MODELLING OF AMMONIA NITROGEN IN RIVER USING ARTIFICIAL INTELLIGENCE TECHNIQUES

CHAI VOON HAO

A project report submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering (Honours) Civil Engineering

Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

MAY 2021

DECLARATION

I hereby declare that this project report is based on my original work except for citations and quotations which have been duly acknowledged. I also declare that it has not been previously and concurrently submitted for any other degree or award at UTAR or other institutions.

Signature : *Chai*

Name : Chai Voon Hao

ID No. : 1601346

Date : 07/05/2021

APPROVAL FOR SUBMISSION

I certify that this project report entitled **"MODELLING OF AMMONIA NITROGEN IN RIVER USING ARTIFICIAL INTELLIGENCE TECHNIQUES"** was prepared by **CHAI VOON HAO** has met the required standard for submission in partial fulfilment of the requirements for the award of Bachelor of Engineering (Honours) Civil Engineering at Universiti Tunku Abdul Rahman.

Approved by,

Signature : 

Supervisor : Dr. Chin Ren Jie

Date : 07/05/2021

ACKNOWLEDGEMENTS

I would like to thank everyone who had contributed to the successful completion of this project. I would like to express my gratitude to my research supervisor, Dr. Chin Ren Jie for his invaluable advice, guidance and his enormous patience throughout the development of the research. Dr. Chin provided me the dataset needed to complete this final year project. Besides, Dr. Chin is also willing to teach and guide me on the usage of MATLAB software to simulate AI models in order for me to generate the results needed by this research. Furthermore, Dr. Chin is also passionate in motivating me along the journey of this final year project, this final year project will not be as smooth if without his encouragement.

Next, I would like to express my deepest thanks to my friend, Mr. Bryan Yap Seng Haw who had offered valuable technical support especially his guidance on RapidMiner software and invaluable suggestions along my venture of final year project. In addition, I would also like to express my gratitude to my loving parents and friends who had helped and given me encouragement throughout the period of this final year project.

ABSTRACT

River with high water quality is essential for the survival of living organisms and human being. Ammonia Nitrogen is one of the water quality parameters or chemical pollutants that severely affect the water quality of rivers in Malaysia. Therefore, a precise estimation on ammonia nitrogen concentration in river is utmost important in various fields including agriculture, water resources and irrigation fields. The aim of this study is to predict the concentration of ammonia nitrogen in river using advance mathematical prediction models with the help of artificial intelligence (AI) techniques. The selected study area in this study is Langat River located in Selangor, Malaysia. Three AI models namely Back Propagation Neural Network (BPNN), Adaptive Neuro-Fuzzy Inference System (ANFIS) and Support Vector Machine (SVM) were developed, and their prediction performance were compared. Five water quality parameters were chosen as the input variables for the training and testing of the AI models. The five water quality parameters included Dissolved solids (DS), turbidity (T), total solids (TS), phosphate (PO_4^{3-}) and nitrate (NO_3^-) were obtained from Department of Irrigation and Drainage (DID) of Malaysia. The water quality parameters mentioned contain 77 dataset which was then utilized as input variables to train and test the AI models. 80% of the 77 dataset were used in the training process while the other 20% of the 77 dataset were used in the training process of the AI models. Min-max normalization was utilized to normalize the ammonia nitrogen concentration values to a range of 0 to 1 prior to the training process. The results generated by the AI models were interpreted and the performance of the AI models were evaluated by statistical analyses comprised of coefficient of determination (R^2), mean squared error (MSE), root-mean-squared error (RMSE), mean absolute error (MAE) and average percentage error. The performance of AI models was intra-compared among their own type and the AI model with the best performance was selected from each developed AI model type. The three best BPNN, ANFIS and SVM models were then compared among one another in term of prediction performance. The comparison result showed that BPNN model trained with log sigmoid function with 4 hidden neurons has the best performance compared to the ANFIS and

SVM models. Therefore, this BPNN model is concluded to be the most suitable AI model to predict ammonia nitrogen concentration in Langat river. One of the recommendations for future research on this topic includes obtaining more dataset for AI model development. Another recommendation is to replace SVM with Support Vector Regression (SVR) as SVR is more effective in solving regression problem.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS / ABBREVIATIONS

AI	artificial intelligence
ANFIS	adaptive neuro-fuzzy inference system
BPNN	back propagation neural network
DO	dissolved oxygen
DS	dissolved solids
INWQS	interim national water quality standard
MAE	mean absolute error
MF	membership function
MSE	mean squared error
NH_3	un-ionized ammonia
NH_4+	ammonium ions
NO_3^-	nitrate
PO_4^{3-}	phosphate
R	correlation coefficient
R^2	correlation of determination
RMSE	root mean squared error
SVM	support vector machine
T	turbidity
TS	total solids
TSS	total suspended solids

CHAPTER 1

INTRODUCTION

1.1 General Introduction

Water has always been an essential component for the survival of living organisms. A country's advancement in economy mostly depends on the quality of river water. As most of the nations are developing tremendously in this era, degradation of water quality in river has become a severe issue that needs to be concerned by government worldwide (Wu, et al., 2017). Different types and intensities of agricultural, anthropogenic and industrial activities are often the main factors that affect the water quality of a river. In all the surface water system, quality of river water was found to be most negatively affected due to their accessible convenience for dumping of waste from other channels such as drains, canals and tributaries (Singh, et al., 2009). Besides, the environmental natural phenomenon such as rainfall, weathering processes and erosion of soil also impact river water quality, but the effect is milder compared to other man-made factors (Wu, et al., 2017). Consequences of river water quality deterioration are a decline in health of aquatic lives and cause partial or complete change of species composition in the polluted river watershed (Ouyang, 2005). Nowadays, a great number of rivers are no longer able to sustain aquatic lives and not suitable for human activities due to worsening water quality. According to one of the research studies on Malaysia rivers, human activities such as agricultural activities and industrialization around the watershed had cause 13 tributaries and 36 rivers to be polluted in the year 1999 and the number of polluted rivers will increase if water quality is left unmonitored (Mokhtar, et al., 2010). Therefore, assessment of river water qualities is utmost important in developing countries as rivers with acceptable water quality will become scarce in future (Pesce, 2000).

Water quality can be branched into three elements which are hydro-morphological, biological and physiochemical quality. In most of the water quality modelling, researchers often choose to use physiochemical water quality variables to determine water quality. Different kind of water quality variables could combine to yield interdependent effects on water quality. For instances,

chemical and physical water variables would be coupled with a few other environmental parameters and finally generate unpredictable results of water quality (Tiyasha, Tung and Yaseen, 2020). However, whether these parameters are monitored independently or in group, only partial information of the total water quality would be retrieved (Pesce, 2000). In common practices, determination of river water quality is usually based on several parameters which include biochemical oxygen demand, dissolved oxygen, nitrogen compounds, solids, chlorophyll-a, water temperature, pH, chemical oxygen demand and electrical conductivity (Tiyasha, Tung and Yaseen, 2020).

1.2 Problem Statement

Ammonia Nitrogen comprises un-ionized ammonia (NH_3) and ammonium ions (NH_4+). Ammonium will dominate ammonia when pH is lower than 8.75, whereas ammonia is the predominant form at pH more than 9.75 (Li, et al., 2020). In the natural condition of river water, ammonia nitrogen occurs as ammonium. However, the portion of un-ionized ammonia builds up as temperature and pH levels inclined (Lin, et al., 2019).

Although the fraction of ammonia is much lower than ammonium in natural water, the toxicity of ammonia is immensely high and dissolved ammonia act as the main contributor to toxicity in ammonia nitrogen. Hence, the concentration of ammonia nitrogen should be studied at all times to maintain the health of rivers. The high toxicity of un-ionized ammonia will cause irreversible negative impacts on aquatic organisms' growth, weight of organs and condition of gill. As a consequence, infancy and survival of fish and other aquatic organisms would be diminished immensely and finally lead to extinction of certain aquatic populations. In return, development of the economy would be affected, and human health will be another concern. Besides, redundant amount of ammonium favours the production of phytoplankton, this will eventually lead to algae bloom and eutrophication in bodies of water. Following the mass death of algae, the appearance of large quantity of dissolved organic matter would significantly reduce the concentration of dissolved oxygen through microbial respiration (Li, et al., 2020). Therefore, survivability of aquatic organism becomes an issue due to hypoxia and environmental issues will also arise such as increase emission of nitrous oxide (Lin, et al., 2019).

As one of the most hazardous water pollution parameters, ammonia nitrogen has dealt tremendous impact on the ecological community by discharging through anthropogenic runoff into aquatic ecosystem (Yu, et al., 2020). Therefore, it is crucial to monitor the concentration of ammonia nitrogen to minimize ammonia nitrogen pollution in river water. However, the relevant study is still limited.

Therefore, it is necessary to develop a reliable model to accurately predict ammonia nitrogen concentration.

1.3 Aim and Objectives

The aim of this study is to predict the concentration of ammonia nitrogen in river with the help of advance predictive tools. The objectives are:

i. To develop the mathematical models for ammonia nitrogen prediction using artificial intelligence (AI) techniques.
ii. To develop back propagation neural network (BPNN), adaptive neuro-fuzzy inference system (ANFIS) and support vector machine (SVM) by using unnormalized and normalized dataset.
iii. To access the developed AI models using statistical analyses.

1.4 Scope and Limitation of the Study

The study area in this research is Langat River in Selangor, Malaysia. The total catchment area of Langat River is roughly around 1815 km^2. Some of the major tributaries of Langat River are Lui River, Semenyih River and Beranang River. Langat River flows through three districts in Selangor which are Sepang, Hulu Langat and Kuala Langat.

Langat Reservoir and Semenyih Reservoir are located inside Langat River basin. The main purpose of Semenyih Reservoir is to supply water to domestic area and industrial area whereas Langat Reservoir is mainly used for generation of electricity for consumers within Langat Valley. Since the basin which Langat River is located experienced fast pace urbanization, new constructions area, expansion of agricultural and industrialization and modern development of road network, hence water quality control is essential. Therefore, this is the main reason Langat River is chosen as the study area.

In this study, three Artificial Intelligence (AI) techniques are selected and used to predict ammonia nitrogen concentration in Langat River. The three AI techniques are Back Propagation Neural Network (BPNN), Adaptive Neuro-Fuzzy Inference System (ANFIS) and Support Vector Machine (SVM). The accuracy of each model is accessed by several statistical analyses which are correlation coefficient (R), correlation of determination (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE) and percentage error.

1.5 Contribution of the Study

This study aims to predict ammonia nitrogen concentration in river by applying Artificial Intelligence (AI) techniques. Unlike the conventional numerical models, AI models are able to conduct self-learning on water quality parameters through complex mathematical models to provide prediction of ammonia nitrogen with an acceptable level of accuracy. In Malaysia, rivers are one of the most important natural resources in agricultural activities, power generation, sustaining human lives, etc. Hence, a computational model that can predict the concentration of ammonia nitrogen with a considerably high level of accuracy is essential for decision-makers and policymakers.

1.6 Outline of the Report

There are a total of 5 chapters containing in this research. In Chapter 1, background of the water quality and ammonia nitrogen is briefly explained. The aim and objectives are also outlined in Chapter 1. Chapter 2 involves the literature review on water quality, AI models and role of AI models in water quality parameters prediction. In Chapter 3, the detail on the study area is provided and the way procedure of data collection and preparation are explained too. Moreover, the methodology in developing the proposed AI models are also demonstrated in Chapter 3. Under Chapter 4, the results of each output values generated by each AI models are presented. Besides, the proposed AI models are compared in terms of prediction accuracy on ammonia nitrogen, performance and effectiveness of the models by evaluation with statistical analyses. In Chapter 5, discussion on the performance of each AI model is carried out and the AI model with the highest performance is determined.

Furthermore, recommendation is also provided to improve the performance of AI models in future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

In this chapter, a brief introduction of a few water quality parameters will be presented. The water quality parameters include those that will be utilized in the development of proposed AI models. Besides, the three proposed AI models namely BPNN, ANFIS and SVM will be described and their contribution and usage in civil engineering and hydrological application will also be reviewed. Lastly, the role of AI models in water quality parameters prediction will be discussed and reviewed based on past researchers' studies.

2.2 Water Quality

Nowadays, the quality of water is utmost important not only for the survival of animals and plants but human also rely on high quality of water to survive without getting health issues. Different country will have different standard on the water quality no matter it is raw water, freshwater or aquaculture water. One of the standards adopted by Malaysia to analyse river water quality is the Interim National Water Quality Standard (INWQS). A national study on "Development of Water Quality Criteria and Standards for Malaysia" was initiated by the government of Malaysia in 1985. INWQS was completed and established by reviewing more than 120 biological and physicochemical parameters in the study. Six classes were termed in the INWQS to classify water quality parameters from class I to V to which class I indicates the parameters are in "best" condition while class V indicates the "worst" condition. INWQS layout that class I to III to be the significant level of water quality suitable to maintain the survivability of majority of the aquatic life depends on the sensitiveness of the aquatic life by considering fish as an indicator. Fish is used due to its high level of economic value. Moreover, class I water supply is drinkable without any treatment to the water needed while class II water supply needs to undergo conventional water treatment before being consumed by humans. Whereas more comprehensive water treatment is essential for water with class III quality. However, class IV water is no longer suitable for consuming, but it can still be

used for irrigation of agricultural crops. Last but not least, water of class V is too contaminated until the extent that it is not suitable for any kind of usage. INWQS play an important role and act as a guideline for decision-makers to make decisions based on the designated beneficial usage. Based on INWQS, water quality of rivers in Malaysia can be identified effectively so that recovery process can be carried out immediately (Zainudin, 2010).

2.2.1 Dissolved Oxygen

Dissolved oxygen is one of the most essential water quality parameters used for characterization and indication of the water quality level in the aquatic ecosystems. Oxygen dissolves into the water in several ways, one of the sources of dissolved oxygen is from the atmosphere through reaeration process. Besides, oxygen is also produced by plantation inside the water body through photosynthesis. However, several activities beneath the water surface contributed to the declination of oxygen content in water body, such as oxidation process by nitrogenous and carbonaceous material, uptake of oxygen by aquatic plants and organisms and oxygen demand required by sediments at the riverbed (Csábrágia, et al., 2017). Nevertheless, the dissolved oxygen concentration also influenced by other factors, for instance, salinity, temperature of water and etc. Dissolved oxygen concentration reflects the overall health in a river and other sources of water bodies, therefore dissolved oxygen often acts as a water quality control parameter at various aquatic systems, to name a few which are wetlands, reservoirs and aquacultures (Olyaie, Abyaneh and Mehr, 2017).

Extremely low concentration of dissolved oxygen in a river that is previously with optimum oxygen content will introduce severe impacts to the river system. The conditions in the river become anaerobic due to lack of dissolved oxygen, therefore large quantity of fishes and aquatic plants will decrease, bad odour will be emitted into the atmosphere and the river's aesthetic will be destroyed, hence render the whole ecosystem in the river become unbalance. To cope with the reduced dissolved oxygen concentration in the river, some of the aquatic animals' breathing mechanisms would be forced to modify while those who can't adapt would eventually go belly up. However, some aquatic animals would reduce the need to carry out activity. The consequences of these actions are the retardation in their development such as deformation of

certain body parts and risen of reproductive issues such as incomplete hatching of egg due to defects of lives during the egg stage. Health, feeding and digestion of fish will be badly influenced when the dissolved oxygen content in the river is lesser than 3mg/L (Cox, 2003).

A research study the dynamics of dissolved oxygen in the River Thames of United Kingdom was carried out over a period of 6 years from 2009 to 2014 by the implementation of a combination of models and direct analysation on the results observed. The research had concluded that the reduction of dissolved oxygen content in the Thames river which is a lowland river is due to a combination of 2 factors. The first factor is due to the authigenic processes where the algal biomass is decomposed by microorganisms where eventually the oxygen demand becomes high and more oxygen is depleted from the water body. The second factor is due to allogenic processes cause by flowing of organic matter into the river water due to extreme summer floods at the Thames river in 2012. The presence of organic matter in the river will cause benthic respiration to be carried out and hence deplete the dissolved oxygen content in the river (Hutchins, et al., 2020).

2.2.2 Total Suspended Solids

Several sources contributed to the Total Suspended Solids (TSS) at coastal areas and in rivers, they are dredging activities, tidal currents, river runoffs and resuspension events. The content of the TSS in the water column includes particles of inorganic and organic matters. The organic matters present in TSS includes living and non-living things, for instance, phytoplankton. While the inorganic matters in TSS encompass suspended materials and clay. High concentration of TSS may cause defective water quality and has a chance to escalate water temperature at the upper layer of water bodies. While TSS flowing along the river, pollutants, heavy metals and nutrients will bind to TSS and eventually turns TSS into one of the major pollutants, hence given rise to detrimental environmental effects in the water bodies by greatly reduce the water quality. The accumulation of a large amount of TSS near the water surface would significantly affect light penetration into deeper layer of water bodies, hence reducing the plants' photosynthesis productivity, disrupting the ecosystem operation and reduce the aesthetics of the water bodies. On the other hand, some

coastal features or areas such as mangrove wetlands, river deltas and marsh required TSS to preserve the accumulation of sediment and also act as protective barriers. When TSS are trapped upstream by reservoirs, rise of sea level and TSS washed away by delta distributary flow, these coastal features would be destroyed. The coastal feature such as mangrove wetland is important and act as a protection to store blue-carbon TSS and act as defensive barriers in opposition to flood and storm surge (Balasubramaniana, et al., 2020).

Besides, due to rapid development at the coastal regions, modification of the land use at particular watershed will generate higher quantity of TSS into the river mouth, hence polluting the estuary. The fine TSS at shallow estuary is easily be washed and re-suspended to the surface of water bodies by waves, tides and wind, hence increase the rate of turbidity (Altunkaynak and Wang, 2011). One of a study by previous researcher shows that increment of settling of organic matter which in his research is algal cells in the shallower water bodies has already been proven to cause increase rate of sediments accumulation (Brezonik and Engstrom, 1998).

The following research by Carlos, 2016 was carried out by assessing the growth performance and effect on L.vannamei, a type of shrimp in different TSS concentrations of Biofloc Technology (BFT) system in a period of 42 days. Five groups of shrimps are put to live in five different BFT system with TSS concentrations of 250 mgL^{-1}, 500 mgL^{-1}, 1000 mgL^{-1}, 2000 mgL^{-1} and 4000 mgL^{-1}. Each BFT tanks contains 200 L of water and the dissolved oxygen (DO) content was maintained beyond 5 mgL^{-1} throughout the research period. At the end of the research, the shrimps L.vannamei growth performance were akin, hence concluded that different TSS concentration does not bring any effect on the growth of the shrimp under the condition of constant DO of above 5 mgL^{-1} being maintained in the BFT tank. Besides, it was observed by the author in this research that the concentration of TSS increased gradually when the culture time progress (Gaona, et al., 2016). From the comment of the author, it was recommended to keep maximum TSS concentration in the range of 500-600 mgL^{-1} in a BFT system. The excessive TSS concentration could interact with other water quality parameters such as nitrogen compounds and possess stress to cultured organisms. Moreover, another study concluded that higher

concentrations of TSS will congests the gills of the shrimp L.vannamei and contribute to respiration problem of the shrimp (Schveitzer, et al., 2013).

The estimation of TSS concentration at the coast and estuary is often difficult as it includes a vast range in temporal and spatial variability. Due to this condition, failure in evaluating TSS concentration at coastal and estuary by applying traditional field sampling methods often arise as these methods are not efficient in spatial and temporal sampling. In one of the studies, the author developed a moderate resolution imaging spectroradiometer (MODIS) 250 m based TSS retrieval model to map the wide range concentration of TSS in estuary and coast of China. The unique advantage of this MODIS 250 m is that it contains ocean colour remote sensing with coverage up to 2 times daily. Moreover, this image sensor could provide 250 m spatial resolutions for the sediment movement at dynamic lake or coast with a clear image. From the result, the author concluded that the values of spectral log-ratio increase with an increase in TSS concentration until the extent of TSS concentration is less than 31 mg/L. When TSS concentrations are higher than 31 mg/L, the spectral log-ratio values then decrease. The spectral log-ratio model is very sensitive to high TSS concentration which has the range from 160 to 577.2 mg/L. Therefore, detection of high temporal variability in TSS exposed in the tidal cycle can be accomplished by applying log-ratio model as it enhances MODIS. With MODIS coupled with log-ratio model, detection of movement of sediment at coastal water can be accomplished with higher accuracy (Chen, et al., 2015).

Despite having TSS originate from upstream and resuspension of sediment from the waterbed, road-deposited sediments (RDS) is also a source of TSS due to the transportation of runoff from rainfall carrying RDS into streams, lakes and coastal water without undergoing any water treatment. A study was carried out based on the relationship of TSS and RDS in rainfall-runoff. The study assessed the effect of heavy metal contained in the RDS on the TSS heavy metal pollutants content due to rainfall runoff and also evaluate the concentration of TSS effluent from industrial areas. According to the research, TSS contains 1.5 to 10.0 times the quantity of heavy metals concentration than RDS where the particle size of the heavy metals falls below 63 μm. However, the heavy metals in RDS which are smaller than 125 μm when washed away by rainfall-runoff, only 22.1% of it is contributed to the heavy

metal's concentration in TSS. So, the other 77.9% of heavy metals concentration in TSS might be originated from effluent of industries within the same watershed instead of RDS from traffic activity. Therefore, it was concluded that the metal by-product from metal processing industries such as milling, cutting, etc. contributed the largest portion of heavy metal pollutants into the stream by runoff and hence increase heavy metals concentration in TSS (Jeong, et al., 2020).

2.2.3 Turbidity

Turbidity is one of the most important water quality parameters used to assess the quality of water bodies. Nevertheless, the concentration of turbidity reflects the load of suspended sediments present in a water body. Turbidity also indicates the clarity of water with turbidity value of zero for pure water. A high amount of sediment loads lead to higher value in concentration of turbidity, hence the penetration of light into the water bodies is limited. Generally, turbidity is not a chemical nor a physical parameter of water quality, it is termed as an optical property of a water body primarily induced by scattering of light and to some degree by the absorption of light of the photons pathlength. To be exact, more light will be scattered by a higher amount of suspended particles hence reduce the clarity in the water column. The distribution size of suspended particles directly affects the turbidity of water. Although the concentration of suspended solids cannot be directly be indicated by turbidity, turbidity is still a satisfactory indicator. Turbidity is a useful water quality parameter to determine coastal morphodynamics, sediment budgets, the dynamics of contaminants and the interaction between contaminants and cohesive sediments.

One of the reasons for high concentration of turbidity is due to algal blooms, but the dominant cause of high turbidity is due to the presence of total suspended sediments. Other sources that cause an increase in turbidity include untreated industrial materials being deposited into rivers and human activities such as agricultural activities, land change and construction by the river. Nevertheless, coastal wind with high speed will cause shear stress to develop at the waterbed hence follow by re-suspension of suspended sediments and cause turbidity value to raise at shallow coastal water. Without neglecting rainfall runoff as an important factor that contributes to the turbidity in river, low level

of rainfall-runoff may reduce the turbidity value as lesser sediments from the watershed flow into the river. High value of turbidity can bring significant impact to the environment and aquatic ecosystem, one of the impacts is the deterioration of dissolved oxygen in the water bodies due to decomposition process by organic matter correlated with high concentration of sediment. Moreover, longevity and infilling of reservoirs and lakes would be indirectly affected by high value of turbidity (Abirhire, et al., 2020). Turbidity usually addressed in different measurements units, for instances Formazin Turbidity Unit (FTU), Nephelometric Turbidity Unit (NTU) and Formazin Nephelometric Unit (FNU). Moreover, the European Union of the Marine Strategy Framework Directive had included turbidity as a compulsory water quality parameter to be measured (Constantin, Doxaran and Constantine, 2016)

According to one of the studies done by previous researcher, the author had made the following conclusion. It was mentioned that during period of high precipitation, the increasing material runoff from the watershed appears to be the dominant factor that contributes to the inclination of turbidity value. On the other hand, during low precipitation period with lower material runoff from the watershed, the material runoff is still a significant factor that affects the turbidity as the lower water level in river causes the material runoff concentration to become more concentrated. However, during medium precipitation period, pollutants discharge into the river dominates the material runoff thus become the main source for high turbidity. The most influencing pollutants addressed here is the ammonium as previous researcher had concluded that it's the most important water quality variable used to predict turbidity. These urban waste pollutants contain nutrients that favour eutrophication and eventually cause algal bloom and reduce the opacity of the water bodies and also the insoluble substances present in the urban waste act as suspended particles that scattered the light that penetrate the water bodies, therefore increasing turbidity (Nieto, et al., 2014).

There are multiple researches done by previous researchers that used different methods to determine and predict turbidity value in different water bodies. One of the research papers suggested the use of a 'river drifter' which is a disposable device that used to determine turbidity in the river as it flows along with the stream current. The measured turbidity data are being transmitted to a

reading device in real-time. From the research, it shows that this river drifter instrument can also identify the sources of sediments and also determine the reason for decreasing sediment loads. The advantages of the river drifter are that the position of the instrument can be tracked using GPS, turbidity ranges from 0-300 NTU can be measured accurately, it is environmentally friendly and the cost of production is cheap. Besides, the river drifter instrument can travel in a large river from 10 km to 100 km. However, some of the tested river drifter instruments were stranded by vegetation at the riverbank, but it is not a big issue as the river drifter instruments can be recovered and deployed on the spot again. Moreover, the GPS signal from the instrument itself would indicate its current location so that the researcher can locate the instrument and deploy it again. If the instrument is stranded at dangerous locations, it is usually left unrecovered as the cost to produce a new river drifter is cheaper than the cost to recover the stranded one (Marchant, et al., 2015).

MODIS red band at 250 m spatial resolution was applied in determining turbidity of coastal water at Danube Delta coastal area. The turbidity data from remote sensing from MODIS were compared with in-situ turbidity data. MODIS is an advantageous remote sensing tool as broad area's turbidity value can be identified and mapped at once. During the days the sky is clear, the turbidity dynamics can be evaluated within a few hours of intervals. From the data retrieved by MODIS, the further the distance of water from the river mouth, the lower the turbidity value was retrieved. This is due to settling of suspended solids on the waterbed as flocculation process occurred and the fast water speed of river decreases as it mixes with the seawater. Suspended particles with larger sizes sink to the bottom of seafloor faster than finer particles at the coastline while finer particles will be transported further away from the shoreline and finally settle progressively. At the estuary, resuspension of inorganic particles and solids discharged from river are two of the main sources that contribute to the turbidity at river mouth. Lower turbidity value is observed far away from offshore and the turbidity at these areas typically affected by phytoplankton which is primarily consisting of organic particles. The research had proven that turbidity data collected by MODIS is accurate to some extent and is capable enough to be applied in coastal turbidity value retrieval and mapping (Constantin, Doxaran and Constantine, 2016).

