py$RATIO)
adf.test(sj_py$RATIO)
```