DEVELOPMENT OF WATER QUALITY INDEX PREDICTION MODEL FOR PENANG RIVERS USING ARTIFICIAL NEURAL NETWORK

ELEENA YASMEEN BINTI MOHD HAMDAN

UNIVERSITI SAINS MALAYSIA

2021

DEVELOPMENT OF WATER QUALITY INDEX PREDICTION MODEL FOR PENANG RIVERS USING ARTIFICIAL NEURAL NETWORK

By

ELEENA YASMEEN BINTI MOHD HAMDAN

Project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Chemical Engineering

JULY 2021

ACKNOWLEDGEMENT

First and foremost, all praises and thanks to Allah, the Almighty, for His showers and blessings throughout my journey to complete this thesis successfully.

I would like to express my deep and sincere gratitude to my final year project supervisor, Dr Norazwan Md Nor, for his guidance in stimulating suggestions and encouragement especially in writing this report. His vision, sincerity and motivation have deeply inspired me which helped me to coordinate my project. He assisted and taught me the methodology to carry out this project and to present the research work as clearly as possible.

Next, I submit my heartiest appreciation to all respected lecturers in the School of Chemical Engineering who indirectly help me in providing guides to prepare this report. Highly thanks and gratitude to Prof. Dr Mohd Roslee Bin Osman for always giving the best in coordinating this final year project report as the coordinator.

Equally important, I am extremely grateful to my parents, Mr. Mohd Hamdan and Mrs. Siti Noor Haziah for their prayers, love, and sacrifices for educating and preparing me for my future. Also, I would like to say thank you to my brothers and sisters for their continuing support and idea for me to complete this project.

Last but not least, I humbly extend my thanks to all my friends and concerned person who cooperated with me in general in completing this project successfully.

Eleena Yasmeen

June 2021

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

Abbreviation	Description
AI	Artificial Intelligence
BOD	Biochemical Oxygen Demand
BR	Bayesian Regularization
DOE	Department of Environment
FANN	Feedforward Artificial Neural Network
kNN	k-Nearest Neighbor
LM	Levenberg-Marquardt
MPCA	Multiway Principal Component Analysis
MSE	Mean Squared Error
PCA	Principal Component Analysis
SCG	Scale Conjugate Gradient
SDG	Sustainable Development Goal
SI	Sub-Index
SVM	Support Vector Machine
WQC	Water Quality Classification
WQI	Water Quality Index

PENGEMBANGAN MODEL RAMALAN INDEKS KUALITI AIR UNTUK SUNGAI DI PULAU PINANG MENGGUNAKAN RANGKAIAN NEURAL BUATAN

ABSTRAK

Sebagai akibat daripada aktiviti perindustrian dan pembandaran, industri berkemungkinan membuang limbah tanpa dirawat ke sungai, tasik dan lautan. Pulau Pinang adalah sebuah bandar kosmopolitan dan secara tidak langsung menimbulkan perhatian terhadap masalah kualiti air. Merujuk kepada masalah yang dihadapi, formulasi indeks kualiti air (IKA) yang dikembangkan oleh Jabatan Alam Sekitar (JAS) dapat membantu pihak berkaitan, kerana ia sangat berguna untuk menilai keadaan kualiti air sungai. Oleh itu, dalam projek ini, model ramalan IKA untuk sungai di Pulau Pinang telah dikembangkan dengan mengaplikasikan rangkaian neural buatan (ANN) di MATLAB. Terdapat 30 parameter kualiti air; dengan sejumlah 1000 sampel diperoleh daripada JAS untuk pelaksanaan model. Untuk mengurangkan lelebih 28 parameter input, analisis komponen utama (PCA) telah diperkenalkan. Hasilnya, prestasi model yang dicadangkan disahkan menggunakan data yang tidak dapat dilihat. Untuk mencapai objektif tersebut, terdapat tiga fasa utama yang diperkenalkan dalam kerangka kerja pengembangan rangkaian; pertama, penentuan pengekstrakan ciri menggunakan analisis komponen utama berbilang cara (MPCA), kedua, pengembangan model ANN serta pemilihan seni binanya untuk analisis BOD dan COD dan ketiga, pemilihan seni bina ANN bagi model ramalan IKA. Bagi pelaksanaan MPCA dalam pengekstrakan ciri untuk BOD dan COD, hanya 4 input yang diperlukan untuk menjelaskan sekurang-kurangnya 99.999% kebolehubahan untuk kedua-dua analisis

BOD dan COD. Secara keseluruhan, untuk analisis BOD, algoritma BR dengan latihan 60% dan 12 nod tersembunyi memberikan R = 0.7825 manakala untuk analisis COD, algoritma BR dengan latihan 70% dan 10 nod tersembunyi menunjukkan R = 0.6716. Dalam proses pembangunan rangkaian, empat senario proses iaitu teknik latihan-pengesahan-ujian untuk meminimumkan model berlebihan dengan 15 nod tersembunyi berdasarkan tiga algoritma terbina iaitu Levenberg-Marquart (LM), Bayesian Regularization (BR) dan Scaled Conjugate Gradient (SCG) telah dihasilkan. Seterusnya, algoritma BR dipilih untuk analisis BOD dan COD kerana ia dapat menghasilkan rangkaian terbaik yang dapat membuat generalisasi dengan cukup baik dengan meminimumkan gabungan ralat. Selepas fasa pemilihan seni bina ANN, tiga sub-senario diandaikan dari bilangan neuron tersembunyi; 15, 30 dan 45 nod telah diperkenalkan ke dalam senario latihan-pengesahan-ujian seperti yang sebelumnya dengan algoritma BR yang dipilih. Secara keseluruhan, algoritma BR dengan latihan 60% dan 30 nod tersembunyi berjaya dikembangkan untuk analisis BOD, sementara itu, latihan 70% untuk analisis COD dengan nilai regresi masing-masing 0.9978 dan 0.9976. Sebelum pengembangan model ramalan WQI berasaskan ANN, algoritma BR dipilih dengan seni bina dua, tiga, empat, lima dan enam-neuron untuk latihan 60% dan 70%. Hasilnya, latihan 60% dengan lima node tersembunyi menunjukkan prestasi terbaik dengan nilai R= 0.827 dan MSE=52.283. Model berasaskan ANN dapat berfungsi sebagai alat yang berguna dalam menganggarkan WQI sungai.