2.2.4 Nitrate

Nitrate presents as a crucial pollutant for surface and ground waters. The source of nitrate comes from nitrogen in water. Nitrogen contained in water goes through the oxidation process and nitrite is generated. Then, the nitrite in the water is converted to nitrate by bacteria in the water by a process by binding oxygen that is available in water. Nitrogen-containing compounds enrich the nutrients content in rivers, streams and reservoirs. Some of the most common main sources of nitrogen in aquatic ecosystem are originated from leakage and disposal materials from septic tanks, animal wastes, industrial wastewater, gases exhausted by vehicles, feedlot discharges and lawn and fertilized field runoff. The nitrate concentration in water bodies such as streams and lakes are especially high during heavy rainfall season (Sulaiman, et al., 2014).

In the agricultural field nowadays, the plantation of vegetables and crops often accompanied by the usage of chemical fertilisers that contain nitrate to boost their growth rate. However, not the whole amount of this chemical fertilisers is used up by crops, the excessive nitrate-containing fertilisers discharge into rivers, lakes, streams and reservoirs. The nitrate then binds with soil particles and also cause air pollution. Moreover, high concentration of nitrate in water may lead to reduction of dissolved oxygen content in the water body and promotes eutrophication. Excessive concentration of nitrate can accelerate the production of plankton and algae which leads to algal bloom. Upon the decease of the plankton and algae, the bacteria undergo decomposition process would consume a large amount of oxygen. Therefore, the dissolved oxygen content in the water body will deplete and eventually no longer sustainable for the survival of aquatic organisms. However, nitrate in small amount is needed by aquatic organisms to sustain their metabolism and growth (Massah and Vakilian, 2019).

It was presumed that nitrate might develop negative impacts on human health. Previous research indicated that dietary nitrate caused critical diseases to the human body such as gastric cancer, birth defects, goitre and methaemoglobinaemia disease of infants. Moreover, there is a probability of development of chronic kidney disease if vegetables that contains immense nitrate concentration is consumed by humans. The Environmental Protection

Agency had imposed that the nitrate concentration in drinking water should be limit to 10 mg/L (Massah and Vakilian, 2019).

Nitrogen is unable to be stored in plant biomass due to deforestation for purposes such as urban development and agricultural activities. Therefore, the nitrogen flows into the stream together with rainfall-runoff. For instances, agricultural land use causes an enormous amount of nitrogen to flow into rivers in watersheds in Changjiang, Mississippi and the Seine. 20% of the nitrogen content in these watersheds are contributed by the agricultural land use. In comparison of different kind of land, streams that are located in urbanised watersheds contain higher amount of nitrate content than streams that are located in watersheds that only have agricultural activities. Lowest nitrate concentration was observed in natural watersheds without any human activities (Muthukrishnan, Lewis and Andersen, 2007).

Previous researches show that reservoir might reduce the effluent nitrate concentration. Research in Illinois shows a 4400-ha reservoir remove 58% of the input nitrate concentration at the outlet. In a research study, the outflow nitrate concentration of the Saylorville Reservoir was modelled by implementing a transfer function approach by taking into account the inflow concentrations of nitrate of the reservoir. The result of this study shows that the downstream nitrate concentration of Saylorville Reservoir can be efficiently modelled by implementing the transfer function that its input variables are the inflow nitrate concentrations from previous month and current month. The transfer function model could predict the nitrate concentration downstream of the reservoir accurately in a time series. This study has also proven that reservoir can reduce the outlet nitrate level. A depletion of nitrate concentration of 22 +- 6% was indicated in downstream water from reservoir. The reason causing the reduction is mainly due to the transfer function model take into account the temporal averaging inherent in the reservoir system (Schoch, Schilling and Chan, 2009).

2.2.5 Ammonia Nitrogen

Ammonia nitrogen is an important water quality parameter in biotechnology, agricultural and clinical industries. Ammonia nitrogen is made up of free ammonium ions (NH_4^+) and ammonia (NH_3) which is equilibrium under natural condition. The concentration of NH_3 increases as the pH and temperature of water body increase. An excessive amount of ammonia nitrogen present in water bodies may lead to death of aquatic organisms, human health, eutrophication, etc. The global nitrogen cycle had been altered by human activities due to the fact that excessive amount of ammonia nitrogen was determined at many of the water bodies around the world that accompanied with urbanization (Lin, et al., 2019).

Ammonia nitrogen found in rivers and lakes commonly present in the wastewater discharge of point and non-point sources from factories and agricultural lands. To provide a safe aquatic environment for aquatic organisms, governments around the world had established guidelines on the limit of ammonia nitrogen allowed to present in water bodies. For instances, in New Zealand and Australia, the maximum concentration of ammonia nitrogen should be limit within 30 μg/L and being maintained at pH of 8.0 in warm water. On the other hand, England and Canada have stricter criteria where the concentration of ammonia nitrogen should be limit within 15 μg/L and 19 μg/L respectively. While in China, to ensure constant and uninterrupted food supplies by protecting the fish reproduction, the Chinese government has set a guideline for ammonia nitrogen to not exceed 20 μg/L in the aquaculture industries (Zhang, et al., 2018).

According to a study conducted by the past researcher, ammonia nitrogen had polluted 43% of the rivers in Malaysia (Zainudin and Mamun, 2013). The Chinese Ministry of Environmental Protection stated that China's yearly emission of ammonia nitrogen into water bodies overshoot 2.3 million tons from 2011 to 2014. Another researcher had concluded that ammonia nitrogen appeared to be the most hazardous and the most toxic of all the nine contaminants that were found in Keelung River in Taiwan (Zhang, et al., 2018). In 2017, Sunai Benut in Malaysia was contaminated with up to 13mg/L of ammonia nitrogen and the source of pollution is leachate flowing out from a landfill into the Simpang Renggam water treatment plant. It was reported that

the contaminated river water had pungent odour (Zainudin, 2017). On another occasion, water of thousands of residents was cut due to ammonia contamination in Semenyih dam. The ammonia was found to be discharged from a factory by the Langat River into Semenyih dam (Perumal, 2016). Besides, on 4 April 2019, a reservoir that contained ammonia had burst and the contaminated water flowed into Sayong River and 17,000 households in Kulai, Johor were affected by water supply disruption (Anon, 2019). On 30 April 2019, 18,076 households in Malacca experienced water disruption due to ammonia contamination in Sungai Batang Melaka which supply raw water to Gadek water treatment plant in Alor Gajah. The pollution occurred due to two catfish breeders discharging polluted water from fishing ponds into the river (Anon, 2019).

Recently, there are a lot of methods and techniques to restore river to its unpolluted state. One of the rising technologies used in water quality restoration is the Bacterial technology Method (BTM). BTM is now commonly applied in wastewater treatment and river water treatment and can be commonly found in usage to treat industrial and domestic wastewater worldwide. BTM is well known for its efficiency in reducing BOD and COD in effluent within a very short time. However, an immense amount of work is needed to be carried out upon evaluation of water quality variables while using usual sampling method during the process of BTM (Kabo-bah and Xie, 2012).

In a study, the author applied mathematical algorithms to test and determine the ammonia nitrogen concentration in Xuxi River, China. Six water quality parameters were used to model ammonia nitrogen concentration in the study, they are water temperature, dissolved oxygen, total phosphorus, transparency, total nitrogen and chemical oxygen demand. The gathering of data was carried out at five sampling points. A weir of 50cm in height was installed at the last sampling site. The purpose of this weir is to restrain the flow in the river during high and low discharges. Therefore, the purification process in the polluted river could be carried out by bacteria under a stable condition. BTM works by injecting microbial accelerator and specific bacteria into the polluted river in order to trigger the native bacteria which is already in the river. The mathematical algorithm used in this study is the Virtual Beach (VB) MLR model. The function of VB model is to aid user in selecting the most appropriate

model in consideration of specific explanatory variables. After the development process, only 6 out of 12 mathematical models were appropriate for the determination of ammonia nitrogen concentration. The data retrieved in the year 2008 was used in the six models for ammonia nitrogen evaluation. However, the prediction errors of all the six models are between +-20% and +-36%. These prediction errors are due to several causes. One of the causes is difficulty in the measurement of ammonia nitrogen due to rapid transformation of ammonia nitrogen into other nitrogen compounds. Besides, the accuracy of final result may be contributed by the accumulation of instrumental errors, observer's mistakes, sampling errors and laboratory errors. The author also thought that involvement of different person during the 2 years measurement period might affect the accuracy of ammonia nitrogen prediction. Improvement of the mathematical models is needed to enhance the prediction accuracy of water quality variables (Kabo-bah and Xie, 2012).

Moreover, analytical methods are used by researchers to determine ammonia nitrogen content. Analytical methods can be categorised into three categories which are fluorometric, spectrophotometric and electrochemical techniques, while each technique also branches into different methods in its category. One of the spectrophotometric techniques is Nessler's reagent method which is already commonly used in the determination of ammonia nitrogen. The reagent is a solution with high alkaline level and contains potassium iodide and mercury (II) iodide. The detection of ammonia nitrogen is through the reaction of Nessler's reagent with ammonia to produce a yellow colour solution. This method has an obvious advantage as its operation is easy due to only one reagent is required throughout the experiment process and heating is also not required. However, Nessler's reagent is toxic to the environment and human as it may potentially cause human health issues. The other downside of this method is that the accuracy of the prediction could be affected by magnesium, calcium and other ions (Lin, et al., 2019).

Another method under spectrophotometric technique is the Indophenol blue method (IPB). The IPB method requires two steps to determine ammonium content through Berthelot reaction. First, in alkaline condition, hypochlorite is reacted with ammonium to produce monochloroamine. Then, IPB was produced through a mixture and reaction of the previous product with phenolic compound

where a catalyst is required to enhance the rate of reaction. In those commonly used IPB methods, phenol is often utilized as the phenolic compound. However, the toxicity and odorous property of phenol make it unsuitable to use. Some researchers had tried to apply the usage of salicylate instead of phenol in IPB method, but the sensitivity of this chemical does not meet the required standard. Hashihama (2015) had improved the IPB method to determine ammonium content in seawater by applying o-phenylphenol (OPP), which is a phenolic compound that possesses lower level of toxicity compared to phenol. This method is well qualified to evaluate the concentration of trace ammonium from a salinity range of 43% to 100% in the seawater (Hashihama, et al., 2015).

Furthermore, o-Phthaldialdehyde (OPA) based fluorometric method is another method under fluorometric technique. An alkaline solution that contains 2-mercaptoethanol act as a solution for the reaction of amino acids and OPA to form a highly fluorescent compound. Sometimes later, researchers enhanced this method by substituting 2-mercaptoethanol with sulphite to yield another OPA based method which is the OPA-sulfite-ammonia that enhanced with better sensitivity and react with ammonia rather than amino acids to determine ammonia concentration. Robert (2017) had established a seawater analyser for on-field application purpose, by using this method nitrogenous compound such as nitrate, nitrite and ammonium concentration could be mapped. The determination of ammonium using OPA based method could obtain a limit of detection (LOD) value from 11 nM to 4000 nM. This low-value range of LOD is sufficient to keep an eye on the ammonium enrichment around the coastal areas (Masserini, et al., 2017). A fluorometric transducer was developed by Cong Wang (2018) to evaluate the amount of ammonium in natural waters coupled with the OPA based method. The advantage of using this method is high accuracy on ammonium concentration prediction could be obtained from two sets of algorithms by adopting atmospheric pressure, temperature, pH and salinity as the input variables (Lin, et al., 2019).

2.3 Artificial Intelligence Model

2.3.1 Back Propagation Neural Network (BPNN)

Back Propagation Neural Network (BPNN) is a branch of Artificial Intelligence (AI) Model from Artificial Neural Network (ANN). ANN was designed as an algorithm that imitates the operation of neural network in the human brain. Generally, like the ANN model, BPNN model consists of three layers which are the input layer, hidden layer and output layer which is shown in Figure 2.1.

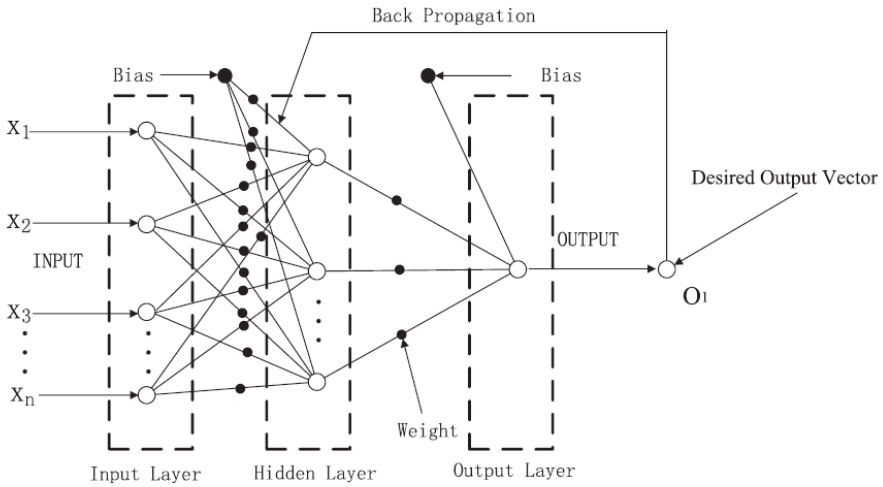

Figure 2.1: Structure of BPNN (Zhang, Ma and Zhang, 2018).

The output result is acquired through the summation of the weighted inputs algebraically as shown in the equation below:

$$O = f\left(\sum_i W_{ij} X_i - \theta_j\right) \tag{2.1}$$

where O_k is the result obtained at the output layer; $f(\cdot)$ indicates the transfer function; W_{ij} denotes the weight between the *i*th neuron of the preceding layer and the *j*th neuron of the present layer; X_i represents the ith input parameter to the neurons at input layer; θ_j is depicted as bias at hidden and output layer neurons (Chen, et al., 2010).

The BPNN comprises of two processes, the initial step is the forward propagation of the information from input layer through data processing in the hidden layer and finally transmit to output layer while the second step is the error back propagation. The input signal from the input layer can be continually

inter-exchangeable between neurons in BPNN. Upon completion of the first learning by the forward process, the output layer will evaluate the result generated. If the actual output result is incompatible with the required output, the back propagation process will transmit the error back to the input and hidden layer through adjustment of the weight in each layer by applying the steepest descent algorithm.

The forward propagation and error back propagation process will be repeated continuously until the error is small enough and falls inside the expected output range or the processes stop when the learning stage of BPNN has achieved the specified learning iterations. At this stage, the BPNN model is said to be fully constructed and the training process is ready to be implemented. As BPNN algorithm was designed as supervised training, the value of input parameter X_i and desired output O must be given by the researcher. The backpropagation process may aid in the continual modification of weight and bias in the layers of BPNN during the learning process, the reduction in error could be achieved after several iterations or a mathematical function can be estimated by the network (Zhou, et al., 2020). In short, the objective of the training process is to minimize the error which can be illustrated by the following equations:

$$E = (1/P) \sum_{P=1}^{P} E_P \tag{2.2}$$

where E represents the global error, P indicates the number of total training times and E_P depicts the training error at *p*th training time. The equation of E_P is illustrated as below:

$$E_P = (1/2) \sum_{k=1}^{N} (O_k - D_k)^2 \tag{2.3}$$

where N represents the total output neurons number, O_k and D_k represent the output values and target output values of the *k*th output neurons respectively (Zhang, Ma and Zhang, 2018).

However, the capability of BPNN to generate accurate data may be governed by three factors. One of the factors is the learning data size while the second factor is the architecture of the BPNN model. The third factor is the level

of complication of the problem that the BPNN model deals with (Chen, et al., 2010).

In this section, literature regarding the application of BPNN in civil engineering and hydrological field will be discussed. Chen, et al. (2010) had proposed a Decision Group Back-Propagation Network (DGBPN) to estimate flood forecasting at Wu-Shi watershed, Taiwan. DGBPN is a combination of multiple BPNN models. DGBPN was established by the author due to the reason that a single BPNN model was not enough to precisely forecast flood event as there exist a lot of uncertainties during the rainfall-runoff process. Therefore, application of DGBPN could prompt the model to choose the most suitable BPNN model at the exact time when the output was estimated by DGBPN. The advantage of using DGBPN model is that the versatility of BPNN model was enhanced in producing effective flood forecasting system in temporal and spatial variation. Besides, DGBPN could produce more accurate results when used to model forecasting events that span over a long period. Moreover, estimation errors could be minimized and eventually reduce the decision making mistakes by decision-makers who obtain forecasting results originated from a BPNN deterministic model. Therefore, flood forecasting could be established more optimally by the application of DGBPN.

In the study of Ghose, Panda and Swain (2010), BPNN and Radial Basis Function Network (RBFN) models were developed to estimate the fluctuations according to water table depth in Sambalpur. The input parameters used in both of the models were precipitation, humidity and temperature while the output was water table depth. The performance of BPNN and RFBN were compared. It was shown that BPNN acquired higher effectiveness in the prediction of water table depth fluctuations for a longer time being in Sambalpur which rainfall intensity is limited. Moreover, although a massive amount of groundwater level data in a short record period was fed to BPNN model as input, adequate prediction results could still be obtained. Although RBFN model has a higher convergence rate, the matter of fact of its high estimation errors could not be neglected. Under circumstances of scanty weather parameters, RBFN is suitable for usage as it is less costly.

Zhang, Ma and Zhang (2018) developed a BPNN model to estimate the spatial development of sea ice in Liaodong Bay. The input parameters for the

BPNN model are wind direction, air temperature, wind duration and wind speed within 9 years that were collected by MODIS were used to train BPNN model. The result showed that the BPNN model achieved high accuracy in spatial development of sea ice of 92% and the estimation error of the sea ice forecast regarding the spatial and area distribution is below 10%. Therefore, the output of BPNN model was in the expected range determined by MODIS. The comparison between Least Square Based method (LSM), logit model and BPNN model regarding the estimation of spatial and area development of sea ice declared that BPNN model performed better than LSM and logit model with higher intensity of accuracy in handling non-linear problem.

In the civil engineering field, Liu, et al. (2020) had applied BPNN and RBFN in the estimation of concrete carbonation depth and the degree of steel corrosion in reinforced concrete structures. From the conclusion made by the author, it was proven that the RBFN model outperforms BPNN model in terms of the accuracy of estimation. Regarding the accuracy in estimating degree of steel corrosion in the reinforced concrete, RBFN was proved practicable and capable by verified engineering test data.

In another research, Ni and Li (2016) established BPNN and generalized regression neural network (GRNN) to reassemble missing data of wind pressure on a 600 m high structure caused by strong typhoon. The reconstruction performance of both BPNN and GRNN were then evaluated. The missing data reconstruction was carried out at several faulty sensor locations by applying BPNN and GRNN while the Bayesian Regularization (BR) technique and the early stopping technique were used to strengthen the ability of BPNN. From the conclusion made by the author, the reconstructed wind pressure data was in the adequate range of actual monitoring data, hence the neural network models were all capable in practical use in reconstruction of wind pressure missing data. Eventually, BPNN coupled with BR technique outperforms other models but required longest training time in the algorithm computation. The author had suggested employing existing monitoring data of wind pressure around intended faulty sensor as the input of the models in order to enhance the intensity of reconstruction accuracy.

2.3.2 Adaptive Neuro-Fuzzy Inference System (ANFIS)

Adaptive Neuro-Fuzzy Inference System (ANFIS) is a popular artificial intelligence technique used frequently by researchers or engineers in hydrological applications. ANFIS is mainly based on the fuzzy logic approach. The reason ANFIS is so popular among researchers is due to it is a technique that combines artificial neural network (ANN) and fuzzy logic (Firat and Gungor, 2007).

The fuzzy inference system happens to be a rule-based method composed of three theoretical elements. The three elements are a database which stipulates membership function, a rule-base which comprise of fuzzy if-then rules and an inference system which fuzzy rules are merged and system outputs are generated. The initial step in fuzzy logic simulation is to establish membership functions of input and output variables, while the second step is to formulate fuzzy rules and the last step is to deduce the system results, output traits and membership function of the output. During the fuzzification process, the input is fuzzified and the input value with an interval of 0 to 1 is mapped with a curved relationship graph. The fuzzified input is then transferred to a decision-making system for reasoning and thinking process, then the output is generated after defuzzification stage (Kermani, et al., 2009). The membership function of input and output variables can be determined by hybrid learning algorithm and backward propagation algorithm. The backward propagation or least squares algorithm are used to train the parameters corresponding to input and output membership functions. These two methods are formulated in ANFIS and ANFIS model apply these two methods in learning process where conditional statements or rules are then established (Firat and Gungor, 2007). ANFIS is suitable to be used as a model to process an enormous amount of data as it applies fuzzy rules and eventually the errors during the training stage would be reduced. Figure 2.2 demonstrates the structure of fuzzy inference system.

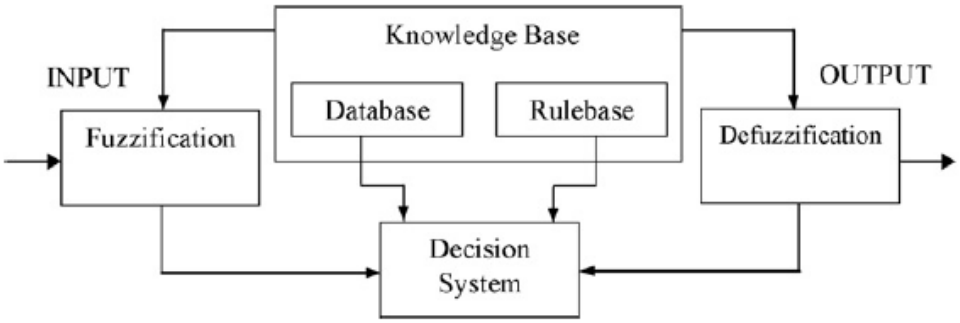

Figure 2.2: General structure of fuzzy inference system. (Firat and Gungor, 2007).

ANFIS consist of a multilayer feed-forward network which utilizes the learning algorithm of ANN technique and fuzzy reasoning where input space is the designated to the output space. The interesting part of ANFIS is that human intelligence can be transformed into fuzzy systems. The input and output relationship in an ANFIS model is determined by ANN in ANFIS which function as a learning mechanism and eventually the fuzzy rules could be established by ANFIS when the input structure has been architectured. While the fuzzy logic of ANFIS is responsible for generating output through its reasoning and thinking ability. There are two types of inference system which are widely used by researchers, they are the zero-order Sugeno inference system and first-order Sugeno inference system. For a general ANFIS, it contains two inputs and one output (Firat and Gungor, 2007). The general structure of ANFIS is demonstrated in Figure 2.3.

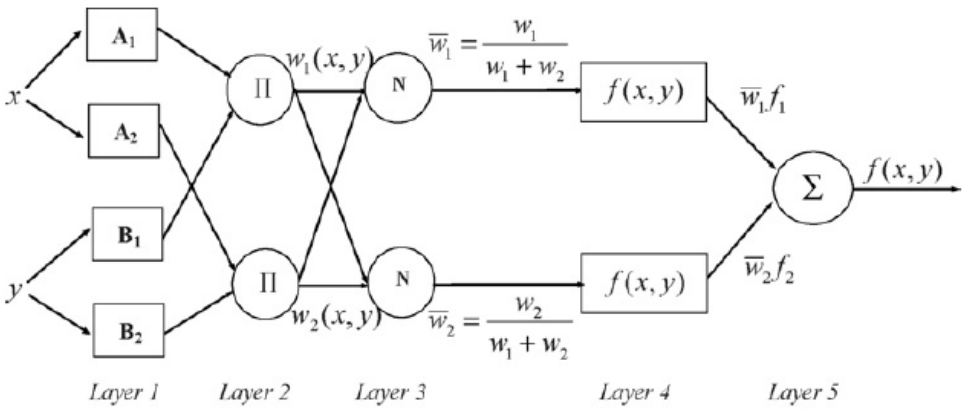

Figure 2.3: General structure of ANFIS (Firat and Gungor, 2007).

Typically, first-order Sugeno inference system consists of two rules as illustrated:

Rule 1: $f_1 = p_1^* x + q_1^* y + r_1$, if x is A_1 and y is B_1

Rule 2: $f_2 = p_2^* x + q_2^* y + r_2$, if x is A_2 and y is B_2

Where parameter x and y are inputs fed to node i; A_i and B_i represent the linguistic labels (e.g. low, medium, high) distinguished by appropriate membership functions; r_i, q_i and p_i indicate the consequence parameters where i is equal to either 1 or 2. By referring to Figure 2.3, ANFIS is made up of 5 layers. The description of each layer of ANFIS model is explained as follows.

Layer 1 of ANFIS is also known as input nodes. Membership grades of the crisp inputs are produced by each note in the first layer by adopting the membership functions. Each of the inputs belongs to each appropriate fuzzy set. The output O_i^1 of each node is illustrated as:

$$O_i^1 = \mu A_i(x) \text{ , i = 1,2;} \qquad O_i^1 = \mu B_{i-2}(y) \text{ , i = 3,4} \tag{2.4}$$

where μA_i and μB_i represent the suitable membership functions for A_i and B_i fuzzy sets respectively.

Layer 2 of ANFIS is also known as rule nodes where one output that indicates the precursor of a fuzzy rule (firing strength) result is obtained with application of AND/OR operator. The firing strength illustrates the satisfactory degrees of the previous part of the rule in layer 1. The incoming signals, π from layer 1 are multiplied with every node in layer 2 and the output is demonstrated as:

$$O_i^2 = w_i = \mu_{Ai}(x)\mu_{Bi}(y) \text{ , i = 1,2} \tag{2.5}$$

Layer 3 of ANFIS is also known as average nodes which *i*th node in this layer is indicated by N in Figure 2.3. The normalized firing strengths, $\overline{w}_i$ is computed as:

$$O_i^3 = \overline{w}_i = \frac{w_i}{w_1 + w_2}, \text{ I = 1,2} \tag{2.6}$$

The nodes in layer 4 are known as consequent nodes whereas the contribution of the *i*th rule for the model output is computed by node *I* of this layer with the formulation:

$$O_i^4 = \overline{w}_i f_i = \overline{w}_i(p_i x + q_i y + r_i), \text{ i = 1,2} \quad (2.7)$$

where $\overline{w}_i$ represents output of layer 3 while p_i, q_i and r_i are the parameter set.

