processes

MDPI

Article

Comparison between Regression Models, Support Vector Machine (SVM), and Artificial Neural Network (ANN) in River Water Quality Prediction

Nur Najwa Mohd Rizal [1], Gasim Hayder [2,*], Mohammed Mnzool [3], Bushra M. E. Elnaim [4], Adil Omer Yousif Mohammed [5] and Manal M. Khayyat [6]

1 College of Graduate Studies, Universiti Tenaga Nasional (UNITEN), Kajang 43000, Malaysia
2 Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Kajang 43000, Malaysia
3 Department of Civil Engineering, College of Engineering, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia
4 Department of Computer Science, College of Science and Humanities in Al-Sulail, Prince Sattam bin Abdulaziz University, Al-Kharj 16278, Saudi Arabia
5 Department of Computer Science, College of Science and Arts, Qassim University, P.O. Box 1162, Al-Bukairiyah 51941, Saudi Arabia
6 Department of Information Systems, College of Computers and Information Systems, Umm Al-Qura University, P.O. Box 7607, Makkah 24382, Saudi Arabia
* Correspondence: gasim@uniten.edu.my

Abstract: Both anthropogenic and natural sources of pollution are regionally significant. Therefore, in order to monitor and protect the quality of Langat River from deterioration, we use Artificial Intelligence (AI) to model the river water quality. This study has applied several machine learning models (two support vector machines (SVMs), six regression models, and artificial neural network (ANN)) to predict total suspended solids (TSS), total solids (TS), and dissolved solids (DS)) in Langat River, Malaysia. All of the models have been assessed using root mean square error (RMSE), mean square error (MSE) as well as the determination of coefficient (R^2). Based on the model performance metrics, the ANN model outperformed all models, while the GPR and SVM models exhibited the characteristic of over-fitting. The remaining machine learning models exhibited fair to poor performances. Although there are a few researches conducted to predict TDS using ANN, however, there are less to no research conducted to predict TS and TSS in Langat River. Therefore, this is the first study to evaluate the water quality (TSS, TS, and DS) of Langat River using the aforementioned models (especially SVM and the six regression models).

Keywords: river; water quality parameters; regression models; ANN; SVM

Citation: Najwa Mohd Rizal, N.; Hayder, G.; Mnzool, M.; Elnaim, B.M.E.; Mohammed, A.O.Y.; Khayyat, M.M. Comparison between Regression Models, Support Vector Machine (SVM), and Artificial Neural Network (ANN) in River Water Quality Prediction. *Processes* **2022**, *10*, 1652. https://doi.org/10.3390/pr10081652

Academic Editor: George Z. Kyzas

Received: 20 July 2022
Accepted: 18 August 2022
Published: 20 August 2022

1. Introduction

River water is one of the prime natural resources that are vital for living beings, especially humans. River water is also considered one of the resources that is at risk even though it is essential for one's life [1]. This is due to the fact that it is also being exploited for several purposes such as electric power generation, agriculture, irrigation, industrialization, and recreation [2–4]. Consequently, the quality of the river will deteriorate and bring harm to the people as well as to the surrounding environment because the contamination occurred either from human activities such as the discharge of effluents from chemical, toxic, and human waste [5] or natural sources of pollution, namely flood and landslides. Therefore, to guarantee that high-quality water is available to be utilized for sundry purposes, it is requisite to have the river water quality is controlled [6].

In the past decades, conventional and statistical techniques that involved in the collection and evaluation of raw data manually have been applied to assess water quality [7].

However, the approaches are time-consuming, expensive, and require a labor-intensive procedure that needed a specialized measuring tool [8,9]. Recently, the usage of Artificial Intelligence (AI) in predicting the river water quality is no more outlandish as AI able to help monitor the condition of the river efficiently compared to statistical and conventional techniques of lab testing [10]. Moreover, AI approaches have been explored by researchers and scholars all around the world and AI has shown great capability in monitoring and forecasting the water quality [11]. Machine learning, as a part of AI has been widely applied in the various fields, especially in hydrology. According to Moubayed [12], there are four types of machine learning algorithms viz. reinforcement learning, unsupervised learning, semi-supervised learning, and supervised learning algorithms. In addition, a supervised learning algorithm is known when a machine learning algorithm obtains the target pattern and the feature vector as an input to develop a model. The developed model can be applied to determine the latest patterns and set output to the model [13].