DEVELOPMENT OF WATER QUALITY INDEX PREDICTION MODEL FOR PENANG RIVERS USING ARTIFICIAL NEURAL NETORK

ABSTRACT

As a consequence of industrialization and urbanization, industries might discharge effluent locally without treatment into rivers, lakes and ocean. Pulau Pinang is a very cosmopolitan city that indirectly raise concern on the issue of water quality problems. Considering the issue encountered, the water quality index (WQI) formulation developed by the Department of the Environment (DOE) might be able to assist the water authorities in some way, as it is useful to assess the river water quality condition. Thereupon, in this study, the WQI prediction model for Penang rivers has been developed by using Artificial Neural Network (ANN) architecture in MATLAB. There are 30 water quality parameters with a total of 1000 samples were obtained from DOE for further model implementation. In order to reduce the redundancy of the input parameters, principal component analysis (PCA) has been introduced. Consequently, the performance of the proposed model was validated using unseen data. To achieve those objectives, there were three main phases implemented in the network development framework; first, the determination of feature extraction using multiway principal component analysis (MPCA), second, the ANN model development and architecture selection for BOD and COD analysis, and third, the ANN architecture selection for WQI prediction model. As for the implementation of MPCA in feature extraction for BOD and COD, there were only 4 inputs required to explain at least 99.999% variability for both analyses. Altogether, for BOD, the BR algorithm with 60% training and 12 hidden nodes gives R=0.7825 whereas for COD, the BR algorithm with

70% training and 10 hidden nodes gives R=0.6716. For ANN prediction model development, four scenarios of the train-validate-test process to minimize model overfitting with 15 hidden nodes based on three built-in algorithms namely Levenberg-Marquart (LM), Bayesian Regularization (BR) and Scaled Conjugate Gradient (SCG) were created. As a result, the BR algorithm was chosen for both BOD and COD analysis as it can generate a good network that generalizes well enough by minimizing the combination of errors and weights. Following the architecture selection phase, three sub-scenarios assumed from the number of hidden neurons; 15, 30 and 45 nodes were introduced into previous train-validate-test scenarios with chosen BR algorithm. On the whole, BR algorithm with 60% training and 30 hidden nodes were successfully developed for BOD analysis, meanwhile, 70% training for COD analysis with the regression values of 0.9978 and 0.9976 respectively. Prior to the development of ANN-based WQI prediction model, the BR algorithm was chosen with two-, three-, four-, five- and six-neuron architectures for 60% and 70% training. As a result, 60% training with five hidden nodes demonstrated the best performance with R- value of 0.827 and MSE value of 52.283. The ANN-based models could serve as reliable and useful tools in estimating the WQI of the river.

CHAPTER 1
INTRODUCTION

Chapter 1 provides an overview of the project area regarding the Water Quality Index (WQI) for Penang rivers. On the whole, this chapter further discusses the research background of WQI, the area of study, the problem statement and the objectives of this final year project.

1.1 Research Background

According to the Department of Statistics Malaysia (2020), despite the closure of our national borders during the Movement Control Order amid the outbreak of the COVID-19 pandemic globally, Malaysia's population in 2020 is estimated at 32.7 million as compared to 32.5 million in 2019 with an annual growth rate of 0.4 percent. Due to that exponential urbanization correspondingly has caused vast quantities of generation of wastes, including residential, agricultural, commercial and transportation waste, which eventually ends up in the lakes and rivers (Huang et al., 2015).

Hence, a significant number of rivers seem to be so infected that somehow resulting in difficulties to rehabilitate the rivers. Henceforth, the water authorities need to overcome a major hurdle to has access to clean and safe water sources to ensure every citizen receives a healthy water supply. Concerning the hardship of the water authorities to manage the water sources, WQI applied by the Department of Environment (DOE) can somehow help them as it is important to assess the river water quality level.

The WQI provides a framework for river risk assessment pertaining to the classification of industrial effluents and the classifying of groups of promising alternatives as supported under the Interim National Water Quality Standard (INWQS) for Malaysia (Nurul Ruhayu et al., 2015). There are few parameters taken into

calculation consideration in the analysis of WQI classification which are biochemical oxygen demand (BOD), chemical oxygen demand (COD), dissolved oxygen (DO), total suspended solid (TSS), ammoniacal nitrogen (NH_3-N) and pH value. The DOE water quality index classification is categorized into five types of classes where each class refers to its distinctive use as listed in **Table 1.1** below.

Table 1.1 Water Classes and Uses

WQI	Class	Uses
>92.7	Class I	Conservation of natural environment. Water Supply I - Practically no treatment necessary. Fishery I - Very sensitive aquatic species.
76.5-92.7	Class IIA	Water Supply II - Conventional treatment. Fishery II - Sensitive aquatic species.
	Class IIB	Recreational use body contact.
51.9-76.5	Class III	Water Supply III - Extensive treatment required. Fishery III – Common, of economic value and tolerant species; livestock drinking.
31.0-51.9	Class IV	Irrigation
<31.0	Class V	None of the above

Corresponding to the above table, we may deduce that a river can be segregated into various categories that denote the beneficial uses to which it can be applied. The classes are based on the permissible limits set by the government for the appropriate polluting standards or conditions. Depending on the water quality of different areas of a river, the river can be zoned according to its suitability and purpose (Mamun et al., 2007). The determination of water quality is important, because water is frequently being used in our daily life. The water quality measurement is aimed to make sure that the content of the water supply is safe to use accordingly.

1.2 Study Area

Pulau Pinang or Penang is a Malaysian state situated on the northwest coast of Peninsular Malaysia. It consists of 2 parts; the island and the mainland as shown in **Figure 1.1**. Every year, thousands of visitors come to Penang to experience the unique cultural heritage and scenery. Not only that, it is also a very cosmopolitan city, perhaps the second busiest in the country after Kuala Lumpur. Hence, things like this could indirectly raise concern on the issue of water quality problems. With this in mind, the water quality data obtained from DOE for Penang rivers will be evaluated in this study.

Figure 1.1 Maps of Pulau Pinang

1.3 Problem Statement

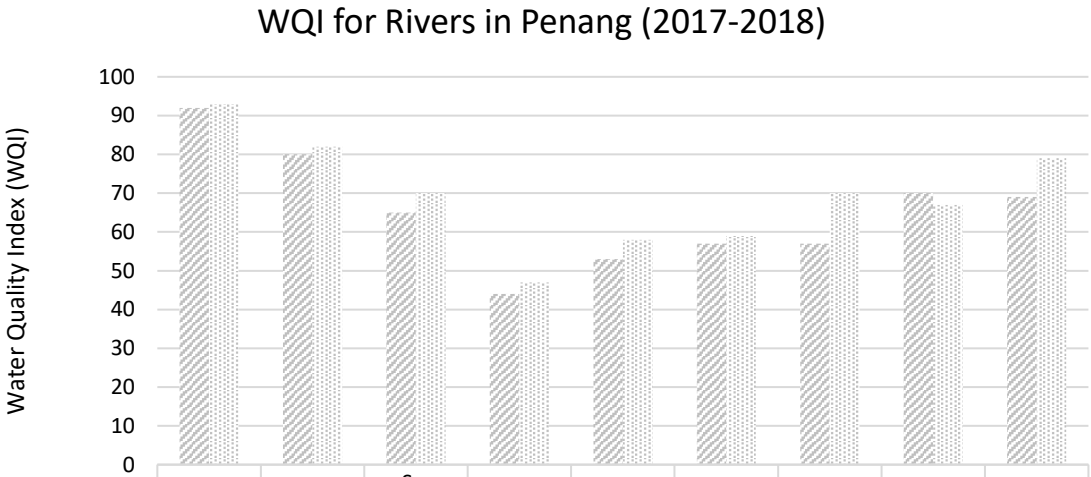
The rivers in Penang were chosen as the analysis of water quality due to the inevitable implication and contribution of the state to the community. In the industrial segment in Penang, the demand for factories and land has been very substantial, therefore they expected by 2020, this segment to expand at a steady pace of 10% growth compared to in 2019. Moreover, the Malaysian Investment Development Authority (MIDA) claimed that Penang has recorded RM13.3 billion in approved manufacturing investment inflows (Hannah, 2020). The improvement in this industrial development calls for the emission of various untreated waste that leads to the water quality of the river has gradually deteriorated.