The node in Layer 5 is termed as output node. Only one single node presents in this layer and it computes the final output of the ANFIS model as: (Kermani, et al., 2009)

$$O_i^5 = \sum_i \overline{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad (2.8)$$

Zhou, et al. (2017) proposed the establishment of ANFIS and ANN models to estimate shear resistance of the reinforced concrete block masonry (RCBM) walls. The input parameters used for training, testing and validation in both of the models are yield strength of transverse and longitudinal reinforcements, axial load, thickness of wall, compressive strength of grouted masonry concrete block, reinforcement ratios in transverse and longitudinal directions, shear span to depth ratio and effective length of wall. The performance of ANFIS and ANN was evaluated by using three different statistical analysis such as mean absolute percentage error (MAPE), root mean squared error (RMSE) and coefficient of determination (R^2) in comparison with the actual data acquired from already established literature. Both ANFIS and ANN models statistical analysis results showed low value of MAPE and RMSE while the R^2 value almost reached unity (= 1), hence these indicated that both models provide satisfactory accuracy and authenticity in the prediction of shear strength of RCBM walls. Besides, performance comparison in respect of R^2 and MAPE between ANFIS and ANN showed that ANFIS model outperformed ANN model in this specific research. Moreover, the author compared ANFIS and ANN with six existing models that mainly made up of empirical expressions from international masonry code commonly used in the evaluation of shear strength in RCBM walls. The results demonstrated that ANFIS and ANN

models outperformed these empirical models: SANZ 2004, GB50003, MSJC, CSA S304.1 standards and models developed by Matsumura and Shing.

Prediction of diamond bit drilling machine penetration rate was studied by Basarir, Tutluoglu and Karpuz (2014) by developing multiple regression methods and ANFIS. The input parameters for both of the models were the classification of rock quality, properties of rock which is the uniaxial compressive strength of different types of rocks and operational parameters of equipment used such as bit rotation and bit load. While the penetration rate was the output of both models. Statistical analysis was performed on both ANFIS and multiple regression models in respect of performance index (PI), variance accounted for (VAF) and RMSE performance indicators. Higher value of VAF indicates that the correctness of the model is higher. The results from the performance assessment of the models showed that ANFIS model possesses higher performance in prediction of penetration rate than multiple regression model as ANFIS achieved the highest value of PI and VAF and lowest RMSE value. Moreover, the excess errors produced in ANFIS was lesser compared to multiple regression model, therefore ANFIS had higher ability in prediction of penetration rate of diamond bit drilling machine.

In another study similar to previous literature carried out by Kucuk, et al. (2011), ANFIS model and traditional multiple linear regression model were developed and compared in respect of the performance of excavator of type impact hammer. The input parameters to the models consisted of geological strength index, power of the impact hammer and strength index of block punch. While the net excavation by impact hammer represents the output of the models. Statistical analysis such as RMSE, VAF and R^2 was performed to evaluate the performance of prediction performance of the traditional multiple linear regression model and ANFIS model. It was confirmed that ANFIS model outperformed traditional multiple linear regression model in terms of prediction capability. At last, the author concluded that ANFIS technique could be applied practically on other excavation machinery in terms of performance prediction.

In a study by Zhou, Wang and Zhu (2016), the compressive strength of un-grouted concrete hollow block masonry prisms was predicted by the development of ANFIS and ANN models. The input parameters utilized in these models were compressive strengths of mortar, height to thickness ratio of the

prisms and the concrete block's compressive strengths. While the output was the compressive strength of the hollow concrete masonry prisms. 72 samples were fed into the training process while 18 samples were utilized for the testing process of ANN and ANFIS models. The performance evaluation of the models was carried out by statistical analysis which utilized mean squared error (MSE), MAPE and R^2. In this study by the author, bell-shaped membership functions were used in the ANFIS model which resulted in high accuracy in value prediction. Furthermore, the statistical analysis demonstrated that ANFIS model prediction performance was somewhat higher than ANN model. Despite only compared the performance between ANFIS and ANN, the author also compared the two models with empirical formulas denoted from three other international masonry design codes which are CSA S304.1, TMS 402 and Eurocode 6. However, the prediction accuracy of the empirical formulas from the design codes did not meet the satisfactory prediction range, while the worse part was that the prediction on compressive strength of the concrete block masonry prisms was underestimated to as high as 20% on an average scale by empirical formulas. ANFIS and ANN models possess advantage of short time requirement in prediction of compressive strength and provided that the results with low error.

Other than that, in the field of hydrological application, several literatures were reviewed. Vijayalaksmia and Babu (2015) had developed ANFIS model to predict the water supply demand required by the consumers based on 4 packages of Hogenakkal Water Supply and Fluorosis Mitigation Project Krishnagiri of Dharmapuri Districts which located at Tamil Nadu, India. A total of six different ANFIS models were established with different input combinations of periods and a distinct number of membership functions for each ANFIS model. The statistical analysis compared the prediction results from the developed ANFIS model with the values of historical observed water demand. The statistical analysis used were correlation coefficient (R), MAPE and RMSE. It was observed that the statistical performance in terms of MAPE, RMSE and R were enhanced with the increment in number of inputs utilized in the development of ANFIS model. Besides, it was indicated that increment in number of membership function will cause the error to increase, the cause was the increment of non-linearity of the network in the ANFIS model. From the

results obtained, the fifth ANFIS model from all the packages with 4 membership functions and with input combinations of Q(t) and Q(t-1) had the highest prediction performance as it generated the least error. Overall, ANFIS is a good artificial intelligence modelling technique for water supply demand forecast.

Chang and Chang (2006) had developed two ANFIS models to forecast the water level 1 to 3 hours ahead of Shihmen reservoir located in Taiwan. The input parameters utilized by ANFIS modelling were 8640 hourly data sets obtained from 132 historical heavy rainfall events and typhoon from the year 1971 to 2001. These hourly data sets were split into three sets, each for training, verification and testing process. The input data for training, verification and testing were 4248, 2064 and 2328 data set respectively. Of the two ANFIS models developed, one of the models had included human decision with an additional input parameter which was the reservoir outflow while the other without. The performance of both ANFIS models was evaluated using statistical analysis which comprised of RMSE, correlation coefficient (CC), G_{bench} which is used to measure the goodness of fit and mean absolute error (MAE). From the observation on statistical analysis results, it was shown that the ANFIS model with additional reservoir outflow as input parameter had better performance than another ANFIS model without the human decision input parameter. Therefore, ANFIS model is a reliable technique to be used in hydrological field as higher accuracy prediction results could be obtained with human decisions are included in the input-output parameters. Despite the human decision, both ANFIS models generated satisfactory and accurate reservoir water level estimation for the next 1 to 3 hours as indicated by the close to unity correlation coefficient value of more than 0.99.

In another study by Chang and Lai (2014), ANFIS models were developed by the author in two distinct scenarios to estimate the changes of shoreline monthly for the following year at seven stations located at Yilan County, Taiwan. In the first scenario which was the lumped scenario, four different ANFIS models with varied input parameters combinations were established and used to estimate the changes at the shoreline. The four different combinations of input parameters included position of the shoreline, current month, information regarding the wave height and location of the station. The

statistical analysis used in evaluation of model performance were correlation coefficient (CC), coefficient of efficiency (CE) and RMSE. The statistical analysis showed that three out of four of the ANFIS models with information of wave height as the input parameter did not exhibit consistency in training, validation and testing stages. However, the one model that did not involve information of wave height as input parameter could provide stable results in all of the three stages. The inconsistency could be explained by the excessive variability of the measurement of wave height. Therefore, with the information of wave height as input parameter, the models provided unpractical and inaccurate results as the RMSE was too large that it exceeds the satisfactory range. Another scenario which was the site-specific scenario, ANFIS models were developed again with different type of input parameters combinations. The input parameters for the ANFIS model in this scenario were the present monthly data. The prediction accuracy and reliability of ARX-based models and Fourier series-based models were outperformed by the ANFIS models in the site-specific scenario due to its higher value of CC and CE and an even smaller value of RMSE. Therefore, the forecasted results by the ANFIS models in the site-specific scenario could be utilized by decision-makers such as government authorities to carry out management works and provide warning to the public regarding shoreline erosion.

2.3.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) had slowly become a popular supervised machine learning algorithm as it could be utilized to deal with classification problems and function estimation on regression-type implementation. The benefits of using SVM is that it possesses better empirical performance. Moreover, the over-generalization and overfitting issues could be rectified by SVM and SVM has the ability to process an enormous amount of data with a simple training process. The principle of SVM is based on structural risk minimization (SRM) and it has superior generalization capability. Whereas an upper bound of the presumed risk is minimized by the SRM. In regression modelling of SVM, several input variables are utilized to generate one output variable.

A model is established by SVM with the utilization of training dataset. During the training stage of SVM, input data is converted into a hyperplane and non-linear structures are converted into linear structures. Figure 2.4 below demonstrated the basis of SVM where the SVM creates uncountable decision functions which are known as hyperplanes. The input data transformed by kernel function in respect of mathematical function and eventually the training dataset are separated. The dataset contains positive and negative data which are denoted by '+' and '*' respectively in Figure 2.4. The positive dataset is placed at one side of the hyperplane while the negative dataset is placed on the other side. The distance from data to the hyperplane is termed as a margin. In this model, the margin between hyperplane and closest positive samples will be maximized and the same goes to the closest negative samples (Olyaie, Abyaneh and Mehr, 2017).

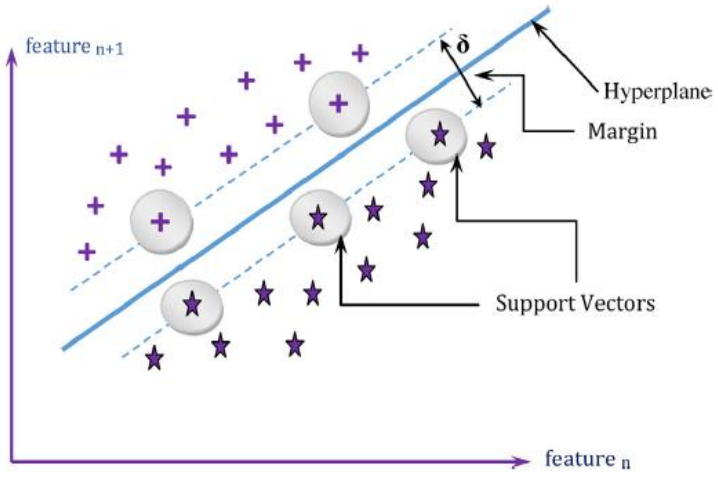

Figure 2.4: Basis of SVM (Raghavendra and Deka, 2014).

The following will be some explanation of equations used in support vector regression. An equation showing N numbers of training data is formed.

$$\{(X_i d_i)\}_i^N \tag{2.9}$$

where X_i indicates the input vector, d_i represents desired value and N indicates the total number of training data. The regression of SVM generated an equation that predicts an outcome, Y of the model.

$$Y = f(x) = W_i \emptyset_i(X) + b \quad (2.10)$$

where W_i is a weight vector, $\emptyset_i$ represents a nonlinear transfer function that plot the input vectors into a high dimensional feature space and b indicates a bias. After some modification and enhancement on the equations of regression modelling, the regression function is expressed as below:

$$f(X, \alpha, \alpha^*) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(X, X_i) + b \quad (2.11)$$

where $K(X, X_i)$ denotes the Kernel function and is expressed as follows: (Olyaie, Abyaneh and Mehr, 2017)

$$K(X, X_i) = \varphi(X_i)^* \varphi(X_j) \quad (2.12)$$

Numerous scholars have made use of SVM due to its better performance compared to traditional statistical models. In the following, application of SVM in hydrological field will be reviewed. A study conducted by Sahana (2020) implemented modified frequency ratio (MFR), conventional frequency ratio (CFR) and SVM to estimate storm surge flood susceptibility at Sundarban Biosphere Reserve in India. The SVM model used by the author was of linear kernel function and modelled with R Studio software. The performance of the models was tested using four different statistical analysis which is the ROC curves that used to determine the overall accuracy of the models during the training and testing stage, the spatially agreed area approach where spatial agreement of flood susceptibility of each model was compared and the seed cell area index (SCAI) which was implemented to compare the area of susceptibility classes and the ratio of area of testing and training data. The evaluation of SCAI generated lowest value for SVM model while CFR had second-lowest value and MFR had the highest value. On the other hand, among all the three models, SVM showed the highest value which was 0.8221 in evaluation of success rate. Besides, the spatially susceptibility agreed area between SVM and MFR was higher than SVM and CFR. Hence, the author concluded that the SVM model

was most practical and had the highest performance in determination of storm surge flood susceptibility.

Khan, et al. (2020) developed ANN, SVM and k-Nearest Neighbour (KNN) models to predict droughts over Pakistan. The author utilized reanalysis data (NCEP/NCAR v1) as the inputs for the three models. Four different statistical indices were used to evaluate the performance of the three models. The statistical indices were the normalized root mean squared error (NRMSE), modified index of agreement (md), R^2 and percentage of bias (Pbias). The results showed that the performance of KNN model was unsatisfactory compared to ANN and SVM models. While SVM model had a higher performance in capturing spatial and temporal characteristics of droughts in Pakistan. However, the machine learning models in this study was lacked of the ability to determine extreme droughts. The author suspected that this happened due to the prevailing generalization skills of SVM model which undertake a global optimum solution as opposed to ANN which can be confined at a local optimum.

In another study by Berbić, et al. (2017), ANN and SVM were developed using Weka software to forecast the significant wave heights for the next 0.5 to 5.5 hours. Three different sets of time step data from wave measuring station were utilized as the input for the models. Sensitivity analysis was carried out to determine the optimal inputs for the models and the optimal parameters for each model function. For SVM, the sensitivity analysis evaluated that the polynomial kernel offered the best results. In the formulation of kernel function, parameter C of 1 and parameter p of 1 were analysed to generate the most optimal results. As the first and second approach of models used only wave heights as inputs, six and three input was adequate to generate accurate results. However, the SVM required four to five times longer to build its model compared to ANN model. In the third approach, it was concluded that additional input of wind velocities could increase the consistency and accuracy of the results. After evaluation of model performance by statistical analysis including relative absolute error (RAE), RMSE, root relative squared error (RRSE), R and MAE, the accuracy of SVM model was found to be superior to ANN model.

Despite having popularity in hydrological application, SVM is also widely used in civil engineering application. Abbas and Suhad (2017)

developed SVM and multivariate non-linear regression models to estimate the compressive strength of lightweight foamed concrete. The input parameters for the SVM model were the values of the mix proportion elements to form the lightweight concrete and compressive strength at 7-day. While the output of the SVM model was 28-day compressive strength of the lightweight foamed concrete. A total of 150 sample size was divided into train samples (111 data) and test samples (39 data). All four types of function in SVM were established by using the 150 sample data. The four types of functions were polynomial, linear, radial basis function and sigmoid. Among the four functions, the radial basis function achieved the highest performance in terms of having the highest correlation coefficient in testing, training and overall data set. Other than that, radial basis function of SVM was tested to achieve the least mean square error and the least standard deviation value among the four functions. Besides, the scattered plot of radial basis function showed that the correlation coefficient was close to unity (0.99) which indicated that radial basis function of SVM had very high precision and practicability in determining the 28-day compressive strength of the lightweight foamed concrete.

Zheng, et al. (2019) proposed ANN and SVM models to estimate the liquefaction-induced uplift displacement of a tunnel. The input variables for both of the models were first going through a relative importance analysis to determine their sensitivity related to the output variable. These input variables included duration, amplitude and frequency of shaking motion; friction angle and relative density of soil; diameter and embedded depth of the tunnel. MAE, RMSE and R^2 were used to assess the performance of ANN and SVM models. The result demonstrated that the ANN model outperformed SVM on the performance of testing stage. This might be due to the difference in number of training epochs and number of hidden nodes of ANN. However, during the centrifuge test, SVM had higher value of R^2 compared to ANN which indicated that SVM showed higher performance in estimation of the centrifuge results. This happened because SVM model was capable to link the relationship of inputs and outputs variables and provide better generalization, unlike ANN model which only fits the data. In short, SVM and ANN models were reliable techniques to predict the liquefaction-induced uplift of tunnels.

2.4 Role of AI Models in Water Quality Parameters Prediction

The prediction of water quality had been a major concern for the past few decades. Researchers for the past few decades used numerical models to determine the quality of water but often the predictions were not accurate enough as the input parameters were not manipulated properly. The manipulation of the numerical models often depends on the experiences of the model user. Therefore, Artificial Intelligence (AI) models were invented and gained popularity recently due to its characteristic that can imitate the thinking behaviour of a human. Like human, the AI models will learn the requisite knowledge in order to produce the optimum results. Several different types of input data will be processed by the AI models and on each time of the learning processes, one or two of the water quality parameters would be altered for model manipulation in order to raise the effectiveness of the model. Therefore, even non-experienced model user is capable of predicting the water quality parameters with the aid of AI model. With the aid of AI model in predicting water quality parameters, decision-makers and policymakers can enhance the water quality planning and water quality control process. From the past 10 to 20 years, a lot of researchers had studied the performance of AI models in forecasting the water quality parameters in different kind of water bodies. The development of different AI models in the prediction of each water quality parameters will be discussed in the following literature review.

2.4.1 Dissolved Oxygen (DO)

Olyaie, Abyaneh and Mehr (2017) compared the performance of four AI models in the prediction of DO concentration in Delaware River situated at Trenton, USA. The AI models developed by the author were two ANN based AI models which were multi linear perceptron (MLP) and radial based function (RBF) while the other two AI models were linear genetic programming (LGP) and SVM. Various input combinations were proposed for each AI models which comprised of river discharge, electrical conductivity, pH and temperature of the river water. Nash-Sutcliffe efficiency coefficient (NS), correlation coefficient statistics (R), root mean squared error (RMSE) and mean absolute relative error (MARE) were used as evaluation criteria to access the performance of the four models and determined the best model for forecasting DO concentration. From

the results of statistical analysis, it was concluded that SVM could generate the most accurate prediction of DO concentration among the four AI models. Following by LGP models which showed better performance than both ANN based models. However, all the AI models could maintain their prediction accuracy during lower value of DO content but not for high values of DO. But an interesting part was that SVM showed notable improvement in predicting the peak DO concentration compared to the two ANN-based models and LGP. Therefore, this research showed that SVM could be practically and satisfactorily employed in forecasting DO concentration as SVM achieved a value of 0.99 in the correlation coefficient analysis (where unity = 1 indicates the perfect accuracy).

In the study carried out by Cao, et al. (2019), a hybrid model which consisted of multi-scale feature extraction and multiple-factor analysis were proposed to estimate the DO concentration in Liyang huangjiadang special aquaculture farms located in Changzhou city, Jiangsu province, China. The proposed hybrid model composed of grey relational degree method, ensemble empirical mode decomposition (EEMD), sample entropy (SE) algorithm and regularized extreme learning machine (RELM). The grey relational degree method was utilized to analyse the relationship between various water parameters with DO and pH and water temperature were determined to have the highest correlation with DO. Then, decomposition of pH, water temperature and DO data into various sub-sequences was carried out by EEMD. After that, the sub-sequences were reconstructed by the SE algorithm to acquire the random component, trend component and detail component. RELM, an AI model was used to predict the DO content from the three components independently. The three components were named as RELM1, RELM2 and RELM3 accordingly. The input parameters for the three components were pH, water temperature and DO. The final result was the superimposed results of the three components. The result was then analysed by MAE, RMSE, MAPE and R^2 including the results from other AI models that used as a comparison with the proposed AI model. In conclusion, the proposed model demonstrated higher prediction accuracy and had higher performance than the other AI models compared such as EEMD-ELM, WD-RELM, Single RELM, Single factor-RELM, EEMD-LSSVM, EEMD-RBF, EMD-RELM and Single factor-EEMD-RELM.

Abba, Hadia and Abudullahi (2017) tested the effectiveness and prediction accuracy of DO concentration at downstream of river Yamuna, Agra city by ANN, multi linear regression (MLR) and ANFIS models. The monthly water parameters data collected at three different locations (upstream, middle stream and downstream) such as pH, water temperature, biological oxygen demand (BOD) and DO were used as inputs for the three models. Different combinations of input parameters were developed for the three models. The first input combination was the data set of a total of 11 parameters excluded DO from upstream, middle stream and downstream while the second input combination was the data set of a total of 7 parameters excluded DO from the middle stream and downstream. The last combination only included 3 input parameters excluded DO from downstream. The data were normalized and divided into calibration and verification phase for DO simulation. The performance of each model was accessed by using RMSE and R^2. The result indicated that a great amount of data set would cause overfitting due to the high complexity of data. During the training phase, ANFIS outperformed ANN and MLR but averagely ANN in determining the DO concentration using the second combination of input parameters showed the best performance followed by ANFIS and MLR. While ANFIS achieved the highest accuracy in the prediction of DO when more data of input parameters were available.

Four AI models namely Multilayer Perceptron Neural Network (MLPNN), General Regression Neural Network (GRNN), Multivariate Linear Regression (MLR) and Radial Bases Function Neural Network (RBFNN) models were developed by Csábrágia, et al. (2017) to predict DO concentration. The input parameters data from 1998-2003 were collected from three different sampling locations namely Mohács, Gyorzámoly and Fajsz located at River Danube, Hungary. The water quality input parameters consisted of runoff, conductivity, temperature and pH. There were four combinations of input data where data of each location represented one combination. While the fourth combination of input data assessed the data from the three sampling locations concurrently. The output was the predicted value of DO concentration for the year 2003 at these three locations. The performance of each model was then evaluated using four different statistical criteria which were MAE, Willmott's index of agreement (IA), R^2 and RMSE. The statistical analysis showed that

GRNN and RBFNN had the best performance and outperformed MLPNN. For all the four input parameters combinations, MLR had the worst performance. The sensitivity analysis on the fourth combinations of input parameters showed that pH value had the highest correlation in determining DO concentration. Generally, the neural network models proposed which were GRNN, MLPNN and RBFNN could efficiently predict DO concentration in rivers and provides information for water quality management.

Heddam and Kisi (2018) carried out a study to compare the performance of three different AI models namely Multivariate Adaptive Regression Splines (MARS), Least Square Support Vector Machine (LSSVM) and M5 Model Tree (M5T) models in prediction of daily DO concentration of three different sampling stations. The data of water quality parameters were obtained by the author from the United States Geological Survey (USGS) where the three stations were located at East Canyon Creek, Summit County, Utah; Fanno Creek at Durham Road, Washington County, Oregon; Delaware River at Trenton, Mercer County, New Jersey. The input water quality parameters were daily pH, river discharge (DI), water temperature (TE) and specific conductance (SC). Six sets of input combinations were developed and applied to the three models. The six different input combinations were: (i) TE, SC, and pH; (ii) SC and TE; (iii) DI and TE; (iv) pH, TE, DI and SC; (v) SC, TE and DI; (vi) pH and TE. The three models were used to model Do concentration of each sampling station separately and then the performance of each model was compared with one another. In the Fanno Creek at Durham Road, MARS model with input combination of TE, pH, SC and DI had the best performance and outperformed LSSVM and M5T1 models. While MARS model which adopted TE and pH as input parameters had better performance than LSSVM and M5T in both test and validation phase at East Canyon Creek. Besides, at the Delaware River, LSSVM model achieved better performance than MARS and M5T which its input parameters containing pH and TE showed the best accuracy among other models in both test and validation phases. However, the models in this study were hard to compare with each other on which model had the best performance in general as the data used for modelling are independently from three different stations. However, a conclusion can be made with TE and pH were the best

input water quality parameters for modelling DO concentration and this eventually showed that pH and TE had a strong correlation to DO.

In the research done by Elkiran, Nourani and Abba (2019), three single AI-based models namely SVM, ANFIS and BPNN models and a traditional linear Auto-Regressive Integrated Moving Average (ARIMA) model were developed to model DO concentration in Yamuna River, India. Additionally, the author also developed three different ensemble techniques namely Weighted Average Ensemble (WAE), Simple Average Ensemble (SAE) and Neural Network Ensemble (NNE) to enhance the AI models. Data of water quality parameters from three different sampling stations of Yamuna River were collected which were BOD, discharge, ammonia (NH3), DO, chemical oxygen demand (COD), pH and water temperature. These water quality parameters were used as input for the models and the three stations were named Hathnikund (SL1), Nizamuddin (SL2) and Udi (SL3). The statistical analysis performed was RMSE and Determination Coefficient (DC). The assessment on the performance of models indicated that ANFIS model had the best performance accuracy in station SL1 and SL2 while SVM model had higher precision in determining DO concentration than ANFIS, ARIMA and BPNN in station SL3. From the result obtained from statistical analysis of ensemble techniques, NNE had the highest performance in increasing an average performance of 14% to single AI models in the verification phase. Therefore, NNE could be practically and efficiently be applied in multi-step ahead modelling of DO concentration due to its capability in dealing with nonlinear processes.

In a study by Rankovic, et al. (2010), a feedforward Neural Network (FNN) model was established to estimate the DO concentration in the Gruza Reservoir located in Serbia. The water quality parameters were collected in three years and were treated as the input variables for the FNN model. The input variables were water temperature, total phosphate, nitrates, iron, electrical conductivity, pH, chloride, nitrites, ammonia and manganese. The input parameters were evaluated by sensitivity analysis to determine their correlation with DO concentration. It was obtained that water temperature and pH were the most effective inputs to determine DO concentration while total phosphate, nitrates and chloride had the least correlation with DO. The predicted DO concentration by FNN was compared with the actual data through evaluation of

MAE, R and Mean Square Error (MSE). From the statistical analysis result, it was found that FNN model having 15 hidden neurons provided the best performance. In a nutshell, the statistical analysis showed that the proposed FNN model in the prediction of DO concentration provided satisfactory results.

