JEE Journal of Ecological Engineering
Volume 22, Issue 1, January 2021, pages 250–257
https://doi.org/10.12911/22998993/129579

Received: 2020.09.10
Revised: 2020.10.19
Accepted: 2020.11.05
Available online: 2020.12.01

Predicting Water Quality Parameters in a Complex River System

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

[1] School of Computing, Telkom University, 40257 Bandung, Indonesia

[2] Research Centre of Human Centric Engineering, Telkom University, 40257 Bandung, Indonesia

[3] Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor Darul Ehsan, Malaysia

[4] Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor Darul Ehsan, Malaysia

[5] College of Graduate Studies, Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor Darul Ehsan, Malaysia

[6] Department of Chemistry, Kaduna State University (KASU), Tafawa Balewa Way, P.M.B. 2339, Kaduna, Nigeria

* Corresponding author's e-mail: hauwa.mustafa@uniten.edu.my

ABSTRACT

This research applied a machine learning technique for predicting the water quality parameters of Kelantan River using the historical data collected from various stations. Support Vector Machine (SVM) was used to develop the prediction model. Six water quality parameters (dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammonia nitrogen (NH_3-N), and suspended solids (SS)) were predicted. The dataset was obtained from the measurement of 14 stations of Kelantan River from September 2005 to December 2017 with a total sample of 148 monthly data. We defined 3 schemes of prediction to investigate the contribution of the attribute number and the model performance. The outcome of the study demonstrated that the prediction of the suspended solid parameter gave the best performance, which was indicated by the highest values of the R2 score. Meanwhile, the prediction of the COD parameter gave the lowest score of R2 score, indicating the difficulty of the dataset to be modelled by SVM. The analysis of the contribution of attribute number shows that the prediction of the four parameters (DO, BOD, NH_3-N, and SS) is directly proportional to the performance of the model. Similarly, the best prediction of the pH parameter is obtained from the utilization of the least number of attributes found in scheme 1.

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

INTRODUCTION

Globally, rivers have been the most utilized natural water source due to their availability and accessibility; this has prompted the growth of civilization and industries close to river banks (Mustafa et al., 2017). However, in the last decades, there has been a high increase in river pollution due to the human-made activities and climate change (Mustafa and Hayder, 2020). Notwithstanding, the research has focused extensively on predicting the river water quality, contaminant classification and risk assessment strategies to formulate more effective management practices and advanced monitoring systems (Tiyasha et al., 2020). Similarly, water quality monitoring and prediction allows a manager to identify a suitable option that satisfies a wide range of conditions. The water parameters such as turbidity, electrical conductivity and dissolved solids in water, for example, describe a complex process controlled by ecological, hydrological and hydrodynamic factors that operate at a wide range of spatiotemporal scales (Najah et al., 2009).

Furthermore, the water quality index (WQI) analysis of rivers is a popular topic in physical sciences, which involves the calculation and description of the water quality parameters and the contamination transmission mechanism. Moreover, the advent of innovative soft computing and artificial intelligence (AI) techniques have led researchers in evaluating the component of water quality and their internal relationship in time series. Recent studies have reported the applications of the artificial intelligence-based methods in the addressing water resources management issues (Slaughter et al., 2017; Tomas et al., 2017; Wu et al., 2018). Similarly, radial basis network (RBF), multilayer perceptron (MLP), and adaptive neuro-fuzzy inference system (ANFIS) were observed to be suitable in predicting the water quality parameters of Karoon River (Emamgholizadeh et al., 2013). Additionally, Najah et al. (2009) used the artificial neural network (ANN) approaches in predicting the water quality parameters of Johor River Basin. The outcome of the study indicated that the performance of the ANN models is efficient, as the mean absolute percentage error of 10% was obtained in the prediction of the water quality parameters. Zhang et al. (2010) proposed a tool for the water allocation schemes analysis of Jiaojiang River basin using the water quantity-quality model. Nikoo and Mahjouri (2013) applied fuzzy inference system and probabilistic support vector machines in estimating the probabilistic water quality of water resources. The outcome of the study indicated that the models could be used in feasibility studies of water conservation projects. (Antanasijević et al., 2014) estimated the dissolved oxygen (DO) concentration in Danube River using a general regression neural network (GRNN) model. The predicted outcome obtained from the study was compared with the output observed from the Monte Carlo simulations. The authors recommended that the GRNN model is an efficient tool for the estimation of the DO concentration in rivers. Heddam (2016a; 2016b) predicted the water quality parameters using ANN in several case studies. He claimed that the AI methods are sufficient for modelling the water quality parameters in time series. Elkiran et al. (2018) estimated the DO concentration of Mathura River in India using feed-forward neural network, multilinear regression, and ANFIS. In the study, DO concentration, biochemical oxygen demand (BOD), temperature and pH parameters of the river were used for the prediction. The findings obtained from the study indicated that the ANFIS models greatly improved the performance over the feed-forward neural network and multilinear regression in the validation step.