Relying on the mentioned six parameters, the WQI formulation approved by the DOE therefore can be evaluated. Most researches have focused on the expression of water quality in a realistic, understandable, and comparative way. And as such, all these vital parameters for determining the water quality are grouped in one expression to alleviate the work from the complexity of evaluating several distinctive parameters. WQI is a single number that can be easily measured and used to characterize the performance of water bodies generally (Tunc Dede et al., 2013).

In like manner, referring to Jun et al. (2019), the previous method in computing WQI is always associated with errors due to the protracted analysis of the water quality parameters with the time involved in gathering and analysing water samples. Not only that, the cost is also very concerning as the experimental testing is very high. Therefore, the development of the WQI prediction model somehow helps as a step towards better management of the rivers' water quality.

Equally important, a study conducted by the DOE in 2018 found that from the 638 controlled rivers across the country, 357 (56%) indicated the standard of clean

water, 231 (36%) were marginally polluted while 50 (8%) were polluted. Comparatively from the provided data by DOE, an observation on the trend of WQI (2017-2018) based on the selected rivers in Penang has been tabulated in **Figure 1.2** below. The trend of the water quality status is slightly higher in 2018 compared to 2017 due to few factors such as the increase in population and the development process as the years go by (The Department of Environment (DOE), 2018).



	Sg. Air Terjun	Sg. Ara	Sg. Bayan Lepas	Sg. Jawi	Sg. Juru	Sg Perai	Sg. Pinang	Sg. Junjong	Sg. Relau
WQI (2017)	92	80	65	44	53	57	57	70	69
WQI (2018)	93	82	70	47	58	59	70	67	79

Figure 1.2 River Water Quality Trend for Rivers in Penang

To emphasize as in previous years, BOD, NH_3-N and TSS have continue to remain as crucial aspects in terms of river pollution. Inefficient handling of sewage or waste from agriculture and industrial sectors had been attributed to high BOD. The key sources of NH_3-N can be traced by animal farming and domestic sewage, while the contributors of TSS are primarily due to inadequate earthworks and land clearing operations. Therefore, it is important to understand dependent and independent variables through this work on the correlation of how each parameter relates to one another through the observation on the emerging trend and pattern.

ANN is a modelling approach that allows learning by example from representative data as well as errors generated to improve the model. A good model with high accuracy gives prediction that are very close to the actual values. The key feature of ANN is making the model generalize and predict on unseen data. Unseen data is a subset of the dataset used to assess the likely future performance of a model. With this intention, this work requires the feature selection as one of the key procedures to eliminate any irrelevant features to test the model's final ability to generalize before deploying to production.

1.4 Objectives

This research aims

a) to determine the suitable feature extraction method to represent the correlation of the water quality parameters;

b) to develop a Water Quality Index (WQI) prediction model by using the Artificial Neural Network (ANN) analysis in MATLAB;

c) to evaluate the performance of the model developed based on model validation using unseen data.

CHAPTER 2
LITERATURE REVIEW

Chapter 2 comprises of literature review on previous discoveries and researches available from credible scientific records and references that are related to the project topic. The supervision of water quality, which consists of data monitoring, may be adjusted to the implementation of the prediction model. The concept of combining the monitoring and modelling models in this study could provide better knowledge and may also assist in evaluating and predicting the future situation of water quality required by different managements.

2.1 WQI prediction model

Since water quality monitoring is a key step in water resources management, several types of artificial intelligence (AI) tools can be used in forecasting the WQI value for proactively manage any pollution that may significantly affect the ecosystem and ensuring that environmental requirements are fulfilled. Among the algorithm tools involved are Artificial Neural Network (ANN) which will be implemented in this project, the Support Vector Machine (SVM) and the k-Nearest Neighbor (kNN).

ANN consists of many processing elements joined together in layers; input, hidden and output, in a unique manner (Khuan et al., 2002). As for SVM, it provides greater potentials compared to ANN in term of nonlinear, high dimensional and other partial elements. SVM may be useful when there are lacking a sufficient number of monitoring stations or ungauged catchment for water quality parameters. A lower value for mean squared error (MSE) usually nearer 0 indicate reliable results and enable higher performance whereby the study result shows the correlation coefficient has maximum values of 0.998 and 0.979 while the MSE value was found to be between

0.004 and 0.681 (Abobakr Saeed et al., 2019). Also, according to Leong et al. (2019), the SVM model provide a better prediction compared to other approaches with the high value of the coefficient of determination, R^2 with 0.9984 and low MSE with 0.0052. Moreover, SVM can automatically map the training data into a featured space. Despite no equation needed to be derived before developing the SVM model for the real process, instead, screening of the data was necessary based on interquartile range rule (Leong et al., 2019). To add more, a study by Kamyab Talesh et al. (2019) found that the forecasting SVM models for the study area for Sefidrud Basin, Iran resulted in R^2 of 0.87 and MSE of 0.061.

Other correlations that has been discussed in the previous study to support the SVM models are the kernel function type; linear, polynomial, radial basis (RBF) and sigmoid functions and, the trial-and-error method were conducted to establish the optimum gamma, γ and capacity, C parameters. Selecting the best ideal values for these parameters would greatly increase the likelihood of achieving a high degree of accuracy in the estimation of the desired data. The data the best kernel function type and optimum values of γ and C parameters are tabulated in the following **Table 2.1**.