In a nutshell, many studies were done on the estimation performance of neural network-based models such as ANN, BPNN, etc. These neural network-based models often provide satisfactory prediction on DO concentration and are reliable AI models to be practically applied in the hydrological field in estimating water quality parameters. Besides, ANFIS models were also frequently utilized by researchers to predict DO concentration in water bodies, generally, ANFIS shows higher accuracy in prediction of DO concentration due to it contains neural network-based technique combined with fuzzy inference system. Another interesting AI model that should be given more attention is SVM as it has a very good generalization ability and does provide high accuracy performance. Also, overfitting problems found in ANN and ANFIS will not be found in SVM as SVM will correct the overfitting issue by itself. Besides, the ability of SVM to process a great amount of input data has shown its advantage in modelling DO concentration.

2.4.2 Total Suspended Solids

As mentioned in chapter 2.2.2, Total Suspended Solids (TSS) consists of inorganic and organic matter and an increase in TSS will eventually increase the value of turbidity in water bodies. Hence, TSS indirectly affect the opacity and light penetration into water bodies and this may lead to serious negative impacts on the aquatic ecosystem. Therefore, studies regarding the modelling of TSS in water bodies using AI models will be reviewed in this section.

Altunkaynak and Wang (2011) investigated the effectiveness of two expert system-related AI models namely fuzzy logic and Geno-Kalman Filtering (GKF) models in estimating the concentration of TSS at Dry Bar and Cat Point in Apalachicola Bay, USA. The field data of wind speed and TSS concentration collected from the two stations at Apalachicola Bay from 1/6/2005 to 30/7/2005 were used as the input parameters for training and testing of the two AI models. Besides, the author also developed traditional hydrodynamic model proposed by other researchers to compare with the

effectiveness of his two proposed AI models. The collected data set was separated into two sections to develop the models. The data collected during June were utilized for the calibration while the data collected during July were utilized for model verification. The predicted TSS concentration values by GKF, fuzzy logic and traditional hydrodynamic models were then compared with the observed data in terms of their prediction accuracy and model performance by using statistical analysis such as coefficient of efficiency (CE), MSE and chi-square(X^2). From the results obtained from statistical analysis, it was found that GKF and fuzzy logic models performed better in terms of prediction accuracy than the traditional hydrodynamic model. On the other hand, GKF model was able to estimate the TSS concentration more accurately than the fuzzy logic model as the fuzzy logic model under-predicted the TSS concentration but still in the opinion of the author, fuzzy logic model could be practically used in the prediction application of TSS concentration in water bodies. However, only local estimation of the fluid domain was able to be effectively performed by GKF and fuzzy logic models. To generate spatial predictions of water quality parameters, the availability of a large amount of data should be provided for the model development of GKF and fuzzy logic models.

In a study by Vicente, et al. (2012), two ANN models namely Multilayer Perceptron Neural Network and Feed Forward Back Propagation Neural Network models were developed to estimate the TSS concentration and oxidability in the water of Monte Novo Reservoir which located in Portugal. The data used as input to train the models were water quality parameters data collected from August 1995 to December 2010 which encompassed 15 years of data. A total of 184 data set of each water quality parameters were recorded. The water quality parameters used as input for modelling were conductivity, water temperature, pH, DO, oxidability, volume of water stored in the reservoir and TSS. The available data were randomly separated into three parts namely, training, testing and validation. 60% of available data were used in training of the models, 25% of the available data were utilized in the testing phase of the models to analyse the performance of the models and the last 15% of the data were applied in the validation phase of the models. 20 repetitions were adopted in each phase. Statistical analysis criteria such as mean absolute deviation, R^2, bias from the modelling phase and MSE were adopted to evaluate the

performance of the two ANN-based models. Other than that, the relationship between the errors and the output values of TSS concentration and oxidability were evaluated to find out the goodness of model fit to the data. From the evaluation result, the Multilayer Perceptron Neural Network and Feed Forward Back Propagation Neural Network models acquired satisfactory performance in prediction of oxidability and TSS concentration as the R^2 values in the training set of both the models had a range of 0.995 to 0.998 while R^2 values for the testing set were ranged from 0.994-0.996 which were near perfect.

Verma, Wei and Kusiak (2013) established five AI models namely K-Nearest Neighbour (KNN), SVM, Multi-layered Perceptron (MLP), Multi-variate Adaptive Regression Spine (MARS) and Random Forest (RF) models to forecast one day-ahead and 7 days ahead of time series estimation for TSS concentration present in wastewater of wastewater treatment plant situated in Des Moines, Iowa. The models were developed and trained by using influent carbonaceous biochemical oxygen demand (CBOD) and influent flow rate to the wastewater treatment plant. The data of the influent of flow rate was recorded every 15 minutes. The five models were then compared against one another in terms of prediction accuracy by evaluation of mean relative error (MRE) and MAE. It turned out that MLP had the highest precision in prediction accuracy and was then selected as the most suitable model to forecast a seven-day-ahead estimation on TSS concentration in the wastewater. Iterative learning was proposed to reduce the prediction error of MLP in the construction of seven-day-ahead prediction of TSS concentration. The constructed seven-day-ahead TSS forecast was evaluated and the result demonstrated that MLP provided a prediction accuracy of 73% on the seven-day-ahead TSS concentration prediction. The author commented that the accuracy of the prediction by the five AI models would increase where adequate data were available.

In a study by Patel, Ruparelia and Barve (2020), the efficiency of the fuzzy inference system (FIS) and mechanistic models in estimating TSS concentration present in the effluent stream from a sedimentation tank of clariflocculator which situated inside a common effluent treatment plant (CETP) located at Ahmedabad, India were compared. The output of the models was forecasted by having influent TSS and influent flow rate as input parameters. The performance of the models was evaluated by performance indexes such as

coefficient of determination (R^2), root mean squared error (RMSE), mean absolute percentage error (MAPE) and percentage mean accuracy. The result showed that the mechanistic and FIS models obtained 12.4% and 14.9% in MAPE respectively. The RMSE were 30 and 48 for mechanistic and FIS models respectively while R^2 were 0.84 and 0.5 respectively. From the statistical evaluation, it was concluded that the mechanistic model had higher prediction accuracy and performance with lower prediction error, also the conclusion was enhanced by 88% of percentage mean accuracy by the mechanistic model. Therefore, the performance of mechanistic model was better than FIS model. Nevertheless, the mechanistic model had the benefit of providing information regarding TSS exist at each layer and in the retentate so the prediction on sludge produced per day can also be predicted.

In conclusion, most of the research paper reviewed are predicting TSS concentration in effluent of wastewater and coastal water. This might be due to wastewater is the main source of pollutant to the river stream and coastal water as the concentration of TSS in wastewater is greater than the surface water. Therefore, great concern was given on the prediction of TSS in wastewater to provide information for decision-makers to control the TSS concentration in the effluent to avoid wastewater with exceeding concentration of TSS is released into nature. From all the AI models previously reviewed to determine TSS concentration, ANN-based AI models could provide satisfactory results especially the Multilayer Perceptron Neural Network and Feed Forward Back Propagation Neural Network models which achieved a near-unity value of coefficient of determination indicates that these two models were almost near to 100% prediction accuracy. However, higher amount of data can increase their accuracy.

2.4.3 Turbidity

Turbidity is one of the most important water quality parameters to determine water quality status of a water body. The level of turbidity is influenced by other water quality parameters. Normally, high level of turbidity indicates high value in other water quality parameters such as nitrate, ammonium, phosphorus and sulphate concentration. Infield sampling, the value of turbidity is obtained by using turbidimeter. Unfortunately, the cost of turbidimeter is high and expensive

in maintenance and easily damageable. Therefore, researchers from the last two decades had passionately participated in the development of AI models to predict turbidity in water bodies. Several research papers have shown some studies on AI models that predict turbidity level with high accuracy.

Alizadeh, et al. (2018) had developed ANN, Extreme Learning Machine (ELM) and SVM to predict three water quality parameters in coastal waters located in Hilo Bay, Pacific Ocean up to 2 hours ahead. The three water quality parameters as the output of the three models were turbidity, salinity and temperature of the coastal water. The input for the three AI models were water quality parameters such as temperature, salinity and turbidity recorded in 1-hour interval from the Wailuku River. Each model was modelled two times and two separate results were obtained; the first modelling included river flow data as input while the second modelling excluded river flow data as input. This was to compare the prediction accuracy of the models with and without river flow data as input parameter. The performance of the predicted water quality parameters output of each model was evaluated based on three performance indexes which were R^2, RMSE and width of the uncertainty band. From the performance evaluation result obtained, it was shown that the models which included river flow data as input parameter had gained accuracy and better performance during the prediction of salinity and turbidity. According to the performance evaluation, the prediction accuracy on turbidity increased significantly with river flow as input parameter while water temperature had the least significant improvement in prediction accuracy. The comparison of the performance of ELM, ANN and SVM models illustrated that all of the models were capable and practicable in predicting water quality parameters. Nevertheless, the statistical analysis showed that the performance and prediction accuracy of the three models were roughly in the same range. However, it was found that the accuracy of these models decreased with an increase in the prediction time ahead.

In a study by Abba, nourani and Elkiran (2019), four AI models namely Hammerstein-wiener (HW), Least Square Support Vector Machine (LSSVM), Non-Linear Autoregressive with Exogenous (NARX) Neural Network and General Regression Neural Network (GRNN) were developed to predict the concentration of several water quality parameters, which were turbidity, suspended solids, pH and hardness. The study area was located at Kano, Nigeria

namely Tamburawa water treatment plant (TWTP). The input parameters for the models' development were collected weekly from the study area, the weekly data included measured treated and raw pH, conductivity, suspended solids, chloride, turbidity, total dissolved solids, hardness and iron. The data set was split to 75% and 25% for the calibration phase and verification phase respectively. Other than that, four black-box models namely HW-E, GRNN-E, LSSVM-E and NARX-E were developed with the non-linear ensemble technique. Which means that each non-linear ensemble model used the predicted output from its original single model as its input layer. The performance of non-linear ensemble models was inter-compared. All the models were evaluated by performance assessment concerning RMSE, MAPE, determination coefficient (DC) and MAE. The comparison results in terms of model performance demonstrated that HW model acquired the best performance in predicting turbidity, hardness and suspended solids while NARX model had the highest accuracy in predicting pH. Among the four single AI models, NARX and HW models had the best overall prediction performance. On the other hand, the non-linear ensemble technique employed on GRNN-E showed it had the highest ability in increasing the accuracy of its original single model with an increment of 30% accuracy for turbidity and hardness, 37% for suspended solids and 34% for pH. In conclusion, the author suggested to couple AI models including extreme learning machine and other hybrid models with non-linear ensemble technique in order to produce higher predictive accuracy.

Teixeira, et al. (2020) investigated the effectiveness of Fuzzy Inference System (FIS) and ANN models in the modelling of turbidity and suspended sediment concentration in four watersheds in the same region located at southern Brazil. The input parameters for FIS and ANN models were turbidity, usual variable discharge and hourly rainfall. The output of the predictive models was then compared with the actual data collected from the study area using performance statistics in respect of percent bias (PBIAS) which average tendency of the predicted data was measured with PBIAS value of 0 as ideal, Nash-Sutcliffe efficiency (NS) and MAE. From the result of statistical analysis, FIS model was found to be the best model to estimate turbidity concentration with a NS of 0.86 in the verification samples. In the simulation of suspended sediment concentration, turbidity and discharge were used as input for

modelling with addition of EWMA caused by the clustering of hourly precipitation. EWMA was known as exponentially weighted moving average which consisted of two parameters which were the original half-life and the time window delay. From statistical performance analysis, FIS model also performed best in modelling suspended sediment concentration. In conclusion, FIS model had a higher precision compared to ANN model and could be practically used in prediction of turbidity and suspended sediment concentration.

In a study by Namu, et al. (2017), ANN models were developed to predict the turbidity of water at inlet and outlet of a settling basin situated in Kinku-keinde irrigation project which was located in Embu Sub-County. Settling time and flow rate were adopted as input parameters to develop the ANN models while turbidity as the output parameter. Different from other research, four best ANN models were selected to determine the best model with the best performance from four hundred developed ANN models. In the training phase, 70% of the data were randomly selected from the available data while 15% of the data were selected for validation phase and the remaining data were utilized in testing phase. The performance of the four best ANN models selected from the four hundred ANN models was evaluated by four statistical measures namely correlation coefficient (R), mean squared error (MSE), root mean squared error (RMSE) and mean absolute error (MAE). The result showed that the ANN model with 1 input neuron, 9 hidden neurons and 1 output neuron achieved the highest performance with R value recorded at 0.9999 and RMSE value recorded at 0.5102. Therefore, it was concluded that the architecture of the best ANN model was developed and determined through trial and error in the training, validation and testing phase.

From the literature review on application of AI models to forecast turbidity, it can be concluded that selection of the correct water quality parameters such as river flow is important in optimizing the effectiveness of the AI models to predict the output. However, it was found that most of the AI models prediction accuracy will decrease as the forecasting period increase. To overcome this some other techniques such as non-linear ensemble technique can be coupled with the single AI model to increase their performance in prediction accuracy. A lot of the research done by researchers on the prediction of turbidity had adopted the usage of ANN model. ANN is an effective technique to predict

turbidity as long as adequate data is supplied for the modelling process and the optimum number of hidden neurons is determined.

2.4.4 Nitrate

Nitrate is often a nutrient or pollutant originated from industrialized factories or agricultural land. The control of nitrate concentration in water bodies such as rivers, lakes, groundwater and ocean is necessary to avoid detrimental effect caused by excessive nitrate concentration such as eutrophication and mass death of aquatic organisms. Therefore, AI models were developed to predict nitrate concentration in water bodies in order for decision-makers to take action immediately to solve and prevent water pollution by nitrate from happening.

Suen, Eheart and ASCE (2003) developed two AI models which were Radial Basis Function Neural Network (RBFNN) and BPNN models to predict nitrate concentration in the Upper Sangamon River located in Illinois. Data from three weather stations located in Upper Sangamon River were collected and used as input for RBFNN and BPNN models. The input parameters were daily highest temperature, cumulative daily rainfall for seven days and daily streamflow. Two sets of RBFNN and BPNN models were developed based on two cases. In case 1, data from the year 1993 to 1996 were utilized in training phase while data from the year 1997 to 2000 were used in validation phase. However, in case 2 the input parameters data of odd year were used in training phase while even year data were utilized in training phase. RBFNN and BPNN were compared in terms of performance to a Multiple Regression Analysis (MRA) and a mechanistic SWAT model. Besides, the author also developed 1 Boolean RBFNN and 1 Boolean BPNN models to compare with the performance BPNN and RBFNN models. The Boolean BPNN and RBFNN models had only 2 output which was 1 and 0, 1 indicated nitrate concentration more than 10 mg/L while 0 indicated nitrate concentration less than 10 mg/L. The statistical analysis used to evaluate the performance of these models were RMSE, false-negative frequency, overall accuracy and false-positive frequency. From the result obtained, Boolean RBFNN, BPNN and Boolean BPNN models trained with case 2 had the best prediction accuracy. However, Boolean ANN models showed higher overall accuracy than the BPNN and RBFNN models while the accuracy of Boolean RBFNN model was higher than Boolean BRNN

model. On the other hand, the comparison between BPNN and RBFNN models showed that BPNN achieved higher overall accuracy and lower RMSE value. This might be due to the information of the data set available was inadequate and was not able to provide sufficient clustering for RBFNN model training. Nevertheless, both BPNN and RBFNN including the Boolean type ANN models trained by case 2 outperformed the MRA and mechanistic SWAT model in terms of overall accuracy, RMSE and false-negative and false-positive frequency. Although RBFNN model had lower overall accuracy than BPNN, its training speed was faster than BPNN model.

In a study by Stamenkovic, Kurilic and Ulnikovic (2020), two AI models based on ANN was proposed to estimate the nitrate concentration in Danube River. Data ranging from 2011 to 2016 were collected from 10 measuring stations in Danube River. These data included 27 water quality parameters which were then utilized as input parameters for the model development. Standard three-layer network and multi-layer perceptron were applied to develop the models. The inputs had the most correlation with nitrate were selected by applying Input Variable Selection (IVS) techniques by the author. The performance of the models was evaluated by MAE, RMSE and R^2. The result showed that both the models performed well in estimating nitrate concentration. Moreover, the adoption of IVS to reduce the input parameters had contributed to the increment of performance to both of the models. Therefore, neural networks were able to provide satisfactory result for water quality parameter prediction and can be practically used.

Ouedraogo, Defourny and Vanclooster (2019) developed a Random Forest Regression (RFR) model and compared the performance of RFR with Multiple Regression (MLR) model in terms of prediction of groundwater nitrate concentration at the African continental scale. 13 different data type were collected and were treated as the input parameters for RFR and MLR. The input parameters were population density, climate class, rainfall class, type of aquifer, unsaturated zone, recharge land cover or land use, application of nitrogen, type of region, depth to groundwater, soil type, topography or slope and hydraulic conductivity. Both the models were evaluated by using statistical analysis in terms of R^2, Nash-Sutcliffe Coefficient of Efficiency (NSE) and RMSE. The result indicated that RFR outperformed MLR by having R^2 of 0.97 while MLR

only achieved R^2 of 0.64. The high performance of RFR was due to its nonparametric nature which means it does not require a normal distribution. Besides, RFR was able to distinguish which input parameter will affect the groundwater due to nitrate pollution. Therefore, RFR model can be practically utilized to forecast groundwater quality parameters as it could provide a worthwhile analysis of complex and non-linear relationship in hydrological studies.

Zare, Bayat and Daneshkare (2011) investigated and compared the efficiency of ANN and Linear Regression (LR) models in predicting groundwater nitrate concentration at Arak Aquifer, Iran. 818 groundwater data were obtained from 53 groundwater wells in Arak Aquifer. The 818-groundwater data were randomly divided into two categories where 70% of the data were applied in training phase of ANN and LR models while the remaining 30% of data were applied in testing phase. The input parameters for ANN and LR models were pH, electrical conductivity, magnesium, sodium, bicarbonate, calcium, sulphate, potassium and chloride concentrations, total hardness and total dissolved solids. The output of the ANN and LR models were compared with the actual data by statistical analysis in terms of RMSE, R and MAE. The result showed that ANN and LR models were able to achieve satisfactory prediction accuracy. However, ANN model had a better performance than LR model with correlation coefficient (R) of 0.87, RMSE of 10.46 mg/L and MAE of 7.77 mg/L while LR model had R of 0.73, RMSE of 14.45 mg/L and MAE of 10.63 mg/L which overall performance was not as good as ANN model. Besides, lesser input parameters were required by ANN model to forecast nitrate concentration than LR model. Therefore, ANN model is more practical to be used in hydrological studies as it is more cost and time effective.

It can be seen from the literature review on prediction of nitrate above that in general ANN based AI model such as BPNN could achieve higher prediction accuracy than the regression models. Some of the studies had adopted techniques such as input variable selection technique to determine the optimal number of input parameters for optimal model development. This will eventually increase the performance of the training models and generate satisfactory results. With the help of AI models, prediction of nitrate concentration could be more precise and accurate. Moreover, control of water

quality could be carried out more effectively if the source and the level of pollution by nitrate are known.

2.4.5 Ammonia Nitrogen

Ammonia nitrogen is the main water quality parameter studied in this research paper. Just like any other water quality parameters, ammonia nitrogen could cause detrimental effect to the aquatic ecosystem if its concentration exceeds the threshold. Therefore, forecasting of ammonia nitrogen is slowly gaining concern by researchers and decision-makers. Development of AI models to predict ammonia nitrogen concentration in water bodies had been studied by several researchers and is presented below.

Kim and Chung (2005) developed ANN and Multiple Regression (MR) models to predict ammonia nitrogen (NH_3-N) concentration at downstream of the dam located at Geum River in Korea. 82 monthly data from January 1993 to December 1999 of Geum River were utilized in the development of MR and ANN models during the training phase. While the data collected between January 1999 and December 2002 were used in verification phase of both models. The input parameters for MR and ANN were river water quality (i.e., alkalinity), dam outflow, concentration of NH_3-N from past records and temperature of river water. The statistical indices used to evaluate the model performance of ANN and MR by comparison of estimated values and actual values of NH_3-N concentration were coefficient of determination (R^2), root means square error (RMSE) and forecast bias. The results from statistical analysis demonstrated that ANN model performed better than MR model in calibration phase with smaller RMSE and bias values and achieved R^2 value higher than 0.99. However, ANN model showed declined prediction accuracy in verification phase and the improvement was mild compared to MR model. On the other hand, during the verification phase, the bias for MR and ANN models showed negative value. This indicated that both of the models overestimated the NH_3-N concentration. Therefore, the author commented on the modification of both MR and ANN models by applying autocorrelation to NH_3-N concentrations of at least 3 lag months to prevent exaggeration during season of low river water flow. In short, ANN model had high capability in

predicting NH_3-N concentration, but its performance was highly reliable on the type of input parameters used.

2.5 Summary

In chapter 2, several water quality parameters were introduced and discussed. The application of several AI models such as BPNN, ANFIS and SVM models in civil engineering and hydrological applications were discussed and reviewed from past research. Moreover, from the review of various research papers, it was found that implementation of AI models could provide satisfactory results in estimating water quality parameters concentration. However, past research on the development of AI models to predict the concentration of ammonia nitrogen in water bodies are limited. More studies have to be carried out to determine the suitable AI models to estimate ammonia nitrogen concentration in water bodies such as rivers and lakes.

CHAPTER 3

METHODOLOGY AND WORK PLAN

3.1 Workflow of Study

The workflow of this study will be discussed in this section. The first step of this research was to have a thorough study on ammonia nitrogen. The generation, sources and impacts of ammonia nitrogen were studied. Problem statement was addressed and AI techniques were suggested to be used in the prediction of ammonia nitrogen in this research. In order to generate a valuable research, contribution of this study was outlined. Therefore, thorough literature review on past research on development of AI models in the prediction of water qualities parameters and ammonia nitrogen was carried out. Besides, the application of the three AI techniques (i.e., BPNN, ANFIS and SVM) in civil engineering and hydrological field was discussed in literature review. Following was the data collection and preparation of training data and testing data for the development of the three AI models. After everything was ready, modelling of the three AI models was carried out using the prepared data. The performance of each model was then evaluated by statistical analyses. Model which prediction had unsatisfactory error should be trained again. Finally, the output which was the ammonia nitrogen concentration was generated and discussed. The workflow of this study is shown in Figure 3.1.

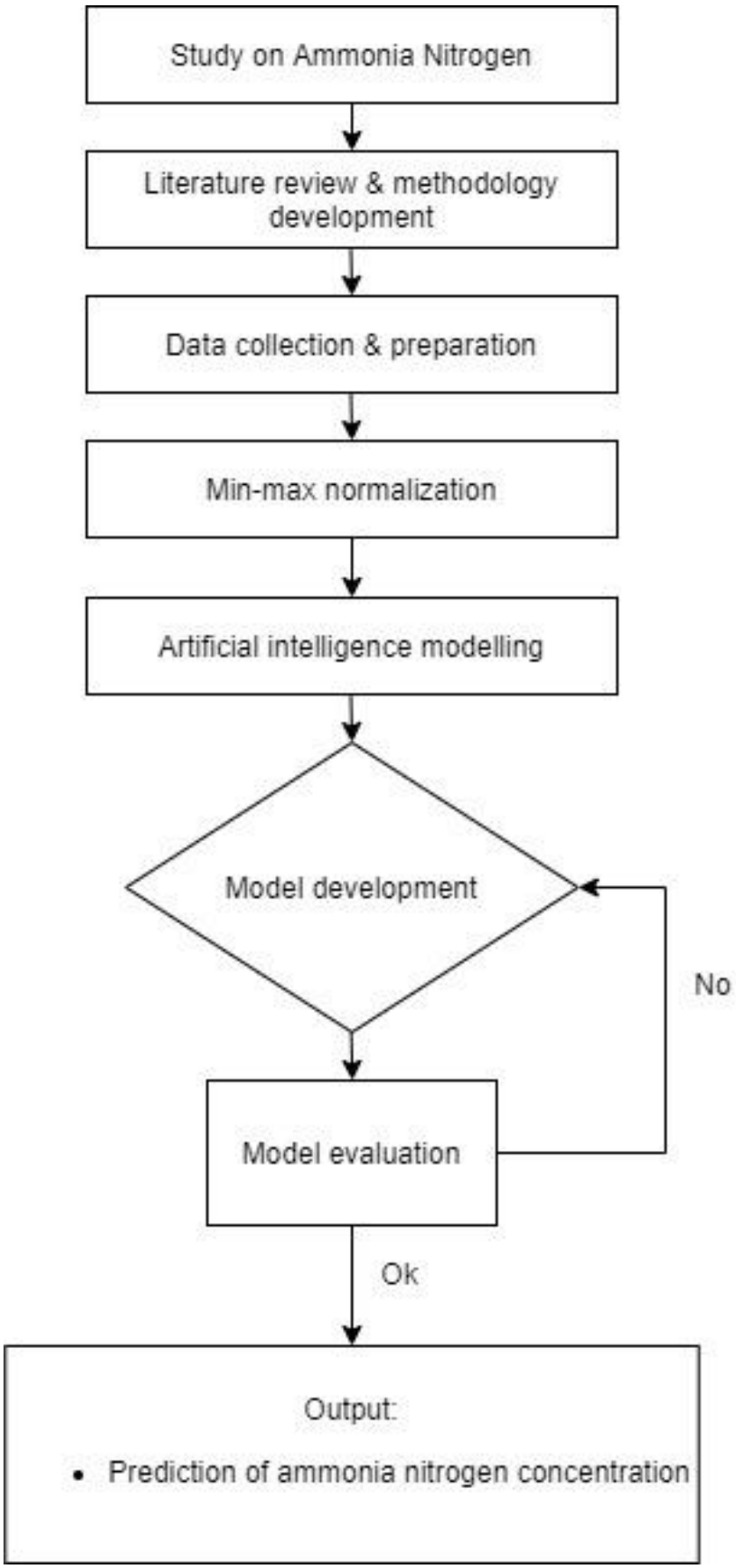

Figure 3.1: Flowchart of workflow of study.

3.2 Study Area

The study area in this research is Langat River located in Selangor, Malaysia. The Langat River is formed by 15 sub-basins which latitudes situated between 2°40'15" N and 3°16'15" N while its longitudes situated between 101°19'20" E and 102°1'10" E. The stream length of the main river is approximately 141km and is located 40km away from east of Kuala Lumpur. The catchment area of Langat Reservoir and Semenyih Reservoir is 54 km^2 and 41 km^2 respectively which built in the year 1981 and 1982 respectively.

