scientific reports

OPEN

Machine learning algorithm as a sustainable tool for dissolved oxygen prediction: a case study of Feitsui Reservoir, Taiwan

Balahaha Fadi Ziyad Sami[1], Sarmad Dashti Latif[2], Ali Najah Ahmed[1], Ming Fai Chow[3]✉, Muhammad Ary Murti[4], Asep Suhendi[4], Balahaha Hadi Ziyad Sami[1], Jee Khai Wong[1], Ahmed H. Birima[5] & Ahmed El-Shafie[6,7]

Water quality status in terms of one crucial parameter such as dissolved oxygen (D.O.) has been an important concern in the Fei-Tsui reservoir for decades since it's the primary water source for Taipei City. Therefore, this study aims to develop a reliable prediction model to predict D.O. in the Fei-Tsui reservoir for better water quality monitoring. The proposed model is an artificial neural network (ANN) with one hidden layer. Twenty-nine years of water quality data have been used to validate the accuracy of the proposed model. A different number of neurons have been investigated to optimize the model's accuracy. Statistical indices have been used to examine the reliability of the model. In addition to that, sensitivity analysis has been carried out to investigate the model's sensitivity to the input parameters. The results revealed the proposed model capable of capturing the dissolved oxygen's nonlinearity with an acceptable level of accuracy where the R-squared value was equal to 0.98. The optimum number of neurons was found to be equal to 15-neuron. Sensitivity analysis shows that the model can predict D.O. where four input parameters have been included as input where the d-factor value was equal to 0.010. This main achievement and finding will significantly impact the water quality status in reservoirs. Having such a simple and accurate model embedded in IoT devices to monitor and predict water quality parameters in real-time would ease the decision-makers and managers to control the pollution risk and support their decisions to improve water quality in reservoirs.

Reservoirs water considers one of the most crucial sources for household needs, irrigation, and other purposes such as industrial needs[1]. However, reservoir's water quality is susceptible to deterioration[2]. The reservoir's water quality status is measured based on three different properties such as physical, chemical, and biological[3,4]. Various water quality parameters are measured for each mentioned property to evaluate water quality. Therefore, there is a need to accurately model these parameters due to their importance for better management and mitigating any risk associated with sustaining the quality within the acceptable level[5]. Dissolved Oxygen (D.O.) is among the most critical parameters in measuring water quality status[6]. Among all the water quality parameters, the Dissolved Oxygen (D.O.) is considered the most representative parameter that showed the class's water quality status, especially in surface water. This is due to the fact that D.O. is vital for the aquatic organisms and fish in the water bodies. The level of dissolved oxygen is a reflection of wind and aerating action. The D.O. level must be within amount to assure the stability of organisms and fish life in the water bodies; the higher the D.O., the better the condition aquatic and fish survival. To indicate the state of any aquatic system, D.O. is used as an indicator, and it is essential for microorganisms when its present in water column[7].

[1]Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor, Malaysia. [2]Civil Engineering Department, College of Engineering, Komar University of Science and Technology, Sulaimany, Kurdistan Region 46001, Iraq. [3]Discipline of Civil Engineering, School of Engineering, Monash University Malaysia, Jalan Lagoon Selatan, 47500 Bandar Sunway, Selangor, Malaysia. [4]School of Electrical Engineering, Telkom University, Bandung, Indonesia. [5]Department of Civil Engineering, College of Engineering, Qassim University, Unaizah, Saudi Arabia. [6]Department of Civil Engineering, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia. [7]National Water and Energy Center, United Arab Emirates University, Al Ain, United Arab Emirates. ✉email: chow.mingfai@monash.edu

Deterministic and stochastic models are used to model the D.O. concentration changes and capture any pattern from the measured data; However, these models require massive data to model the D.O. pattern and consider very complex[8]. Other models, such as the statistical model introduced to overcome the conventional models. Since many factors impact the concentration of D.O. in the reservoir, which can cause to nonlinearity pattern, the statistical model fails to capture it since it assumes that the relationship between D.O. and other parameters is linear[9]. Alternatively, Machine Learning (ML) techniques have been proposed as an other technique to capture the nonlinearity in any complex system[10,11].

Artificial Neural Network (ANN) methods were used in conjunction with numerical simulation models to boost the simulation results[12]. Recently, ML techniques have been used intensively in modeling complex parameters related to water resources, such as predicting sea-level rise[13–15], rainfall prediction[16,17], reservoir water level prediction[18,19], and streamflow forecasting[11,20,21]. Inspired by the robust performance of ML in capturing the nonlinearity patterns in most of the engineering systems, different algorithms of ML have been adopted to predict the water quality parameters. Predicting the class Water Quality Index (W.Q.I.) has been carried out using different ML algorithms by many researchers[22–24]. Artificial Neural Network (ANN) has been used to predict total nitrogen and phosphorus in the United States (U.S.) lakes[25]. At the same time, a support vector machine (SVM) was developed to predict the concentration of biological oxygen demand (B.O.D.) at the Johor river, Malaysia[26].

Regarding dissolved oxygen concentrations, an adaptive neuro-fuzzy inference system (A.N.F.I.S.) was proposed to predict D.O. at the Johor river, Malaysia[8]. However, the limitations of the A.N.F.I.S. model were reported by Ahmadlou et al.[27]. These drawbacks are that it is not very accurate and cannot find the best parameters; it is also prone to get stuck in a local minimum, contributing to its lack of prediction abilities.

A model was developed by Heddam[28] to predict dissolved oxygen concentration using an optimally pruned extreme learning machine (O.P.E.L.M.). The study found that O.P.E.L.M. provided a reasonable estimate of D.O. However, Sánchez-Monedero et al.[29] found that O.P.E.L.M. tends to degrade too many neurons, which results in noticeable performance degradation in some data sets.

The least-squares support vector (L.S.S.V.R.) has been proposed by Liu et al.[30] to predict the amount of dissolved oxygen in intensive anaerobic ponds. However, It has been found that L.S.S.V.R. performance depends heavily on selecting the kernel coefficient and regularization coefficient, which are necessary for the optimization process and the final L.S.S.V.R. model. Regrettably, there is no unique, perfect method to specify the given parameters in the L.S.S.V.R. model. Extreme machine learning was developed by[31] and used to predict dissolved oxygen in urban rivers. In addition to that, recently, the concentration of dissolved oxygen in fishery pond was predicted using a gated recurrent unit[32].

To overcome the inherent limitations established by standalone models, hybrid models have been proposed to optimize these algorithms' hyper-parameters by augmenting them with different optimization algorithms. For instance, different hybrid models have been developed and used to predict dissolved oxygen concentration[33–35]. Teaching–learning-based optimization algorithm (T.L.B.O.) is used to predict dissolved oxygen[36]. Various regression equations were optimized, including quadratic, exponential, logarithmic, and linear using T.L.B.O. Then the findings from T.L.B.O. compared with an artificial bee colony (A.B.C.) optimizer. Better results were obtained by hybridizing the quadratic regression equation with T.L.B.O. Besides the hybridized model's complexity, the authors used many parameters (twenty parameters) as inputs to develop the model. One of the drawbacks of such a model is to have access to a significant amount of available and reliable water quality parameters data, which is challenging.