Table 2.1 Comparison of the best kernel function and optimum parameter of γ and C

Best kernel function type chosen by reference			**γ**	**C**	**References**
	R^2	**MSE**			
RBF	0.998	0.117	[0.001,10]	8.5	(Abobakr Saeed et al., 2019)
Polynomial	0.9184	1.6454	-	-	(Leong et al., 2019)
RBF	-	-	0.1	1	(Kamyab Talesh et al., 2019)

Unlike the ANN and the SVM that used to predict WQI, the kNN is a machine learning algorithm to estimate the water quality classification (WQC). The kNN classifies by finding the given points nearest N-neighbors and assigns the class of majority of N-neighbors to it. According to Ahmed et al. (2019), kNN is not

recommended for large datasets because all the processing takes place while testing, and it iterates through the whole training data. Correspondingly, Theyazn et al. (2020) also discussed on kNN model where the K value is used to find the closest points in the feature vectors. To compare the accuracy of the kNN model in predicting WQC, a table on the accuracy, sensitivity, precision and F-score evaluations is illustrated in **Table 2.2** below. Different approaches where n = 5 configuration and Euclidean distance function (Di) expression were taken by Ahmed et al. and Theyazn et al. respectively.

Table 2.2 Comparison on the kNN model performance

Model	Accuracy (%)	Sensitivity (%)	Precision (%)	F-score (%)	References
kNN	72.7	47.83	47.34	47.5	(Ahmed et al., 2019)
	83.63	84.73	87.5	85.84	(Theyazn et al., 2020)

2.2 Artificial Neural Network (ANN) for WQI prediction

From the works of literature, there are several approaches in developing the ANN-based WQI prediction models complying with various researchers. A study by Rahim et al., (2017) has work on two feedforwards artificial neural network (FANN) model predictions for WQI. Two parameters which are BOD (NET1) and COD (NET2) are selected as the output variables with the model development using TSS, NO_3, Kalium, NH_3-NL, total solid, zinc and turbidity as the input set. NET3 work as the WQI predictor and the network input was based on the selected output parameters. The input sets for WQI model prediction and FANN modelling structure are shown in **Table 2.3** and **Figure 2.1** respectively (Rahim et al., 2017).

Table 2.3 Input sets for WQI model prediction

FANN WQI Prediction NET1 and NET2	
Input	Output
TSS NO_3 Kalium NH_3-NL Total solid Zinc Turbidity	BOD (NET1) COD (NET2)
FANN WQI Prediction NET3	
Input	Output
BOD COD	WQI (NET3)

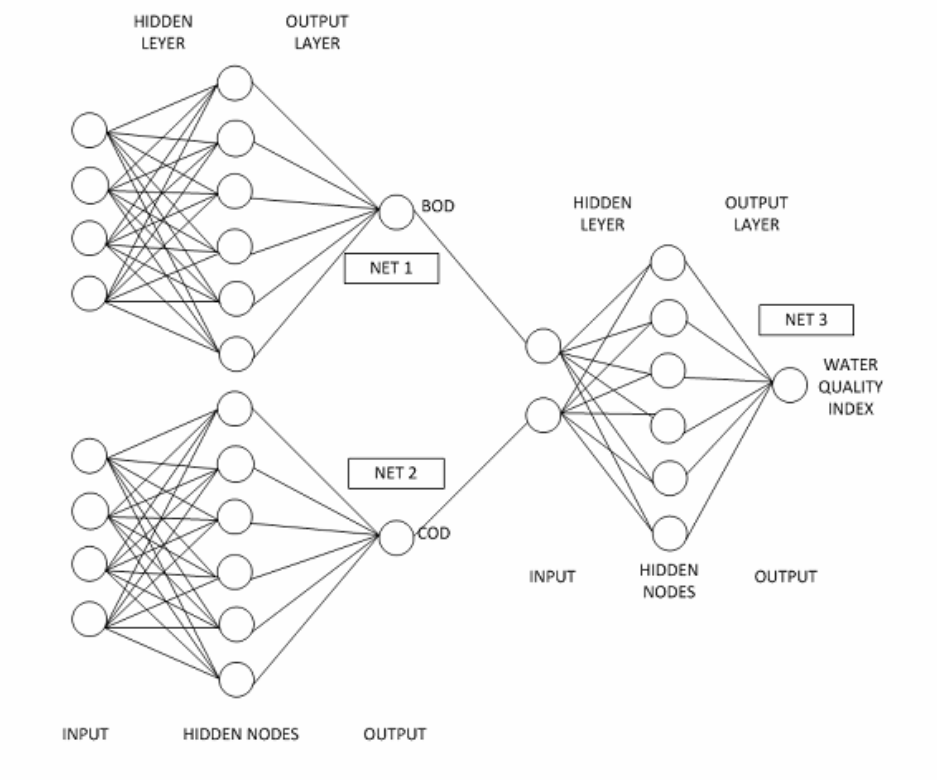


Figure 2.1 FANN modelling structure for WQI prediction (Redraw from Rahim et al., 2017)

Correspondingly, as mentioned by Haghiabi et al., (2018), the structure of ANN consists of networks of neurons that work together in parallel with correlated

mathematical operations of neurons. They reported that some measures to alleviate the trial and error process should be taken into account for the implementation of the ANN. They clarified that for the preliminary design of the ANN model, throughout the first stage, one hidden layer comprised of numbers of neurons comparable to input data is considered after the database division. With the use of few selected input parameters such as pH, TDS and others (mentioned in **Figure 2.2**), it is notable that the structure of ANN declared by the researcher consisted of two hidden layers where its first and second hidden layers included eight and three neurons, respectively as per **Figure 2.2** below. Also, in this study, the size of the network is modified to increase the precision of the developed model where it was proved by the increment of the number of hidden layers (Haghiabi et al., 2018).

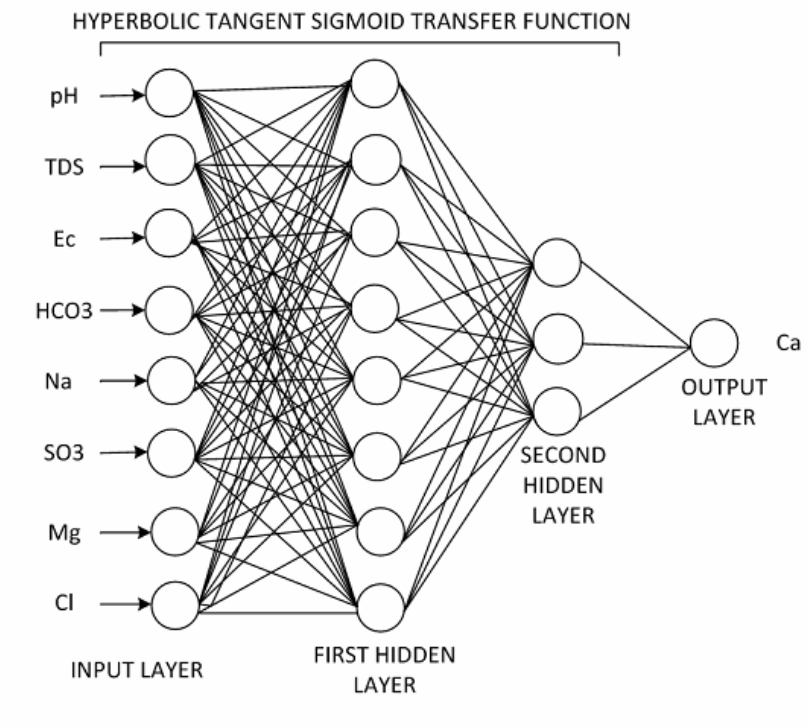


Figure 2.2 Structure of ANN model for water quality prediction (Redraw from Haghiabi et al., 2018)

