Water Quality Classification Using an Artificial Neural Network (ANN)

Khadijah Sulaiman1*, Lokman Hakim Ismail1, Mohd Adib Mohammad Razi1, Mohd Shalahuddin Adnan1 and Rozaida Ghazali2

1 Faculty of Civil and Environmental Engineering, University Tun Hussein Onn Malaysia
2 Faculty of Computer Sciences and Information Technology, University Tun Hussein Onn Malaysia
*Corresponding author: khadijahhumaira94@gmail.com

Abstract. Malaysia is currently a rapidly developing country to achieve a 2020 vision. However the development that has been carried out contributed to a negative impact on the environment especially on water quality. Due to the deterioration of water quality, serious management efforts on water quality has been taken. Thus, the aim of this study is to investigate a technique that can automatically classify the water quality. The technique is based on the concept of Artificial Neural Network (ANN). Since the greater part of their methodologies depend on the idea of ‘pattern recognition’. Thus, it is convenient to inspect its ability in classify water quality. There are six environmental data were used in this study such as pH, total suspended solids (TSS), dissolved oxygen (DO), chemical oxygen demand (COD), biological oxygen demand (BOD), and ammonia. The data was obtained by in-site measurement and laboratory analysis. Then, the data was used as the feeder of input variables in the ANN database system. After training and testing the network of ANN, the result showed that 80.0% of accuracy classification with 0.468 of root mean square error (RMSE). This showed the encouraging results for classification.

1. Introduction
Poor administration of water quality can be disastrous to human civilisation as appeared by occurrences of the disease all over the world [1]. Because of the effects on human wellbeing, most nations have set-up the frameworks for water quality administration to guarantee the quality of water. The extension of water quality administration began in 1985, with the development of water quality criteria and benchmarks based on the water quality index [2].

The evaluation of water quality based on the index system has been used for over the past decades. However, there is still a limitation on the index system because there are many used data cannot correlate with an index [2]. In addition, the model requires a series of calculation before a final index value and class grade value can be obtained.

In response to these limitations, this research aims to investigate a technique that can efficiently and easily to classify the water quality. One of the approaches that conceivable in the evaluation of water quality is the Artificial Neural Networks (ANN), where there is a growing tendency of soft computing techniques and primarily data-driven approaches. Its function is to complement or even replace the knowledge-based model, which may deliver assessment systems that are operationally straightforward, adaptable and comprehensive, where the outcomes are effortlessly justifiable [3]. Hence, many types of researches have been studied by including the modelling of water quality [4-7].
2. Materials and Methods

2.1. Study Area and Water Quality Data
The Straits Malacca is subjected to a great variety of pollutants due to its strategic location as a major international shipping lane and the concentration of agriculture, industry and urbanisation which predominate on the west coast of Peninsular Malaysia. According to [8], Strait Malacca have a serious environmental problems because of many river discharge and due to agricultural and industry along the Strait Malacca.

The data set used in this study was generated through a data collection at site and laboratory analysis. Three locations of this study were chosen which are Pontian River, Batu Pahat River and Muar as shown in figure 1.

![Figure 1. The location of sampling site.](image)

The water quality parameters were collected from river and estuaries. The sampling site spread along the river including the river mouth and downstream of the river. Grab water samples were collected for 30 checkpoints for Pontian River and 35 checkpoints for each Batu Pahat River and Muar River. The two in-situ parameters, pH and DO were collected using HORIBA water quality monitor, meanwhile water samples were collected for lab analysis using HACH DR 6000. Parameters of lab analysis were BOD, COD, ammonia nitrogen and TSS.

2.2 Water Quality Standard
In Malaysia, the existing methodology for river water quality classification and monitoring is quite extensive. At the moment, Malaysia has over 1000 manual and automatic river water quality monitoring stations in 146 basins maintained by the Department of Environmental (DOE) alone [9]. These exclude other agencies such as Department of Irrigation and Drainage (DID) as well as the respective state level agencies.