There are several studies that have utilized machine learning models to predict water quality parameters. For instance, a study conducted by Niroobakhsh et al. [14] have developed two different artificial neural network (ANN) models to predict total dissolved solids (TDS) for Jajrood River in Iran. The type of ANN models used were radial basis function (RBF) and a multilayer perceptron (MLP). Two evaluation metrics (i.e., R^2 and RMSE) were used to evaluate the performance of the model. Based on the study conducted, the authors concluded that RBF model performed better than the developed MLP model, where the values of R^2 were 0.9362 compared to MLP model that achieved R^2 = 0.8968. Furthermore, Talib and Amat [15] proposed an ANN model to forecast chemical oxygen demand (COD) concentration for Dondang River in Penang, Malaysia. A total of nine water quality parameters i.e., phosphate, temperature, nitrate, biochemical oxygen demand (BOD), total solids (TS), dissolved solids (DO), suspended solids, pH, and ammonia were used as input parameters for the modeling. From the results obtained, the developed ANN model has obtained R^2 and R values of 0.83 and 0.94, respectively.

On the other hand, Najah et al. [16] have developed several machine learning models such as ensemble ANN, MLP-ANN, and also support vector machine (SVM) models to predict three water quality parameters (i.e., BOD, DO, and COD) in Johor River, Malaysia. From the study, the authors concluded that SVM model with five input parameters outperformed the other developed models by having the mean square error (MSE) and correlation of efficiency (CE) values of 0.07 and 0.95, 0.07 and 0.91, and also 0.12 and 0.93 for COD, BOD, and DO predictions, respectively. Moreover, Zhou et al. [17] predict the sulphate content of lakes in China using different Kernel functions of Gaussian process regression (GPR) model, such as Matern 5/2, rational quadratic, squared exponential, and exponential functions. The authors also tested several machine learning models (i.e., support vector regression models, bagging tree model, boosted tree model, and decision tree model) and compare the mentioned models with the GPR models. According to the study, the authors concluded that exponential GPR model outperformed other models, where the values of RMSE, R^2, and mean absolute error (MAE) obtained were 7.269, 0.72, and 5.046, respectively.

Other previous studies such as [9] and [18] also applied machine learning models and have successfully predicted the water quality parameters with high accuracy. Thus, it has shown that machine learning models are suitable in predicting river water quality parameters with high degree of robustness and accuracy. Therefore, this study aims to predict DS, TS, and TSS of Langat River, Malaysia by using nine different machine learning models i.e., six different regression models, two SVMs, and an ANN model. In order to determine the best and optimum machine learning model in predicting the aforementioned water quality parameters, the outcomes of the models were compared. This study also able to contribute in monitoring the water quality of Langat River as there are less to none research that have predicted TS and TSS of Langat River. Hence, this is the first study to evaluate the water quality (TSS, TS, and DS) of Langat River using machine learning models (especially using SVM and the six regression models).

The study area and the methodology are explained in the next section. The results of the study and the discussion regarding the results of the models are discussed in Section 3. Lastly, the conclusion of the study is provided in Section 4.

2. Methodology

2.1. Study Area

The study area chosen for the study was the Langat River in Malaysia. Langat River is located in the State of Selangor and it originated from the peak of Mount Nuang (Gunung Nuang). From there, the river flows southward towards the Straits of Malacca [19,20]. In addition, The Langat River basin is the second biggest basin in the State of Selangor, where it has an approximate catchment area of 1815 km^2 and is 141 km long [21,22].

The daily historical data of 24 water quality parameters (Station No. 2917601) which spans from January 1891 until March 2019 have been obtained from the Department of Irrigation and Drainage (DID), Malaysia [23]. The historical data yields 161 available data points for each parameters. BOD_5, potassium, manganese, iron, phosphate, sulphate, silica, chemicals, magnesium, TS, DS, solids, chloride, fluoride, ammonia, nitrate, sodium, pH, colour, turbidity, conductivity, hardness, alkalinity, and calcium, including TSS that is calculated by Equation (1) are the 25 water quality parameters used as inputs for the modeling.

$$\text{Total suspended solid} = \text{Total solids} - \text{Total dissolved solids} \quad (1)$$

The unit for TSS, TS, and TDS in the equation is mg/L [24]. In addition, the statistical analysis for water quality parameters based on the raw data is shown in Table 1 and Figure 1.