Despite the acceptable performances these models achieved, however, there are few limitations associated with the hybridization of ML. One of these limitations is the complexity and complicated architecture and the difficulties in initializing the input parameters for these hybrid models[37]. Kumar et al.[38] found that the artificial neural network's prediction performance can be enhanced by improving the training approach without hybridizing it with optimization algorithms. In addition to that, a recent study highlighted the importance of the input combinations of ML algorithms' output accuracy, where the optimal input combinations can lead to a high level of accuracy without the need to augment ML with optimizers[39,40].

Therefore, this study's chief aim is to propose an artificial intelligence model with simple architecture and a high-performance level to predict dissolved oxygen concentrations. This study will use historical data recorded for 29 years from the Fei-Tsui Reservoir to train the model to accomplish this goal. The number of neurons will be optimized in order to obtain the desired results. Different input combinations will be investigated and examined to enhance the model's performance. Recently many researchers have been developing AI models with a few inputs[41]. For example, Moghadam et al. used four input parameters and DO concentration to predict DO concentration in three different lead times[42]. Therefore, in this study few input parameters will be investigated.

Sensitivity analysis and uncertainty analysis will be carried out to validate the proposed model. Different statistical indices will be introduced to inspect the proposed model performance. For better visualization, Taylor's diagram, violin plot, and percentage of relative error between the projected data and the observed one have been implemented in this study.

Methodology

Study area and data description.

Located in Taiwan's north region, the Fei-Tsui reservoir serves a 300 km^2 catchment area approximately, as shown in Fig. 1[43]. Since the 1980's, for Taipei city, the Fei-Tsui reservoir is considered the primary source of drinking water. One hundred fifty days is the approximate duration when the water resides in the reservoir. Since 1987, monthly measurements have been conducted to examine the reservoir's water status based on different water quality parameters. The water quality samples have been collected at the outlet of the five inflow tributaries of the reservoir and another seven sampling locations at the reservoir's

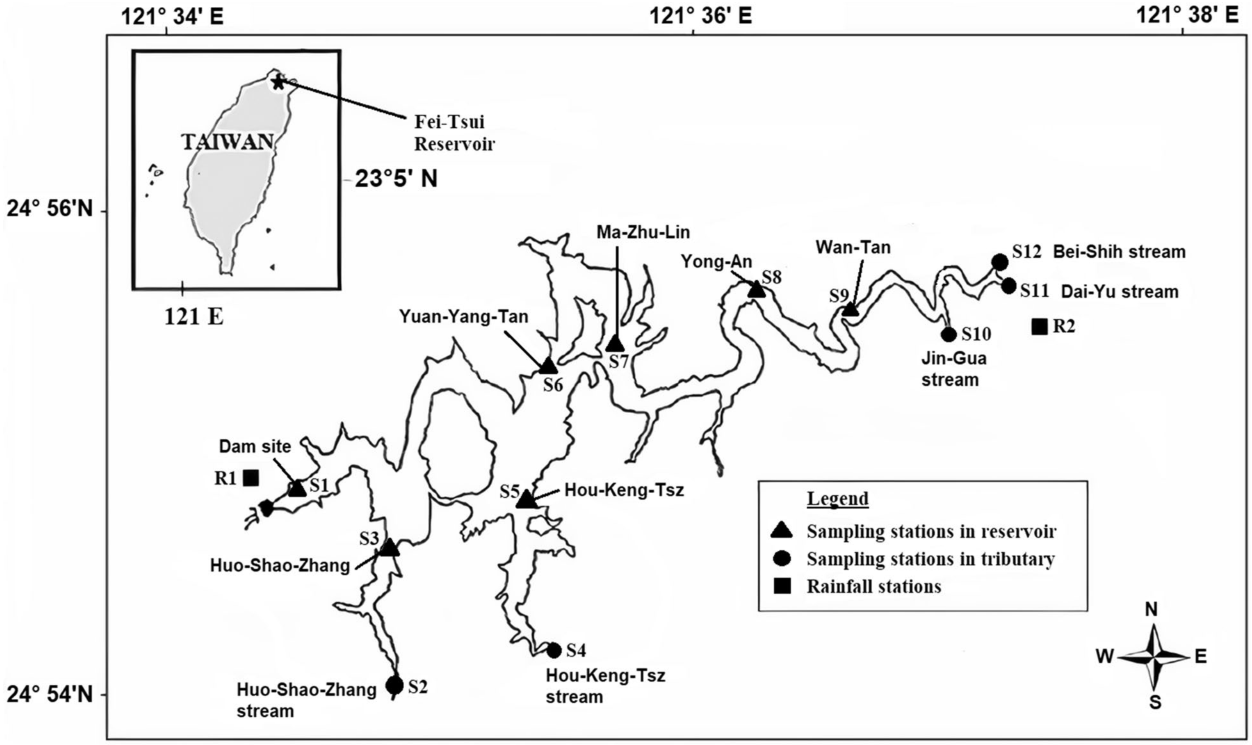

Figure 1. Location of Fei-Tsui Reservoir and sampling sites.

Mean	7.91
Standard error	0.10
Median	8
Mode	8
Standard deviation	0.58
Sample variance	0.34
Kurtosis	15.27
Skewness	− 3.35
Range	3.41
Minimum	5.27
Maximum	8.68
Sum	229.66

Table 1. Descriptive analysis of the observed dissolved oxygen (D.O.)

main lake[44]. The data was obtained from the administration office of the Taipei Fei-Tsui Reservoir. Table 1 shows the descriptive analysis of the measured Dissolved Oxygen (D.O.) concentrations.

Model development. An Artificial Neural Network with a single hidden layer was proposed to predict the dissolved oxygen concentration in the Fei-Tsui reservoir. The architecture of the proposed model can be seen in Fig. 2[45]. The proposed model consists of an input layer, which presents the input parameters that will be used to develop the model. In contrast, the output layer presents the model's output, which is the dissolved oxygen concentrations. Weights and biases connect the input and output layers to the hidden layer.

The hidden layer consists of several neurons. In this study, a different number of neurons will be investigated. In the beginning, the number of neurons will be set to equal five, then ten, fifteen, and finally, twenty. The predicted dissolved oxygen concentration will be compared with the observed concentration to choose the best model with the best-optimized number of neurons that give the lowest error. 29 years of monthly water quality data (348) will be used in developing the proposed model. 80% of the data will be used to train the model, while 20% will be used to test the model's accuracy. The pre-processing step was carried out by scaling the dataset between 0 to 1. Different types of activation functions and transfer functions will be explored and optimized.

