Biointerface Research in Applied Chemistry
Platinum Open Access Journal (ISSN: 2069-5837)

Article
Volume 11, Issue 2, 2021, 9285 - 9295
https://doi.org/10.33263/BRIAC112.92859295

Implementation of Machine Learning Methods for Monitoring and Predicting Water Quality Parameters

Gasim Hayder [1,2], Isman Kurniawan [3,4], Hauwa Mohammed Mustafa [5,6,*]

[1] Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor Darul Ehsan, Malaysia; gasim@uniten.edu.my (G.H.);
[2] Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor Darul Ehsan, Malaysia
[3] School of Computing, Telkom University, Bandung, Indonesia 40257; ismankrn@telkomuniversity.ac.id (I.K.);
[4] Research Centre of Human Centric Engineering, Telkom University, Bandung, Indonesia
[5] College of Graduate Studies, Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor Darul Ehsan, Malaysia hauwa.mustafa@uniten.edu.my (H.M.M.);
[6] Department of Chemistry, Kaduna State University (KASU), Tafawa Balewa Way, P.M.B. 2339, Kaduna, Nigeria
* Correspondence: hauwa.mustafa@uniten.edu.my;

Scopus Author ID 57217195204

Received: 9.08.2020; Revised: 7.09.2020; Accepted: 8.09.2020; Published: 11.09.2020

Abstract: The importance of good water quality for human use and consumption can never be underestimated, and its quality is determined through effective monitoring of the water quality index. Different approaches have been employed in the treatment and monitoring of water quality parameters (WQP). Presently, water quality is carried out through laboratory experiments, which requires costly reagents, skilled labor, and consumes time. Thereby making it necessary to search for an alternative method. Recently, machine learning tools have been successfully implemented in the monitoring, estimation, and predictions of river water quality index to provide an alternative solution to the limitations of laboratory analytical methods. In this study, the potentials of one of the machine learning tools (artificial neural network) were explored in the predictions and estimation of the Kelantan River basin. Water quality data collected from the 14 stations of the River basin was used for modeling and predicting (WQP). As for WQP analysis, the results obtained from this study show that the best prediction was obtained from the prediction of pH. The low kurtosis values of pH indicate that the appearance of outliers give a negative impact on the performance. As for WQP analysis for each station, we found that the WQP prediction in station 1, 2, and 3 give the good results. This is related to the available data of those stations that are more than the available data in other stations, except station 8.

Keywords: machine learning; water quality parameters; turbidity; suspended solids; Kelantan River.

1. Introduction

Globally, river water at an initial state is believed to be free from contaminations and the safest water source for human consumption and applications. However, a high rate of industrialization, population growth, and urban development has led to the deterioration of their sustainability [1]. The effective approach of river clean up involves the detection and measurement of both natural and anthropogenic influences and an increased understanding of the sources of pollutants, which are important factors in planning, prevention, and cleaning processes [2]. Moreover, water quality from rivers has significant importance in a number of areas such as agriculture, hydropower plants, domestic and residential water supplies, recreation, tourism, and transportation [3]. Water quality is currently estimated through a series of experimental and statistical analysis, that requires costly reagents, time, and skilled labor [4]. Consequently, the disastrous effects of water pollution require a cheaper and more efficient alternative. Furthermore, ongoing urban river pollution issues in Malaysia have called for more studies to be carried out in order to gain insight into new approaches to technologies, particularly in the river water treatment plant (RWTP) [2].