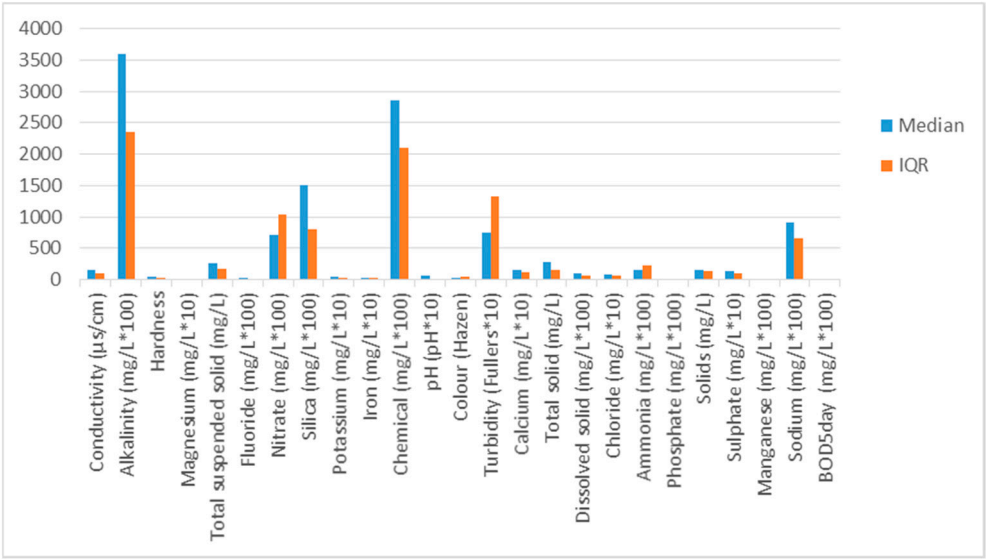

Figure 1. The statistical analysis for water quality parameters based on the raw data.

Due to the tremendous development being built along the Langat River Basin, the river has experienced several flood incidents within the basin [22,25]. Therefore, total suspended solids (mg/L), dissolved solids (mg/L × 10), and total solids (TS) were the chosen parameters to be predicted in the study.

Table 1. The statistical analysis for water quality parameters based on the raw data.

Variable	Median	IQR
Conductivity (μs/cm)	150	99
Alkalinity (mg/L × 100)	3600	2350
Hardness	44	30
Magnesium (mg/L × 10)	10	5
Total suspended solid (mg/L)	266.81	165.92
Fluoride (mg/L × 100)	30	12
Nitrate (mg/L × 100)	710	1035
Silica (mg/L × 100)	1500	800
Potassium (mg/L × 10)	46.5	21.25
Iron (mg/L × 10)	34.5	28.5
Chemical (mg/L × 100)	2850	2100
pH (pH × 10)	65	5
Colour (Hazen)	30	50
Turbidity (Fullers × 10)	750	1320
Calcium (mg/L × 10)	160	115.5
Total solid (mg/L)	284	154
Dissolved solid (mg/L × 100)	94	71
Chloride (mg/L × 10)	82	55
Ammonia (mg/L × 100)	160	228
Phosphate (mg/L × 100)	10	0
Solids (mg/L)	158	134
Sulphate (mg/L × 10)	130	97.75
Manganese (mg/L × 100)	12	11
Sodium (mg/L × 100)	920	650
BOD_{5day} (mg/L × 100)	3	6

2.2. Data Pre-Processing

The historical data were cleaned and pre-processed before being used as inputs in the modelings. According to Chen et al. [26], the practice of directly erasing the missing data is not recommended even though most researchers have applied it in their research. Therefore, in this study, the missing values from the data have been cleaned by replacing the missing values with a constant value (zero value) since the data obtained are limited. Although zero is known as a meaningless value, but according to Chollet [27], to insert missing values as zero are acceptable (i.e., with neural networks). The model will ignore the zero values since it has been trained that zero values are equal to missing data [27]. Thus, zero values have no effect on the neural network, since it cancels the corresponding weight after multiplication. However, due to limited data obtained, zero values are important to preserve the time series of the data.

Then, the data have been normalized in the range between 0 and 1 since few data have a high value which may give errors in the modeling. Next, the data have been divided into three where 70% of data was used for the training set, 15% of data was used for the testing set, and the remaining 15% for the validation set. MATLAB 2020b has been used to develop all models.

2.3. Models Development

The regression models used in the study were fine tree, medium tree, boosted tree, bagged tree, rational quadratic GPR, and lastly, exponential GPR models. The aforementioned models have been developed in MATLAB using Regression Learner Application and therefore, the models have been grouped as the regression models. Moreover, cross-validation factors were applied in the modeling and have been set to 10-folds.

Furthermore, two SVM models namely fine Gaussian SVM and medium Gaussian SVM were also applied in the study. SVM is known as kernel-based AI model as it consists of a kernel function, regression model complexity, and also regularization [28]. Both of the SVM models have been developed using Regression Learner Application. The cross-validation factors for the SVM models have also been set to 10-folds.

As for the artificial neural network (ANN) model, it was developed using Neural Network Fitting app in MATLAB and a two-layer feed forward network with two different transfer functions were used, where the sigmoid transfer function was applied in the hidden neurons while the linear transfer function was applied at the output neurons. The number of hidden neurons has been set to 1, 5, and 10, however, the ANN model with 10 hidden neurons exhibited the best performance. Thus, 10 hidden neurons have been chosen in the study. Moreover, the ANN model was trained by Levenberg-Marquardt backpropagation algorithm. The architecture of the ANN model is shown in Figure 2 while Table 2, shows the criteria selected for all models for the water quality prediction.