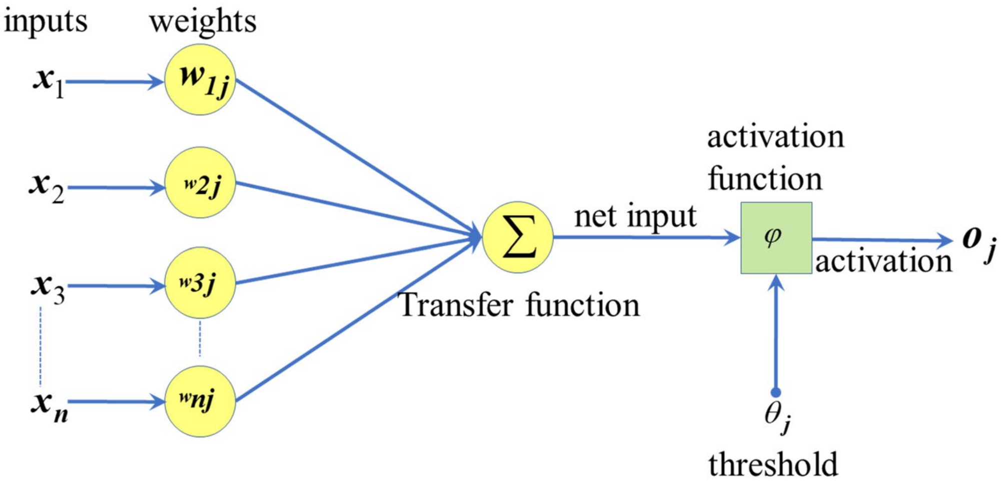

Figure 2. Structure of the proposed model.

Parameters		Water temperature ℃	BOD mg/L	Iron mg/L	Total organic carbon mg/L
DO	Average	24.14	0.70	0.09	1.05
	Min	23.32	0.37	0.03	0.72
	Max	25.13	1.43	0.49	2.18
	Standard deviation (SD)	0.44	0.23	0.08	0.35
	Coefficient of variation (CV)	1.83	33.72	97.77	34.30
	Coefficient of correlation	− 0.49	0.46	0.17	0.27

Table 2. Statistical analysis and coefficient of correlation between the input and the output parameters.

Similarly, different training algorithms will be investigated to develop a model with a high level of precision. Finally, cross-validation with different k-fold (3, 5, 7 and 9) will be carried out to minimize the risk of overfitting. This validation procedure has been implemented using the trails-and-errors procedure, each trail with different value of the k-fold until achieving the one with minimal possibility of experienced overfitting. MATLAB Programming language was used to develop the proposed model.

One of the primary reasons for developing a model to predict the D.O. from other surface water parameters is that the D.O. is relatively costly and time-consuming to acquire and monitor. On the other hand, the main reason for selecting the Water Temperature, Biological Oxygen Demand, Iron, and Total Organic Carbon as a predictor for the D.O. is first because of the availability of these parameters. Secondly, there is a direct relationship between all these parameters and the D.O.; for example, the greater the amount of Biological Oxygen Demand in the water stream, the more rapidly is the depletion of the D.O. in water. Similarly, for the temperature (T), the more the T, the less the D.O. in the water stream will have occurred. Iron could critically consume the D.O. because of D.O. will be consumed as an oxidant for the Iron concentration. Hence, the D.O. concentration could dramatically reduce its amount in water stream. Finally, the Total Organic Carbon is the measuring indicator for how pure is the water stream considering the organisms' life, which is indirectly affected by the level of D.O. in the water stream. Therefore, these parameters have been considered as predictors for D.O. in the current research. Table 2 shows the statistical analysis and the correlations between these parameters and D.O.

Three different statistical indices will be applied to measure how the proposed model predicts dissolved oxygen concentration. These indices are Root Mean Square Error (R.M.S.E.), Coefficient of Correlation (Correlation), and Coefficient of Determination (R-squared). The formulas for these indices with comprehensive explanation can be found in study carried out by Najah et al.[46]. In addition to that, Taylor's diagram and violin plots will be performed to assess the correlation between observed and predicted data. Sensitivity and uncertainty analysis will be carried to validate the proposed model's reliability. Figure 3 demonstrates the flow of the proposed method in this study. As can be seen from the flowchart, after the secondary data is collected, a pre-processing step was carried out to normalize the dataset before building the models. Then, different models will be built using different algorithms, and each model optimized by tuning the hyper-parameters of each model. In addition to that, a comparison between the proposed model and the developed models in literature will be carried out to highlight the contribution of this research.