In view of the past research work mentioned, a comparative study on the implementation of the AI techniques using different software packages is necessary to improve the accuracy level and its applications. However, several data analysis programs do not involve comprehensive modification in the implementation of the AI techniques. Hence, this research study explores the ratification of one of AI approaches, namely support vector machines (SVM), for monitoring and predicting river water quality parameters.

METHODOLOGY

Collection of Data

In this study, the historical dataset of water quality parameters (dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), pH, ammonia nitrogen (NH_3-N), and suspended solids (SS)) was used. The dataset was obtained from the measurement of 14 stations of Kelantan River from September 2005 to December 2017 with a total sample of 148 monthly data. Missing values were found in the dataset, since no measurements were performed on that day. The missing values were filled by using the interpolation method. The location of the measurement station along the river is presented in Figure 1.

Prediction Model Scheme

Figure 1 depicted the water flow of the river to the area of measurement station 1 (area 1). Hence, the water properties in area 1 are affected by the water quality in other areas. This is because water from all areas gathers and flows to area 1. However, to improve the measurement efficiency, we can replace the conventional water quality measurement in area 1 by using a prediction model. The model for six parameters was developed by using the data of water quality in other areas as input variables. In this case, we considered the prediction of a parameter that is affected by the value of other parameters. For example, the value of the COD parameter is utilized to predict the value of the BOD parameter. Therefore, the number of input variables is equal to the multiplication of the number of areas by six parameters. The

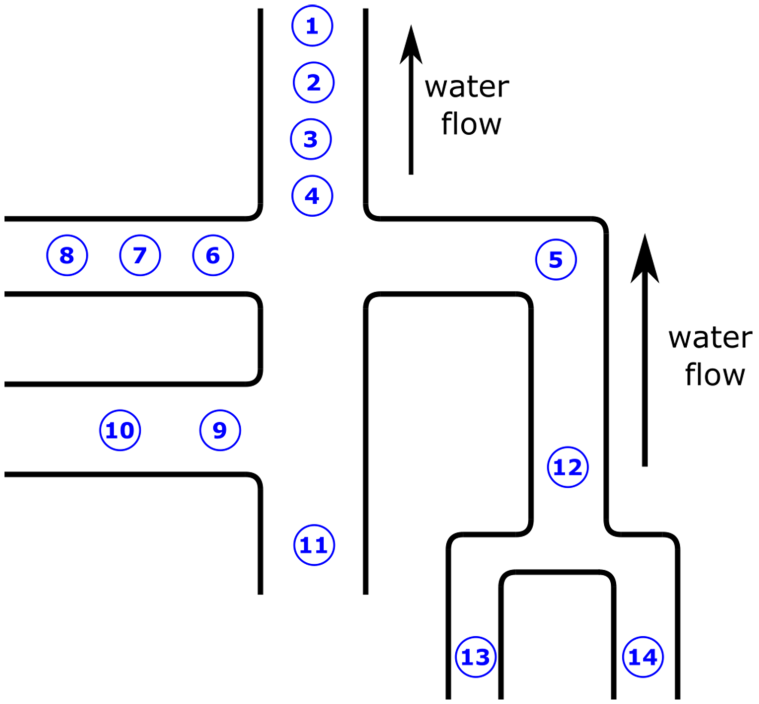

Figure 1. The map of measurement station along the Kelantan River

development of the model was performed using 3 schemes. Those schemes differ by the number of areas that are considered to affect the water quality in area 1. In the first scheme, we considered the edge areas only, i.e. area 8, 10, 11, 13, 14, with the total number of input variables of 30. Meanwhile, the second scheme was conducted by considering the edge and branch areas, i.e. area 4, 8, 10, 11, 12, 13, 14, with the total number of input variables as 42. In the third scheme, we considered all the remaining areas, i.e. area 1–13, with the total number of input variables of 78. The detailed information of those prediction model schemes can be seen in Table 1.