The climate within Langat River catchment is made up with constant temperature all year long, high humidity and high rainfall intensity (Juahir, et al., 2010). The average precipitation rate on Langat River is 2316mm annually and normally falls within the range of 1800-3000mm yearly (Abidin, Sulaiman and Yusoff, 2017). The weather in the study area is influenced by the South West monsoon across the Straits of Malacca. Typically, April to November is of wet season while January to March is of drier season. The South West monsoon that drifts across the Straits of Malacca is the primary factor that affects the weather in Langat River basin. Hence, higher rainfall intensity is observed during this period and flooding has become a common phenomenon in this study area (Juahir, et al., 2010).

The gradient of riverbed at upstream of Langat River is steep and the flow velocity of the river is high. The riverbed of Langat River is mainly formed by stones and rocks, while the riverbanks have gentle slope and are protected by vegetation. It is reported that minor bank erosions and sedimentation are problems that occurred at downstream of Hulu Langat town. Although bank erosions problems may happen in any rivers in Malaysia, it is not the primary reason Langat River is chosen to be the study area in this research. Langat River is chosen as the area of interest in this research is because it flows through a highly urbanised area which is between Bangi and Cheras. In this area, issues of bank erosions are serious. Moreover, Langat River that flows through Bangi area is affected by dredging and sand mining activities which then cause drastic depositions and sedimentations problems (Abidin, Sulaiman and Yusoff, 2017). These are the reasons Langat River is chosen as the study area. If evaluation of water quality in Langat River is not carried out, then the water quality of Langat

River will eventually worsen and become irreversible. One of the examples of critically polluted river in Malaysia is Klang River. Figure 3.2 shows the Langat River basin.

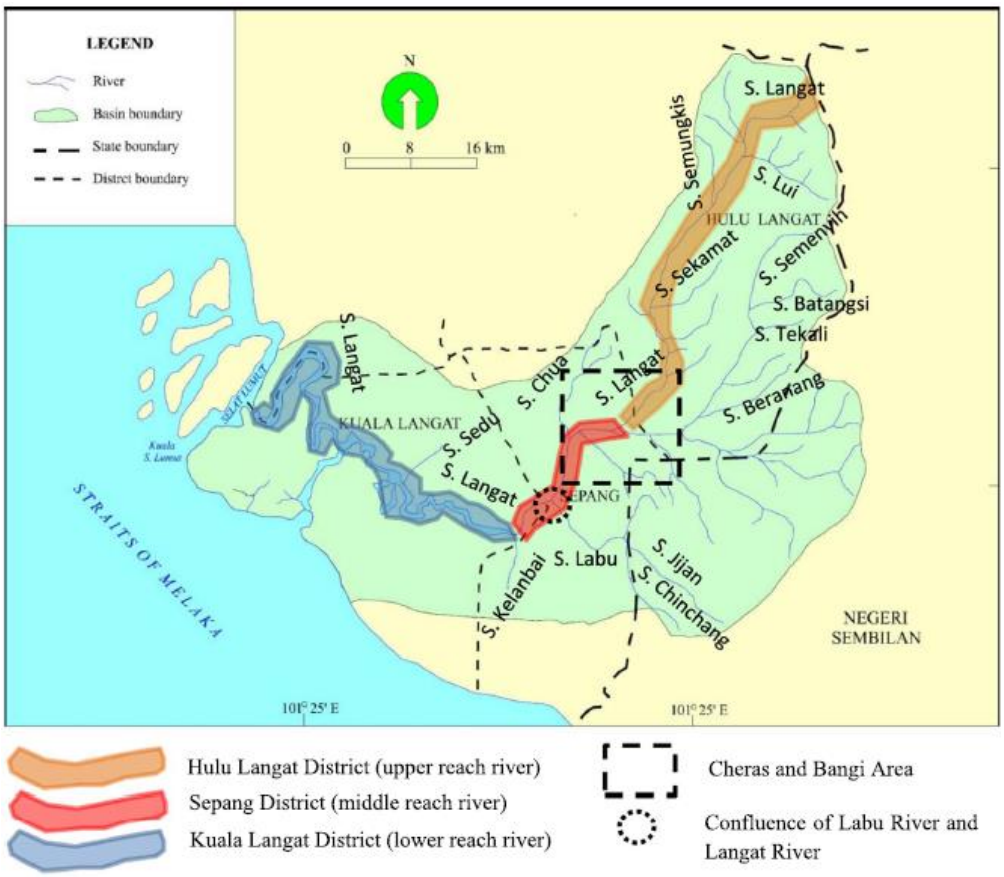

Figure 3.2: Langat River basin (Abidin, Sulaiman and Yusoff, 2017).

3.3 Data Collection and Preparation

The water quality data in this study is provided by Department of Irrigation and Drainage (DID) of Malaysia, obtained from a station with site number 2816641 located at Langat River at Dengkil for a duration from 17 July 2012 to 22 February 2018.

In this research, BPNN, ANFIS and SVM techniques were proposed for the prediction of ammonia nitrogen concentration. Dissolved solids (DS), turbidity (T), total solids (TS), phosphate (PO_4^{3-}) and nitrate (NO_3^-) were selected as the input parameters for the BPNN, ANFIS and SVM models because they were recognized to have higher correlation with ammonia nitrogen. 77 sample datasets were divided into 2 sets for training and testing phase. 80%

(62 datasets) of the datasets were utilized in the training phase of the 3 AI models while 20% (15 datasets) of the datasets were used to test and validate the 3 AI models (Chin, et al., 2018).

To optimize the performance of BPNN, ANFIS and SVM models, identification of training and testing ratio is a crucial step that cannot be neglected. If adequate input-output patterns were used to train the AI models and adequate amount of unseen data were utilized in testing the AI models, the predicted result will achieve better prediction accuracy. According to a research paper, the most popular range of training to testing data ratio used by past researchers were 60%:40% to 80%:20% (Chin, et al., 2018). Since there is a lacking information about training to testing ratio used by past researchers in determining ammonia nitrogen concentration, therefore the upper limit of 80%:20% for the training to testing ratio was selected to supply the AI models with the most feasible input-output patterns.

3.4 Min-Max Normalization

Normalization is an important pre-processing technique used in artificial intelligence (AI) modelling to transform raw input data into suitable form for AI model training. It was found in certain case that utilizing raw data in AI modelling may cause problems to raise and generate inaccurate results. Therefore, by normalizing raw input data, bias within the AI model can be reduced and most importantly the data will be scaled down to the same range. By utilizing the normalized data, training time of AI model can be accelerated (Aksu, Güzeller and Eser, 2019).

In this research, min-max normalization was chosen to normalize the raw input data. Only ammonia nitrogen concentration data were normalized using min-max normalization method as the value range of ammonia nitrogen concentration which ranges from 120 mg/l to 4900 mg/l are too large to be trained efficiently. Therefore, the original ammonia nitrogen concentration value range were rescaled to fall between range of 0 and 1. The formula of min-max normalization used is as follows:

$$x' = \frac{(x_i - x_{min})}{(x_{max} - x_{min})} \tag{3.1}$$

where x' is the data after normalized, x_i indicates input value, x_{min} indicates minimum value and x_{max} indicates maximum value of actual ammonia nitrogen concentration (Aksu, Güzeller, and Eser, 2019). Min-max normalization only changes the value range but will not alter the relationship between among the data. Training process of AI models was then carried out using the normalized dataset.

3.5 Model Structure

3.5.1 Development of BPNN Model

A three-layer BPNN model was implemented to predict ammonia nitrogen concentration. The architecture of the BPNN model is shown in Figure 3.3. In the input layer of the BPNN model, each neuron denoted one input water quality parameter where the input water quality parameters were dissolved solids (DS), turbidity (T), total solid (TS), phosphate (PO_4^{3-}) and nitrate (NO_3^-). While in the output layer, only one neuron represented the output parameter was predicted by the BPNN model which was the ammonia nitrogen concentration.

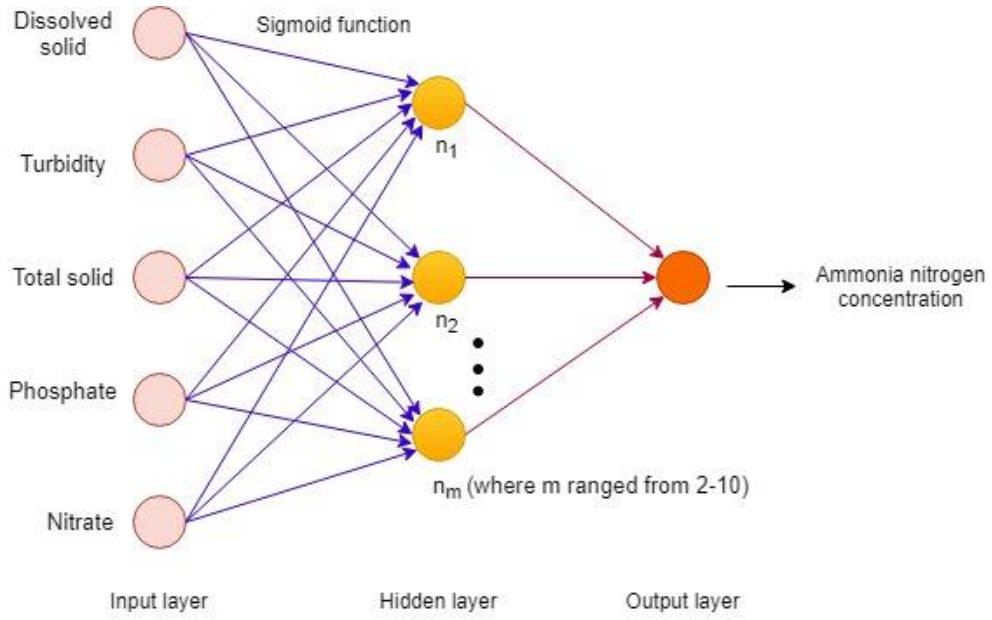

Figure 3.3: BPNN architecture for ammonia nitrogen prediction.

In this study, the architecture of the BPNN model contained only one hidden layer. The transfer function implemented between the input layer and

hidden layer was sigmoid function. The reason sigmoid function was selected is due to its capability to provide a prediction model with greater performance. The training algorithm employed in BPNN model was the Levenberg-Marquardt (trainlm). When dealing with function fitting (non-linear regression) problems, Levenberg-Marquardt (trainlm) exhibited higher performance (Chin, et al., 2018).

In developing BPNN model, it is critical to determine the optimal number of neurons in the hidden layer. Due to the fact that neural network models are sensitive to number of neurons in the hidden layer, under-fitting issue may occur when there are too less neurons while over-fitting issue occurs when the number of neurons is redundant. Therefore, these issues may cause the prediction accuracy of BPNN model to be lowered and reduce in its reliability.

In this study, trial and error method was implemented to determine the optimal number of hidden neurons to prevent the over-fitting and under-fitting issues. Initially, the hidden neurons were assigned at a number of 2 hidden neurons and slowly incremented to 10. The BPNN model at each iteration of trial and error with increasing number of hidden neurons were then trained and tested in order to identify the optimal number of hidden neurons required to generate the best performance of the BPNN model. The development of BPNN model was carried out by using MATLAB.

3.5.2 Development of ANFIS Model

ANFIS model was another AI model employed to predict ammonia nitrogen concentration. Same as the BPNN model, the five water quality parameters (i.e., dissolved solids (DS), turbidity (T), total solid (TS), phosphate (PO_4^{3-}) and nitrate (NO_3^-)) were employed as the input parameters to develop ANFIS model. The output parameter predicted by the ANFIS model was ammonia nitrogen concentration. The architecture of ANFIS model is shown in Figure 3.4.

The optimal number of membership function (mf) in developing the ANFIS model was determined through trial-and-error method. In this study, the adjustment of membership functions in the ANFIS model during the learning phase was carried out by utilizing gradient descent method so that the premise parameters could be deduced. In order to achieve lower value of root-mean-

squared error, the resultant parameters were amended by implementing least mean squared method so that the predicted output will have a higher precision compared to the actual output.

Takagi-Sugeno fuzzy rules were implemented in the ANFIS model in this study. In the development of ANFIS model, R rules were constructed to map several variables (i.e., x_1, x_2, x_3 and x_4) to a distinct output y. However, same output membership function could not be generated by dissimilar rules.

Prior to determining the type of optimal membership function that will be used in training the ANFIS model, 8 types of membership functions were firstly implemented in the trial-and-error process. The 8 membership functions were trapezoidal membership function, Gaussian membership function, product of two sigmoid membership function, difference of two sigmoid membership function, triangular membership function, generalized bell membership function and Pi-shaped curved membership function. Table 3.1 shows the different types of membership function type of inputs used in developing the ANFIS models. The single optimal membership function type was then selected to develop the ANFIS model (Chin, et al., 2018).

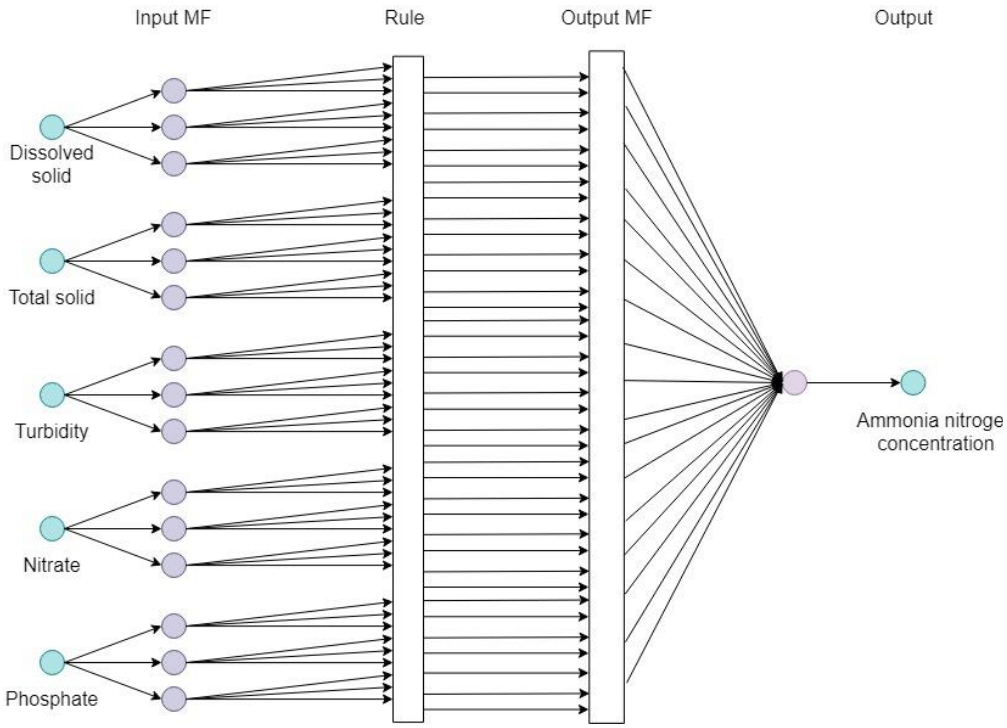

Figure 3.4: ANFIS architecture.

Table 3.1: ANFIS with Varied Input and Output Membership Functions.

No. of model	Input of membership function (MF) type	Output of membership function (MF) type
1	Triangular MF	Constant
2	Triangular MF	Linear
3	Trapezoidal MF	Constant
4	Trapezoidal MF	Linear
5	Generalized bell MF	Constant
6	Generalized bell MF	Linear
7	Gaussian MF	Constant
8	Gaussian MF	Linear
9	Gaussian 2 MF	Constant
10	Gaussian 2 MF	Linear
11	Pi-shaped curved MF	Constant
12	Pi-shaped curved MF	Linear
13	Difference of two sigmoid MF	Constant
14	Difference of two sigmoid MF	Linear
15	Product of two sigmoid MF	Constant
16	Product of two sigmoid MF	Linear

3.5.3 Development of SVM Model

Rapid Miner was implemented to develop the SVM model. In this study, kernel function was used in developing the SVM models as it provides good performance when handling nonlinear data. The input data was mapped and derived by equations shown below where K represents the kernel function:

$$\emptyset(x)^T \emptyset(y) = K(x, y) \tag{3.2}$$

The output of the SVM model was demonstrated in the following equation:

$$f(x) = w^T \emptyset(x_i) + b = \sum_{i=1}^{N} a_i y_i K(x_i, x) + b \quad (3.3)$$

In this study, the selection of kernel function to be used in developing the SVM models was done through trial-and-error method. The five water quality parameters (i.e., dissolved solid, total solid, turbidity, nitrate and phosphate) are selected as the input variables to train and test each of the SVM models. The SVM models were developed using ANOVA, Radial Basis Function (RBF), dot, polynomial, neural, epachnenikov and multiquadric kernel functions as these functions could deal with regression problems. After the training and testing process, the predicted ammonia nitrogen concentration by each SVM models developed using different kernel functions were then evaluated by using statistical analyses in order to determine which kernel function can achieves the best performance in SVM model. Figure 3.5 below illustrates the architecture of the SVM model.

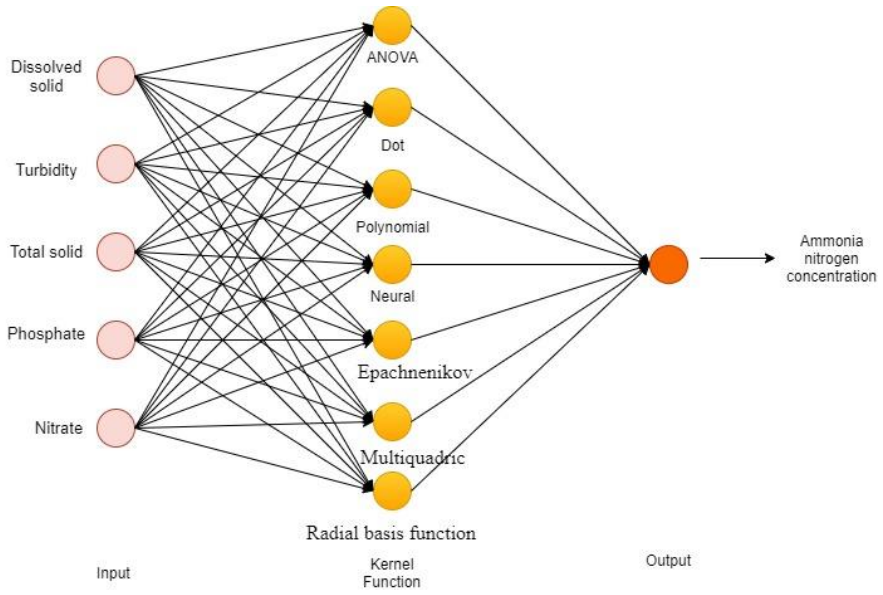

Figure 3.5: SVM Architecture.

3.6 Statistical Analyses

After the training process of the models had done, evaluation on the performance of BPNN, ANFIS and SVM models were carried out. In this study, the evaluation was carried out by using statistical analyses which comprised of the coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), root-mean-squared error (RMSE) and average percentage error. The coefficient of determination (R^2) indicates the percentage of variability between two variables. The range of R^2 is 0 to 1 where 0 indicates the worst while 1 indicates the best. The higher the coefficient, the better the goodness of fit (Chin, et al., 2018). The formula of R^2 is shown below:

$$R^2 = \left[\frac{n\sum_{i=1}^{n} o_i y_{pi} - (\sum_{i=1}^{n} o_i)(\sum_{i=1}^{n} y_{pi})}{\sqrt{[n\sum_{i=1}^{n} o_i^2] \times [n\sum_{i=1}^{n} y_{pi}^2 - (\sum_{i=1}^{n} y_{pi})^2]}}\right]^2 \quad (3.4)$$

The mean squared error (MSE) indicates how close a set of points is to a regression line. The smaller the value of MSE means closer to having a line of best fit (Chin, et al., 2018). The formula of MSE is shown as follow:

$$MSE = \frac{1}{n}\sum(y_i - x_i)^2 \quad (3.5)$$

By using mean absolute error (MAE), problem of error cancelling each other can be avoided, hence actual prediction error can be determined accurately (Chin, et al., 2018). The formula of MAE is:

$$MAE = \frac{1}{n}\sum|y_i - x_i| \quad (3.6)$$

With the utilization of root mean squared error (RMSE), deviation between the truth and observed value can be measured. RMSE is sensitive to outliers and can better indicate the precision of measurement. If the difference between a predicted value and actual value is large, then RMSE will be higher in value (Chin, et al., 2018). The following shows the formula of RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \quad (3.7)$$

The average percentage error determines the percentage of error between the actual value and the predicted value.

$$Average\ Percentage\ error = \frac{|True\ value - Predicted\ value|}{True\ value} \quad (3.8)$$

Where n indicates the number of data pairs, x indicates the actual output value and y indicates the estimated output value.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Parameter Tuning of Adaptive Neuro-Fuzzy Inference System (ANFIS) Model

In this study, several parameters had been tuned for developing the ANFIS models that were used in training process. A total of 16 ANFIS models by utilizing 8 different membership functions types were developed hence one constant and one linear output were produced by each membership function type. The optimum type of membership function was found by trial-and-error method by training the data for every type of membership function types.

It is complex to choose the number of membership function for the input parameters to be adopted in the ANFIS structure, hence the number of membership function were fixed at three for all the membership functions types as this number was most commonly used. Other than that, in the design of ANFIS architecture, hybrid optimization method which includes a combination of least squares-type and backpropagation methods were adopted to tune the patterns in the ANFIS architecture. Besides, the error tolerance in each training process was set to 0.00001 while the epochs of each training process was fixed at 1000 epochs which in other word known as the training iterations.

4.1.1 Model Performance of ANFIS with Unnormalized Datasets

In this section, unnormalized datasets were used in the training and testing process which were the 77 original datasets given by Department of Irrigation and Drainage (DID) of Malaysia. The model performance evaluation of each ANFIS models with unnormalized dataset is summarized in Table 4.1.

From the statistical analysis generated, all the 16 ANFIS models have a relatively low value of correlation coefficient, R. The value of correlation coefficient normally ranges between -1.0 to 1.0. The R values of the developed ANFIS models shown in Table 4.1 ranges between 0.014 to 0.349, indicating that there is a positive correlation between the experimental values and the modelled values. However, the low R value indicates that the relationship

between the actual value and the predicted value is weak. The relationship between the two variables can only be termed as strong which in other word higher accuracy when the correlation coefficient value is more than 0.9. By comparing the developed ANFIS models, Model U4 and Model U12 exhibit the strongest relationship between the predicted values and the experimental values which are 0.349 and 0.334 respectively.

Other than that, the performance of the developed ANFIS models with unnormalized dataset can also be determined from mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE) where smaller value of the error indicates the model is better in terms of its performance. From the results shown in Table 4.1, the ANFIS models with constant type output membership function tabulated a smaller error than the linear type output membership function. In brief, model U3 has the lowest MAE while model U5 has the highest MAE which are 1131.263 and 16175.873 respectively for constant output membership function. For linear output membership function, model U4 has the lowest MAE while model U16 has the highest MAE which are 11673.393 and 8410319.163 respectively. Whereas the least RMSE recorded for constant output membership function is 1432.362 while 25937.834 for linear output membership function. Most of the RMSE of the developed ANFIS models have exceed the tolerable scale, hence model U11 turn out to be the model with the best performance among the other models. Although model U11 has a moderate R value among the 16 models but it achieves the lowest MAE, MSE and RMSE values among the other models.

Besides, the applicability of the developed ANFIS models with unnormalized dataset were also further analysed using coefficient of determination, R^2 and average percentage error. Model with higher value of R^2 shows a higher percentage variation in the predicted output which can be explained by the actual output and vice versa. In contrast, the lower the average percentage error the higher the model prediction accuracy.

Table 4.1: Statistical Analysis of the ANFIS Models with Unnormalized Dataset.

Models	R	R^2	MAE	MSE	RMSE
U1	0.160	0.025	3201.235	51906889.083	7204.644
U2	0.146	0.021	263025.869	869028744031.522	932217.112
U3	0.223	0.050	1131.263	2058770.864	1434.842
U4	0.349	0.122	11673.393	672771241.656	25937.834
U5	0.117	0.014	16175.873	2176400015.540	46651.903
U6	0.014	0.000	3787556.250	91479273843625.800	9564479.800
U7	0.149	0.022	9426.093	790557630.066	28116.857
U8	0.193	0.037	1109060.361	7478862842461.410	2734750.965
U9	0.151	0.023	3947.625	108401921.612	10411.624
U10	0.153	0.023	6477856.622	617441519958131.000	24848370.569
U11	0.225	0.051	1134.818	2051661.918	1432.362
U12	0.334	0.112	15679.298	1668389231.565	40845.921
U13	0.156	0.024	10639.502	1370870896.630	37025.274
U14	0.160	0.026	448234.874	1472626629414.290	1213518.286
U15	0.156	0.024	10639.502	1370870903.304	37025.274
U16	0.153	0.023	8410319.163	101680315110 9080.000	31887350.958

Figure 4.1 reveals the plots of predicted output against the actual output of the ANFIS models with unnormalized dataset. From the scatter plots shown in Figure 4.1, ANFIS model U4 and U12 have the best R^2 value among the other ANFIS models which are 0.122 and 0.112 respectively. There are more than half of the points that falls close to the linear regression line for both model U4 and U12. In addition, the scatter plot style in Figure 4.1(a) and Figure 4.1(b) are likely similar. However, it was noticed that there is a predicted value which has a large deviation from the actual output, and it is a negative value for both model U4 and U12. The worst ANFIS model is model U6 with only a R^2 value of

0.0002 which indicates that there is a very weak relationship between the actual and predicted output hence the prediction accuracy of model U6 is unreliable and in other words is very low which is near to zero.

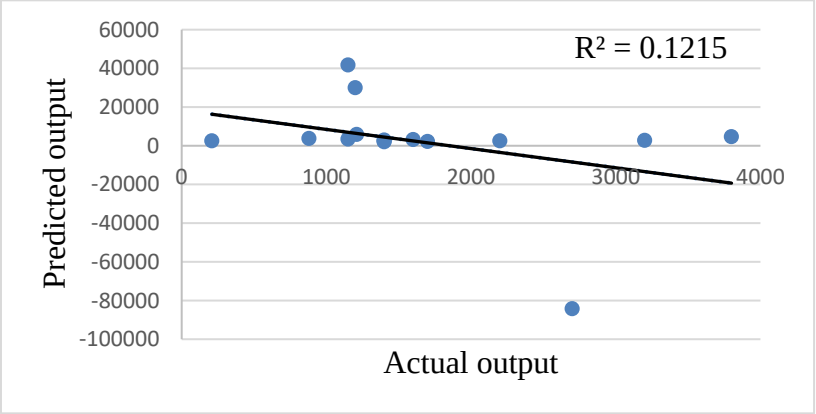

Figure 4.1(a): Relationship in Terms of Coefficient of Determination between Predicted and Actual Outputs of Model U4.