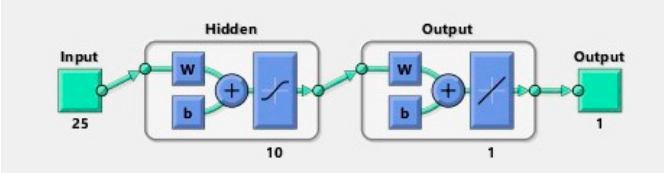

Figure 2. The architecture of the ANN model.

Table 2. The criteria selected for the water quality parameters modeling for all models.

Type of Model	Parameters	Default Value
Regression Tree -Fine Tree Model	Minimum leaf size:	4
Regression Tree -Medium Tree Model	Minimum leaf size:	12
Ensemble Tree -Boosted Tree Model	Minimum leaf size: Number of learners: Learning rate:	8 30 0.1
Ensemble Tree -Bagged Tree Model	Minimum leaf size: Number of learners:	8 30
Gaussian Process Regression -Exponential GPR	Basis function: Kernel function:	Constant Exponential
Gaussian Process Regression -Rational Quadratic GPR	Basis function: Kernel function:	Constant Rational quadratic
Artificial Neural Network	Training algorithm: Epoch:	Levenberg-Marquardt 0–100 epochs
Support Vector Machine -Fine Gaussian SVM	Kernel function: Kernel scale:	Gaussian 1.2
Support Vector Machine -Medium Gaussian SVM	Kernel function: Kernel scale:	Gaussian 4.9

2.4. Model Performance Evaluation

Root mean square error (RMSE), mean square error (MSE), and determination of coefficient (R^2) were the statistical indicators used to assess the performance of the developed models. The indicators are defined as follows [24,29]:

$$\mathrm{MSE} = \frac{1}{n}\sum(y' - y)^2 \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum(y' - y)^2} \tag{3}$$

$$R^2 = \left(\frac{n(\sum yy') - (\sum y)(\sum y')}{\sqrt{\left[n\sum y^2 - (\sum y)^2\right]\left[n\sum y'^2 - (\sum y')^2\right]}} \right)^2 \tag{4}$$

where y is the observed value, and y' is the predicted value. The n is defined as the number of data samples.

3. Results and Discussion

Table 3 shows the prediction analysis for all models in the testing dataset. Based on the table the ANN model has outperformed all models by obtaining high values of R^2 which were 0.9988, 0.9835, and 0.9880 in predicting TSS, DS, and TS, respectively. The ANN model also achieved low values of RMSE and MSE in predicting the water quality parameters. Thus, it has been shown that the ANN model has predicted the parameters accurately compared to the regression and SVM models. In the test dataset, fine Gaussian SVM and medium tree model were the worst models for predicting the parameters.

Table 3. Prediction analysis of all parameters (testing dataset).

Parameters	Type of Models	R2	MSE	RMSE
TSS (mg/L)	Fine Tree	0.8200	0.0004	0.0195
	Medium Tree	0.0000	0.0021	0.0462
	Boosted Tree	0.8100	0.0004	0.0203
	Bagged Tree	0.7000	0.0006	0.0254
	Rational Quadratic GPR	0.9600	7.6799×10^{-5}	0.0088
	Exponential GPR	0.7200	0.0006	0.0243
	Fine Gaussian SVM	0.0000	0.0021	0.0461
	Medium Gaussian SVM	0.4900	0.0011	0.0328
	Artificial Neural Network	0.9988	2.7904×10^{-5}	5.2824×10^{-3}
DS (mg/L × 100)	Fine Tree	0.5400	2.5927×10^{-6}	0.0016
	Medium Tree	0.0000	5.6248×10^{-6}	0.0024
	Boosted Tree	0.5400	2.5839×10^{-6}	0.0016
	Bagged Tree	0.4700	2.9763×10^{-6}	0.0017
	Rational Quadratic GPR	0.6200	2.1625×10^{-6}	0.0015
	Exponential GPR	0.4400	3.1423×10^{-6}	0.0018
	Fine Gaussian SVM	0.0500	5.3231×10^{-6}	0.0023
	Medium Gaussian SVM	0.2800	4.0532×10^{-6}	0.0020
	Artificial Neural Network	0.9835	6.2799×10^{-8}	2.5060×10^{-9}

Table 3. *Cont.*

Parameters	Type of Models	R2	MSE	RMSE
TS (mg/L)	Fine Tree	0.7800	0.0005	0.0213
	Medium Tree	0.0000	0.0021	0.0457
	Boosted Tree	0.8200	0.0004	0.0193
	Bagged Tree	0.5700	0.0009	0.0299
	Rational Quadratic GPR	0.9600	8.5497×10^{-5}	0.0092
	Exponential GPR	0.7500	0.0005	0.0228
	Fine Gaussian SVM	0.0000	0.0021	0.0456
	Medium Gaussian SVM	0.5400	0.0010	0.0319
	Artificial Neural Network	0.9880	0.8484×10^{-5}	5.3370×10^{-3}