Several research works have been conducted in the field of water resource management with the sole aim of improving the water quality of rivers. The application of artificial intelligence (AI) that provides a given system with the ability to automatically learn and improve from experience without being explicitly programmed is known as Machine learning (ML) [5]. However, computing tasks where designing and programming algorithms with good performance is difficult or infeasible were successfully achieved through the application of ML [5]. The procedures involved in ML necessitates scrutiny through data to look for patterns and to adjust program action accordingly. By using these techniques, model accuracy can be enhanced [6]. Machine learning models can suitably present conceptions of the hydrological process but requires a large set of data. Although in these techniques, outputs are obtained without in-depth observation of the physical details of the water resource being investigated [7]. For instance, Nouranin et al. [8] applied artificial intelligence based on non-linear models (feed-forward neural network (FFNN), adaptive neuro-fuzzy inference system (ANFIS), support vector machine (SVM), and classical multi-linear regression (MLR) method) for prediction of water quality indicators (biochemical oxygen demand (BOD), chemical oxygen demand (COD) and nitrogen (TN) of Nicosa wastewater treatment plant (NWWTP). Also, genetic programming (GP) technique was used to forecast the rainfall runoff in Sungai Sayong, Johor [4]. Furthermore, in order to improve river suspended sediment predictions, [7] combined the application of the Artificial Intelligence Model with Metaheuristic Algorithms by applying adaptive neuro-fuzzy system (ANFIS) and multilayer feed-forward neural network (MFNN) by the bat algorithm (BA) and weed algorithm (WA). However, human activities in the Kelantan river basin have great importance for socio-economic life in the north-eastern state of peninsular Malaysia, covering about 4.4% of Malaysia with a total area of 15,099 km^2 [9]. The aforementioned human activities have a vital impact on the geographical environment, specifically the quality of the water resources. Therefore, this research is carried out on the dataset collected from 14 stations of Kelantan River, located in Malaysia. Machine learning techniques were applied on the collected dataset for the water quality parameters such as (dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), pH, ammonia nitrogen (NH_3-NL), suspended solids (SS).

2. Materials and Methods

2.1. Study area and data collection.

Kelantan River catchment and its tributaries are occupying about 80% of Kelantan state's surface area with a total catchment area of 11,900 km^2, forming the Kelantan River Basin, which is one of the major basins in Malaysia [9]. The 248 km long Kelantan River originates in the Tahan mountain range and flows northward into the South China Sea. It is a common tropical basin that receives precipitation throughout the year [10]. Seven major sub-catchments such as Nenggri, Gullimard, Pergau, Kuala Krai, Galas, Lebir, and Kota Bahru forming the main rivers which cover about 13170 km^2 of drainage area [11]. The estimated average annual precipitation of the Kelantan River Basin is about 2500mm, and the maximum annual rainfall (1750 mm) occurs during the northeast monsoon season between mid-October and mid-January [12,13]. About 0.5 million of the population are able to sustain as the main river in Kelantan state is adequate to supply most of the water for drinking, agricultural activities, plantation irrigation, small scale fishing industries, sand mining activities, and also contributing a way of transportation and channel of receiving wastewater treatment effluents [14].

According to Yen and Rohasliney, [15] the sand mining activities on the Kelantan River, physicochemical parameters such as total suspended solids (TSS), turbidity, and nitrate concentration had increased to incredibly significant levels that exceed the standards of the Malaysian Interim National Water Quality Standard (INWQS). The Kelantan River Basin is frequently affected by monsoon floods with the occurrences of an increase in flooding from once in 50 years to once in 15 years [16]. Rahman et al. [9] mentioned that after a severe flood in 2014, chemical contaminants such as heavy metals and carcinogens were found in floodwater, and exposure to pathogen or toxicants can lead to a long-term negative environmental impact. Meanwhile, in a recent study [14], the Kelantan River is found to fall under Class I according to the Malaysian National Water Quality Standard (NWQS) in terms of anthropogenic metals. Unfortunately, in terms of heavy metals, total suspended solids (TSS) pollution is up to 291 mg/L, which was considered somewhat polluted under the Malaysian Department of Environment-Water Quality Index (DOE-WQI).

2.2. Dataset.

In this study, the dataset used is the time series data of six water quality parameters (WQP), i.e., dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), pH, ammonia nitrogen (NH_3-NL), and suspended solids (SS) that were collected from 14 stations in Kelantan River. The data size for each station is different depending on the availability of the data. The number of data collected in 4 stations (out of the total 14 stations), i.e station number 1, 2, 3 and 8, is 250 monthly data, which is comprised of the monthly WQP from February 1997 to December 2017.

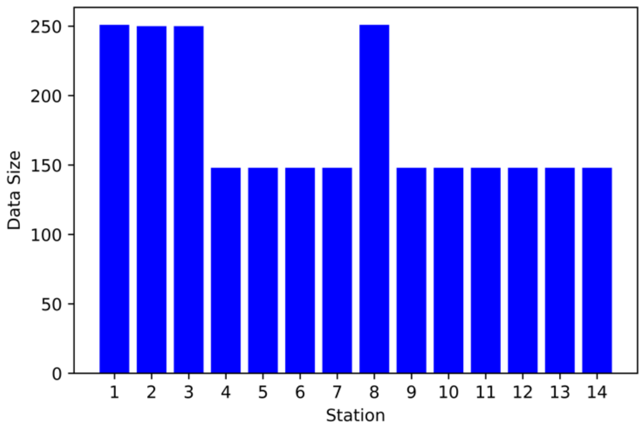

Figure 1. The number of WQP data for each site.