To compare the performance of the ANN model, data validation can be done by determining the coefficient of determination (R^2) and mean squared error (MSE) value. The data is tabulated as following **Table 2.4**:

Table 2.4 Coefficient Determination and Mean Squared Error for WQI Model Validation

Stage	R^2	MSE	References
Training	0.92	0.153	(Rahim et al., 2017)
Testing	0.93	0.133	
Training	0.92	0.238	(Haghiabi et al., 2018)
Testing	0.84	0.295	

2.3 Principal Component Analysis (PCA)

Abdi & Williams (2010) defined PCA as a viable strategy for analyzing a data set with numerous quantitative dependent variables all of which are interconnected. The focus of this approach is to extract relevant information from the data table and express it via a series of new possible orthogonal variables known as principal components (PCs). Meanwhile, Chen et al., (2020) mentioned that PCA is a dimensionality reduction method that reduces the scale of the input data set to avoid redundancy. According to Mahapatra et al., (2012), this dimensionality reduction can be broadly classified into two categories notably R-mode and Q-mode. To explain more, R-mode is referred to PCA that is utilized to establish a framework among the variables whereas Q-mode indicates PCA used to group cases. Each variable will be associated with one of the elements, with each element having a high correlation with a restricted number of variables. Generally, in recent years, the application of Q-mode PCA has been widely used by researchers particularly in the prediction of water quality.

In like manner, this technique has eliminated the correlation between evaluation indicators, as well as considerably reduced the workload of indicator selection and

calculation. In 2020, Yang et al. has conducted a study on the surface water quality variables in terms of spatial and temporal distribution across eight monitoring stations at Xin'anjiang River in Huangshan, China. PCA technique is applied and eighteen water quality indexes had been reduced to three important principal components (PCs) notably PC1 (49.54%), PC2 (24.03%) and PC3 (13.67%) with each PC represent oxygen-consuming pollutants, heavy metals and water sample pH respectively. To sum up, the total variance of the original data set is 87.24%.

Also, a study by Sahoo et al., (2015) observed that four PCs out of twelve input parameters with a total variance of 65.159% are sufficient in the analysis with PC1, PC2, PC3 and PC4 account for 28.485%, 16.496%, 10.458% and 9.719% of total variation respectively. Another analysis has been conducted in 2001 by Petersen et al., by the same token observed the water quality at 14 stations along the Elbe river in Germany which is divided into two sections namely Riverine and Estuarine part. Petersen et al. in their study agreed upon PCA is a standard method applied to obtain a reduction in input data by considering only PCs of the original data. This work consists of eight parameters and it was found that two PCs are sufficient to describe nearly 60% of the observed total variance data respectively; PC1 and PC2 of 39% and 20% for Riverine part and 37% and 21% for Estuarine part. The comparison of the observed PCs from mentioned researchers are tabulated in the following **Table 2.5**,

Table 2.5 Correlation between original data and principal component

Location		No. of original data	No. of principal component(s)	Total variance (%)	Reference
Xin'anjiang River, Chine		18	3	87.24	(Yang et al., 2020)
Brahmani River, India		12	4	65.159	(Sahoo et al., 2015)
Elbe River, Germany	Riverine part	8	2	59	(Petersen et al., 2001)
	Estuarine part			58	

CHAPTER 3
METHODOLOGY

This chapter provides an overview of the project implementation. It includes the general research flow diagram, ANN architecture selection and the WQI prediction model development.

3.1 Research Methodology

Generally, this paper provides a study on the prediction of the water quality index model by Artificial Neural Network (ANN) based framework that is generated using MATLAB simulation. Next, the performance of the model developed will further analysed. The overall work of this final year project is illustrated in **Figure 3.1** below.

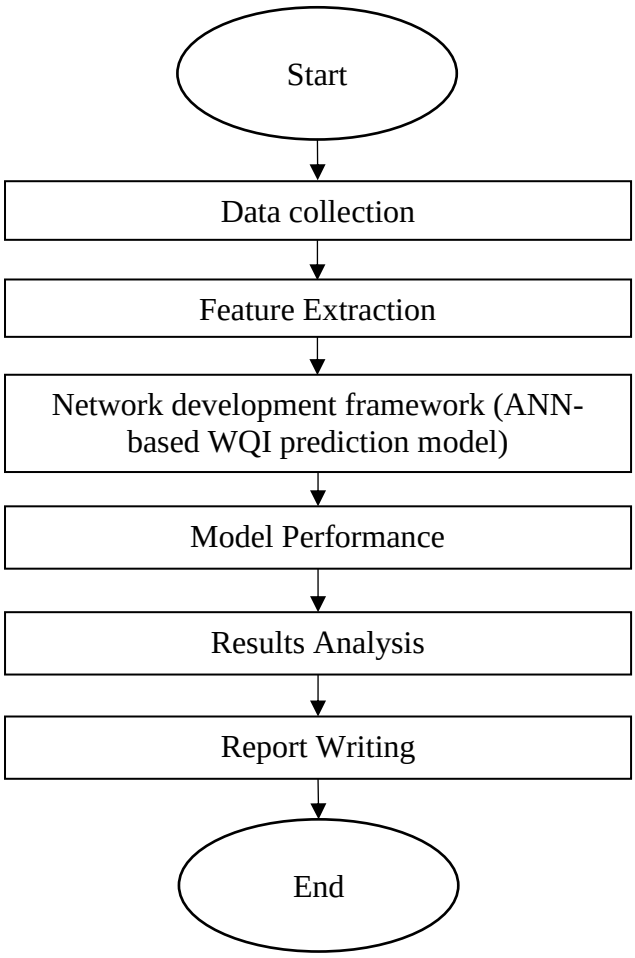


Figure 3.1 Flow diagram of the simulation project on ANN-based WQI prediction model

In brief, to kick start of the project, a literature search on few journals and articles was done and consequentially come up with specific aims and objectives of my study area. In this study, referring to the obtained data from the DOE, there were six parameters required; BOD, COD, TSS, DO, NH_3-N and pH for further analysis of WQI determination. The procedure consists of 3 steps as follows and the formula for the WQI calculation (Equation 1) is as written in (Nurul Ruhayu et al., 2015):

1. Identify sub-index (SI) equation based on the value of the parameter.
2. Calculate the sub-index (SI) of every parameter.
3. Calculate the water quality index (WQI).

$$WQI = 0.22SI_{DO} + 0.19SI_{BOD} + 0.16SI_{COD} + 0.16SI_{TSS} + 0.15SI_{NH3-N} + 0.12SI_{pH} \quad \text{(Equation 1)}$$

3.2 Data Collection

As has been noted, the sample and data collection were carried by the Department of Environment, Malaysia. There are 30 water quality variables that were monitored and collected from Penang rivers with a total of 1000 samples were used for this study. The parameters that were involved were summarized in **Table 3.1**.