The prediction analysis for all models in the overall datasets is shown in Table 4. From the table, it showed that rational quadratic GPR model has achieved the highest values of R^2 (1.00) in predicting TSS, DS, and TS. On the other hand, the rest of the models have exhibited good to fair performances in predicting the water quality parameters, where the values of $R^2 > 0.61$ for all three predictions, MSE < 0.02, and RMSE < 0.20. However, fine Gaussian SVM model has showed poor performance by having $R^2 < 0.20$, and higher values of RMSE and MSE than the other machine learning models in predicting TSS and TS. For DS prediction, medium Gaussian SVM achieved the lowest R^2 value (−0.03) and the highest RMSE and MSE values. In addition, Figure 3 shows the scatter plots for ANN model based on the respective water quality parameters.

Table 4. Prediction analysis of all parameters (overall dataset).

Parameters	Type of Models	R^2	MSE	RMSE
TSS (mg/L)	Fine Tree	0.9200	0.0017	0.0418
	Medium Tree	0.8100	0.0042	0.0649
	Boosted Tree	0.9200	0.0018	0.0429
	Bagged Tree	0.8600	0.0031	0.0556
	Rational Quadratic GPR	1.00	3.2522×10^{-7}	0.0006
	Exponential GPR	0.9500	0.0012	0.0350
	Fine Gaussian SVM	0.1600	0.0189	0.1376
	Medium Gaussian SVM	0.6100	0.0088	0.0937
	Artificial Neural Network	0.9998	4.1676×10^{-6}	0.0020
DS (mg/L × 100)	Fine Tree	0.7500	0.0040	0.0636
	Medium Tree	0.5800	0.0063	0.0826
	Boosted Tree	0.7900	0.0035	0.0590
	Bagged Tree	0.5700	0.0069	0.0832
	Rational Quadratic GPR	1.00	4.8138×10^{-8}	0.0002
	Exponential GPR	0.9000	0.0017	0.0411
	Fine Gaussian SVM	0.7500	0.0040	0.0636
	Medium Gaussian SVM	−0.0300	0.0167	0.1292
	Artificial Neural Network	0.9953	0.0001	0.0104

Table 4. *Cont.*

Parameters	Type of Models	R^2	MSE	RMSE
TS (mg/L)	Fine Tree	0.9200	0.0017	0.0409
	Medium Tree	0.8300	0.0037	0.0605
	Boosted Tree	0.9300	0.0016	0.0395
	Bagged Tree	0.8600	0.0030	0.0546
	Rational Quadratic GPR	1.00	3.118×10^{-7}	0.0006
	Exponential GPR	0.9400	0.0013	0.0358
	Fine Gaussian SVM	0.1800	0.0177	0.1331
	Medium Gaussian SVM	0.6300	0.0080	0.0895
	Artificial Neural Network	0.9970	6.5419×10^{-5}	0.0081