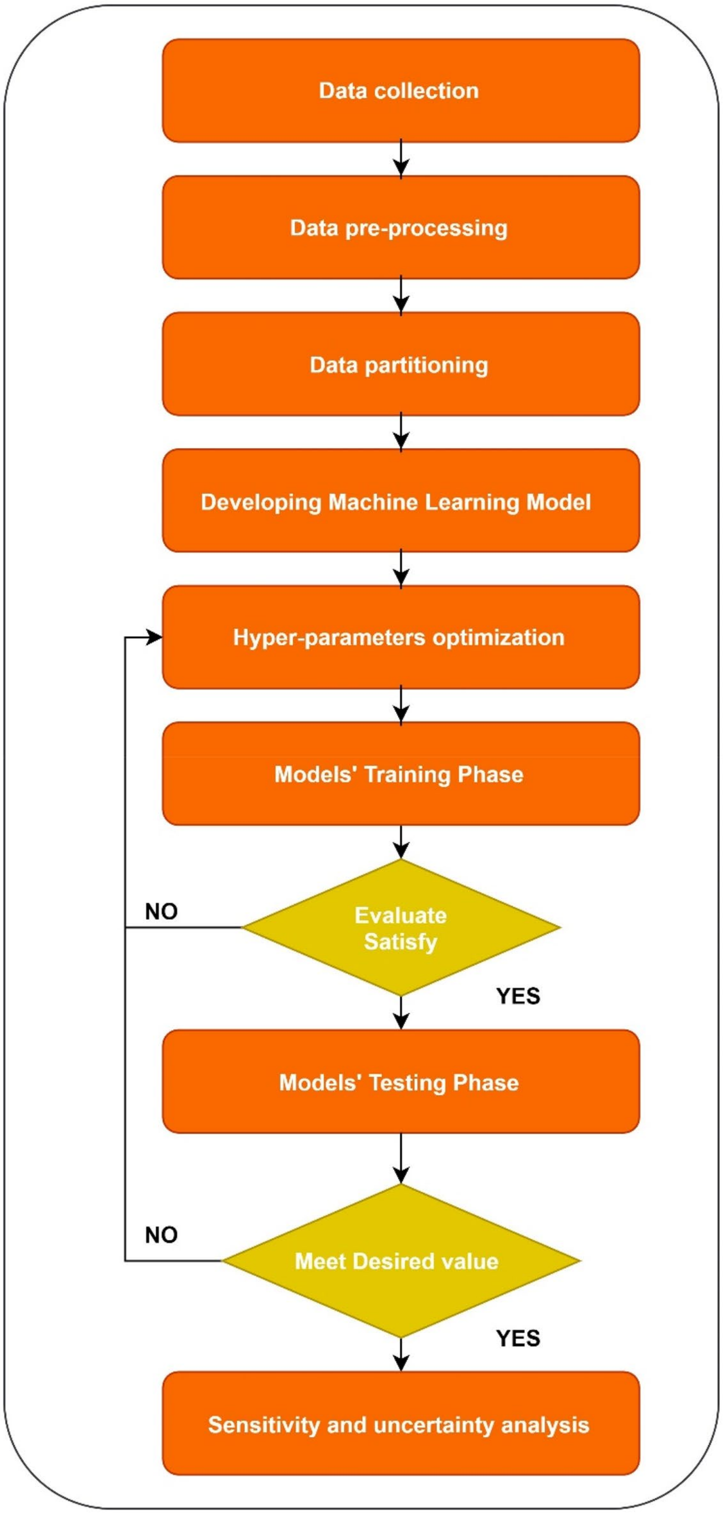

Figure 3. Flowchart of the study.

Results and discussion

One of the main characteristics of defining artificial neural network models is choosing the number of neurons for the hidden layer. An insufficient number of neurons can cause the model not to capture the data's nonlinearity[47]. However, if more neurons are introduced, that might increase the model's time and lead to overfitting. Therefore, to overcome such issues, in this study, different models were developed to find the optimum number of neurons where the number of neurons set to be 5, 10,15, and 20. In addition to that, identifying the right input combination is one of the vital factors that need to be considered in optimizing the performance of

Number of neurons	10	15	5	15	10
Models	**M.1**	**M.2**	**M.3**	**M.4**	**M.5**
Correlation	0.929	0.971	0.865	**0.988**	0.598
R-squared	0.857	0.940	0.731	**0.980**	0.336
RMSE	0.240	0.143	0.293	**0.136**	0.487

Table 3. Performance of each developed model based on the optimum number of neurons. Significant values are in bold.

	Actual	**M.1**	**M.2**	**M.3**	**M.4**	**M.5**
Max	**8.680**	8.560	8.747	8.680	**8.386**	8.772
Mean	**7.920**	8.008	7.939	7.890	**7.816**	8.024
Min	**5.270**	5.270	5.302	5.972	**5.149**	7.509
SD	**0.577**	0.601	0.596	0.535	**0.579**	0.235

Table 4. Comparison between the proposed model and the actual DO for the testing dataset. Significant values are in bold.

the artificial neural network model[5,26,48]. In this study, five different models with different input combinations have been introduced. The first model (M.1) was developed using one input parameter (water temperature), while the second model (M.2) was developed by introducing another input parameter (water temperature and biological oxygen demand). A third model (M.3) used three input parameters (biological oxygen demand, iron, and total organic carbon). The fourth model (M.4) was developed using a different combination of three parameters (water temperature, biological oxygen demand, and iron). And finally, model five (M.5) was developed using all four parameters as input (water temperature, biological oxygen demand, iron, and total organic carbon). Twenty models have been developed with different inputs and numbers on neurons to find the best model for predicting dissolved oxygen concentrations changes. Table 3 presents the performance of each developed model for the optimum number of neurons. It can be seen that the best number of neurons falls in the range of 5 to 15 for each model. It can be seen, in all the proposed models, the number of neurons contributes significantly to the improvement of the models' accuracy, except in model five (M.5). It has been noticed that the poor performance in M.5 is associated with the input combinations, not with the number of neurons. Such findings reveal the input combinations' importance in developing a robust model. In addition to that, it can be observed from Table 3 that the best performance model (M.4) can be achieved when the number of neurons is equal to 15 and the combination of the input is water temperature, biological oxygen demand, and iron. This is followed by M.2 with the same number of neurons but with two input combinations (water temperature and biological oxygen demand). It can also be observed that when the total organic carbon was introduced as input in M.3 and M.5, the artificial neural network models' accuracy dropped. This indicates that the total organic carbon should not be considered input in developing reliable models to predict dissolved oxygen concentration changes. Feed-forward MLP model is used in this study. Regarding the training algorithm, three different algorithms were investigated, namely Levenberg–Marquardt, Bayesian regularization and Scaled conjugate gradient. The best results obtained by using the latter training algorithm. Scaled conjugate gradient is powerful training algorithm where there is no need for much memory. It is also proved to be faster in convergence compared to the other two used training algorithms. With regards to activation function, it was found that tanh (hyperbolic) function is best among the different inspected functions.

It can be seen from Table 4 the performance of each developed model with a different input combination. It can be observed that the fourth model (M.4) outperforms all other models in predicting the dissolved oxygen and manages to capture the peak and low concentration of the dissolved oxygen. Moreover, it can be seen that the mean of the predicted data is close to the actual observed data.

To test the proposed model's reliability and to determine the model's validity, Taylor's diagram is recommended by many researchers and is commonly used[19,49]. It can be seen from Fig. 4 the relation between the correlation and the standard deviation for the actual and the predicted concentration of dissolved oxygen for the five models. It can be seen that M.4 is outperforming all other models where the distribution of standard deviation for the predicted data is close to the actual one, which suggests that the proposed model is consistent in capturing the observed data pattern.

The average percentage of relative error for each model has been computed to examine the error percentage between the predicted and the actual observed data, as shown in Fig. 5. The value confirmed this study, where M.4 indicates the lowest error compared to other developed models. While M.2 is ranked second, and the highest error observed with M.5.

A Violin plot is used to demonstrate the difference between the actual and predicted data from each model, as shown in Fig. 6. This plot helps to understand the probability distribution of the data. It can be seen that the best model is M.4, which its predicted data have similar distribution compared with the actual data.