Table 3.1 Water quality parameters for Penang rivers

Independent variables (Input)			**Dependent variables (Output)**
Dissolved oxygen	Chlorine	Zinc	Biological oxygen
Suspended solid	Oil and Grease	Calcium	demand
Ammoniacal	Methyl Blue-	Iron	Chemical oxygen
nitrogen	Activated Substances	Potassium	demand
Temperature	Phosphate	Magnesium	
Conductivity	Arsenic	Sodium	
Turbidity	Mercury	pH	
Dissolved solid	Cadmium	Salinity	
Total solid	Chromium	Fasecal coliform	
Nitrate	Plumbum	Total coliform	

3.3 Feature Extraction for ANN development

In this part, a feature extraction method based on multiway principal component analysis (MPCA) have been applied to the pre-processing task. The feature extraction will reduce the dimensional of the input data before the ANN starts to train the input to imitate the target. Similar to the PCA which is a powerful method used to analyse a data set that consists of several intercorrelated quantitative dependent variables, the MPCA technique was chosen to represent the correlation of the 28 water quality input parameters. The results of the approach implemented in this study is the combination of MPCA with ANN for WQI prediction model.

The MPCA method essentially divides a large and complex data space into a series of column sub-data spaces, where the sub-statistical models in each-sub data are established. The specific algorithm of MPCA is to add a data matrix expansion module before the PCA algorithm. Then, the other parts of MPCA are the same as the standard PCA algorithm.

3.4 ANN prediction model with MATLAB

3.4.1 ANN architecture development and selection

3.4.1(a) BOD and COD analysis

On the whole, the workflow of the network development framework for the ANN-based water quality prediction model by using MATLAB was developed. There were two parts involved in this model implementation specifically development and selection. Notably, the method applied was illustrated in the following **Figure 3.2**,

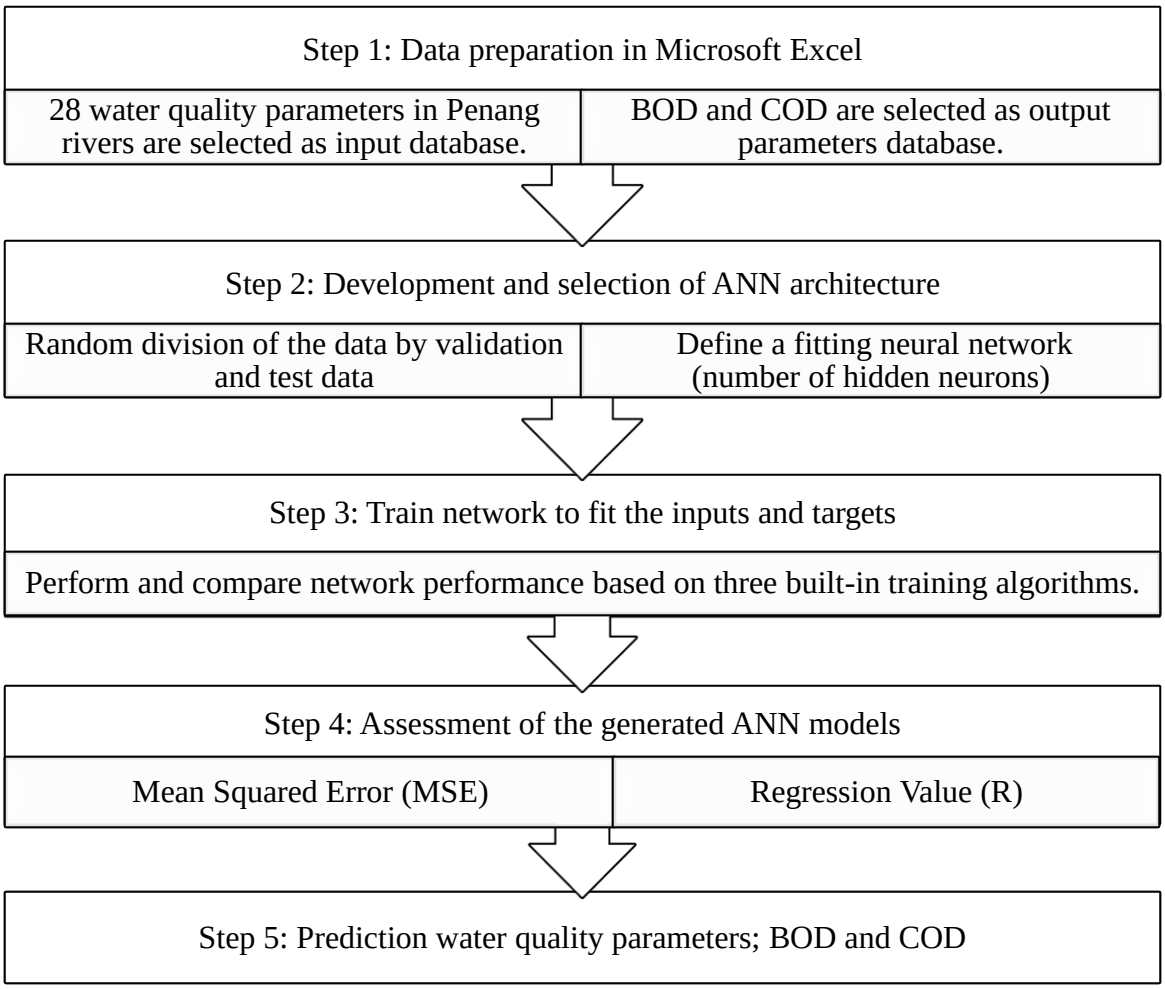


Figure 3.2 Brief workflow of the applied method for ANN approach

To demonstrate, each of the applied methodologies in the workflow will be further described accordingly. First, in **Step 1**, available data of water quality parameters for Penang rivers obtained from DOE was prepared in Microsoft Excel. There were 28 water quality parameters chosen as input database while BOD and COD were chosen as single target, respectively. Consequently, there were 1000 samples of data for each water quality target.

Next, several trials were created in **Step 2** to obtain the best ANN architecture for each output run. Four scenarios based on random division; training, validation and testing set of the 1000 samples and number of hidden neurons defined as 15 nodes were developed. To emphasize, mentioned scenarios are summarized in **Table 3.2** below.