Meanwhile, the number of data collected in the remaining other 10 stations, i.e. station number 4, 5, 6, 7, 9, 10, 11, 12, 13 and 14, is 148 monthly data, which is comprised of the monthly WQP from September 2005 to December 2017. The distribution of data size for each station is presented in Figure 1. The missing value found in the data was treated by using the interpolation method. By using this method, the missing value at a certain time was determined by considering the value before and after the time.

The statistical properties of WQP in station 1 are provided in Table 1, while the properties of the parameters in the rest stations are provided in Supporting Information. Here, the variability of the WQP was described by the value of minimum, maximum, mean, standard deviation, skewness, and kurtosis. We found that the skewness of DO and pH in station 1 have a low skewness coefficient. Meanwhile, the rest of the parameters have a large skewness coefficient, indicates the large difference of mean and median in those parameters. Also, we found that COD, NH_3-NL, and SS have a large kurtosis coefficient that indicates the appearance of the outliers.

Table 1. Statistical properties of WQP in station 1.

Parameter	Min	Max	Mean	Std. Dev.	Skewness	Kurtosis
DO (% Sat)	57.90	99.50	83.65	7.12	-0.62	0.87
BOD (mg/l)	0.50	9.00	2.75	1.71	1.16	0.91
COD (mg/l)	0.90	129.00	17.26	13.47	4.47	28.81
pH (unit)	5.69	8.17	7.03	0.41	-0.52	1.21
NH_3-NL (mg/l)	0.01	2.54	0.11	0.22	7.40	69.39
SS (mg/l)	9.00	1510.00	144.20	167.84	4.62	27.74

For the overall evaluation, the average statistical properties of WQP of all stations are presented in Table 2. Similar to the statistical properties of station 1, we found that only DO and pH parameters have a low skewness coefficient. This indicates that, in the case of skewness, the characteristic of data for each station is quite similar. We also found that the value of the kurtosis coefficient of NH_3-NL and SS is significantly larger than the coefficient of other parameters. This indicates that the appearance probability of outlier in both parameters is higher, and thus the prediction of both parameters will be more difficult. From the total number of data for each station, we split the data before training, validation, and test data in the ratio of 3:1:1. The trained data was used to develop the model, while the validation data was used to adjust the value of the model parameter. The test set is used as the external data to verify the acceptability of the model.

Table 2. The average of statistical properties of WQP.

Parameter	Min	Max	Mean	Std. Dev.	Skewness	Kurtosis
DO (% Sat)	46.50	148	89.04	9.72	-0.20	2.23
BOD (mg/l)	0.50	36	3.33	2.63	3.11	22.49
COD (mg/l)	0.90	136.20	17.74	11.05	3.09	20.81
pH (unit)	4.30	8.73	7.02	0.51	-0.72	1.33
NH_3-NL (mg/l)	0.01	4.08	0.08	0.18	9.49	144.10
SS (mg/l)	0.90	3380.00	126.77	202.45	6.21	67.52

2.3. Artificial neural network.

An artificial neural network (ANN) is a method inspired by the work of the nervous system in the human body. The architecture of ANN is comprised of an input layer, a hidden layer, and an output layer. In the process of ANN, the hidden node will process the value fed in the input layer and transmit the prediction to the output layer. From several architectures of ANN, the three-layer feed-forward backpropagation network is the ANN architecture that is very commonly used [17]. This architecture comprises of one hidden layer with nonlinear transfer functions and one output layer with linear transfer functions. The architecture of ANN can be adapted for time series analysis by using the previous values as the input to predict the following values. A schematic diagram of the ANN network for the time series analysis is provided in Figure 2, where X_t represents the value of variable X at time$t$.