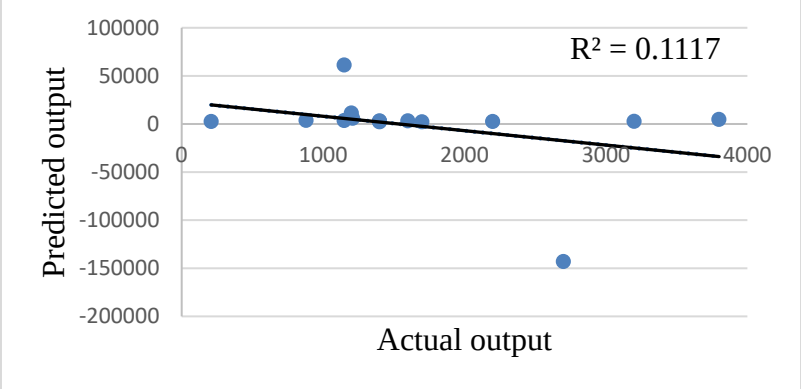

Figure 4.1(b): Relationship in Terms of Coefficient of Determination between Predicted and Actual Outputs of Model U12.

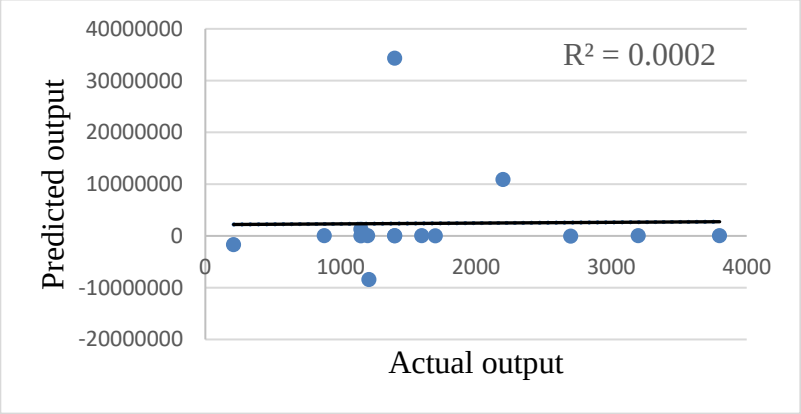

Figure 4.1(c): Relationship in Terms of Coefficient of Determination between Predicted and Actual Outputs of Model U6.

Figure 4.2 is the plot of bar chart showing the average percentage error of the developed ANFIS model with unnormalized dataset. It is observed that all of the developed ANFIS models have exceeded 100% of average percentage error which in terms indicated that all the ANFIS models have very bad accuracy in predicting ammonia nitrogen concentration. In other words, the deviation is very large between the predicted output and the actual output. The low prediction accuracy of the developed ANFIS models is likely due to insufficient dataset used in the training process. In spite of the large average percentage error, model U3 and U11 have achieved the lowest percentage error among the 16 developed ANFIS models which are both 142%.

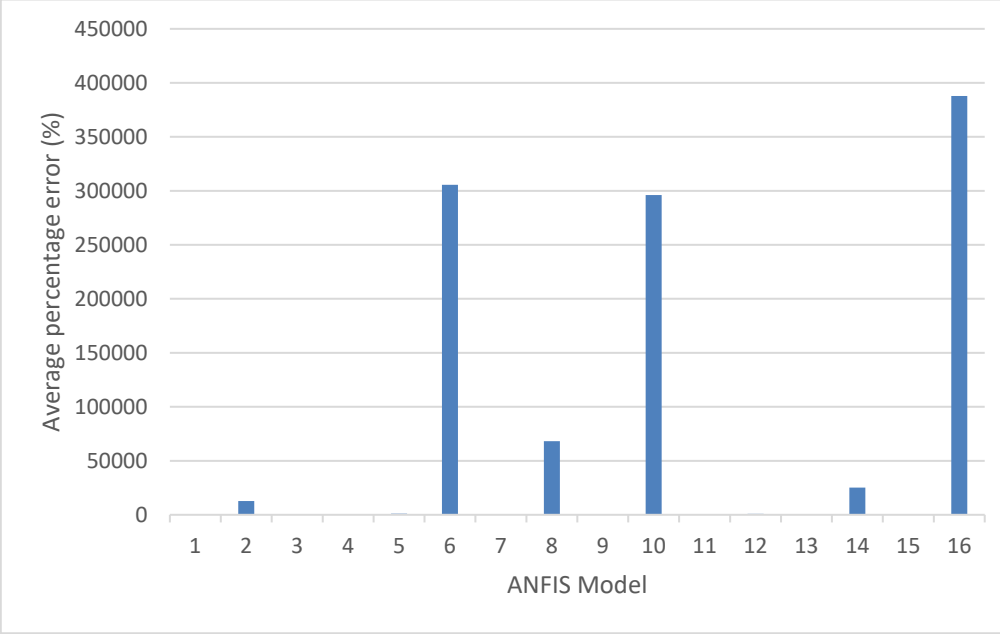

Figure 4.2: Average Percentage Error of the ANFIS Models with Unnormalized Dataset.

4.1.2 Model Performance of ANFIS with Normalized Datasets

The model performance of ANFIS models utilizing unnormalized dataset is not satisfactory, hence in order to improve the performance of the ANFIS model, normalization is introduced. According to Moeeni and Bonakdari, 2017, normalization of data is vital in improving the accuracy of statistical model prediction result. The previous statement is further affirmed in the authors' result and discussion which being stated that normalization is effective in improving the model output accuracy. In this section, the 77 datasets were first normalized with min-max normalization before the datasets were utilized in developing the ANFIS models. Only the ammonia nitrogen values are transformed using the min-max normalization into values between 0 and 1. However, one of the disadvantages of min-max normalization is it cannot handle outliers efficiently. The model performance evaluation of each ANFIS models with normalized dataset is summarized in Table 4.2.

The correlation coefficient, R of the 16 developed ANFIS models with normalized dataset are the same as what is obtained in developed ANFIS models with unnormalized dataset. Apparently, the low R values of all the 16 ANFIS models with unnormalized dataset denoted that the accuracy of predicted value

to the actual value is low. The R value of all the 16 models is below 0.400, hence only model N4 and N12 show strongest relationship between actual values and predicted values among the other ANFIS models.

It can be seen from Table 4.2 that the values of MAE, MSE and RMSE of the developed ANFIS models with normalized dataset had reduced significantly after utilizing min-max normalization method on the training and testing datasets. From the tabulated results in Table 4.2, the error for the constant type of output membership function is smaller than the linear type output membership function for each type of membership function developed. For constant type output membership function, model N3 has the lowest MAE while model N5 has the highest MAE which are 0.231 and 3.290 respectively. Whereas for linear type output membership function, model N4 has the lowest MAE while model N10 has the highest MAE which are 2.430 and 1327.513 respectively.

Furthermore, model N11 achieved the lowest RMSE value which is 0.292 among the other constant typed output membership function. On the other hand, the lowest RMSE recorded for linear type output membership function belongs to model N4 with an RMSE value of 5.436. The models with the lowest MSE for constant and linear type output membership function are model N11 and N4 respectively as well because MSE is just squared of RMSE. In short, model N11 has the highest prediction accuracy with lowest MAE, MSE and RMSE which are 0.232, 0.085 and 0.292 respectively. Most of the models still have large errors which mainly due to the insufficient training data utilized.

Moreover, the ANFIS models with normalized dataset were also analysed using R^2 and average percentage error. The R^2 values for all the 16 developed ANFIS models with normalized dataset are the same as the R^2 values of the develop ANFIS models with unnormalized dataset, therefore the best and worst scatter plot of relationship between actual and predicted outputs can refer to Figure 4.1. Similar to ANFIS model with unnormalized dataset, ANFIS model with normalized dataset of model N4 and N12 have the best R^2 value among the other ANFIS models which are 0.122 and 0.11 respectively while model N6 has the lowest prediction accuracy.

Table 4.2: Statistical Analysis of the ANFIS Models with Normalized Dataset.

Models	R	R^2	MAE	MSE	RMSE
N1	0.160	0.025	0.653	2.160	1.470
N2	0.146	0.021	53.636	36126.122	190.069
N3	0.223	0.050	0.231	0.086	0.293
N4	0.349	0.122	2.430	29.446	5.426
N5	0.117	0.014	3.290	90.237	9.499
N6	0.014	0.000	773.556	3819216.746	1954.282
N7	0.149	0.022	1.921	32.812	5.728
N8	0.193	0.037	228.300	313799.499	560.178
N9	0.151	0.023	0.806	4.521	2.126
N10	0.153	0.023	1327.513	25931125.532	5092.261
N11	0.225	0.051	0.232	0.085	0.292
N12	0.334	0.112	3.201	69.524	8.338
N13	0.156	0.024	2.172	57.117	7.558
N14	0.160	0.026	284.050	855075.565	924.703
N15	0.156	0.024	2.172	57.117	7.558
N16	0.153	0.023	1143.872	12206700.336	3493.809

Figure 4.3 shows the plot of average percentage error of each developed ANFIS model with normalized dataset. From Figure 4.3, it is identified that model N6, N8, N10, N14 and N16 have extremely large average percentage error compared to other ANFIS models. Although other ANFIS models do not have extreme values of average percentage error, their average percentage error still exceeds the 100% limit. This has suggested that the ANFIS models with normalized dataset have considerable low accuracy in predicting the ammonia nitrogen concentration with the current datasets available. The high average percentage error indicates that the prediction accuracy of the predicted output is no way close to the actual output. On the whole, model N3 and N11 have similar and the lowest average percentage error which is 142%.

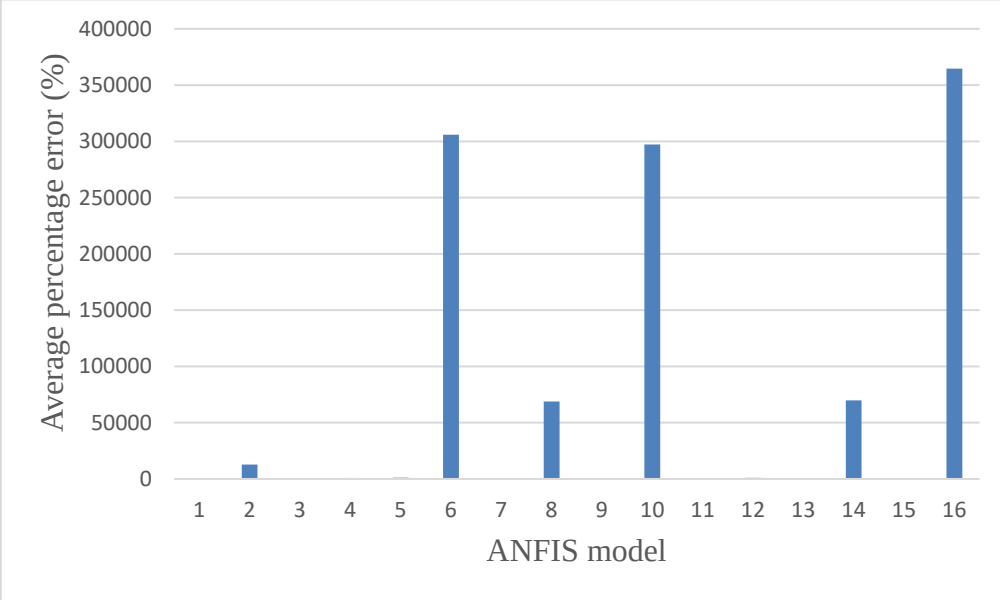

Figure 4.3: Average Percentage Error of the ANFIS Models with Normalized Dataset.

4.1.3 Selection of the Most Suitable ANFIS Prediction Model

The performance of the developed ANFIS models were evaluated by utilizing different statistical analyses. It is observed that the ANFIS model with normalized dataset provide better results than the ANFIS model with unnormalized dataset. Therefore, model with the best performance is selected from the ANFIS model with normalized dataset.

According to the result obtained, model N11 obtained a relatively low MAE, MSE and RMSE values which are 0.232, 0.085 and 0.292 respectively. Although model N11 R and R^2 values are lower than that of model N4 and N12, but its lowest MAE, MSE and RMSE values are superior to the effect of R and R^2. Which in other words, the prediction accuracy of model N11 has higher priority over the defined relationship between the actual and predicted output. The R and R^2 values of model N11 are 0.225 and 0.051 respectively. Other than that, model N11 also acquired the lowest average percentage error which is 142%. Hence, model N11 is chosen to be the most suitable ANFIS model to be utilized in the prediction of the ammonia nitrogen concentration in Langat River. Following is the description of design architecture for model N11:

- Number of membership function: 3

- Membership function type input: Pi-shaped curved membership function
- Membership function type output: constant.
- Optimization method: Hybrid (combination of back-propagation and least square methods)

4.2 Parameter Tuning of Support Vector Machine (SVM) Model

The SVM models were developed using a software known as RapidMiner. While developing the SVM models, a few parameters had been tuned. A total of seven different types of SVM models each utilizing a different kernel function had been developed. In the design of SVM models, the max training iterations for each SVM models were set to 1000 iterations. Whereas the other parameters for each type of kernel function was left unchanged and followed the software provided parameters.

4.2.1 Model Performance of SVM with Unnormalized Datasets

In this section, unnormalized datasets were utilized in the training and testing process of the SVM models. The model performance evaluation of the seven different SVM models with unnormalized dataset is specified in Table 4.3.

The results generated from statistical analysis shows that all the SVM models have relatively low R values. Especially for neural, anova, epachnenikov and multiquadric kernel function SVM models, the R values achieved by these SVM models are zero. Which indicates that there is no relationship at all between the predicted output with the actual output and is unimportant. On the other hand, the SVM model which utilized radial kernel function recorded the highest R value while SVM model which utilized polynomial kernel function recorded the lowest R value. This indicates that SVM model with radial kernel function has the strongest relationship between actual output and predicted output. However, with a R value of 0.068, the radial kernel function type SVM model is still not satisfactory to be utilized practically to predict ammonia nitrogen concentration in the river.

Unlike the R value, five kernel functions have R^2 value of zero where these kernel functions are polynomial, neural, anova, epachnenikov and

multiquadric. SVM model with radial kernel function has the highest R^2 of 0.005 while SVM model with dot kernel function has lower R^2 value of 0.001.

The errors tabulated in Table 4.3 are large because the data have not been normalized. From the tabulated results, all the kernel function SVM models have a close range of MAE values which ranges from 888.678 to 908.580. In fact, SVM model with multiquadric kernel function achieved the lowest MAE value which is 888.678 while SVM model with neural kernel function achieved the highest MAE value which is 908.580. Similarly, all the SVM models have close range RMSE value ranges from 1073.628 to 1076.440 except SVM model with multiquadric kernel function which has the lowest RMSE value of 1040.673. Hence, SVM model with multiquadric kernel function has the best performance among the other model in terms of MAE, MSE and RMSE.

Table 4.3: Statistical Analysis of the SVM Model with Unnormalized Dataset.

Kernel Function	R	R^2	MAE	MSE	RMSE
Dot	0.031	0.001	907.693	1155704.551	1075.037
Radial	0.068	0.005	906.825	1153166.709	1073.856
Polynomial	0.005	0.000	906.401	1152677.082	1073.628
Neural	0.000	0.000	908.580	1158723.074	1076.440
Anova	0.000	0.000	906.837	1153652.143	1074.082
Epachnenikov	0.000	0.000	907.212	1154316.021	1074.391
Multiquadric	0.000	0.000	888.678	1083000.293	1040.673

The plot of average percentage error of each SVM models with unnormalized dataset is shown in Figure 4.4. From the recorded average percentage error, all the seven SVM models have low performance in predicting ammonia nitrogen concentration in Langat River as their average percentage error have exceeded 100%. As a matter of fact, the accuracy of all the SVM models is extremely low and have exceeded the tolerable scale. Although having high value of average percentage error, SVM model with multiquadric kernel

function still achieved the lowest average percentage error among the other SVM models which is 104.18%.

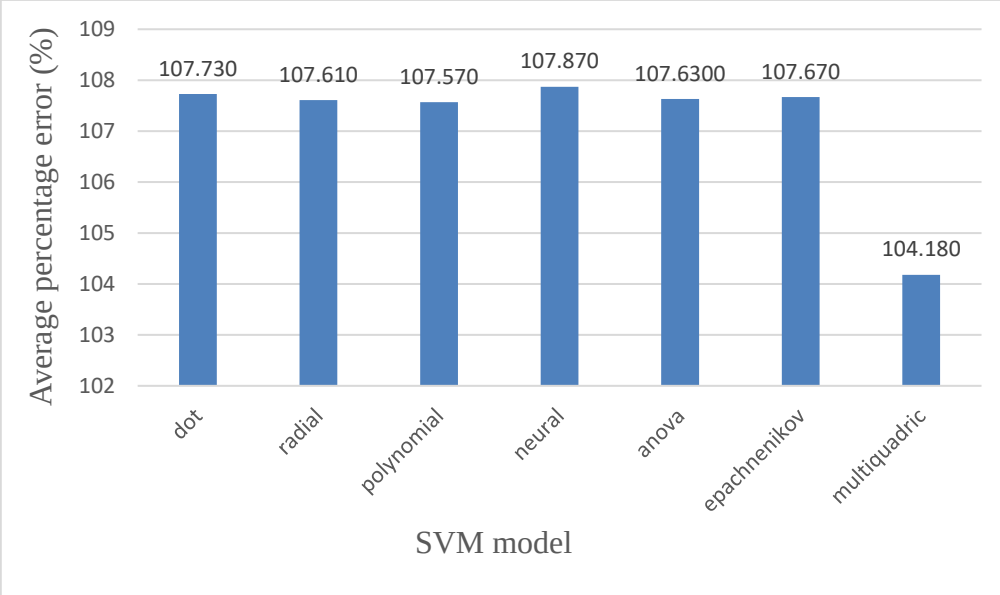

Figure 4.4: Average Percentage Error of SVM Models with Unnormalized Dataset.

4.2.2 Model Performance of SVM with Normalized Datasets

In this section, the 77 datasets were first normalized with min-max normalization before the datasets were utilized in the training and testing process of SVM models. Take note that only the ammonia nitrogen values are normalized using the min-max normalization into values between 0 and 1. Table 4.4 summarizes the model performance evaluation results of SVM models with normalized dataset.

As observed from Table 4.4, All the SVM models have R value of zero except dot kernel function SVM model with a R value of 0.088. Similar for R^2, all the normalized SVM models have zero value except dot kernel function SVM model with a R^2 value of 0.008.

Furthermore, the MAE, MSE and RMSE values of the seven SVM models were compared. The multiquadric kernel function SVM model recorded the lowest MAE value of 0.181 while the neural kernel function SVM model recorded the highest MAE value of 3.393. Multiquadric kernel function SVM model also achieved the lowest MSE and RMSE value of 0.045 and 0.212 respectively. Whereas neural kernel function SVM model has the highest MSE

and RMSE value recorded at 15.928 and 3.991 respectively. Hence, SVM model with multiquadric kernel function has the best performance in terms of MAE, MSE and RMSE.

Table 4.4: Statistical Analysis of the SVM Model with Normalized Dataset.

Kernel Function	R	R^2	MAE	MSE	RMSE
Dot	0.088	0.008	0.203	0.058	0.240
Radial	0.000	0.000	0.241	0.091	0.301
Polynomial	0.000	0.000	0.232	0.086	0.294
Neural	0.000	0.000	3.393	15.928	3.991
Anova	0.000	0.000	0.269	0.102	0.319
Epachnenikov	0.000	0.000	0.268	0.114	0.337
Multiquadric	0.000	0.000	0.181	0.045	0.212

The plot of average percentage error of each SVM models with normalized dataset is shown in Figure 4.5. From the recorded results, all the seven normalized SVM models have low performance in predicting ammonia nitrogen concentration in Langat River as their average percentage error have exceeded 100%. As a matter of fact, the accuracy of all the SVM models is extremely low and have exceeded the tolerable scale. Although having high value of average percentage error, SVM model with multiquadric kernel function still achieved the lowest average percentage error among the other SVM models which is 104.18%. Whereas neural kernel function SVM model is the outlier among the seven normalized SVM models, reaching an average percentage error of 1408.32%.

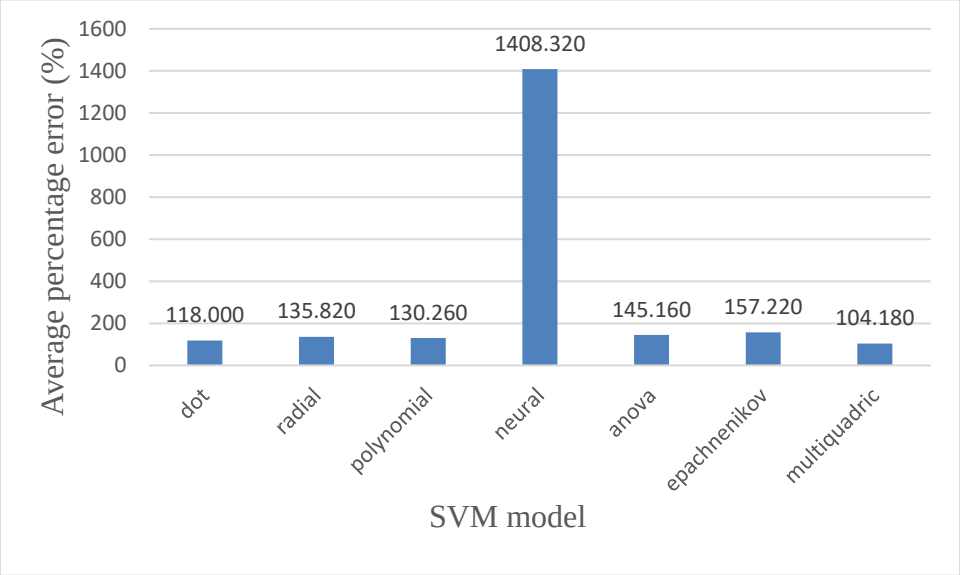

Figure 4.5: Average Percentage Error of SVM Models with Normalized Dataset.

4.2.3 Selection of the Most Suitable SVM Prediction Model

The performance of each SVM models were explained earlier, therefore SVM model with the best performance has to be selected in this section. The SVM model with normalized dataset provides better results compared to the SVM model with unnormalized dataset. Hence, SVM model with the best performance is selected from the SVM model with normalized dataset.

According to the recorded statistical analysis results, dot kernel function SVM model is selected to be the most suitable SVM prediction model. Although its MAE, MSE and RMSE values are lower than that of multiquadric kernel function SVM model, it is the only developed SVM model that has R and R^2 values. SVM models with zero R and R^2 values are not in the selection consideration as their predicted and actual output have no relationship which makes it unmeaningful. The R, R^2, MAE, MSE and RMSE values for dot kernel function SVM model are 0.088, 0.008, 0.203, 0.058 and 0.240 respectively. The average percentage error of dot kernel function SVM model is 118%.

4.3 Parameter Tuning of Back Propagation Neural Network (BPNN) Model

During the development of BPNN models, several parameters had been tuned and set for the training process. A total of 9 BPNN models were developed by using hidden neurons ranges from 2 to 10. The optimum number of hidden neurons was determined by trial-and-error method.

Following are the parameters tuned for the training process of the BPNN models. Firstly, the number of hidden layers was set to 1. Then, the error tolerance and epochs were set to 0.00001 and 1000 respectively. Furthermore, unnormalized and normalized dataset were utilized in the training of BPNN models for later comparison. Two different training function were used in the training process of the BPNN models with normalized dataset which consists of log sigmoid and tangent sigmoid training functions. While only log sigmoid function alone was used in the training process of BPNN models with unnormalized dataset.

4.3.1 Model Performance of BPNN with Unnormalized Datasets

In this section, statistical analyses result of BPNN models based on the unnormalized datasets will be discussed. The statistical analyses result is demonstrated in Table 4.5 and each BPNN models are compared to one another to determine the most suitable model for the prediction of ammonia nitrogen concentration in Langat River.

From the statistical analyses result tabulated in Table 4.5, the BPNN model with 4 hidden neurons recorded the highest R and R^2 values at 0.424 and 0.179 respectively. This indicate that BPNN model with 4 hidden neurons has the strongest relationship between the actual output and predicted output or to put it another way, higher portion of the predicted result is close to the actual result. Whereas BPNN model with 3 hidden neurons recorded the lowest R and R^2 values at 0.146 and 0.021 respectively. This in turn denote that the relationship between the actual output and the predicted output of BPNN model with 3 hidden neurons is the weakest among the other BPNN models. From a clearer point of view, the comparison of R^2 value between BPNN model with 3 and 4 hidden neurons is shown in Figure 4.6(a) and Figure 4.6(b). It illustrates

that BPNN model having 4 hidden neurons has points which are closer to the linear regression line than the BPNN model with 3 hidden neurons.

Table 4.5: Statistical Analysis of the BPNN Model with Unnormalized Dataset.