Table 3.2 Scenario for ANN architecture trials

	Scenario		No. of hidden neurons	Input database	Output database
	Training (%)	**80**			
1	Validation (%)	10	15	28 Parameters	BOD / COD
	Testing (%)	10			
	Training (%)	**70**			
2	Validation (%)	15	15	28 Parameters	BOD / COD
	Testing (%)	15			
	Training (%)	**60**			
3	Validation (%)	20	15	28 Parameters	BOD / COD
	Testing (%)	20			
	Training (%)	**50**			
4	Validation (%)	25	15	28 Parameters	BOD / COD
	Testing (%)	25			

After creation of trial scenarios, for instance 80% of total data samples was selected as training set; 10% of total data samples as validation set and final 10% of total data samples as test data, henceforth, 1000 data will be divided randomly with selected data test percentage. According to literatures, 50-80% of training data was chosen by most of the researchers' studies. The illustration of the network architecture was demonstrated in the following **Figure 3.3**,

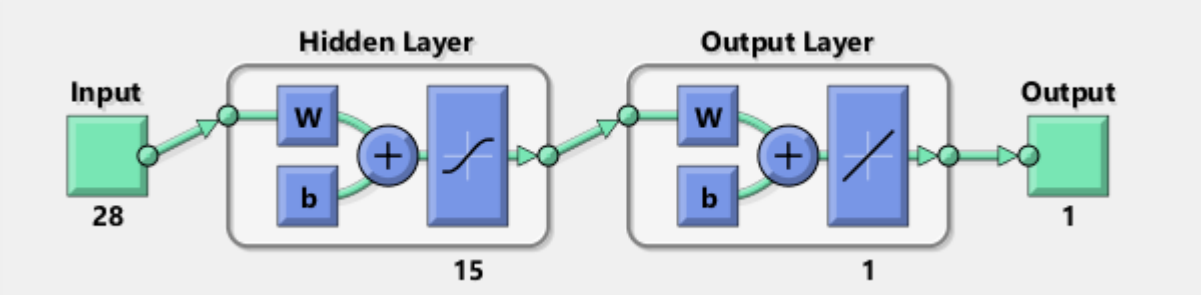


Figure 3.3 Overview of the 3-layer ANN architecture (28-15-1)

Under those circumstances, **Step 3** is performed to decide the best network architecture based on the validation and test data as well as the number of hidden neurons. For this reason, the best scenario is chosen by comparing the network performance considering three training algorithms namely Levenberg-Marquardt (LM), Bayesian Regularization (BR) and Scaled Conjugate Gradient (SCG). Training multiple times will generate different results due to different initial conditions and sampling. Therefore, **Step 4** was implemented by assessing the generated ANN models by comparing the MSE and R values. The network architecture was exemplified in the following **Figure 3.4**.

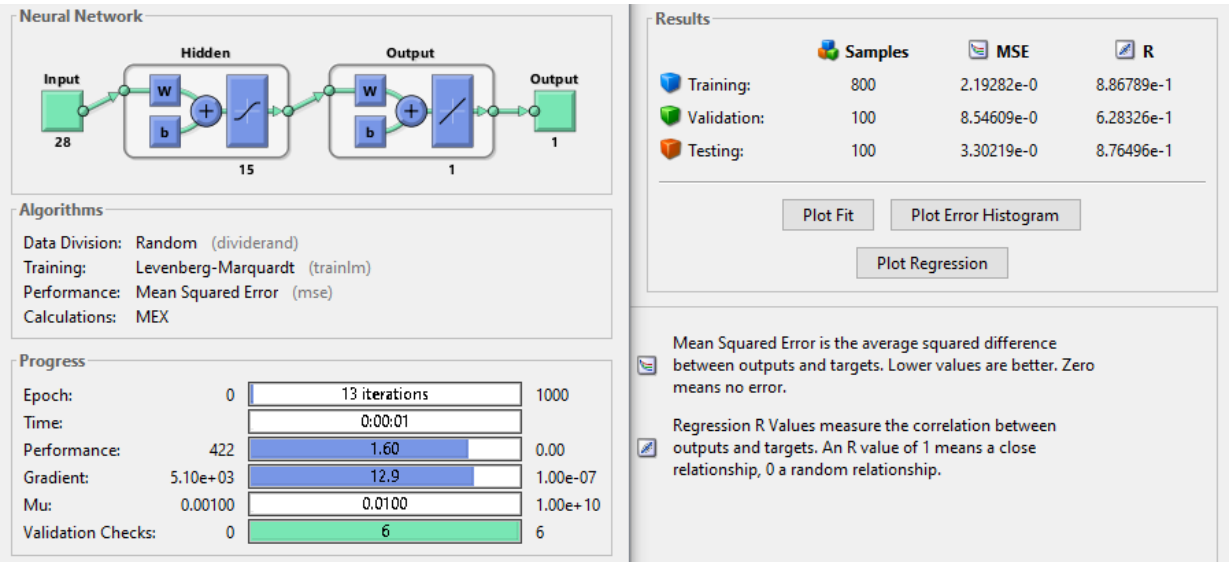


Figure 3.4 Overview of performance for 28-15-1 architecture for LM algorithm

Again, **Step 2** and **Step 4** were repeated for the selection of the best ANN architecture based on the performance analysis obtained from the previous work. For this reason, three sub-scenarios assumed from the number of hidden neurons in the fitting network's hidden layer are constructed. The training algorithm chosen for this work was Bayesian Regularization. To enumerate, **Table 3.3** was prepared.

Table 3.3 ANN architecture selection

Scenario	Training Algorithm	Sub-scenario (Hidden nodes)		
1 2 3 4	BR	15	30	45

In the long run, **Step 5** was achieved for the selection of the best architecture with the best performance criteria between the 2 criteria plots; MSE and R values by considering line fitting with R values reaching 1.0 and MSE values approaching 0. The best network architecture was chosen after further optimization analysis was done.

3.4.1(b) WQI analysis

In the final analysis, a work on the architecture development for the WQI prediction model by implementing ANN in MATLAB was developed. The process flow of the ANN-based WQI prediction model was almost similar as in previous water quality parameters prediction model; BOD and COD (refer **Figure 3.2**). Given these point, BOD and COD data from previous work were now chosen as the input and WQI was set as the output. Accordingly, several trials were created to obtain the best ANN architecture for WQI prediction. Two scenarios based on random division dataset and number of hidden neurons were applied. To emphasize, **Table 3.4** was prepared.

Table 3.4 ANN architecture development and selection for WQI

Scenario	**Input**	**Output**	**Training Algorithm**	**Number of hidden neurons**				
1	BOD	WQI	BR	2	3	4	5	6
2	COD							

Next, the performance of the model developed was analyzed based on the best R and MSE values.