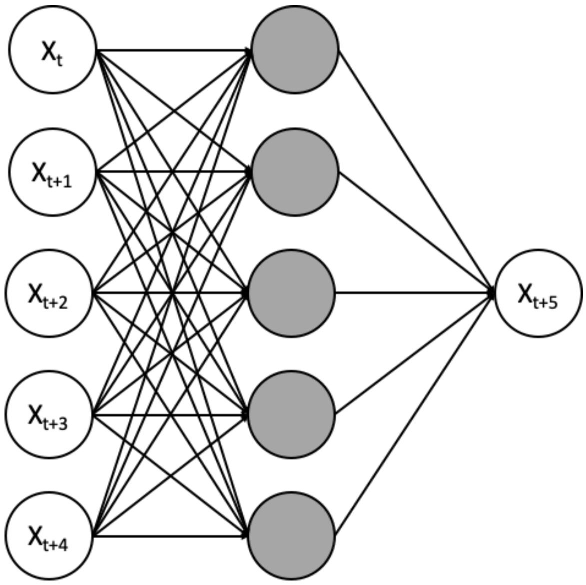

Figure 2. The architecture of ANN for time series prediction.

The training of the ANN network was performed by adjusting the weight values to obtain the output with the lowest error. After the training process, the performance of the ANN must be validated after using the external dataset. The adjustment of the weight values was performed by using a backpropagation algorithm with the steepest descent direction as the reference. From the viewpoint of the optimization process, the ANN training process is equivalent to the process of minimizing a multivariable error function as a function of the network weights [17]. The backpropagation algorithm adjusts the weight values by evaluating the gradient of the error function in relation to the weight values at each iteration by using Equation 1.

$$w^{k+1} = w^k + \alpha \times f(w^k) \tag{1}$$

Where w^k, α, and $f(w^k)$ represent weight at time k, learning rate, and gradient of error. The learning rate determines how fast the algorithm will learn to obtain the correct output. This parameter gives a contribution to the performance of the learning algorithm. There are several algorithms used in gradient descent optimization, such as RMSProp and Adam.

In this study, we utilized three-layer ANN architecture in which the number of hidden nodes is equal to the number of input. The activation function in the hidden and output layer ReLU and Linear function, respectively. Adam optimizer was used to adjust the weight values in the backpropagation algorithm. The default epochs number used to train the model is 500. However, we optimized the epochs number for each model by implementing an early stopping algorithm with the patience value of 10. Therefore, the iteration will be stopped if the loss value of the validation set does not improve after 10 epochs. Regarding the time series analysis, the value of look back is very important and gives a contribution to accuracy. This value indicated the number of previous data that was used to predict the following data. The too-large value of look back will lead to overfitting, while the too small value of look back will lead to underfitting. To obtain a more suitable number of look back, we performed a tuning procedure by considering the lookback value range from 3 to 10. The detail of the ANN parameter used in this study is provided in Table 3. The values of the dataset were normalized to a range of 0 to 1 to speed up the learning process and lead to a faster convergence state.

Table 3. The parameter values of ANN architecture.

Parameter	Values
No. of layer	3
No. of nodes	Equal to input number
Activation function (hidden layer)	ReLU
Activation function (output layer)	Linear
Optimizer	Adam
Loss function	Mean square error (MSE)
Default epochs number	500
Patience	10
Look back ranges	[3, 4, 5, …, 10]

2.4. Model validation.

The performance of the ANN model was validated by using three validation parameters, i.e., root mean square of error (RMSE), the mean absolute percentage error (MAPE), and Pearson correlation coefficient (PCC). Those parameters were used to determine the ability of the ANN model to predict the water quality parameters (WQP) in each station. RMSE and MAPE, which represent the degree of error obtained from the prediction, were calculated by using Equations 2 and 3, respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - P_i)^2} \quad (2)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - P_i}{A_i}\right| \times 100 \quad (3)$$

Where A_i and P_i represent the actual and predicted monthly WQP values, respectively, and n represents the number of data. Since the RMSE values depend on the values range, the parameter was calculated by using the scaled dataset to allow the comparison of the values amongst the WQP. The third validation parameter is PCC, that represents a linear correlation between actual and predicted values. The range values of this parameter are from -1 to +1, where the PCC values of +1, 0, and -1 represent positive linear correlation, no correlation, and negative linear correlation, respectively. PCC was calculated by using Equation 4. We consider the PCC of the test set as an overall parameter to determine the performance of the models.

$$PCC = \frac{n\sum A_i P_i - \sum A_i \sum P_i}{\sqrt{n\sum A_i^2 - (\sum A_i)^2}\sqrt{n\sum P_i^2 - (\sum P_i)^2}} \quad (4)$$

3. Results and Discussion

The result of optimized values of look back and epochs number for each parameter and station was provided in Table 4. We found the variation of those values that indicate the different tendencies of water quality parameters (WQP) in each station. The high values of look back indicate that the model required more input values to give a good prediction, while the low values of look back indicate that the few data is enough for the model to make a good prediction. This corresponds to the relation between the previous and the following data. We also found the variation of epoch numbers for each parameter and station. The variation corresponds to the epoch number required by the model to obtain satisfied loss value of validation data.