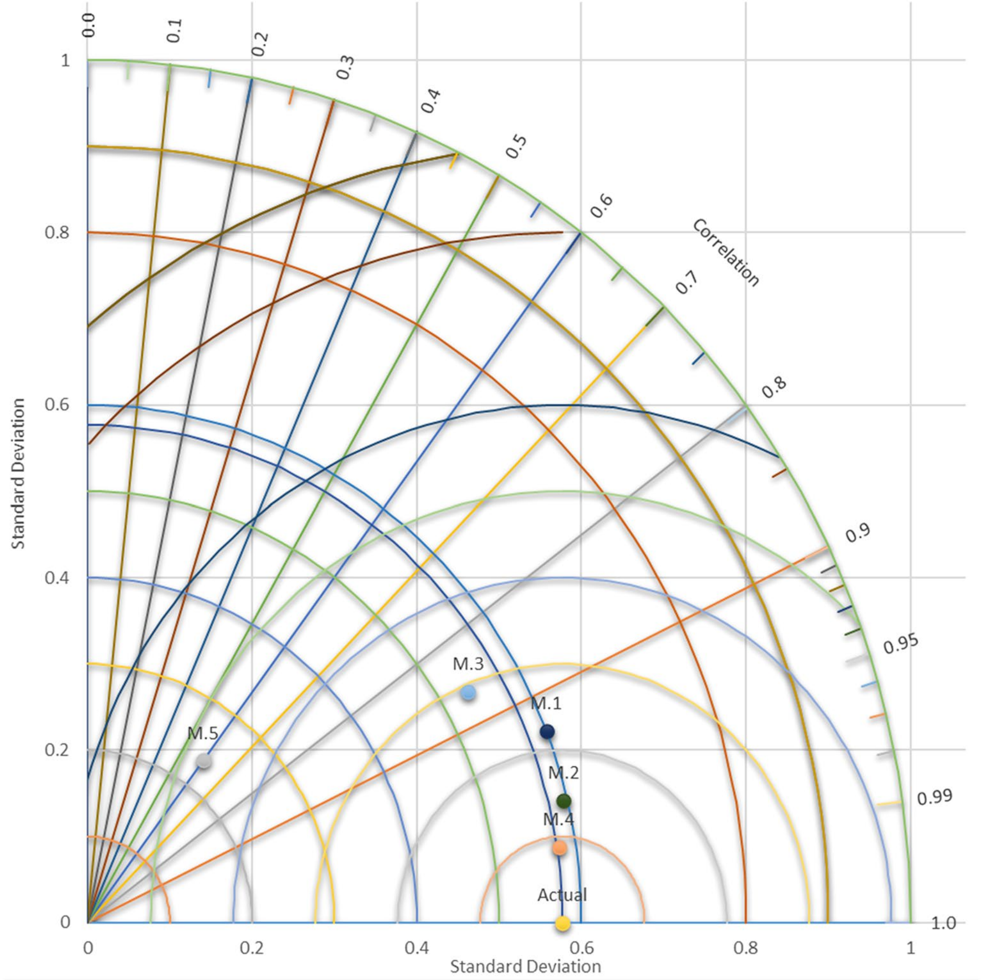

Figure 4. Taylor diagram for the proposed five models.

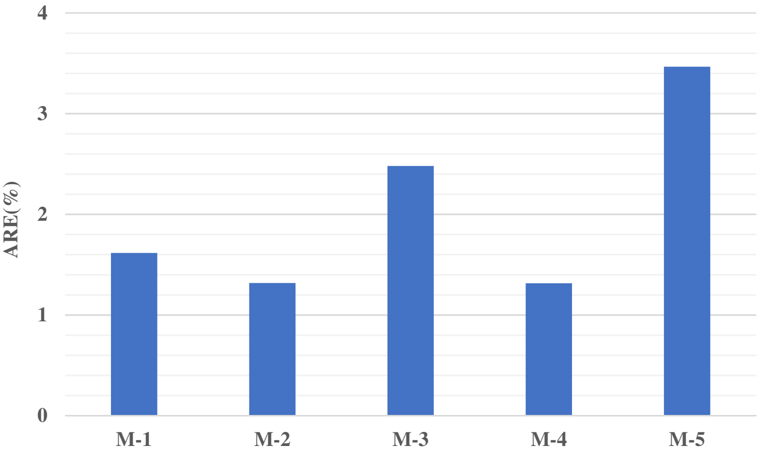

Figure 5. Average Relative Error of each proposed model.

To measure the proposed models' performance when new data are introduced, the d-factor value is used for this purpose. When the d-factor values close to zero mean that the model can still perform well if a new data set is introduced[50]. This study uses the following equations to calculate d-factor values:

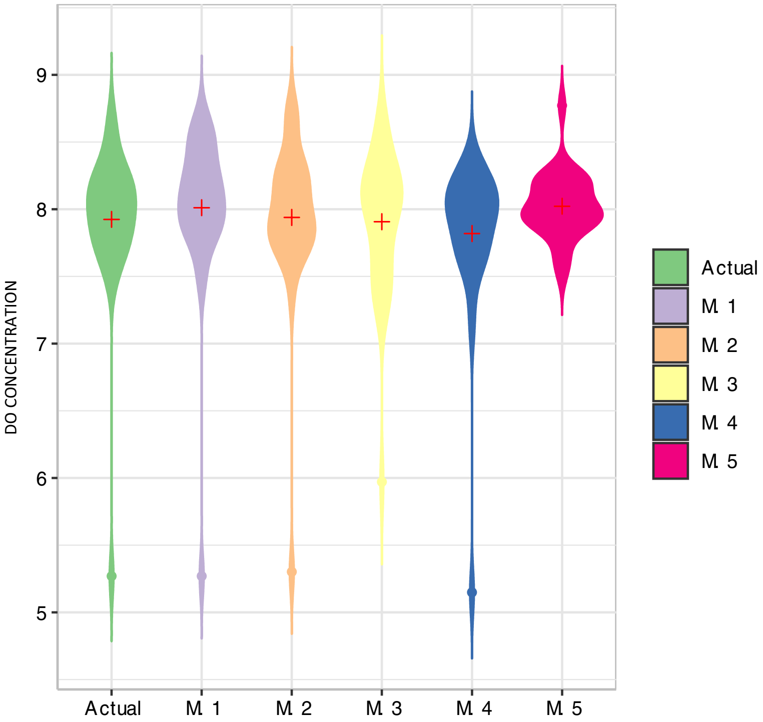

Figure 6. Violin plot between actual and proposed models.

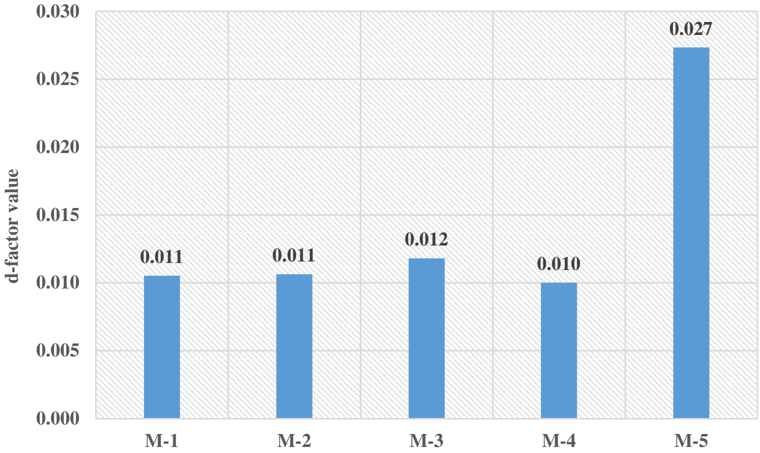

Figure 7. d-factor values for each proposed model.

$$\text{d-factor} = \frac{\overline{d}_x}{\sigma_x} \tag{1}$$

$$\overline{d}_x = \frac{1}{N}\sum_{i=1}^{N}(X_U - X_L) \quad i = 1, \ldots, N \tag{2}$$

σ_x represents the standard deviation of actual data x and $\overline{d}_x$ represents the average distance between the upper X_U (the value that is greater than or equal to every element in the dataset) and lower X_L(the value that is less than or equal to every element in the dataset), i denotes the order of the record in the time series data (i = 1,…,N), while N represents the number of the observed dataset. It can be seen from Fig. 7 that M.4 shows the lowest d-factor value, which indicates this model architecture is reliable to be adopted when a new set of data used and can perform with a high level of accuracy.