No. of Hidden Neurons	R	R^2	MAE	MSE	RMSE
2	0.165	0.027	1414.893	2583835.971	1607.431
3	0.146	0.021	1142.081	1894892.129	1376.551
4	0.424	0.179	1909.363	5210337.946	2282.616
5	0.228	0.052	935.686	1117479.922	1057.109
6	0.206	0.042	1167.848	2023507.609	1422.500
7	0.219	0.048	1372.124	3049410.602	1746.256
8	0.173	0.030	1527.451	4056730.355	2014.133
9	0.248	0.062	1591.159	4497183.205	2120.656
10	0.285	0.081	2229.400	6445077.949	2538.716

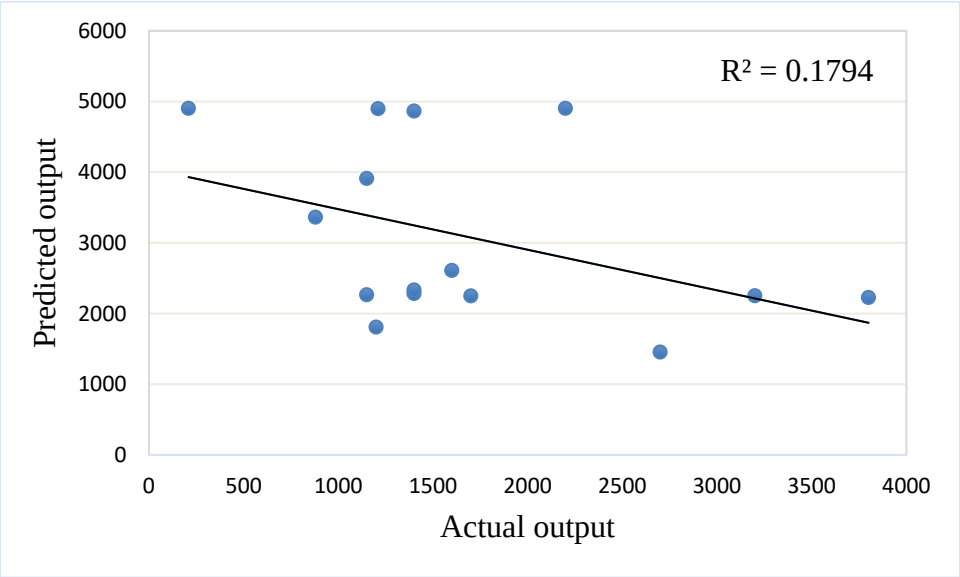

Figure 4.6(a): Relationship in Terms of Coefficient of Determination between Predicted and Actual Outputs of the BPNN Model with 4 hidden neurons.

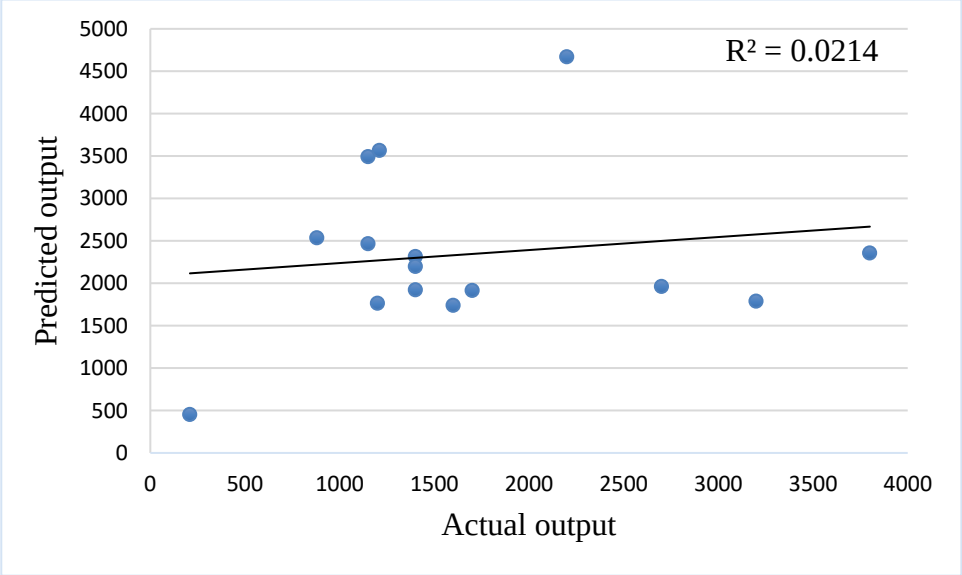

Figure 4.6(b): Relationship in Terms of Coefficient of Determination between Predicted and Actual Outputs of the BPNN Model with 3 hidden neurons.

Besides, the BPNN model with the lowest MSE, RMSE and MAE values is BPNN model with 5 hidden neurons. The corresponding MAE, MSE and RMSE values for this model are 935.686, 1117479.922 and 1057.109 respectively. On the other hand, BPNN model with 10 hidden neurons has the highest MAE, MSE and RMSE values recorded at 2229.400, 6445077.949 and 2538.716 respectively. This denote that BPNN model with 5 hidden neurons has the best performance in terms of error as its prediction error is the lowest among all the developed BPNN models.

Furthermore, Figure 4.7 illustrates the average percentage error of the developed BPNN models for each number of hidden neurons. In short, BPNN model with 5 hidden neurons has the least average percentage error recorded at 70.67% while the BPNN model with 10 hidden neurons recorded the highest average percentage error at 262.31%. Hence, BPNN model with 5 hidden neurons has the best performance in term of average percentage error which in turn determined to be the BPNN model with the best performance.

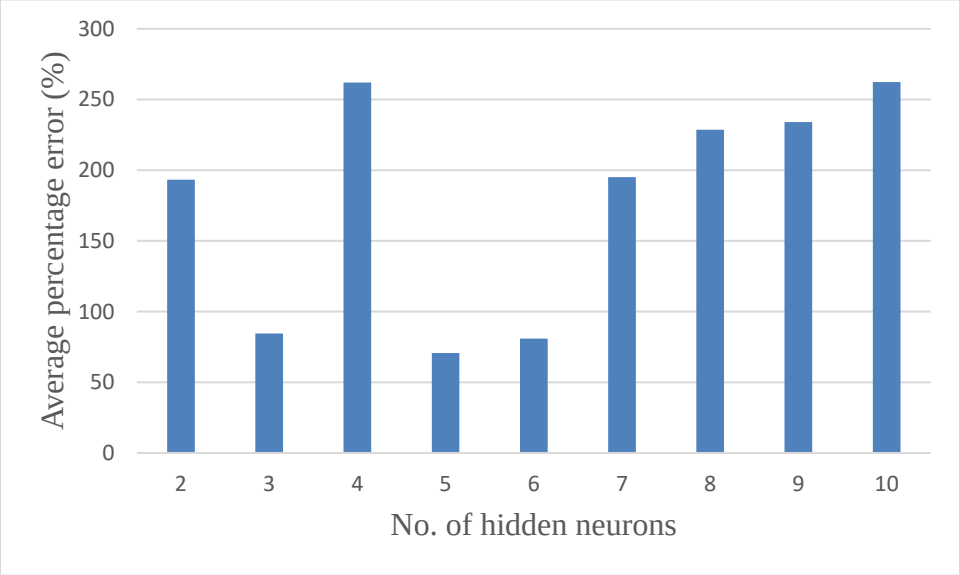

Figure 4.7: Average Percentage Error of BPNN Models with Unnormalized Dataset.

4.3.2 Model Performance of BPNN with Normalized Dataset

In this section, the BPNN models were developed using normalized datasets where only ammonia nitrogen values were normalized. Whereas two transfer functions which are log sigmoid and tangent sigmoid transfer function were utilized in training the 9 BPNN models separately. Then, the BPNN model having the best performance will be each selected from one transfer function trained BPNN models.

4.3.2.1 Model Performance of BPNN with Normalized Dataset Trained with Log Sigmoid Transfer Function

The statistical analyses result of the developed BPNN models with normalized dataset trained with log sigmoid transfer function are presented in Table 4.6 to select the most optimal model for the prediction of ammonia nitrogen concentration in Langat River. Comparisons are made on the BPNN models with different number of hidden neurons ranges from 2 to 10 hidden neurons.

From the statistical analyses result tabulated in Table 4.6, BPNN model with 4 hidden neurons has the highest R and R^2 values recorded at 0.465 and 0.217 respectively. This indicates that there is a strong relationship between the actual output and the predicted output as the R value of this model is closer to 1 as compared to the other model. The higher R^2 value of BPNN model with 4

hidden neurons indicates that there is a higher percentage variation in the predicted output which can be explained by the actual output when compared to the other model. Whereas BPNN model with 2 hidden neurons achieves the lowest R and R^2 values among the BPNN models trained with log sigmoid transfer function which is recorded at 0.030 and 0.001 respectively. Figure 4.8(a) and Figure 4.8(b) illustrate the relationship between the predicted output and actual output of BPNN model with 2 and 4 hidden neurons respectively in term of coefficient of determination. In Figure 4.8(a), there are three outliers from the line of regression while in Figure 4.8(b) there are only two outliers from the line of regression. Therefore, BPNN model with 4 hidden neurons is the best performance model in term of coefficient of determination.

Besides, the performance of the BPNN models can be further determined from MAE, MSE and RMSE values. As tabulated in Table 4.6, BPNN model with 4 hidden neurons has the lowest MAE, MSE and RMSE values among the other models recorded at 0.153, 0.030 and 0.172 respectively. Hence, BPNN model with 4 hidden neurons turns out to be the model with the highest prediction accuracy and achieves the highest performance among the other models. In contrast, BPNN model with 10 hidden neurons achieves the highest value of MAE, MSE and RMSE which is recorded at 0.413, 0.247 and 0.497 respectively. This indicates that BPNN model with 10 hidden neurons is the model with the lowest prediction accuracy and performance.

Table 4.6: Statistical Analysis of the BPNN Model with Normalized Dataset Trained with Log Sigmoid Transfer Function.

No. of Hidden Neurons	R	R^2	MAE	MSE	RMSE
2	0.030	0.001	0.262	0.090	0.299
3	0.032	0.001	0.255	0.094	0.307
4	0.465	0.217	0.153	0.030	0.172
5	0.032	0.001	0.299	0.141	0.376
6	0.150	0.023	0.267	0.122	0.349
7	0.064	0.004	0.255	0.097	0.311
8	0.118	0.014	0.279	0.120	0.346
9	0.204	0.042	0.285	0.128	0.357
10	0.139	0.019	0.413	0.247	0.497

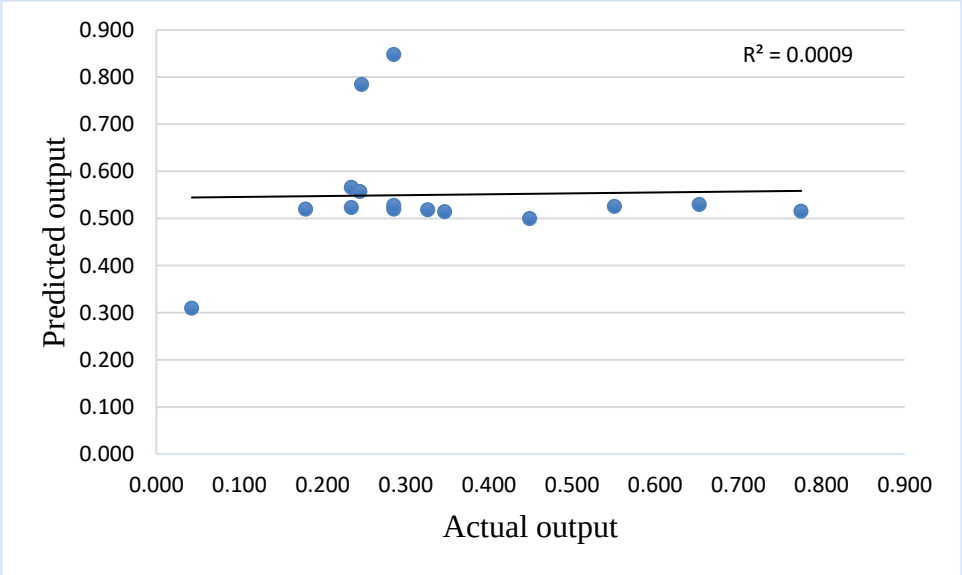

Figure 4.8(a): Relationship in Term of Coefficient of Determination between Predicted and Actual Outputs of the BPNN Model with 2 hidden neurons.

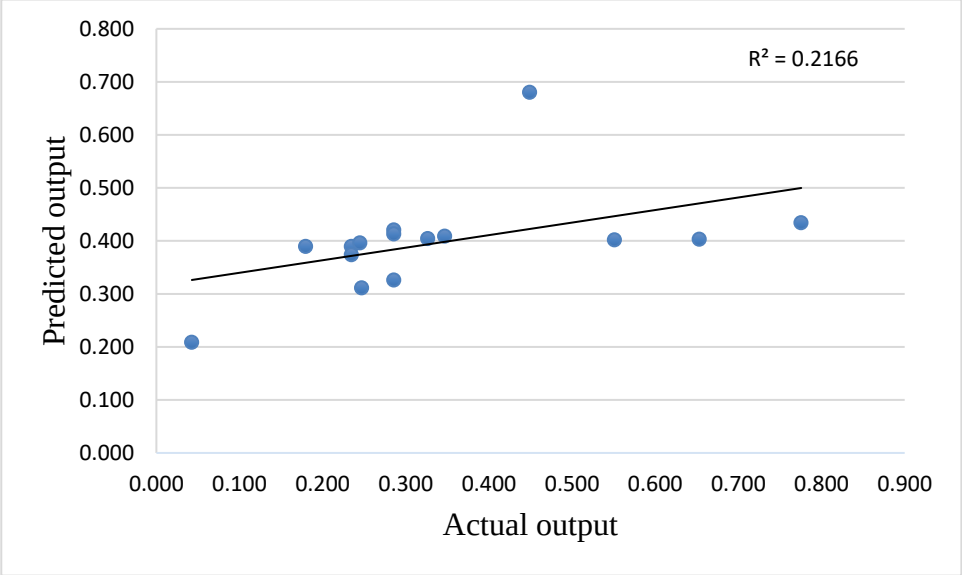

Figure 4.8(b): Relationship in Term of Coefficient of Determination between Predicted and Actual Outputs of the BPNN Model with 4 hidden neurons.

Furthermore, Figure 4.9 illustrates the average percentage error of the BPNN models trained with log sigmoid transfer function. It is observed that BPNN model with 4 hidden neurons has the lowest average percentage error recorded at 68.50% which makes it the most reliable model in predicting ammonia nitrogen concentration. In other words, BPNN model with 4 hidden neurons has the best performance among the other models. On the other hand, BPNN model with 9 hidden neurons achieves the highest average percentage error of 187.71% which indicates it has an extremely weak prediction accuracy on the ammonia nitrogen concentration in Langat River. Therefore, BPNN model with 4 hidden neurons is selected as the most suitable log sigmoid trained BPNN model for predicting the ammonia nitrogen concentration in Langat River.

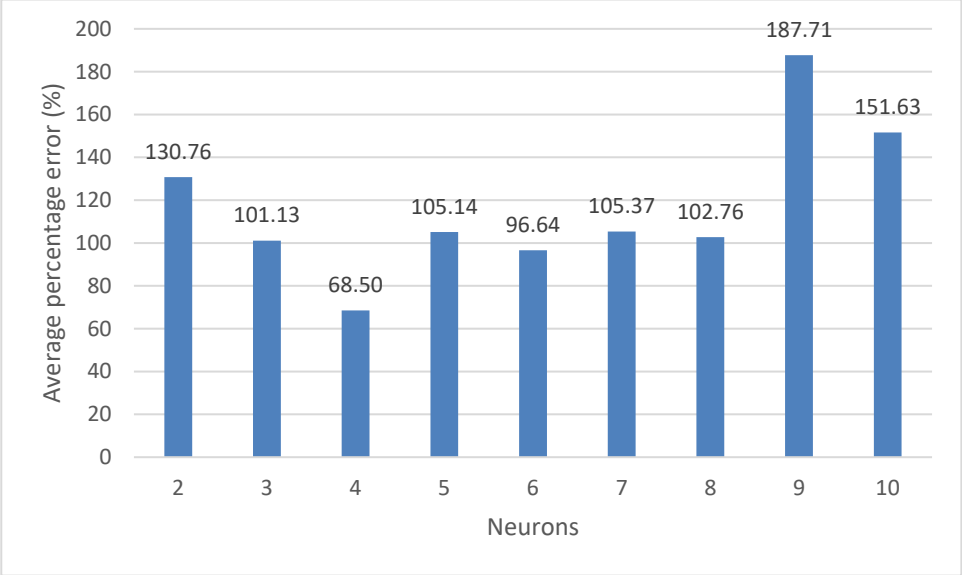

Figure 4.9: Average Percentage Error of BPNN Models with Normalized Dataset Trained with Log Sigmoid Transfer Function.

4.3.2.2 Model Performance of BPNN with Normalized Dataset Trained with Tangent Sigmoid Transfer Function.

In this section, the model performance of the developed BPNN models with normalized dataset trained with tangent sigmoid transfer function are presented in Table 4.7. Comparison is then made on the BPNN models with different number of hidden neurons ranges from 2 to 10 hidden neurons.

From the statistical analyses result tabulated in Table 4.7, the BPNN model having the highest R and R^2 values is model with 6 hidden neurons which recorded at 0.434 and 0.188 respectively. On the other hand, the model having the lowest R and R^2 values is model with 2 hidden neurons which recorded at 0.043 and 0.002 respectively. This shows that model with 6 hidden neurons has the strongest relationship between the predicted output and the actual output while vice versa for the BPNN model having 2 hidden neurons. The relationship between predicted output and actual output in term of coefficient of determination is illustrated in Figure 4.10. In Figure 4.10(a), the scatter plot illustrates that more points are plotted closer to the linear regression line which in other words, bigger portion of the predicted output can be explained by the variation of actual output. However, in Figure 4.10(b), there are a few extreme outliers illustrates in the scatter plot which means lesser predicted output is explained by the variation of actual output. Therefore, BPNN model with 6

hidden neurons is the model with the best performance in term of R and R^2 among the BPNN models trained with tangent sigmoid transfer function.

Nevertheless, the performance of the BPNN models can be further determined by utilizing MAE, MSE and RMSE. The lower the values of these errors, the better the performance of the models. According to Table 4.7, BPNN model with 4 hidden neurons recorded the lowest MAE, MSE and RMSE values tabulated at 0.244, 0.098 and 0.313 respectively. While BPNN model with 7 hidden neurons recorded the highest MAE, MSE and RMSE values tabulated at 0.471, 0.287 and 0.536 respectively. Hence, BPNN model with 4 hidden neurons achieves the best performance in term of MAE, MSE and RMSE as its prediction error is the lowest.

Table 4.7: Statistical Analysis of the BPNN Model with Normalized Dataset Trained with Tangent Sigmoid Transfer Function.

No. of Hidden Neurons	R	R^2	MAE	MSE	RMSE
2	0.043	0.002	0.277	0.105	0.324
3	0.089	0.008	0.296	0.125	0.353
4	0.132	0.017	0.244	0.098	0.313
5	0.191	0.037	0.271	0.114	0.337
6	0.434	0.188	0.306	0.143	0.378
7	0.077	0.006	0.471	0.287	0.536
8	0.115	0.013	0.267	0.115	0.339
9	0.104	0.011	0.292	0.122	0.349
10	0.066	0.004	0.255	0.090	0.300

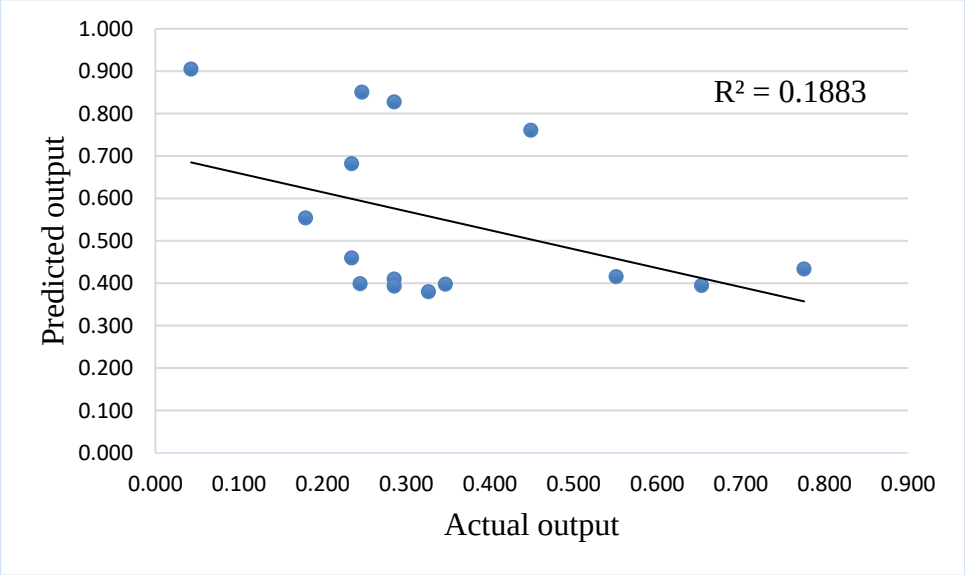

Figure 4.10(a): Relationship in Term of Coefficient of Determination between Predicted and Actual Outputs of the BPNN model with 6 hidden neurons.

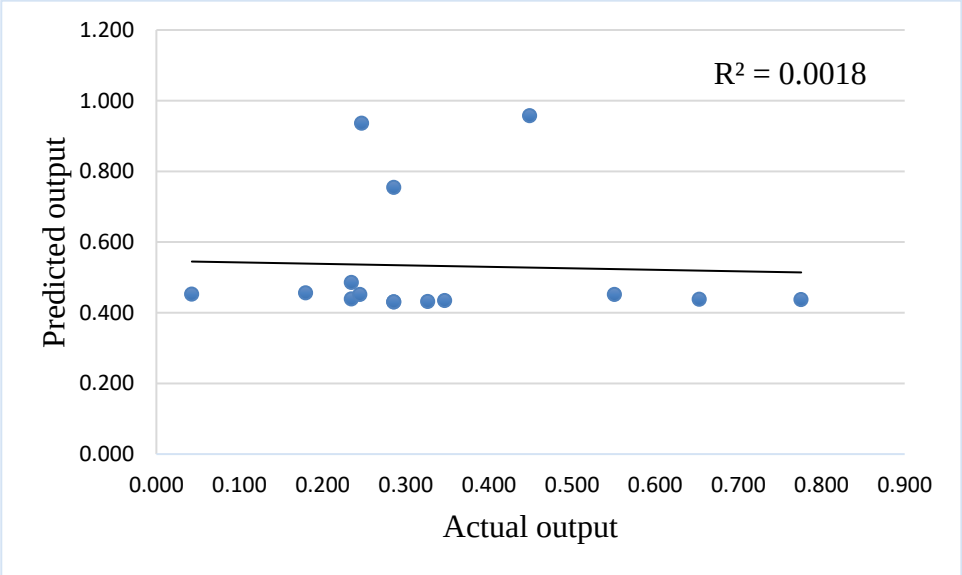

Figure 4.10(b): Relationship in Term of Coefficient of Determination between Predicted and Actual Outputs of the BPNN model with 2 hidden neurons.

Last but not least, the performance of the BPNN models in this section can also be identified by average percentage error. As illustrated in Figure 4.11, BPNN model with 8 hidden neurons achieves the lowest average percentage error recorded at 94.13% while the other models are all having average percentage error greater than 100%. Hence, BPNN model with 8 hidden neurons

is the optimal model as it achieves the highest prediction accuracy among the other BPNN models trained with tangent sigmoid transfer function.

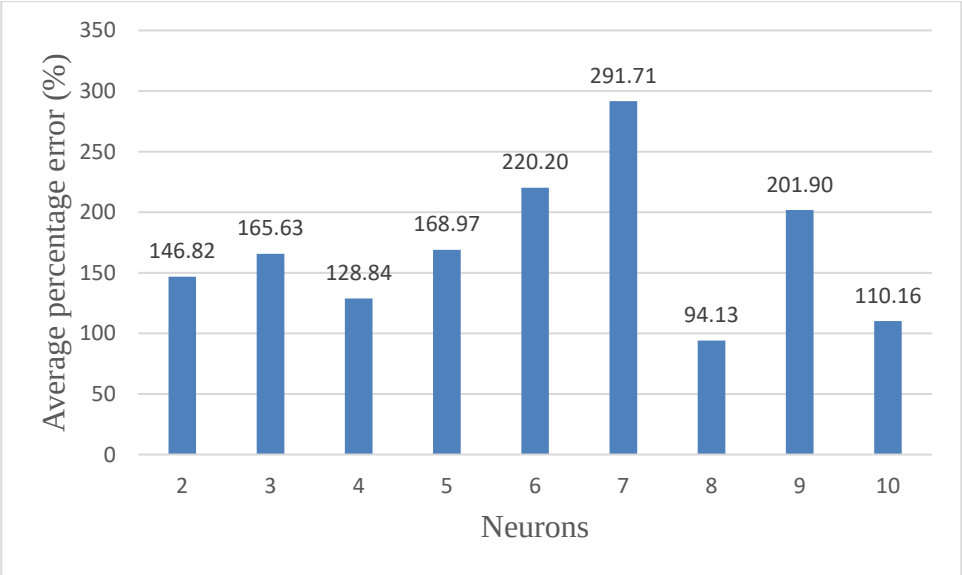

Figure 4.11: Average Percentage Error of BPNN Models with Normalized Dataset Trained with Tangent Sigmoid Transfer Function.

4.3.3 Selection of the Most Suitable BPNN Prediction Model

Three of the best BPNN models with the optimal performance from each type of BPNN models are tabulated in Table 4.8 and will be compared in the following context.

First of all, BPNN model with unnormalized dataset is first compared with the BPNN model with normalized dataset that have the same transfer function which is the log sigmoid transfer function. It is clearly shown that BPNN model with normalized dataset possess a better representation of MAE, MSE and RMSE values than the BPNN model with unnormalized dataset which in other words easy to study by researcher. Hence, BPNN model with unnormalized dataset is out of consideration.

Next, comparison is made between the two BPNN models with normalized dataset trained with log sigmoid and tangent sigmoid transfer functions. The BPNN model trained with log sigmoid transfer function with 4 hidden neurons establish higher R and R^2 values than the BPNN model trained with tangent sigmoid transfer function with 8 hidden neurons. Other than that, the BPNN model trained with log sigmoid function with 4 hidden neurons

generally has lower MAE, MSE, RMSE and average percentage error compared to the BPNN model trained with tangent sigmoid transfer function with 8 hidden neurons. Therefore, BPNN model trained with log sigmoid transfer function with 4 hidden neurons is selected to be the most optimal BPNN model to predict ammonia nitrogen concentration in Langat River.

Table 4.8: Comparison between the Three Optimal BPNN Models.

Type of BPNN models	Unnormalized	Normalized	Normalized
Transfer function	Log sigmoid	Log sigmoid	Tangent sigmoid
No. of neurons	5	4	8
R	0.229	0.465	0.115
R^2	0.052	0.217	0.013
MAE	935.686	0.153	0.267
MSE	1117479.922	0.030	0.115
RMSE	1057.109	0.172	0.339
Average Percentage error (%)	70.670	68.500	94.130

4.4 Comparison on the Effectiveness of Different AI Models

In this section, comparison will be made between the best BPNN, ANFIS and SVM models in term of their model performance. Table 4.9 outlines the comparison between the three AI models. Each of the AI models is the most effective model of its kind. It is clear that BPNN model has the highest R and R^2 values among the three AI models which recorded at 0.465 and 0.217 respectively. Besides, in terms of MAE, MSE and RMSE, BPNN model also tabulated the lowest values which is 0.153, 0.030 and 0.172 respectively. In comparison of the average percentage error, BPNN model still achieves the lowest value which is recorded at 68.50%. It can be concluded that BPNN model has the best performance among the three AI models and is most suitable to be utilized to predict ammonia nitrogen concentration in Langat River. However, ANFIS and SVM models are not suitable to be used as the prediction model to predict ammonia nitrogen concentration in Langat River.