Table 4. The optimized values of look back and epochs.

	Station 1						Station 2					
Parameter	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**
Look back	7	7	5	5	7	7	5	8	5	5	8	7
Epochs	325	228	189	500	132	279	285	257	266	477	290	165

	Station 3						Station 4					
Parameter	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**
Look back	4	5	5	5	4	7	6	5	5	4	7	4
Epochs	227	212	411	235	209	132	11	163	410	303	283	73

	Station 5						Station 6					
Parameter	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**
Look back	4	5	5	5	3	8	4	4	7	4	7	8
Epochs	380	225	274	492	267	31	303	500	412	500	104	17

	Station 7						Station 8					
Parameter	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**
Look back	4	5	7	4	8	8	5	7	5	3	6	5
Epochs	390	350	461	470	214	18	331	163	247	500	242	178

	Station 9						Station 10					
Parameter	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**
Look back	4	3	5	4	3	8	4	4	8	8	8	8
Epochs	492	233	500	500	20	68	463	368	13	358	13	37

	Station 11						Station 12					
Parameter	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**
Look back	8	8	8	4	9	7	4	5	9	4	9	8
Epochs	185	190	175	459	132	74	500	323	229	500	115	12

	Station 13						Station 14					
Parameter	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**	**DO**	**BOD**	**COD**	**pH**	**NH_3**	**SS**
Look back	7	3	5	4	4	3	4	8	8	4	4	5
Epochs	14	372	500	382	500	259	436	125	211	500	228	233

The learning curve of the ANN training process of WQP in station 1 is presented in Figure 3. The curve represents the performance of the training process and can be used as an indicator of whether the model is a good fit, underfit, or overfit. The underfit condition is indicated by the large gap between train and validation loss. Meanwhile, the overfit condition is indicated by the increase of validation loss and the decrease of train loss at the end of the training process.

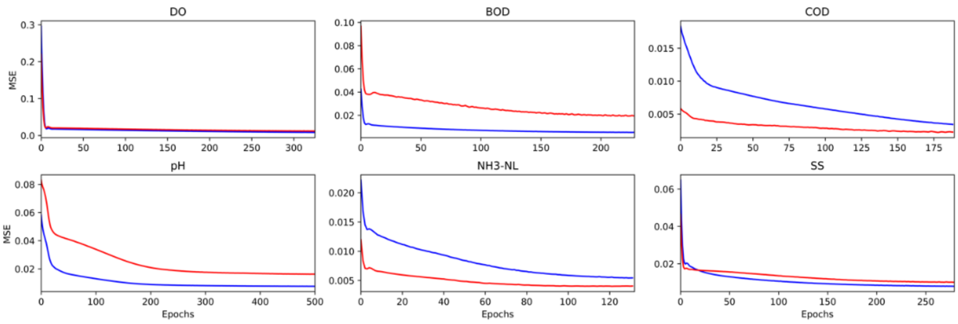

Figure 3. Learning curve of ANN process for WQP in station 1. The blue and red line represent train and validation loss, respectively.

According to Figure 3, we found that the training was stopped after both of the train and validation loss coincidentally decrease. Besides, the losses values are found to be quite stable for several epochs before the training process finishes. We also found that the gap between train and validation loss is not too large. Those conditions indicate that the models are a good fit. However, we found that several models in other stations need to improve because of the indication of underfitting or overfitting. Hyperparameter tuning can be performed to overcome these problems.

To verify the model prediction, the time series data of the WQP was predicted by using the model. The comparison of the actual values with the predicted values of the WQP in station 1 is provided in Figure 4, while the comparison of those values in the rest stations are provided in Supporting Information. We found that the predicted values for the parameters in station 1 are very close to the actual one. The small deviation of the predicted values of the test set specifies the ability of the model to predict the external dataset. The linear correlation of the actual and predicted values of the parameters in station 1 is presented in Figure 5. We found that the majority of data points are gathered around the straight line, while few data points have a small deviation from the line. The deviation indicates the degree of error of the predicted values compared to the actual one. The results show the ability of the model to recognize the pattern of the time series data.