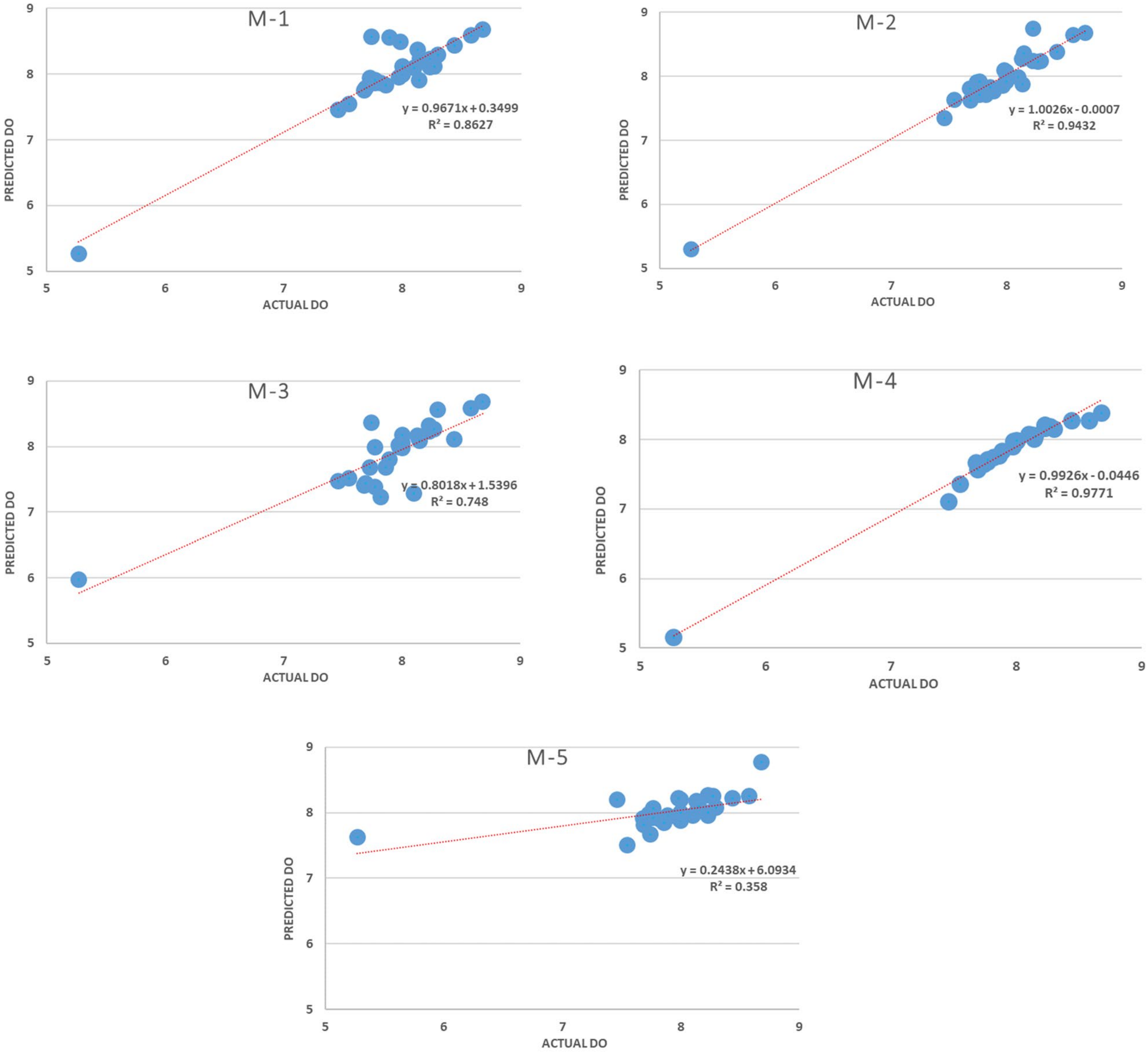

Figure 8. Predicted vs. actual scatter chart.

Figure 8 shows the scatter plots between the predicted and actual D.O. for the five developed models where it can be seen M.4 outperformed all other models with different input combinations.

To sum up, the lowest accuracy reported when M.5 used to predict D.O. It should be noted that M5. model exhibits acceptable accuracy in capturing the maximum and average values of D.O. concentrations. However, for the minimum values of D.O. concentrations, the M.5 model is unable to capture it where the absolutes relative error percentage ranged between 42 to 59%. Model M.3 ranked four among the five developed models with an average relative error percentage of 5%. M.1 ranked three among the five other models. It performs better than M.3 and M5 where the average absolute relative error percentage equals 1.7%. And finally, M.2 ranked second, outperforming M.1, M.3, and M.5. However, M.2 was unable to capture the extreme concentration values of D.O. as M.4.

A comparison was conducted between the proposed model and other literature models to compare the current study findings to other studies. Kisi et al.[51] proposed a Bayesian model averaging (B.M.A.) model to predict the concentration of dissolved oxygen. And the findings were compared with different data-driven methods, including an extreme learning machine (E.L.M.), classification and regression tree (CART), and adaptive neuro-fuzzy inference system (A.N.F.I.S.). The r-squared for testing ranged from 0.718 to 0.836 (B.M.A. (0.836), E.L.M. (0.822), A.N.F.I.S. (0.831) and CART(0.718)) for one of the stations used in the study. Four input parameters were used as inputs to develop the four models. In the current study, only three parameters were used as input to the model. As mentioned earlier, choosing the right combinations has a crucial impact on the performance of the model. In the current study, when water temperature, biological oxygen demand, and iron were used as input, the artificial neural network achieved a high accuracy level where r-squared equals 0.98. Multi-layer perceptron neural network developed to predict dissolved oxygen concentration in Malaysia's Johor River[52]. In this study, five different input combinations were used to develop the model. The performance of the model was acceptable were r-squared was equal to 0.95. Compared with the proposed model in the current study, a high accuracy level has been achieved where r-squared equals 0.98 with a smaller number of input combinations. It can be concluded that the current proposed model is more accurate and can be adopted as a tool to predict the changes in the concentration of dissolved oxygen. A point that can be raised from the earlier comparison is that the studies were conducted using different datasets in different countries. Therefore, for future works, a more valid comparison should be performed to consider these algorithms in predicting dissolved oxygen concentrations at the Fei-Tsui reservoir.