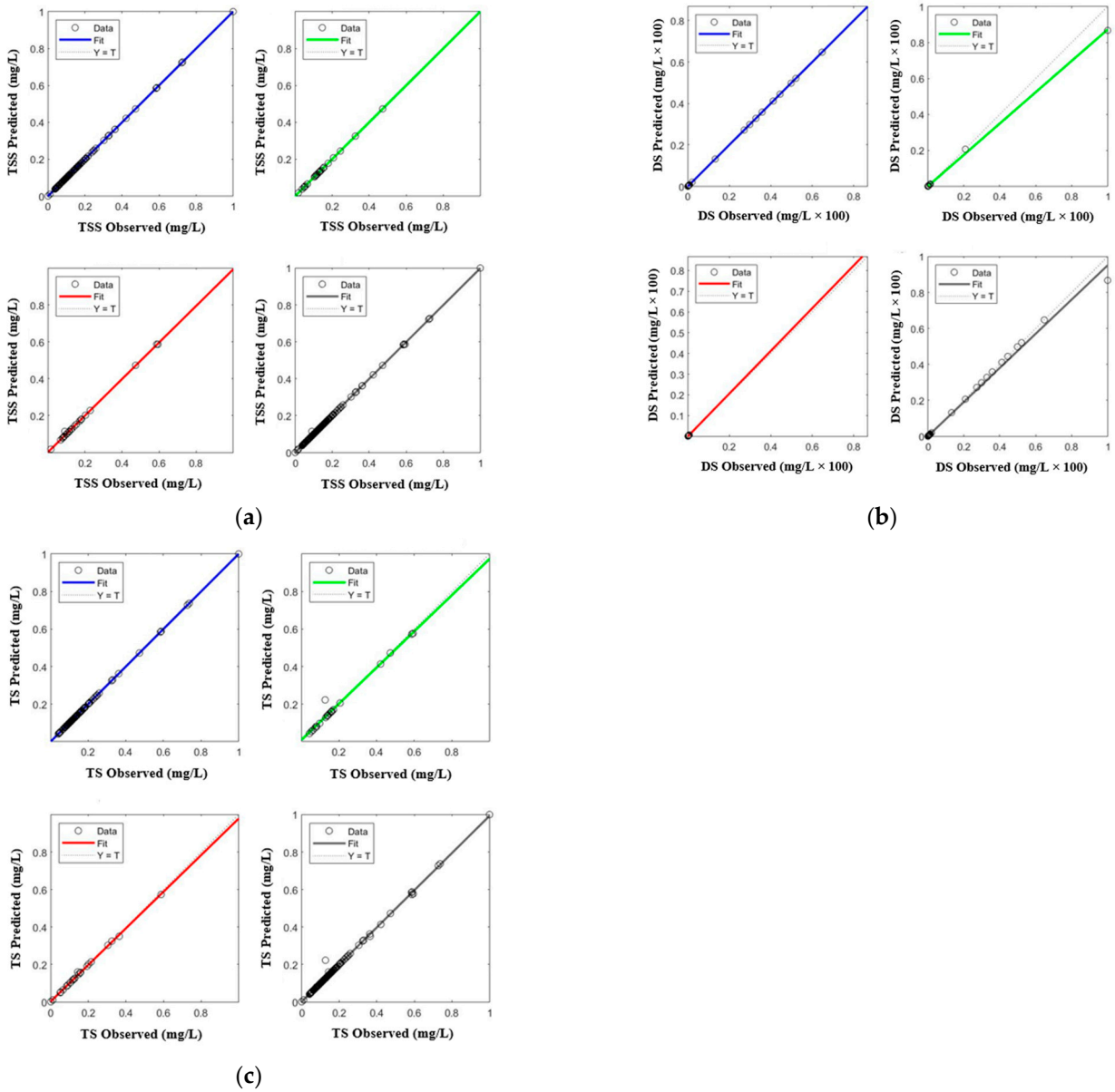

Figure 3. The scatter plots of ANN models in predicting (**a**) TSS, (**b**) DS and (**c**) TS.

Comparisons between Regression, SVM, and ANN models

Based on Tables 3 and 4, ANN model has proven to be good in predicting TS, TSS, and DS with high degree of accuracy and robustness by obtaining high values of R^2 and lower values of RMSE and MSE. Moreover, rational quadratic GPR model also exhibited similar performance, however, over-fitting might occurred during modeling since the model obtained R^2 = 1.00 for overall datasets in all three predictions. A research conducted by Zhou et al. [17] also used GPR models to predict water quality parameters, but it does not achieved results that is too perfect like the one obtained in this study. Similarly, the SVM models also did not performed well as it always exhibited poor performances. Past researches done by Najah et al. [16] and also several studies performed by [30] and [31] have proven that SVM model able to predict many kinds of water quality parameters. Contrary, the developed SVM models in this study did not achieved the same outcome. The reason was probably that SVM model is much suitable to be used to identify subtle patterns in a complex datasets [32] or it is because SVM model performed too well in a small datasets. Therefore, it can be concluded that ANN models were stable in forecasting the water quality parameters. In addition, fine tree, boosted tree, bagged tree, and exponential GPR models also able to be used to predict TSS, TS, and DS in Langat River.

4. Conclusions

This study has used nine different machine learning models, which consist of six regression models, two support vector machine models, and one artificial neural network model to predict dissolved solids (DS), total solids (TS), and total suspended solids (TSS) in Langat River, Malaysia. A total number of 25 water quality parameters that have been retrieved from the Department of Irrigation and Drainage (DID) have been applied as inputs in the modelling. Furthermore, ANN and rational quadratic GPR models turn out to be excellent in predicting TS, DS, and TSS as the models successfully achieved high accuracy and low errors but, the GPR model exhibited the characteristic of over-fitting since this study used small datasets. Similarly, SVM models also obtained poor accuracy in predicting the water quality parameters and over-fitting might occurred in the modeling. As for the remaining models, the models have shown fair to poor performances in predicting the water quality parameters. Therefore, ANN model was the best machine learning model in predicting DS, TSS, and TS of Langat River. Lastly, future research can be conducted to overcome this problem by changing the type of kernel function used for SVM and GPR models, or applying more complex datasets for the modeling.