Support Vector Machine

In this present study, the model prediction was generated using Support Vector Machine (SVM). SVM is a branch of machine learning (ML) technique developed using the theory of statistical learning. The basic principle of the SVM implementation in pattern recognition is the mapping of the input vectors into a possibly higher dimension of feature space, either linearly or non-linearly. The mapping process is controlled by the type of kernel function. Then, an optimal hyperplane is constructed to obtain the maximal separation of two classes, or extended to multi-class. The SVM training is performed by seeking a globally optimized solution and managing the over-fitting problem. Therefore, the SVM method has an advantage in processing a large number of features (Vapnik, 1998). SVM is also known as the largest margin classifier, since this method tries to find an optimal hyperplane that results in the largest margin. The representation of the hyperplane and margin used in SVM is presented in Figure 2.

The main goal of SVM is to construct a classifier from the available samples by avoiding misclassifying in future predictions. The separating hyperplane used in the classifier is expressed as $\overrightarrow{W} \cdot \vec{x} + b = 0$, which refers to the formulation of $y_i(\overrightarrow{W} \cdot \vec{x} + b) \geq 1, i = 1, \ldots, N$. During the training, SVM will look for an optimal separating hyperplane by minimizing $(1/2)\|\overrightarrow{W}\|^2$ subject to the constraint. In this case, $\|\overrightarrow{W}\|^2$ represents the Euclidean norm of $\vec{w}$, which maximizes the distance between the hyperplane and support

Table 1. The detail information of 3 prediction model schemes

Scheme	Considered Areas	No. of input variables
1	Edge areas: Area 8, 10, 11, 13, 14	5 x 6 = 30
2	Edge and branch areas: Area 4, 8, 10, 11, 12, 13, 14	7 x 6 = 42
3	All the remaining areas: Area 1–13	13 x 6 = 78

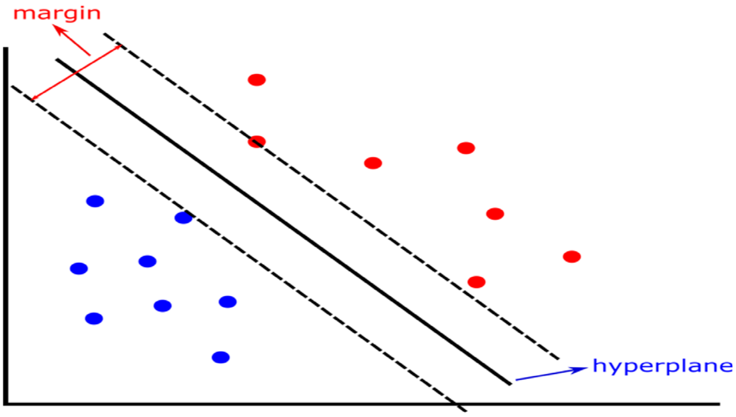

Figure 2. The representation of hyperplane and margin in SVM

vectors. The training procedure of SVM is converted into convex Quantum Programming (QP) problem by utilizing Lagrange multipliers. The solution of the QP problem is represented as a global optimal expressed as:

$$\vec{W} = \sum_{i=1}^{N} y_i \alpha_i \cdot \vec{x}_i \tag{1}$$

where: $\vec{x}_i$ represents support vector when $\alpha_i > 0$. After the training process, the decision function used in prediction is formulated as:

$$f(\vec{x}) = sgn\left(\sum_{i=1}^{N} y_i \alpha_i \cdot \vec{x} \cdot \vec{x}_i + b\right) \tag{2}$$

where: *sgn()* represents the given sign function.

Moreover, to allow errors during the training, slack variable (ζ) with $\zeta_i > 0, i = 1, \ldots, N$ were introduced by Cortes and Vapnik (Vapnik, 1995). This technique is known as a soft margin, which is effective in preventing overfitting. By considering the slack variable, the relaxed separation constraint is formulated as

$$y_i(\vec{W} \cdot \vec{x} + b) \geq 1 - \zeta_i, i = 1, \ldots, N \tag{3}$$

and the optimal hyperplane is obtained by minimizing

$$\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N} \zeta_i \tag{4}$$

where: C represents a regularization parameter that controls a trade-off between the optimal margin and training error. Similarly, to obtain an optimal hyperplane, the input vector was mapped into a higher dimensional Hilbert space, in which the process is controlled by the kernel function. The kernel functions that are commonly used in the SVM model are RBF, linear, and polynomial kernel function. The polynomial kernel function can be expressed as:

$$K(x, y) = (\langle x, y \rangle + 1)^E \tag{5}$$

where: E represents the exponent value. In the case of the linear kernel, the value of the exponent value is 1. Meanwhile, the RBF kernel function can be expressed as:

$$K(x, y) = e^{-(\gamma \cdot \langle x-y, x-y \rangle^2)} \tag{6}$$

Hyperparameter Tuning

The performance of the SVM model was improved by performing a hyperparameter tuning procedure. This process aims to obtain the optimal parameter that will be used in model development. The SVM parameter that is tuned in this step consists of a regularization parameter (C), kernel coefficient (gamma), and kernel function. The option of parameter values used in the hyperparameter tuning is presented in Table 2.

Model Validation

The performance of the SVM model was measured by calculating two validation parameters, i.e. coefficient correlation (R2) and mean square of error (MSE). The parameters were used as a reference to determine the validity of the model for each scheme and parameter. These parameters were formulated as:

$$R2 = 1 - \frac{\sum_{i=1}^{n}(A_i - P_i)^2}{\sum_{i=1}^{n}(A_i - \bar{A})^2} \tag{7}$$

Table 2. Parameter and values option used in hyperparameter tuning

Parameter	Values Option
C	[0.1, 1, 10, 100]
Gamma	[“auto”, “scale”]
Kernel	[“rbf”, “poly”, “sigmoid”]

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(A_i - P_i)^2 \tag{8}$$

where: A_i, $\bar{A}$ and P_i represent the actual values in i-th month, the average of actual values and predicted values, respectively, while n represents the number of data. In the case of MSE, we calculated those parameters by using a scaled dataset to allow the comparison of the results amongst the water quality parameters.

RESULTS AND DISCUSSIONS

Hyperparameter tuning

The performance of the SVM model was improved by conducting hyperparameter tuning for each scheme and parameter. The optimized parameter of the SVM model for schemes 1, 2 and 3 are presented in Tables 3, 4 and 5, respectively. We found that the sigmoid kernel function is not suitable for our study, as this function was not chosen from the hyperparameter tuning results. The chosen optimized kernel function for all parameters is the RBF function, except for the parameter for NH_3-N. The optimized values of the regularization parameter (C) are varied for each water quality parameter. This is related to the tolerance level of the SVM model to accept errors during the training. The variation of the C parameter reflected the different characteristics of the dataset of water quality parameters.

Model validation

The SVM models developed by the optimized hyperparameter were evaluated by comparing the predicted values with the actual ones. The plot of predicted values against the actual ones of scheme 1 is presented in Figure 3. According to Figure 3, we found that all of the data points were close to the straight diagonal line, except the BOD parameter, indicate low values of error. We also found that the deviation of the data points of the BOD parameter is quite large compare to other parameters.

The results of the validation parameter, i.e. R2 and MSE, for schemes 1, 2 and 3 are presented in Tables 6, 7 and 8, respectively. As for scheme 1, we found that the R2 score of train data for all water quality parameters is more than 0.80, which signifies a satisfactory result in predicting the train data. However, the true quality of the model is evaluated according to the ability in predicting the external data as represented by the R2 score of test data. We found that the prediction of the SS parameter gave the best performance with an R2 score of 0.901. Meanwhile, the worst performance was found in the prediction of the COD parameter with an R2 score of 0.241. This indicates that the data set of the COD parameter is more complex than others. In this case, the number of used attribute seems not enough to reveal the pattern of the COD data set.

As for scheme 2, we found that the R2 of train data for all the water quality parameters is satisfactory, as all the R2 values were observed to be more than 0.90. However, the R2 of test data is different for each parameter. The best performance is obtained from the prediction of the SS parameter with an R2 score of 0.940. Meanwhile, the prediction of COD gives the worst performance with an R2 score of 0.499. By comparing the results of COD prediction in scheme 1, we found that the addition of attribute in scheme 2 improves the R2 score from 0.241 to 0.449. Even

Table 3. Optimized SVM parameter used in scheme 1

Water Parameter	SVM Parameter		
	C	Gamma	Kernel
DO	100	auto	rbf
BOD	10	scale	rbf
COD	100	scale	rbf
pH	10	auto	rbf
NH_3-N	1	scale	poly
SS	100	auto	rbf