Data on water quality is used to determine the status by examining whether it is clean, slightly polluted or polluted. The quality of river water has been monitored by the Department of Environment (DOE) since the late seventies. The primary aims were to establish and detect any changes in the water quality status and to identify the pollution sources. The classification has been continued on an annual basis by putting rivers in Class I, II, III, IV, or V based on the Interim National Water Standards for Malaysia (INWQS) and based on the guidelines for water use as shown in table 1.
Table 1. Guidelines for Water Use [10]

| Class | Definitions |
|-------|-------------|
| I     | • Conservation of the natural environment.  
        • Water supply I - Practically no treatment necessary (except by disinfection or boiling only).  
        • Fishery I - Very sensitive aquatic species. |
| II A  | • Water supply II – Conventional treatment required.  
        • Fishery II - Sensitive aquatic species. |
| I I B  | • Recreational use with body contact. |
| I I I  | • Water supply III – Extensive treatment required.  
        • Fishery III - Common of economic value, and tolerant species; livestock drinking. |
| I V    | Irrigation. |
| V     | None in above |

2.3 Artificial Neural Network
The terminology of ANN is developed from the biological model of the brain. A neural network consists of a set of connected cells, which is called the neurons. The neurons receive impulses from either input cells or other neurons and perform some kind of transformation of the input and transmit the outcome to the neurons or to output cells. The neural networks are built from layers of neurons connected so that one layer receives input from the preceding layer of neurons and passes the output on the subsequent layer. According to [11], the data fed through the input layer is scaled by an initial weighting through the connecting between neurons. The relationship between inputs (x), weights (w) and outputs (y) as shown in figure 2.

![Computation Process of Neuron in ANN](image)

3. Experiments and Results

3.1 Input Selection
The total of 100 data set from excel file was imported from excel file to develop the ANN model. Six parameters that had been collected were used in this input selection and five categories classes based on water quality standard were set as a target variables to produce an output.

3.2 Data Pre-processing
Normalization procedure before presenting the input data to the network is required since mixing variables with large magnitudes and small magnitudes will confuse the learning algorithm on the
The importance of each variable and may force it to finally reject the variable with the smaller magnitude. Normalization will scale the minimum value to 0 and maximum value to 1.

3.3 Training and testing network
In order to train and test the network, experimental data was categorized into two sets which were for training and testing sets. The training set was used to identify the biases and weight of the network. Meanwhile the testing set was used for calibration to prevent the overtraining of the network and could be used to measure the performance of the network. According to the [12], generally testing set should consist of 10-40% from the data set.

3.4 Artificial Neural Network model
The ratio of 80:20 was used in this experimental which means from 100 data sets, 80 data set was used for training and 20 data was used for testing. The algorithm of the neural network was used to train and test the data set. All the testing was recorded into the network to make it learn the potential relationship between water quality parameter and their corresponding categories. Six parameters represented the input layer nodes of the ANN model, while the five output layer nodes of the ANN model were represented the five different class categories. Percentages of classification accuracy and root mean square error (RMSE) was used to represent the performance of ANN as shown in Equation (1). The lower the RMSE, the more accurate is the estimation. In order to speed up the training process and maintain the error reduction, the hidden layer, momentum and learning rate were adjusted. From the trial and error of experimental, there were 1 hidden layer, 0.2 momentums, 0.3 learning rate of the network produced 0.468 of RMSE after 1000 iterations (cycle).

$$RMSE = \sqrt{\frac{1}{N}(X_{\text{observed}} - X_{\text{predicted}})^2}$$ (1)

Where, $N$ is total number of observation in the data set.