For comparison purposes, the performance of the developed model (M.4) was compared with two other models, namely Random Forest (R.F) and Boosted Tree (B.T) regressions. The comparison was carried out using maximum and average relative percentage error. It has been observed that the maximum relative percentage error for M.4 is equal to 4.7%, while the maximum relative percentage error for B.T and R.F is 46% and 49%, respectively. At the same time, the average relative percentage error for M.4 is 1.3% which is the lowest than both B.T (4.1) and R.F (4.6).

Conclusion

The study focuses on predicting dissolved oxygen concentration as crucial water quality parameters in the Fei-Tsui reservoir in Taiwan using an artificial neural network model with simple architecture. Twenty-nine years of historical data provided the basis for development of the model. To test the model's reliability and optimize the algorithm, different numbers of neurons were used. Various numbers of input combinations were used to enhance the model's accuracy. Statistical indices were used to validate the accuracy of the model. The results reveal that the best number of neurons equals fifteen, while the best input combinations are three input parameters. These parameters are water temperature, biological oxygen demand and iron. The proposed model exhibits a high level of accuracy in predicting dissolved oxygen concentration changes where the r-squared is equal to 0.98. Taylor's diagram shows that the proposed model (M-4) displays a high consistency and accuracy level. Further investigation in implementing the proposed model in this research can predict other water quality parameters and be applied at locations with different climatic conditions for generalization purposes. There is a need to investigate machine learning models' integration with sensing technologies to efficiently monitor and predict water quality parameters for a smart early warning system. In addition, although the proposed optimization of the hyper parameters of the ANN modeling approach could provide proper prediction accuracy for DO, the accuracy could be improved by implementing the optimization of the hyper parameters of other AI model such as Random Forest and Boosted Tree algorithm.

Received: 7 November 2021; Accepted: 24 January 2022
Published online: 07 March 2022

References

1. Latif, S. D., Azmi, M. S. B. N., Ahmed, A. N., Fai, C. M. & El-Shafie, A. Application of artificial neural network for forecasting nitrate concentration as a water quality parameter: A case study of Feitsui Reservoir, Taiwan. *Int. J. Des. Nat. Ecodynamics* **15**, 647–652 (2020).
2. Najah, A., Elshafie, A., Karim, O. A. O. A. & Jaffar, O. Prediction of johor river water quality parameters using artificial neural networks. *Eur. J. Sci. Res.* **28**, 422–435 (2009).
3. Parsaie, A. & Haghiabi, A. H. Numerical routing of tracer concentrations in rivers with stagnant zones. *Water Sci. Technol. Water Supply* **17**, 825–834 (2017).
4. Mansour-Bahmani, A., Haghiabi, A. H., Shamsi, Z. & Parsaie, A. Predictive modeling the discharge of urban wastewater using artificial intelligent models (case study: Kerman city). *Model. Earth Syst. Environ.* https://doi.org/10.1007/s40808-020-00900-z (2020).
5. Najah Ahmed, A., El-Shafie, A. A., Karim, O. A. O. A. & El-Shafie, A. A. An augmented Wavelet De-noising Technique with Neuro-Fuzzy Inference System for water quality prediction. *Int. J. Innov. Comput. Inf. Control* **8**, 7055–7082 (2012).
6. Ay, M. & Kişi, Ö. Estimation of dissolved oxygen by using neural networks and neuro fuzzy computing techniques. *KSCE J. Civ. Eng.* **21**, 1631–1639 (2016).
7. Chen, W. B. & Liu, W. C. Artificial neural network modeling of dissolved oxygen in reservoir. *Environ. Monit. Assess.* **186**, 1203–1217 (2014).
8. Najah, A., El-Shafie, A., Karim, O. A. & El-Shafie, A. H. Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring. *Environ. Sci. Pollut. Res.* **21**, 1658–1670 (2014).
9. Tarmizi, A., Ahmed, A. N. & El-Shafie, A. Dissolved oxygen prediction using support vector machine in Terengganu river. *Middle-East J. Sci. Res.* **21**, 2182–2188 (2014).
10. Ehteram, M. *et al.* reservoir operation by a new evolutionary algorithm: Kidney algorithm. *Water Resour. Manag.* **32**, 4681–4706 (2018).
11. Ehteram, M. *et al.* Assessing the predictability of an improved ANFIS model for monthly streamflow using lagged climate indices as predictors. *Water* **11**, 1130 (2019).
12. Parsaie, A., Emamgholizadeh, S., Azamathulla, H. M. & Haghiabi, A. H. ANFIS-based PCA to predict the longitudinal dispersion coefficient in rivers. *Int. J. Hydrol. Sci. Technol.* **8**, 410–424 (2018).
13. Khan, F. A. F. A. *et al.* Complex extreme sea levels prediction analysis: Karachi coast case study. *Entropy* **22**, 549 (2020).
14. Muslim, T. O. *et al.* Investigating the influence of meteorological parameters on the accuracy of sea-level prediction models in Sabah, Malaysia. *Sustainability* **12**, 1193 (2020).
15. Lai, V. *et al.* Modeling the nonlinearity of sea level oscillations in the Malaysian coastal areas using machine learning algorithms. *Sustainability* **11**, 4643 (2019).
16. Ridwan, W. M. W. M. *et al.* Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. *Ain Shams Eng. J.* https://doi.org/10.1016/j.asej.2020.09.011 (2020).
17. El-Shafie, A., Mukhlisin, M., Najah, A. A. A. & Taha, M. R. R. Performance of artificial neural network and regression techniques for rainfall-runoff prediction. *Int. J. Phys. Sci.* **6**, 1997–2003 (2011).
18. Hipni, A. *et al.* Daily forecasting of dam water levels: Comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS). *Water Resour. Manag.* **27**, 3803–3823 (2013).