Table 4.9: Comparison Between the Best BPNN, ANFIS and SVM Models.

Model	BPNN	ANFIS	SVM
R	0.465	0.225	0.088
R^2	0.217	0.051	0.008
MAE	0.153	0.232	0.203
MSE	0.030	0.085	0.058
RMSE	0.172	0.292	0.240
Average Percentage Error (%)	68.50	142	118

CHAPTER 5

CONCLUSION & RECOMMENDATION

5.1 Conclusion

Three mathematical prediction models namely BPNN, ANFIS and SVM were developed using AI technique to predict the ammonia nitrogen concentration in Langat River and their performance were evaluated by statistical analyses in this research. Five water quality parameters comprising of DS, T, TS, PO_4^{3-} and NO_3^- collected from Langat River, Dengkil were utilized as input variables by the three AI models for ammonia nitrogen concentration prediction. 77 datasets of the selected water quality parameters were implemented as inputs for the AI models training and testing processes. Before the training process, parameter tuning of each AI models were carried out. Each AI model was trained and tested with unnormalized and normalized dataset where ammonia nitrogen is the water quality parameter which being normalized. 8 types of membership functions were used to develop 16 ANFIS models and the number of membership function was fixed at 3. On the other hand, seven SVM models were developed using seven types of kernel functions. Furthermore, 9 BPNN models were developed by utilizing 2 to 10 hidden neurons with 1 hidden layer. Two different transfer functions namely log sigmoid and tangent sigmoid transfer functions were implemented in the training process of BPNN models with normalized dataset as an effort to try to achieve a better result. The results predicted by the AI models were intra-compared and inter-compared with one another and their performance were evaluated in terms of R, R^2, MAE, MSE, RMSE and average percentage error.

From the statistical analyses results, it was determined that AI models developed by normalized dataset had better performance than AI models developed with unnormalized dataset. Each AI models were intra-compared among their type. ANFIS model N11 with Pi-shaped curved membership function input, 3 membership function and constant membership function output achieved the best performance among the other ANFIS models and is selected to be the most optimal ANFIS model. On the other hand, SVM model

with dot kernel function shows better performance than the other SVM models as it is the only model which has R and R^2 values. Therefore, SVM model with dot kernel function is the optimum SVM model to predict ammonia nitrogen concentration. Lastly, BPNN model with trained with log sigmoid function with 4 hidden neurons outperform other BPNN models with lowest MAE, MSE, RMSE and average percentage error.

While inter-comparing the performance of BPNN, ANFIS and SVM models, BPNN model trained with log sigmoid function with 4 hidden neurons achieve the best performance with the highest R and R^2 value and the lowest MAE, MSE, RMSE and average percentage error. Hence, it is the most suitable AI model to predict ammonia nitrogen concentration in Langat River.

5.2 Recommendations

The data used in this research were obtained from Department of Irrigation and Drainage (DID) of Malaysia. One of the problems is that the 77 dataset recorded by DID is insufficient for satisfactory prediction by using the AI models. Another problem arise is there are some missing data in the water quality parameters data provided by DID hence causing limitation in choosing suitable water quality parameters as input variables for AI models and this will surely affect the prediction performance of AI models. Therefore, it is suggested that DID should record more data in between shorter interval of time from each station along Langat River in order for future research on prediction of water quality to be carried out in Langat River.

SVM can deal better with classification problems but not as good in regression problems. Hence, Support Vector Regression (SVR) is suggested to solve regression problem in this research. SVR is an effective mathematical predictive tool for real-value function prediction. SVR has an important advantage which is its computational complexity is independent of the dimensionality of input space. Aside from that, SVR superb generalization ability leads to its high prediction accuracy (Awad and Khanna, 2015).

REFERENCES

Abba, S.I., Hadia, S.J. and Abdullahi, J., 2017. River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques. *Procedia Computer Science,* Volume 120, pp. 75-82.

Abba, S.I., Nourani, V. and Elkiran, G., 2019. Multi-parametric modeling of water treatment plant using AI-based non-linear ensemble. *Journal of Water Supply: Research and Technology- AQUA.*

Abbas M.A. and Suhad M.A., 2017. Modelling the strength of lightweight foamed concrete using support vector machine (SVM). *Case Studies in Construction Materials,* Volume 6, pp. 8-15.

Abidin, R.Z., Sulaiman, M.S. and Yusoff, N., 2017. Erosion risk assessment: A case study of the Langat River bank in Malaysia. *International Soil and Water Conservation Research,* Volume 5, pp. 26-35.

Abirhire, O., Davies, J.M., Guo, X. and Hudson, J., 2020. Understanding the factors associated with long-term reconstructed turbidity in Lake Diefenbaker from Landsat-imagery. *Science of the Total Environment,* Volume 724, p. 138222.

Aksu, G., Güzeller, C. O. & Eser, M. T. , 2019. The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model. *International Journal of Assessment Tools in Education,* 6(2), pp. 170-192.

Alizadeh, M.J., Kavianpour, M.R., Danesh, M., Adolf, J., Shamshirband, S. and Chau, K.W., 2018. Effect of river flow on the quality of estuarine and coastal waters using machine learning models. *Engineering Applications of Computational Fluid Mechanics,* Volume 12, pp. 810-823.

Altunkaynak, A. and Wang, K.H., 2011. A comparative study of hydrodynamic model and expert system related models for prediction of total suspended solids concentrations in Apalachicola Bay. *Journal of Hydrology,* Volume 400, pp. 353-363.

Amos T.K. and Xie Y., 2012. Numerical models for predicting the fate of ammonia-nitrogen under bacterial technology. *Journal of Applied Sciences in Environmental Sanitation,* Volume 7, pp. 183-192.

Andrea, L.S., Keith, E.S. and Chan, K. S., 2009. Time-series modeling of reservoir effects on river nitrate concentrations. *Advances in Water Resources,* Volume 32, pp. 1197-1205.

Anon., 2019. *Ammonia pollution in Sungai Sayong disrupts water supply to 17,000 households in Malaysia's Kulai,* Johor Bahru: Channel News Asia.

Anon., 2019. *Ammonia pollution left 18,000 homes in Melaka without water supply,* Melaka: Malay mail.

Awad, M. and Khanna, R., 2015. Support vector regression. In *Efficient learning machines* (pp. 67-80). Apress, Berkeley, CA.

Balasubramaniana, S.V., Pahlevana, N., Smitha, B., Binding, C., Schalles, J., Loisel, H., Gurlin, D., Greb, S., Alikas, K., Randla, M., Bunkei, M., Moses, W., Nguyễn, H., Lehmann, M.K., O'Donnell, D., Ondrusek, M., Han, T.H., Fichotr, C.G. and Moore,T., 2020. Robust algorithm for estimating total suspended solids (TSS) in inland and nearshore coastal waters. *Remote Sensing of Environment,* Volume 246, p. 111768.

Basarir, H., Tutluoglu, L. and Karpuz, C., 2014. Penetration rate prediction for diamond bit drilling by adaptive neuro-fuzzy inference system and multiple regressions. *Engineering Geology,* Volume 173, pp. 1-9.

Berbić, J., Ocvirk, E., Carević, D. and Lončar, G., 2017. Application of neural networks and support vector machine for significant wave height prediction. *Oceanologia,* Volume 59, pp. 331-349.

Brezonik, P.L. and Engstrom, D.R., 1998. Modern and historic accumulation rates of phosphorus in Lake Okeechobee, Florida. *Journal of Paleolimnology,* Volume 20, pp. 31-46.

Cao,W.J., Huan, J., Liu, C., Qin, Y.L. and Wu, F., 2019. A combined model of dissolved oxygen prediction in the pond based on multiple-factor analysis and multi-scale feature extraction. *Aquacultural Engineering,* Volume 84, pp. 50-59.

Chang, F.J. and Chang, Y.T., 2006. Adaptive neuro-fuzzy inference system for prediction of water level in reservoir. *Advances in Water Resources,* Volume 29, pp. 1-10.

Chang, F.J. and Lai, H.C., 2014. Adaptive neuro-fuzzy inference system for the prediction of monthly shoreline changes in north eastern Taiwan. *Ocean Engineering,* Volume 84, pp. 145-156.

Chang, F.J.,Tsai, Y.H., Chen, P.A., Coynel, A. and Vachaud, G., 2015. Modeling water quality in an urban river using hydrological factors e Data driven approaches. *Journal of Environmental Management,* Volume 151, pp. 87-96.

Chen, C. S., Chen, B. P. T., Chou, F. N. F. and Yang, C. C., 2010. Development and application of a decision group Back-Propagation Neural Network for flood forecasting. *Journal of Hydrology,* Volume 385, pp. 173-182.

Chen, S., Han, L., Chen, X., Li, D., Sun, L. and Li, Y., 2015. Estimating wide range Total Suspended Solids concentrations from MODIS 250-m imageries: An improved method. *ISPRS Journal of Photogrammetry and Remote Sensing,* Volume 99, pp. 58-69.

Chin, R.J., Lai, S.H., Ibrahim, S., Jaafar, W.Z.W. and Elshafie, A.H.K.A., 2018. New approach to mimic rheological actual shear rate under wall slip condition. *Engineering with Computers,* Volume 35, pp. 1409-1418.

Constantin, S., Doxaran, D. and Constantine, Ș., 2016. Estimation of water turbidity and analysis of its spatio-temporal variability in the Danube River plume (BlackSea) using MODIS satellite data. *Continental Shelf Research,* Volume 112, pp. 14-30.

Cox, B.A., 2003. A review of dissolved oxygen modelling techniques for lowland river. *the Science of the Total Environment,* Volume 314-316, pp. 303-334.

Csábrágia, A., Molnára, S., Tanosa, P. and Kovács, J., 2017. Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube. *Elsevier,* Volume 100, pp. 63-72.

Elkiran, G., Nourani, V. and Abba, S.I., 2019. Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. *Journal of Hydrology,* Volume 577, p. 123962.

Engstrom, Patrick, L.B. & Daniel, R., 1998. Modern and historic accumulation rates of phosphorus in Lake Okeechobee, Florida. *Journal of Paleolimnology,* Volume 20, pp. 31-46.

Firat M. and Gungor, M., 2007. River flow estimation using adaptive neuro fuzzy inference system. *Mathematics and Computers in Simulation,* Volume 75, pp. 87-96.

Fulazzaky, M.A., Seong, T.W. and Masirin, M.I.M., 2009. Assessment of Water Quality Status for the Selangor River in Malaysia. *Water, Air, and Soil Pollution ,* Volume 205, pp. 63-77.

Gaona, C.A.P., Serra, F.D.P., Furtado, P.S., Poersch, L.H. and Wasielesky Jr. W., 2016. Effect of different total suspended solids concentrations on the growth performance of Litopenaeus vannamei in a BFT syste. *Aquacultural Engineering,* Volume 72-73, pp. 65-69.

Ghose, D.K., Panda, S.S. and Swain, P.C., 2010. Prediction of water table depth in western region, Orissa using BPNN and RBFN neural networks. *Journal of Hydrology,* Volume 394, pp. 296-304.

Hashihama, F., Kanda, J., Tauchi, A., Kodama, T., Saito, H. and Furuya, K., 2015. Liquid waveguide spectrophotometric measurement of nanomolar ammonium in seawater based on the indophenol reaction with o-phenylphenol (OPP). *Elsevier,* Volume 143, pp. 374-380.

Heddam, S. and Kisi, O., 2018. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *Journal of Hydrology,* Volume 559, pp. 499-509.

Huang, D.M. He, S.Q., He, X.H. and Zhu, X., 2017. Prediction of wind loads on high-rise building using a BP neural network combined with POD. *Journal of Wind Engineering & Industrial Aerodynamics,* Volume 170, pp. 1-17.

Hutchins, M.G., Harding, G., Jarvie, H.P., Marsh, T.J., Bowes, M.J. and Loewenthal, M., 2020. Intense summer floods may induce prolonged increases in benthic respiration rates of more than one year leading to low river dissolved oxygen. *Journal of Hydrology X,* Volume 8, p. 100056.

Jeong, H., Choi, J.Y., Lee, J., Lim, J. and Ra K., 2020. Heavy metal pollution by road-deposited sediments and its contribution to total suspended solids in rainfall runoff from intensive industrial areas. *Environmental Pollution,* Volume 265, p. 115028.

Jin, T., Cai, S.B.,Jiang, D.X. and Liu, J., 2019. A data-driven model for real-time water quality prediction and early warning by an integration method. *Environmental Science and Pollution Research,* Volume 26, pp. 30374-30385.

Juahir, H., Zain, S.M., Yusoff, M.K., Hanidza, T.I.T., Armi, A.S.M., Toriman, M.E. and Mokhtar, M., 2010. Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. *Environmental Monitoring and Assessment,* Volume 173, pp. 625-641.

Kabo-bah, A.T. and Xie, Y., 2012. Numerical models for predicting the fate of ammonia-nitrogen under bacterial technology. *Journal of Applied Sciences in Environmental Sanitation,* Volume 7, pp. 183-192.

Kannel, P.J., Lee, S.H., Lee, Y.S., Kanel, S.R. and Khan, S.P., 2007. Application of Water Quality Indices and Dissolved Oxygen as Indicators for River Water Classification and Urban Impact Assessment. *Environmental Monitoring and Assessment ,* Volume 132, pp. 93-110.

Kermani, M.Z., Beheshti, A.A., Ashtiani, B.A. and Yazdi, S.R.S., 2009. Estimation of current-induced scour depth around pile groups using neural network and adaptive neuro-fuzzy inference system. *Applied Soft Computing,* Volume 9, pp. 746-755.

Khan, N., Sachindra, D.A., Shahid, S., Ahmed, K., Shiru, M.S. and Nawaz, N., 2020. Prediction of droughts over Pakistan using machine learning algorithms. *Advances in Water Resources,* Volume 139, p. 103562.

Kim, S.W. Chung and J.H., 2005. Development of water quality models for supporting NH3-N control in a dam regulated river. *Water Science & Technology,* Volume 52, pp. 83-90.

Kucuk, K., Aksoy, C.O., Basarir, H., Onargan, T., Genis, M. and Ozacar, V., 2011. Prediction of the performance of impact hammer by adaptive neuro-fuzzy inference system modelling. *Tunnelling and Underground Space Technology,* Volume 26, pp. 38-45.

Kunwar P. Singh, Ankita Basant, Amrita Malik, Gunja Jain, 2009. Artificial neural network modeling of the river water quality—A case study. *Elsevier*, pp. 888-895.

Li, D. L., Xu, X. B., Li, Z., Wang, T. and Wang, C., 2020. Detection methods of ammonia nitrogen in water: A review. *Trends in Analytical Chemistry*, Volume 127, p. 115890.

Lin, K., Zhu, Y., Zhang, Y. and Lin, H., 2019. Determination of ammonia nitrogen in natural waters Recent advances and applications. *Trends in Environmental Analytical Chemistry*, Volume 24.

Lin, K., Zhu, Y., Zhang, Y.B., Lin, H., 2019. Determination of ammonia nitrogen in natural waters: Recent advances and applications. *Trends in Environmental Analytical Chemistry*, Volume 24.

Liu, Q.L., Sun, P.X., Fu, X.Y., Zhang, J., Yang, H., Gao, H.G. and Li, Y.Z., 2020. Comparative analysis of BP neural network and RBF neural network in seismic performance evaluation of pier columns. *Mechanical Systems and Signal Processing*, Volume 141, p. 106707.

Mamun, A.A., Idris, A., Sulaiman, W.N.A. and Muiby, S.A., 2007. A revised water quality index proposed for the assessment of surface water quality in malaysia. *Research Gate*, Volume 26, pp. 523-529.

Marchant, R., Reading, D., Ridd, J., Campbell, S. and Ridd, P., 2015. A drifter for measuring water turbidity in rivers and coastal oceans. *Marine Pollution Bulletin*, Volume 91, pp. 102-106.

Massah, J. and Vakilian, K.A., 2019. An intelligent portable biosensor for fast and accurate nitrate determination using cyclic voltammetry. *Biosystems Engineering*, Volume 77, pp. 49-58.

Masserini, R.T., Fanning, K.A., Hendrix, S.A. and Kleiman, B.M., 2017. A Coastal Surface Seawater Analyzer for nitrogenous nutrient mapping. *Continental Shelf Research*, Volume 150, pp. 48-56.

Moeeni, H., & Bonakdari, H., 2017. Impact of Normalization and Input on ARMAX-ANN Model Performance in Suspended Sediment Load Prediction. *Water Resources Management*, 32(3), pp. 845-863.

Mohamed, I., Othman, F., Ibrahim, A.I.N., Alaa-Eldin, M.E. and Yunus, R.M., 2014. Assessment of water quality parameters using multivariate analysis for Klang River basin, Malaysia. *Environmental Monitoring and Assessment*, Volume 187, p. 4182.

Mokhtar, Mazlin, Hafizan Juahir · Sharifuddin M. Zain · Mohd Kamil Yusoff ·T. I. Tengku Hanidza · A. S. Mohd Armi and Mohd Ekhwan Toriman , 2010. Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. *SpringerLink*, Volume 173, pp. 625-641.

Muthukrishnan, S., Lewis, G.P. and Andersen, C.B., 2007. Relations among land cover, vegetation index, and nitrate concentrations in streams of the Enoree River Basin, piedmont region of South Carolina, USA. *Developments in Environmental Science,* Volume 5, pp. 1-761.

Namu, P.N., Raude, j.M., Mutua, B.M. and Wambua, R.M., 2017. Prediction of Water Turbidity using Artificial Neural Networks: A Case Study of Kiriku-Kiende Settling Basin in Embu County, Kenya. *American Journal of Water Resources,* Volume 5, pp. 54-62.

Ni, Y.Q. and Li, M., 2016. Wind pressure data reconstruction using neural network techniques: A comparison between BPNN and GRNN. *Measurement,* Volume 88, pp. 468-476.

Nieto, P.J.G., Gonzalo, E.G., Fernández, J.R.A. and Muñiz, C.D., 2014. Hybrid PSO–SVM-based method for long-term forecasting of turbidity in the Nalón river basin: A case study in Northern Spain. *Ecological Engineering,* Volume 73, pp. 192-200.

Olyaie, E., Zare Abyaneh, H., & Danandeh Mehr, A., 2017. A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in Delaware River. *Geoscience Frontier,* Volume 8, pp. 517-527.

Othman, F., M. E., A. E., and Mohamed, I., 2012. Trend analysis of a tropical urban river water quality in Malaysia. *Journal of Environmental Monitoring,* Volume 14, pp. 3164-3173.

Ouedraogo, I., Defourny, P. and Vanclooster, M., 2019. Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. *Hydrogeology Journal,* Volume 27, pp. 1081-1098.

Ouyang, Y., 2005. Evaluation of river water quality monitoring stations by principal component analysis. *Water Research,* Volume 39, pp. 2621-2635.

Palansamy, Y. C. E. a. T. B., 2019. *Two incidents of pollution in Pasir Gudang affected thousands: Here's what we know so far,* Kuala Lumpur: Malay Mail.

Patel, N., Ruparelia, J. and Barve, J., 2020. Prediction of total suspended solids present in effluent of primary clarifier of industrial common effluent treatment plant: Mechanistic and fuzzy approach. *Journal of Water Process Engineering,* Volume 34, p. 101146.

Perumal, E., 2016. *Numerous contamination incidents along Sg Semenyih,* Kajang: The Star.

Pesce, S., 2000. Use of water quality indices to verify the impact of coâ rdoba city (argentina) on suquia river. *Elsevier,* 34(11), pp. 2915-2926.

Raghavendra, S.N. and Deka, P.C., 2014. Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing,* Volume 19, pp. 372-386.

Rajaee, T., Khani, S. & Ravansalar, M., 2020. Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review. *Chemometrics and Intelligent Laboratory Systems,* Volume 200.

Rankovic, V., Radulovic, J., Radojevic, I., Ostojic, A. and Comic, L., 2010. Neural network modeling of dissolved oxygen in the Gruˇza reservoir, Serbia. *Ecological Modelling,* Volume 221, pp. 1239-1244.

Sahana, M., Rehman, S., Sajjad, H. and Hong, H., 2020. Exploring effectiveness of frequency ratio and support vector machine models in storm surge flood susceptibility assessment: A study of Sundarban Biosphere Reserve, India. *Catena,* Volume 189, p. 104450.

Schoch, A.L., Schilling, K.E. and Chan, K.S., 2009. Time-series modeling of reservoir effects on river nitrate concentrations. *Advances in Water Resources,* Volume 32, pp. 1197-1205.

Schveitzer, R., Arantes, R., Costódio, P.F.S., Espírito, C.M., Santo, Arana, L.V., Seiffert, W.Q. and Andreatta, E.R., 2013. Effect of different biofloc levels on microbial activity, water quality and performance of Litopenaeus vannamei in a tank system operated with no water exchange. *Elsevier,* Volume 56, pp. 59-70.

Stamenkovic´, L.J., Kurilic´, S.M. and Ulnikovic, V.P., 2020. Prediction of nitrate concentration in Danube river water by using artificial neural networks. *Water Science & Technology: Water Supply.*

Suen, J.P., Eheart, J.W. and ASCE, M., 2003. Evaluation of Neural Networks for Modeling Nitrate Concentrations in Rivers. *Journal of Water Resources Planning and Management,* Volume 129, pp. 505-510.

Sulaiman, R., Ismail, Z., Othman, S.Z., Ramli, A.H. and Shirazi, S.M., 2014. A comparative study of trends of nitrate, chloride and phosphate concentration levels in selected urban rivers. *Measurement,* Volume 55, pp. 74-81.

Teixeira, L.C., Mariani, P.P., Pedrollo, O.C., Castro, N.M.R. and Sari, V., 2020. Artificial Neural Network and Fuzzy Inference System Models for Forecasting Suspended Sediment and Turbidity in Basins at Different Scales. *Water Resources management,* Volume 34, pp. 3709-3723.

Tiyasha, Tung, T.M. and Yaseen Z.M., 2020. A survey on river water quality modelling using artificial intelligence models: 2000–2020. *Journal of Hydrology,* Volume 585, p. 124670.

Verma, A.,Wei, X.P. and Kusiak, A., 2013. Predicting the total suspended solids in wastewater :A data-mining approach. *Engineering ApplicationsofArtificialIntelligence,* Volume 26, pp. 1366-1372.

VICENTE, H., COUTO, C., MACHADO, J., ABELHA, A. and NEVES, A., 2012. Prediction of water quality parameters in a reservoir using artificial neural networks. *Journal of Design & Nature and Ecodynamics,* Volume 7, pp. 310-319.

Vijayalaksmia, D.P. and Babu, J.K.S., 2015. Water Supply System Demand Forecasting Using Adaptive Neuro-Fuzzy Inference System. *Aquatic Procedia,* Volume 4, pp. 950-956.

Wu, Z.S., Wang, X.L., Chen,Y.W., Cai, Y.J. and Deng, J.C., 2017. Assessing river water quality using water quality index in Lake Taihu Basin, China. *Science of the Total Environment,* Volume 612, pp. 914-922.

Yousif, M., Al-Suhaili, R.H. and Yousif, W.M.S.K., 2008. Prediction of turbidity in tigris river using artificial neural networks. *Journal of Engineering,* Volume 14.

Yu, H.H., Yang, L., Li, D.L. and Chen, Y.Y., 2020. A hybrid intelligent soft computing method for ammonia nitrogen prediction in aquaculture. *INFORMATION PROCESSING IN AGRICULTURE.*

Zainudin, Z. and Mamun A.A. , 2013. Sustainable river water quality management in malaysia. *IIUM Engineering Journal,* 14(1).

Zainudin, Z., 2017. *Protecting our drinking water,* Kluang: New Straits Times.

Zainudin, Z., 2010. Benchmarking River Water Quality in Malaysia. *Research Gate.*

Zare A.H., Bayat V.M., and Daneshkare A.P., 2011. Forecasting nitrate concentration in groundwater using artificial neural network and linear regression models. *International Agrophysics,* Volume 25, pp. 187-192.

Zhang, L., Xu, E.G., Li, Y., Liu, H.L., Vidal-Dorsch, D.E. and Giesy, J.P., 2018. Ecological risks posed by ammonia nitrogen (AN) and un-ionized ammonia (NH3) in seven major river systems of China. *Chemosphere,* Volume 202, pp. 136-144.

Zhang, L., Xu, E.G., Li, Y.B., Liu, H.L., Vidal-Dorsch, D.E. and Giesy, J.P., 2018. Ecological risks posed by ammonia nitrogen (AN) and un-ionized ammonia (NH3) in seven major river systems of China. *Chemosphere,* Volume 202, pp. 136-144.

Zhang, N., Ma, Y. and Zhang, Q., 2018. Prediction of sea ice evolution in Liaodong Bay based on a back-propagation neural network model. *Cold Regions Science and Technology,* Volume 145, pp. 65-75.

Zheng, G., Zhang, W.B., Zhang, W.G., Zhou, H.Z. and Yang, P.b., 2019. Neural network and support vector machine models for the prediction of the liquefaction-induced uplift displacement of tunnels. *Underground Space.*

Zhou, Q., Wang, F.L. and Zhu, F., 2016. Estimation of compressive strength of hollow concrete masonry prisms using artificial neural networks and adaptive neuro-fuzzy inference systems. *Construction and Building Materials,* Volume 125, pp. 417-426.

Zhou, Q., Zhu, F., Yang, X., Wang, F.L., Chi, B. and Zhang, Z.M., 2017. Shear capacity estimation of fully grouted reinforced concrete masonry walls using neural network and adaptive neuro-fuzzy inference system models. *Construction and Building Materials,* Volume 153, pp. 937-947.

Zhou, Y.W., Zheng, S.B., Huang, Z.Y., Sui, L.L. and Chen, Y., 2020. Explicit neural network model for predicting FRP-concrete interfacial bond strength based on a large database. *Composite Structures,* Volume 240, p. 111998.