CHAPTER 4
RESULTS AND DISCUSSION

This chapter presents the result obtained where all the data and results are thoroughly discussed to meet the outline of the research objectives. The results simulated from MATLAB are tabulated in figure and table form to be further analysed.

4.1 Feature Extraction with ANN Prediction Model

Results for feature extraction when the MPCA were implemented to the ANN prediction model has been discussed in the following part, where the results on the effects of MPCA as a feature extraction method in the ANN prediction model will be compared. As BR algorithm can minimize or eliminate the need for lengthy cross-validation (Burden & Winkler, 2008), it was chosen in regards with the implementation of MPCA meanwhile 60% and 70% training dataset was applied for BOD and COD analysis respectively.

After performing the MPCA for BOD, only 4 inputs were required as the components to explained at least 99.999% variability. **Figure 4.1** showed the R values obtained using the BR algorithm and 60% training with two-, four-, six-, eight-, ten-, and twelve hidden neurons. The number of hidden neurons was adjusted accordingly as the input database decreased from 28 parameters initially to only four principal components. From the illustration, 12 hidden neurons gave the highest R-value (0.7825) compared to others.

Likewise, the same approaches were applied in the development of MPCA for the target COD, and similarly only 4 inputs were required out of 28 initially. **Figure 4.2** showed the R values obtained using the BR algorithm and 70% training with 2-12 hidden neurons. From the study, 10 hidden neurons give the highest R-value; 0.6716.

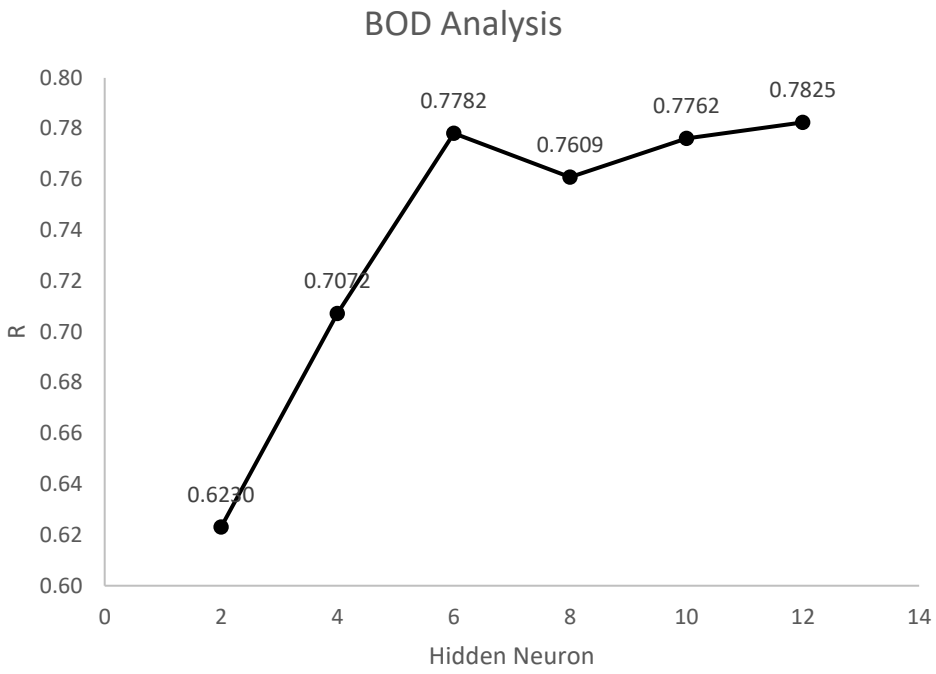


Figure 4.1 R values obtained using BR Algorithm and 60% Training

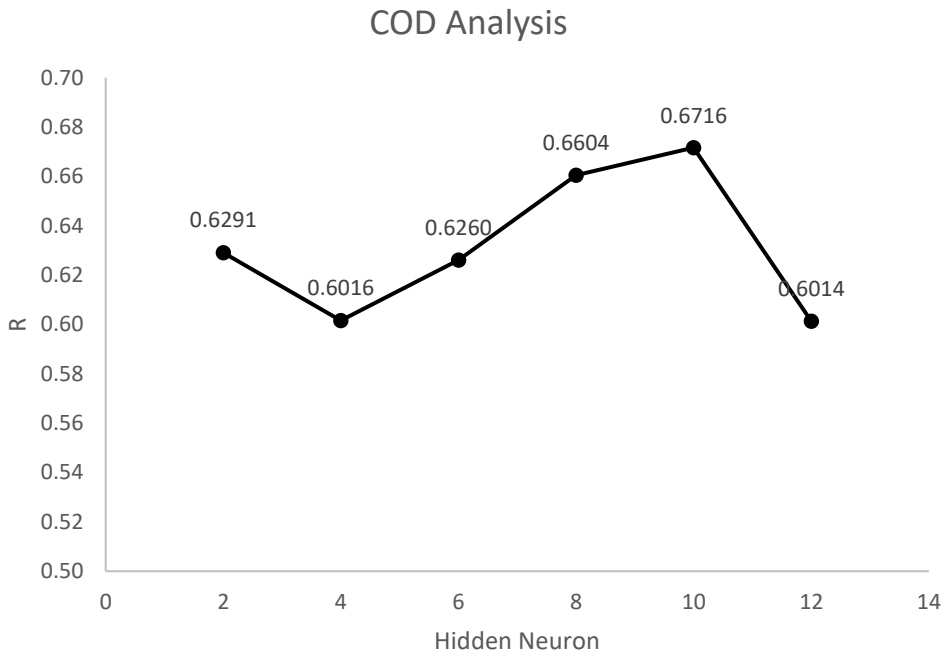


Figure 4.2 R values obtained using BR Algorithm and 70% Training

From the simulations, the chosen ANN architecture for BOD and COD after MPCA was introduced was emphasized in **Table 4.1** and **Table 4.2** respectively.

Table 4.1 Chosen network architecture for BOD analysis after MPCA

Input	4 Parameters
Output	BOD
Training (%)	60
Validation (%)	20
Testing (%)	20
Hidden Neurons	12
Training Algorithm	Bayesian Regularization (BR)

Table 4.2 Chosen network architecture for COD analysis after MPCA

Input	4 Parameters
Output	COD
Training (%)	70
Validation (%)	15
Testing (%)	15
Hidden Neurons	10
Training Algorithm	Bayesian Regularization (BR)

To sum up, the MPCA did simplified and reduced the data redundancy to only 4 input required, however, the performance of the prediction model with chosen network architecture for BOD and COD was not very well. As compared to the other work using 28 input databases, the ANN architecture could achieve almost 0.99 of regression values, unlike the analysis after MPCA was implemented, the regression