Table 4. Optimized SVM parameter used in scheme 2

Water Parameter	SVM Parameter		
	C	Gamma	Kernel
DO	10	auto	rbf
BOD	10	auto	rbf
COD	100	scale	rbf
pH	10	scale	rbf
NH_3-N	10	scale	poly
SS	100	auto	rbf

Table 5. Optimized SVM parameter used in scheme 3

Water Parameter	SVM Parameter		
	C	Gamma	Kernel
DO	10	auto	rbf
BOD	10	auto	rbf
COD	100	auto	rbf
pH	10	scale	rbf
NH_3-N	100	scale	poly
SS	100	auto	rbf

Table 6. The results of the validation parameter for scheme 1

Parameter	R2		MSE	
	Train	Test	Train	Test
DO	0.992	0.710	0.008	0.290
BOD	0.992	0.603	0.008	0.397
COD	0.992	0.241	0.008	0.759
pH	0.992	0.857	0.008	0.143
NH_3-N	0.886	0.636	0.820	0.874
SS	0.992	0.901	0.008	0.099

though the R2 score is still low, the improvement indicates that the number of attributes contributed to the R2 score of COD prediction.

As for scheme 3, we found that the R2 score of the train data for all water quality parameters is good with the score of more than 0.90. According to the R2 score of test data, we found that the prediction of the SS parameter gives the best result with an R2 score of 0.936. Meanwhile, the worst performance is obtained from the prediction of COD with an R2 score of 0.490. The value of the R2 score of COD prediction in scheme 3 is not significantly different compared to the value in scheme 2. This indicates that the addition of attribute in scheme 3 failed to improve the results of COD prediction. Generally, the best and worst results of all schemes were obtained from the prediction of the SS and COD parameters, respectively. This indicate that the data quality of the