Author Contributions: Conceptualization, N.N.M.R. and G.H.; methodology, N.N.M.R. and G.H.; software, N.N.M.R.; validation, N.N.M.R.; formal analysis, N.N.M.R. and G.H.; investigation, N.N.M.R. and G.H.; resources, N.N.M.R., G.H., M.M., B.M.E.E., A.O.Y.M. and M.M.K.; data curation, N.N.M.R..; writing—original draft preparation, N.N.M.R.; writing—review and editing, N.N.M.R. and G.H.; visualization, N.N.M.R., G.H., M.M., B.M.E.E., A.O.Y.M. and M.M.K.; supervision, G.H.; project administration, G.H., M.M., B.M.E.E., A.O.Y.M. and M.M.K.; funding acquisition, G.H., M.M., B.M.E.E., A.O.Y.M. and M.M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Grant Scheme (FRGS) with the project code 20190105FRGS from the Ministry of Higher Education Malaysia, and Universiti Tenaga Nasional (UNITEN) BOLD Fund 2022, and the Deanship of Scientific Research at Umm Al-Qura University under Grant Code: (22UQU4400271DSR09).

Data Availability Statement: The data presented in this study are available on reasonable request from the corresponding author.

Acknowledgments: The authors would like to express their gratitude to the Ministry of Higher Education Malaysia for the Fundamental Research Grant Scheme (FRGS) and Universiti Tenaga Nasional (UNITEN) for the BOLD Fund, and the Deanship of Scientific Research at Umm Al-Qura University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Varol, M. Use of water quality index and multivariate statistical methods for the evaluation of water quality of a stream affected by multiple stressors: A case study. *Environ. Pollut.* **2020**, *266*, 115417. [CrossRef]
2. Mekuria, D.M.; Kassegne, A.B.; Asfaw, S.L. Assessing pollution profiles along Little Akaki River receiving municipal and industrial wastewaters, Central Ethiopia: Implications for environmental and public health safety. *Heliyon* **2021**, *7*, e07526. [CrossRef] [PubMed]
3. Hayder, G.; Solihin, M.I.; Mustafa, H.M. Modelling of river flow using particle swarm optimized cascade-forward neural networks: A case study of Kelantan River in Malaysia. *Appl. Sci.* **2020**, *10*, 8670. [CrossRef]
4. Al-Sulttani, A.O.; Al-Mukhtar, M.; Roomi, A.B.; Farooque, A.A.; Khedher, K.M.; Yaseen, Z.M. Proposition of new ensemble data-intelligence models for surface water quality prediction. *IEEE Access* **2021**, *9*, 108527–108541. [CrossRef]
5. Jagaba, A.H.; Kutty, S.R.M.; Hayder, G.; Baloo, L.; Abubakar, S.; Ghaleb, A.A.S.; Lawal, I.M.; Noor, A.; Umaru, I.; Almahbashi, N.M.Y. Water quality hazard assessment for hand dug wells in Rafin Zurfi, Bauchi State, Nigeria. *Ain Shams Eng. J.* **2020**, *11*, 983–999. [CrossRef]
6. Fathi, E.; Zamani-Ahmadmahmoodi, R.; Zare-Bidaki, R. Water quality evaluation using water quality index and multivariate methods, Beheshtabad River, Iran. *Appl. Water Sci.* **2018**, *8*, 1–6. [CrossRef]
7. Kang, G.; Gao, J.Z.; Xie, G. Data-driven water quality analysis and prediction: A survey. In Proceedings of the 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), Redwood City, CA, USA, 6–9 April 2017; pp. 224–232.
8. Alqahtani, A.; Shah, M.I.; Aldrees, A.; Javed, M.F. Comparative Assessment of Individual and Ensemble Machine Learning Models for Efficient Analysis of River Water Quality. *Sustainability* **2022**, *14*, 1183. [CrossRef]
9. Hmoud Al-Adhaileh, M.; Waselallah Alsaade, F. Modelling and prediction of water quality by using artificial intelligence. *Sustainability* **2021**, *13*, 4259. [CrossRef]
10. Koranga, M.; Pant, P.; Kumar, T.; Pant, D.; Bhatt, A.K.; Pant, R.P. Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand. *Mater. Today Proc.* **2022**, *57*, 1706–1712. [CrossRef]
11. El Bilali, A.; Taleb, A. Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment. *J. Saudi Soc. Agric. Sci.* **2020**, *19*, 439–451. [CrossRef]
12. Moubayed, A. Optimization Modeling and Machine Learning Techniques towards Smarter Systems and Processes. Ph.D. Thesis, The University of Western Ontario, London, ON, Canada, 2018; p. 5573.