19. Sapitang, M., M Ridwan, W., Faizal Kushiar, K., Najah Ahmed, A. & El-Shafie, A. Machine learning application in reservoir water level forecasting for sustainable hydropower generation strategy. *Sustainability* **12**, 6121 (2020).
20. Tikhamarine, Y., Souag-Gamane, D., Najah Ahmed, A., Kisi, O. & El-Shafie, A. Improving artificial intelligence models accuracy for monthly streamflow forecasting using grey Wolf optimization (GWO) algorithm. *J. Hydrol.* **582**, 124435 (2020).
21. Osman, A. *et al.* Adaptive Fast Orthogonal Search (FOS) algorithm for forecasting streamflow. *J. Hydrol.* **586**, 124896 (2020).
22. Ho, J. Y. J. Y. *et al.* Towards a time and cost effective approach to water quality index class prediction. *J. Hydrol.* **575**, 148–165 (2019).
23. Abba, S. I. *et al.* Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index. *Environ. Sci. Pollut. Res.* **27**, 41524–41539 (2020).
24. Rezaie-Balf, M. *et al.* Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach. *J. Clean. Prod.* **271**, 122576 (2020).
25. Sinshaw, T. A., Surbeck, C. Q., Yasarer, H. & Najjar, Y. Artificial neural network for prediction of total nitrogen and phosphorus in US lakes. *J. Environ. Eng.* **145**, 1–11 (2019).
26. Ahmed, A. N. & El-Shafie, A. An application Support Vector Machine model (SVM) technique for Biochemical Oxygen Demand (BOD) prediction. In *Int. Conf. Artif. Intell. Pattern Recognition, AIPR 2014, Held 3rd World Congr. Comput. Inf. Technol. WCIT* 209–212 (2014).
27. Ahmadlou, M. *et al.* Flood susceptibility assessment using integration of adaptive network-based fuzzy inference system (ANFIS) and biogeography-based optimization (BBO) and BAT algorithms (BA). *Geocarto Int.* **34**, 1252–1272 (2019).
28. Heddam, S. Use of optimally pruned extreme learning machine (OP-ELM) in forecasting dissolved oxygen concentration (DO) several hours in advance: A case study from the Klamath River, Oregon, USA. *Environ. Process.* **3**, 909–937 (2016).
29. Sánchez-Monedero, J. *et al.* On the suitability of Extreme Learning Machine for gene classification using feature selection. In *Proc. 2010 10th Int. Conf. Intell. Syst. Des. Appl. ISDA'10* 507–512 (2010). https://doi.org/10.1109/ISDA.2010.5687215.
30. Liu, S. *et al.* A hybrid WA-CPSO-LSSVR model for dissolved oxygen content prediction in crab culture. *Eng. Appl. Artif. Intell.* **29**, 114–124 (2014).
31. Zhu, S. & Heddam, S. Prediction of dissolved oxygen in urban rivers at the Three Gorges Reservoir, China: Extreme learning machines (ELM) versus artificial neural network (ANN). *Water Qual. Res. J.* **55**, 106–118 (2020).
32. Li, W. *et al.* Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (GRU). *Inf. Process. Agric.* **8**, 185–193 (2021).
33. Nacar, S., Bayram, A., Baki, O. T., Kankal, M. & Aras, E. Spatial forecasting of dissolved oxygen concentration in the eastern Black sea basin, Turkey. *Water* **12**, 1041 (2020).
34. Zhang, Y.-F., Fitch, P. & Thorburn, P. J. Predicting the trend of dissolved oxygen based on the kPCA-RNN model. *Water* **12**, 585 (2020).
35. Wang, Y., Yuan, Y., Pan, Y. & Fan, Z. Modeling daily and monthly water quality indicators in a canal using a hybrid wavelet-based support vector regression structure. *Water* **12**, 1476 (2020).
36. Bayram, A., Uzlu, E., Kankal, M. & Dede, T. Modeling stream dissolved oxygen concentration using teaching–learning based optimization algorithm. *Environ. Earth Sci.* **73**, 6565–6576 (2015).
37. Tikhamarine, Y. *et al.* Rainfall-runoff modelling using improved machine learning methods: Harris hawks optimizer vs. particle swarm optimization. *J. Hydrol.* **589**, 125133 (2020).
38. Kumar, P. *et al.* Optimised neural network model for river-nitrogen prediction utilizing a new training approach. *PLoS ONE* **15**, e0239509 (2020).
39. Ehteram, M. *et al.* Performance improvement for infiltration rate prediction using hybridized Adaptive Neuro-Fuzzy Inferences System (ANFIS) with optimization algorithms. *Ain Shams Eng. J.* https://doi.org/10.1016/j.asej.2020.08.019 (2020).
40. Othman, F. *et al.* Efficient river water quality index prediction considering minimal number of inputs variables. *Eng. Appl. Comput. Fluid Mech.* **14**, 751–763 (2020).
41. Heddam, S. *Intelligent Data Analytics Approaches for Predicting Dissolved Oxygen Concentration in River: Extremely Randomized Tree Versus Random Forest, MLPNN and MLR* 89–107 (2021). https://doi.org/10.1007/978-981-15-5772-9_5.
42. Moghadam, S. V. *et al.* An efficient strategy for predicting river dissolved oxygen concentration: Application of deep recurrent neural network model. *Environ. Monit. Assess.* **193**, 1–18 (2021).
43. Latif, S. D. *et al.* Development of prediction model for phosphate in reservoir water system based machine learning algorithms. *Ain Shams Eng. J.* https://doi.org/10.1016/J.ASEJ.2021.06.009 (2021).
44. Chow, M. *et al.* Long term trends and dynamics of dissolved organic carbon (DOC) in a subtropical reservoir basin. *Water* **9**, 545 (2017).
45. Banadkooki, F. B. *et al.* Suspended sediment load prediction using artificial neural network and ant lion optimization algorithm. *Environ. Sci. Pollut. Res.* **27**, 38094–38116 (2020).
46. Najah, A., El-Shafie, A. & Karim, O. *Prediction of Water Quality Parameters Using Artificial Intelligence: Case study-Johor River Basin* (LAP Lambert Academic Publishing, 2011).
47. Najah Ahmed, A. *et al.* Machine learning methods for better water quality prediction. *J. Hydrol.* **578**, 124084 (2019).
48. Abobakr Yahya, A. S. *et al.* Water quality prediction model based support vector machine model for ungauged river catchment under dual scenarios. *Water* **11**, 1231 (2019).
49. Jumin, E. *et al.* Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction. *Eng. Appl. Comput. Fluid Mech.* **14**, 713–725 (2020).
50. Nur Adli Zakaria, M. *et al.* Application of artificial intelligence algorithms for hourly river level forecast: A case study of Muda River, Malaysia. *Alexandria Eng. J.* **60**, 4015–4028 (2021).
51. Kisi, O., Alizamir, M. & Docheshmeh Gorgij, A. R. Dissolved oxygen prediction using a new ensemble method. *Environ. Sci. Pollut. Res.* **27**, 9589–9603 (2020).
52. Najah, A. A., El-Shafie, A., Karim, O. A. & Jaafar, O. Integrated versus isolated scenario for prediction dissolved oxygen at progression of water quality monitoring stations. *Hydrol. Earth Syst. Sci.* **15**, 2693–2708 (2011).

Acknowledgements

The authors would like to appreciate the open access publication payment supported by the School of Engineering, Monash University Malaysia. The authors would also like to appreciate the financial support received from the 2020106TELCO grant by the Innovation & Research Management Center (iRMC), Universiti Tenaga Nasional (UNITEN), Malaysia.

Author contributions

Conceptualization: B.F.Z.S., B.H.Z.S., Methodology: M.A.M., A.S., A.H.B., formal analysis and investigation: S.D.L., A.N.A., writing—original draft preparation: S.D.L., A.N.A., A.E., writing—review and editing: A.N.A., A.E., supervision: W.J.K., C.M.F., A.N.A.

Funding

This research received no external funding.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.F.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.