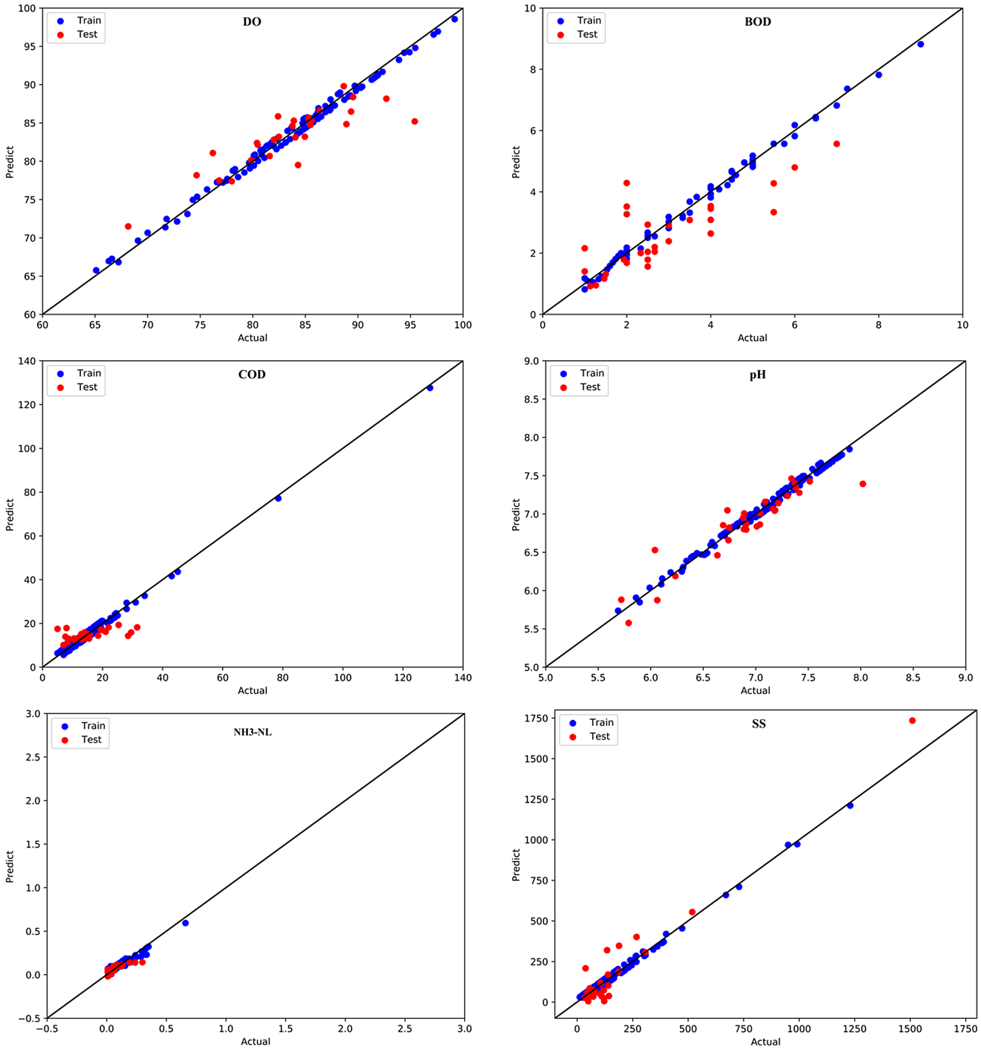

Figure 3. The plot of actual and predicted values of the water quality parameters obtained from scheme 1

Table 7. The results of the validation parameter for scheme 2

Parameter	R2		MSE	
	Train	Test	Train	Test
DO	0.992	0.742	0.008	0.258
BOD	0.993	0.668	0.007	0.332
COD	0.992	0.499	0.008	0.501
pH	0.993	0.834	0.007	0.166
NH_3-N	0.992	0.810	0.008	0.190
SS	0.992	0.940	0.008	0.060

Table 8. The results of the validation parameter for scheme 3

Parameter	R2		MSE	
	Train	Test	Train	Test
DO	0.992	0.773	0.008	0.227
BOD	0.992	0.691	0.008	0.309
COD	0.993	0.490	0.007	0.510
pH	0.993	0.815	0.007	0.185
NH_3-N	0.992	0.843	0.008	0.157
SS	0.993	0.936	0.007	0.064

SS and COD parameters is quite similar for each measurement site.

The contribution of the attribute number in each scheme on the model performance was investigated by comparing the R2 score of test data for all water quality parameters, as presented in Figure 4. The number of the attribute from scheme 1 to scheme 3 is increased and leads to the increasing of the model complexity. The positive correlation was found in the R2 scores of the DO, BOD, NH_3-N and SS parameters. In these parameters, the increase of the attribute number leads to an increasing in the R2 score. This shows that the attribute number can improve the performance of the model. Conversely, the R2 score of the pH parameter decreases as the addition of the attribute number increases. This point out that the increasing of attribute number lead to too complex model and caused overfitting state. In the case of the prediction of the COD parameter, we found that the best R2 score was obtained from scheme 2. However, the difference in the R2 score between scheme 2 and scheme 3 is not significant. The overall results reveal the importance of the attribute number to obtain satisfying results. Moreover, we found that no scheme that gives the best performance for all parameters.

CONCLUSION

The values of six water quality parameters, i.e. dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), pH, ammonia nitrogen (NH_3-N) and suspended solids (SS) of station 1 were predicted by using the SVM model. The prediction was performed by defining 3 schemes according to the number of attributes used for model development. Amongst the water quality parameters, the prediction of the SS parameter gave the best results with the highest values of the R2 score for both the train and test data. Meanwhile, the worst results were obtained from the prediction of the COD parameter. Regarding the contribution of attribute number in each scheme, we found that the prediction of four parameters, i.e. the DO, BOD, NH_3-N and SS parameters, were improved as the contribution of the attribute number increases. Conversely, the best prediction of the pH parameter was obtained from scheme 1 with the least number of attributes.

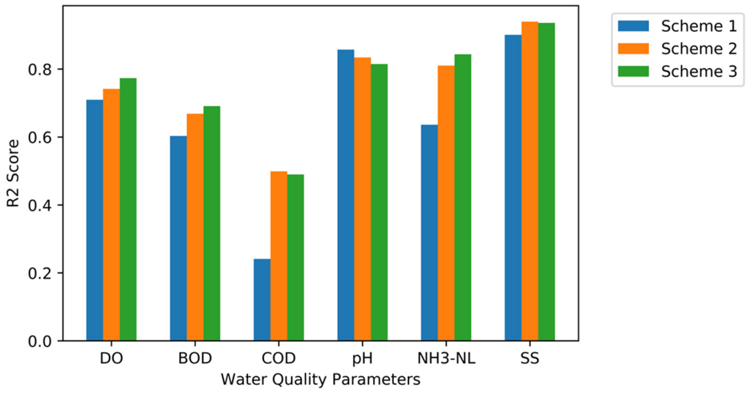

Figure 4. The comparison of R2 score of test data of water quality parameters calculated by using different schemes

Acknowledgement

The authors would like to acknowledge the support from the Indonesia-Malaysia Grant (Telkom University-Universiti Tenaga Nasional (UNITEN)) for the financial support under TELCO Grants 2019 with reference number 2020104TELCO.