13. Gakii, C.; Jepkoech, J. A classification model for water quality analysis using decision tree. *Eur. J. Comput. Sci. Inf. Technol.* **2019**, *7*, 1–8.
14. Niroobakhsh, M.; Musavi-Jahromi, S.H.; Manshouri, M.; Sedghi, H. Prediction of water quality parameter in Jajrood River basin: Application of multi layer perceptron (MLP) perceptron and radial basis function networks of artificial neural networks (ANNs). *Afr. J. Agric. Res.* **2012**, *7*, 4131–4139. [CrossRef]
15. Talib, A.; Amat, M.I. Prediction of chemical oxygen demand in Dondang River using artificial neural network. *Int. J. Inf. Educ. Technol.* **2012**, *2*, 259–261. [CrossRef]
16. Najah, A.; El-Shafie, A.H.A.H.; Karim, O.A.; Jaafar, O.; El-Shafie, A.H. An application of different artificial intelligences techniques for water quality prediction. *Int. J. Phys. Sci.* **2011**, *6*, 5298–5308. [CrossRef]
17. Zhao, J.; Guo, H.; Han, M.; Tang, H.; Li, X. Gaussian process regression for prediction of sulfate content in lakes of China. *J. Eng. Technol. Sci.* **2019**, *51*, 198–215. [CrossRef]
18. Dogan, E.; Sengorur, B.; Koklu, R. Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. *J. Environ. Manag.* **2009**, *90*, 1229–1235. [CrossRef]
19. Basheer, A.O.; Hanafiah, M.M.; Abdulhasan, M.J. A study on water quality from Langat River, Selangor. *Acta Sci. Malays. (ASM)* **2017**, *1*, 1–4. [CrossRef]
20. Soo, E.Z.X.; Jaafar, W.Z.W.; Lai, S.H.; Othman, F.; Elshafie, A.; Islam, T.; Srivastava, P.; Hadi, H.S.O. Evaluation of bias-adjusted satellite precipitation estimations for extreme flood events in Langat river basin, Malaysia. *Hydrol. Res.* **2020**, *51*, 105–126. [CrossRef]
21. Hassim, M.; Yuzir, A.; Razali, M.N.; Ros, F.C.; Chow, M.F.; Othman, F. Comparison of rainfall interpolation methods in Langat River Basin. *IOP Conference Series: Earth Environ. Sci.* **2020**, *479*, 012018. [CrossRef]
22. Saudi, A.S.M.; Kamarudin, M.K.A.; Ridzuan, I.S.D.; Ishak, R.; Azid, A.; & Rizman, Z.I. Flood risk index pattern assessment: Case study in Langat river basin. *J. Fundam. Appl. Sci.* **2017**, *9*, 12–27. [CrossRef]
23. Rizal, N.N.M.; Hayder, G.; Yussof, S. River water quality prediction and analysis—Deep learning predictive models approach. In *Advances in Science, Engineering, and Technology (ASTI)*; (Submitted for review-in press); Springer: Berlin/Heidelberg, Germany.
24. Rizal, N.N.M.; Hayder, G.; Yusof, K.A. Water quality predictive analytics using an artificial neural network with a graphical user interface. *Water* **2022**, *14*, 1221. [CrossRef]
25. Ahmed, M.F.; Mokhtar, M.; Alam, L.; Ta, G.C.; Lee, K.E.; Khalid, R.M. Recognition of local authority for better management of drinking water at the Langat River Basin, Malaysia. *Int. J. Eng. Technol.* **2018**, *7*, 148–154. [CrossRef]
26. Chen, Y.; Song, L.; Liu, Y.; Yang, L.; Li, D. A review of the artificial neural network models for water quality prediction. *Appl. Sci.* **2020**, *10*, 5776. [CrossRef]
27. Chollet, F. *Deep Learning with Python*; Manning: Shelter Island, NY, USA, 2021.

28. Tiyasha; Tung, T.M.; Yaseen, Z.M. Asurvey on river water quality modellinh using artificial intelligence models: 2000–2020. *J. Hydrol.* **2020**, *585*, 124670. [CrossRef]
29. Ewusi, A.; Ahenkorah, I.; Aikins, D. Modelling of total dissolved solids in water supply systems using regression and supervised machine learning approaches. *Appl. Water Sci.* **2021**, *11*, 1–16. [CrossRef]
30. Li, X.; Sha, J.; Wang, Z.L. A comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen. *Hydrol. Res.* **2017**, *48*, 1214–1225. [CrossRef]
31. Abobakr Yahya, A.S.; Ahmed, A.N.; Binti Othman, F.; Ibrahim, R.K.; Afan, H.A.; El-Shafie, A.; Fai, C.M.; Hossain, M.S.; Ehteram, M.; Elshafie, A. Water quality prediction model based support vector machine model for ungauged river catchment under dual scenarios. *Water* **2019**, *11*, 1231. [CrossRef]
32. Rajaee, T.; Jafari, H. Two decades on the artificial intelligence models advancement for modeling river sediment concentration: State-of-the-art. *J. Hydrol.* **2020**, *588*, 125011. [CrossRef]