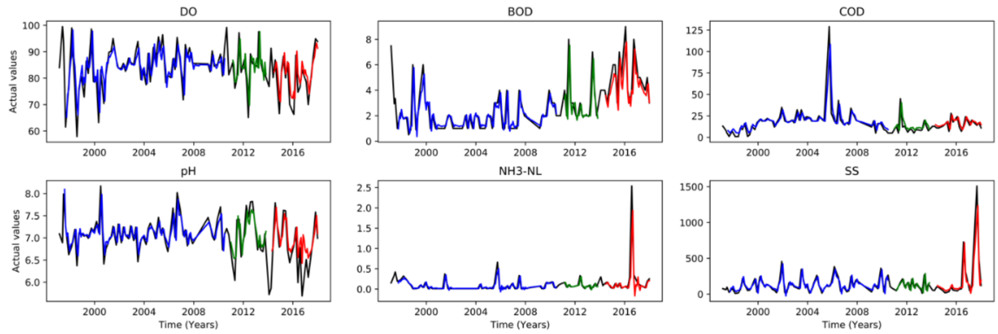

Figure 4. The plot of time series prediction against the actual data of WQP in site 1. The black, blue, green and red line represent actual data, prediction of train, validation, and test data, respectively.

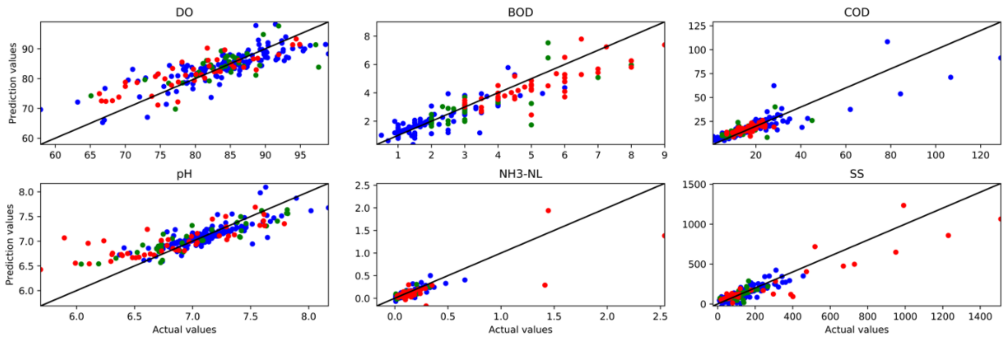

Figure 5. The actual vs predicted data of WQP in site 1. The blue, green, and red points represent train, validation, and test data, respectively.

The validation parameter, i.e., RMSE, MAPE, and PCC is represented in Table 5. To obtain the general insight, we calculate the average values of the validation parameter of each WQP and each station as provided in Tables 5 and 6, respectively. In the case of WQP, we found that the COD parameter presents the lowest RMSE values for the train, and test set, while the SS parameter presents the lowers RMSE values for the validation set. According to MAPE calculation, we found that the pH parameter presents the lowest MAPE values for train, validation, and test set. From PCC analysis, we found that the parameter with the highest PCC values for the train set is COD, while the highest values for validation and test set is pH. By considering the PCC of the test set as the overall performance, we found that the best results were obtained from the prediction of pH. This is related to the statistical properties of the pH that has the smallest value of kurtosis. This indicates that the small appearance of outliers in the pH dataset leads to good performance. We also found that the average value of PCC for all parameters is more than 0.60, which indicates the validity of the ANN models.

Table 5. The average of validation parameter for each WQP.