REFERENCES

1. Antanasijević, D., Pocajt, V., Perić-Grujić, A., Ristić, M., 2014. Modelling of dissolved oxygen in the danube river using artificial neural networks and Monte carlo simulation uncertainty analysis. J. Hydrol. 519, 1895–1907. https://doi.org/10.1016/j.jhydrol.2014.10.009
2. Elkiran, G., Nourani, V., Abba, S.I., Abdullahi, J., 2018. Artificial intelligence-based approaches for multi-station modelling of dissolve oxygen in river. Glob. J. Environ. Sci. Manag. 4, 439–450. https://doi.org/10.22034/gjesm.2018.04.005
3. Emamgholizadeh, S., Bateni, S.M., Jeng, D.S., 2013. Artificial intelligence-based estimation of flushing half-cone geometry. Eng. Appl. Artif. Intell. 26, 2551–2558. https://doi.org/10.1016/j.engappai.2013.05.014
4. Heddam S. 2016a. Multilayer perceptron neural network-based approach for modeling phycocyanin pigment concentrations: case study from lower Charles River buoy, USA. Environ. Sci. Pollut. Res. 23, 17210–17225. https://doi.org/10.1007/s11356–016–6905–9
5. Heddam S. 2016b. Generalized regression neural network-based approach as a new tool for predicting total dissolved gas (TDG) downstream of spillways of dams: a case study of columbia river basin dams, USA. Environ. Process. 4, 235–253. https://doi.org/10.1007/s40710–016–0196–5
6. Mustafa, A., Sulaiman, O., Shahooth, S., 2017. Application of QUAL2K for Water Quality Modeling and Management in the lower reach of the Diyala river. Iraqi J. Civ. Eng. 11, 66–80.
7. Mustafa, H.M., Hayder, G., 2020. Recent studies on applications of aquatic weed plants in phytoremediation of wastewater: A review article. Ain Shams Eng. J. https://doi.org/10.1016/j.asej.2020.05.009
8. Najah, A., Elshafie, A., Karim, O.A., Jaffar, O., 2009. Prediction of johor river water quality parameters using artificial neural networks. Eur. J. Sci. Res. 28, 422–435.
9. Nikoo, M.R., Mahjouri, N., 2013. Water Quality Zoning Using Probabilistic Support Vector Machines and Self-Organizing Maps. Water Resour. Manag. 27, 2577–2594. https://doi.org/10.1007/s11269–013–0304–5
10. Slaughter, A.R., Hughes, D.A., Retief, D.C.H., Mantel, S.K., 2017. A management-oriented water quality model for data scarce catchments. Environ. Model. Softw. 97, 93–111. https://doi.org/10.1016/j.envsoft.2017.07.015
11. Tiyasha, Tung, T.M., Yaseen, Z.M., 2020. A survey on river water quality modelling using artificial intelligence models: 2000–2020. J. Hydrol. 585, 124670. https://doi.org/10.1016/j.jhydrol.2020.124670
12. Tomas, D., Čurlin, M., Marić, A.S., 2017. Assessing the surface water status in Pannonian ecoregion by the water quality index model. Ecol. Indic. 79, 182–190. https://doi.org/10.1016/j.ecolind.2017.04.033
13. Vapnik, V., 1998. Statistical Learning Theory. Wiley-Interscience, New York.
14. Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, Berlin.
15. Wu, Z., Wang, X., Chen, Y., Cai, Y., Deng, J., 2018. Assessing river water quality using water quality index in Lake Taihu Basin, China. Sci. Total Environ. 612, 914–922. https://doi.org/10.1016/j.scitotenv.2017.08.293
16. Zhang, W., Wang, Y., Peng, H., Li, Y., Tang, J., Wu, K.B., 2010. A coupled water quantity-quality model for water allocation analysis. Water Resour. Manag. 24, 485–511. https://doi.org/10.1007/s11269–009–9456–8