3.5 Results
Based on table 2, it illustrated the confusion matric of the actual class (true class) and output class (predict class). From the illustration of confusion matric its shows 20 data of testing set. Vertically reading, there are 7 data in class II but only 6 data were correctly classified with 85.71% accuracy and 12 data in class III but only 9 data were classified correctly with 75.00% accuracy. Meanwhile 1 data in class IV were classified correctly with 100% accuracy. There no record matches the class I and V. The overall data correctly classified 16 out 20 data set with 80.0% of classification accuracy as shown in table 3.

| Table 2. Confusion matric for classification |
|---------------------------------------------|
| Predict | True I | True II | True III | True IV | True V |
| Predict I | 0 | 0 | 0 | 0 | 0 |
| Predict II | 0 | 6 | 3 | 0 | 0 |
| Predict III | 0 | 1 | 9 | 0 | 0 |
| Predict IV | 0 | 0 | 0 | 1 | 0 |
| Predict V | 0 | 0 | 0 | 0 | 0 |
| Class accuracy | 0 % | 85.71 % | 75.00 % | 100 % | 0 % |
Table 3. Overall results of the testing set

| Number of record | Number of correctly classified | Accuracy percentage |
|------------------|-------------------------------|---------------------|
| 20               | 16                            | 80.0%               |

4. Conclusion
In this study, it was identified that water quality classification using ANN produced encouraging results of classification. As a conclusion, it can be concluded that, ANN is a model that can easily classify the water quality with the justifiable output. This provided performance results can be used as a reference to managing a water quality at study area.

5. References
[1] Fredrick T, Ponnaiah M, Murhekar M V, Jayaraman Y, David J K, Vadivoo S and Joshua V 2015 Cholera outbreak linked with lack of safe water supply following a tropical cyclone in Pondicherry, India *Journal of Health, Population and Nutrition* **33** 31-38
[2] Tyagi S, Sharma B, Singh P and Dobhal R 2013 Water Quality Assessment in Terms of Water Quality Index, *American Journal of Water Resources* **1** 34-38
[3] Ibrahim D, 2016 An Overview of Soft Computing, *12th International Conference on Application of Fuzzy Systems and Soft Computing (Vienna, Austria)* vol 102 (Vienna, Austria : Elsevier) pp 34-38
[4] Wechmongkhonkon S, Poomtong N and Areerachakul S 2012 Application of Artificial Neural Network to Classification Surface Water Quality *World Academy of Science, Engineering and Technology* **6** 199-203
[5] Gazzaz N M, Mohd Kamil Y, Ahmad Z A, Hafizan J and Mohammad F R 2012 Artificial neural network Modelling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors *Marine Pollution Bulletin* **64** 2409-2420
[6] Rao V 2013 A Classical Way of Finding Water Pollution by Using Artificial Neural Network, *School of Computing Science and Engineering* **3** 449-457
[7] Gour S and Gour M 2014 Neural Network Approach in Water Quality Data Analysis for the River Narmada *Binary Journal of Data Mining & Networking* **4** 49-53
[8] Alzahrany A A, Elias S and Siong W B 2007 Determination of Heavy Metal Pollution In Depth Profile Of Marine Sediment Samples From The Strait Of Malacca *Fresenius Environmental Bulletin* **16** 1279-1287
[9] Department of Environment Malaysia (DOE) 2006 *Malaysia Environmental Quality Report* In: Chapter 3: River Water Quality (Malaysia: Sasyaz Holdings Sdn Bhd,) pp 24
[10] Ismail Z, Sulaiman R and Karim R 2014 Evaluating Trends of Water Quality Index of Selected Kelang River Tributaries *Environmental Engineering and Management Journal* **13** 61-72
[11] Mansourkhaki A, Berangi M and Haghiri M 2018 Comparative Application of Radial Basis Function and Multilayer Perceptron Neural Networks to Predict Traffic Noise Pollution in Tehran Roads, *Journal of Ecological Engineering* **19** 113-121
[12] Ali N, Ahmed E, Othman A K and Othman J 2009 Prediction of Johor River Water Quality Parameter Using Artificial Neural Networks *European Journal of Scientific Research* **28** 422-435

Acknowledgement
The authors would like to thank the Ministry of Education Malaysia (MOE) and University Tun Hussein Onn (UTHM) for financial and technical support. This study has been supported by the Fundamental Research Grant Scheme (FRGS), VOT 1524.