	RMSE			MAPE			PCC		
Parameter	**Train**	**Val.**	**Test**	**Train**	**Val.**	**Test**	**Train**	**Val.**	**Test**
DO	0.119	0.189	0.204	3.165	5.592	5.377	0.783	0.755	0.703
BOD	0.090	0.265	0.178	23.455	25.408	19.237	0.835	0.777	0.749
COD	**0.089**	0.167	**0.090**	29.131	27.649	18.135	**0.886**	0.731	0.678
pH	0.094	0.133	0.160	**2.109**	**3.344**	**4.217**	0.872	**0.871**	**0.860**
NH_3-NL	0.094	0.090	0.187	132.383	82.324	64.321	0.768	0.733	0.753
SS	0.117	**0.066**	0.114	127.356	164.008	152.258	0.669	0.739	0.806

In the case of the validation parameter for each station, we found the lowest RMSE of WQP for train, validation, and test set are found in the prediction of WQP in stations 2, 3, and 10, respectively. As for MAPE analysis, we found that the prediction of WQP in station 6 gives the lowest value of MAPE for a train set, while the prediction of WQP in station 1 gives the lowest MAPE for validation and test set. As for PCC, the highest value of PCC for train, validation, and test set found in the prediction of WQP in stations 2, 3, and 1, respectively. The PCC values of the test set pointed out that the best prediction of WQP was obtained from the prediction of WQP in station 1. However, the prediction of WQP in station 2 and 3 also gave a satisfying result. This relates to the number of data available for those stations that are more than the available data in the other station, except station 8. This indicates that the number of data used in the development of the model gave a contribution to the performance of the model.

Table 6. The average of validation parameter for each station.

	RMSE			MAPE			PCC		
Station No	**Train**	**Val.**	**Test**	**Train**	**Val.**	**Test**	**Train**	**Val.**	**Test**
1	0.079	0.098	0.203	34.380	**27.455**	**23.287**	0.862	0.747	**0.821**
2	**0.069**	0.116	0.235	32.515	36.341	36.873	**0.911**	0.811	0.794
3	0.074	**0.067**	0.134	49.283	43.052	25.968	0.857	**0.858**	0.786
4	0.152	0.180	0.270	35.993	70.654	49.920	0.652	0.760	0.691
5	0.106	0.123	0.163	51.034	57.100	28.857	0.780	0.709	0.683
6	0.096	0.088	0.130	**31.384**	40.040	24.781	0.798	0.761	0.703
7	0.105	0.158	0.123	33.835	42.988	25.588	0.821	0.815	0.752
8	0.086	0.113	0.108	53.082	61.179	76.238	0.851	0.745	0.783
9	0.109	0.117	0.111	98.792	88.351	76.126	0.673	0.763	0.802
10	0.110	0.091	**0.096**	42.748	38.675	29.682	0.800	0.631	0.710
11	0.113	0.328	0.156	56.893	62.463	71.087	0.832	0.790	0.818
12	0.101	0.108	0.155	87.769	42.240	33.855	0.745	0.820	0.727
13	0.106	0.277	0.164	69.656	66.386	68.750	0.801	0.758	0.729
14	0.103	0.261	0.128	63.699	42.503	43.928	0.850	0.777	0.817

4. Conclusions

An artificial neural network (ANN) with three hidden layers was used to predict time series data of water quality parameters (WQP), i.e., dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), pH, ammonia nitrogen (NH3-NL), suspended solids (SS). The WQP were collected from 14 stations in the Kelantan River, Malaysia. The ANN model was trained by using the optimized value of look back and epochs number. The validation was performed by calculating the root mean square of error (RMSE), mean absolute percentage of error (MAPE), and Pearson correlation coefficient (PCC). As for WQP analysis, we found that the best results were obtained from the prediction of pH parameter. The lowest kurtosis values of pH indicated that the appearance of outliers had an impact on the model. As for WQP for each station, we found that the good prediction was obtained from the prediction of WQP in stations 1, 2, and 3. The more available data in those stations indicated that the number of data is important in the prediction of time series data.

Funding

This research was funded by Indonesia-Malaysia Grant, (Telkom University-Universiti Tenaga Nasional (UNITEN)), grant number 2020104TELCO.

Acknowledgments

The authors would like to acknowledge the financial support from Universiti Tenaga Nasional (UNITEN) and Universitas Telkom under the matching international grant.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Abba, S.I.; Hadi, S.J.; Abdullahi, J. River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques. *Procedia Comput. Sci.* **2017**, *120*, 75–82, https://doi.org/10.1016/j.procs.2017.11.212.
2. Mohiyaden, H.A.; Sidek, L.M.; Hayder, G.; Noh, M.N. Water quality assessment Klang river water treatment plants. *Int. J. Eng. Technol.* **2018**, *7*, 639–642, http://dx.doi.org/10.14419/ijet.v7i4.35.23075.
3. Venkatramanan, S.; Chung, S.Y.; Lee, S.Y.; Park, N. Assessment of river water quality via environmentric multivariate statistical tools and water quality index: A case study of Nakdong River Basin, Korea. *Carpathian J. Earth Environ. Sci.* **2014**, *9*, 125–132.
4. Ahmed, U.; Mumtaz, R.; Anwar, H.; Shah, A.A.; Irfan, R.; García-Nieto, J. Efficient water quality prediction using supervised machine learning. *Water (Switzerland)* **2019**, *11*, 1–14, https://doi.org/10.3390/w11112210.
5. Bagheri, M.; Akbari, A.; Ahmad, S. Advanced control of membrane fouling in filtration systems using artificial intelligence and machine learning techniques : A critical review. *Process Saf. Environ. Prot.* **2019**, *123*, 229–252, https://doi.org/10.1016/j.psep.2019.01.013.
6. Hassanpour, F.; Sharifazari, S.; Ahmadaali, K.; Mohammadi, S.; Sheikhalipour, Z. Development of the FCM-SVR Hybrid Model for Estimating the Suspended Sediment Load. *KSCE J. Civ. Eng.* **2019**, *23*, 2514–2523, https://doi.org/10.1007/s12205-019-1693-7.
7. Ehteram, M.; Ghotbi, S.; Kisi, O.; Ahmed, A.N.; Hayder, G.; Fai, C.M.; Krishnan, M.; Afan, H.A.; EL-Shafie, A. Investigation on the potential to integrate different artificial intelligence models with metaheuristic algorithms for improving river suspended sediment predictions. *Appl. Sci.* **2019**, *9*, 1–25, https://doi.org/10.3390/app9194149.
8. Nourani, V.; Elkiran, G.; Abba, S.I. Wastewater treatment plant performance analysis using arti ficial intelligence – an ensemble approach. **2018**, 2064–2076.
9. Rahman, H.A.A.; Ibrahim, N.A.; Abdul Hamid, A.A.; Abdul Hamid, T.H.T. Effect of 2014 massive flood on well water qualities: A case study on Kelantan River basin, Malaysia. *J. Water L. Dev.* **2018**, *38*, 127–136, https://doi.org/10.2478/jwld-2018-0049.
10. Tan, M.L.; Ibrahim, A.L.; Cracknell, A.P.; Yusop, Z. Changes in precipitation extremes over the Kelantan River Basin, Malaysia. *Int. J. Climatol.* **2017**, *37*, 3780–3797, https://doi.org/10.1002/joc.4952.
11. Pradhan, B.; Youssef, A.M. A 100-year maximum flood susceptibility mapping using integrated hydrological and hydrodynamic models: Kelantan River Corridor, Malaysia. *J. Flood Risk Manag.* **2011**, *4*, 189–202, https://doi.org/10.1111/j.1753-318X.2011.01103.x.
12. Perera, E.D.P.; Lahat, L. Fuzzy logic based flood forecasting model for the Kelantan River basin, Malaysia. *J. Hydro-Environment Res.* **2015**, *9*, 542–553, https://doi.org/10.1016/j.jher.2014.12.001.
13. Nashwan, M.S.; Ismail, T.; Ahmed, K. Flood susceptibility assessment in Kelantan river basin using copula. *Int. J. Eng. Technol.* **2018**, *7*, 584–590, https://doi.org/10.14419/ijet.v7i2.8876.
14. Hee, Y.Y.; Suratman, S.; Aziz, A.A. Water Quality and Heavy Metals Distribution in Surface Water of the Kelantan River Basin (Malaysia). *Orient. J. Chem.* **2019**, *35*, 1254–1264, https://doi.org/10.13005/ojc/350402.
15. Yen, T..; Rohasliney, H. Status of Water Quality Subject to Sand Mining in the Kelantan River, Kelantatan. *Trop. Life Sci. Res.* **2013**, *24*, 19–34.

16. Tan, M.L.; Ibrahim, L.; Cracknell, P.; Yusop, Z. Changes in precipitation extremes over the Kelantan River Basin , Malaysia. **2016**, https://doi.org/10.1002/joc.4952.
17. Haykin, S. Neural networks: a comprehensive foundation by Simon Haykin. *Knowl. Eng. Rev.* **1999**, *13*, 409–412.