APPLICATIONS OF
MODELLING AND SIMULATION
http://arqiipubl.com/ams eISSN 2600-8084 VOL 7, 2023, 132-145

Application of Long-Short Term Memory for Accurate Biochemical Oxygen Demand Prediction in Rivers through Water Quality Parameters

Norashikin M. Thamrin[1]*, Azhar Jaffar[2], Megat Syahirul Amin Megat Ali[1,3], Muhammad Farid Misnan[1], Ahmad Ihsan Mohd Yassin[1,3], Noorolpadzilah Mohamed Zan[2] and Nik Nor Liyana Nik Ibrahim[4]

[1]School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, Shah Alam, Malaysia
[2]Department of Electrical Engineering, Politeknik Ungku Omar, Ipoh, Malaysia
[3]Microwave Research Institute, Universiti Teknologi MARA, Shah Alam, Malaysia
[4]Department of Chemical and Environmental Engineering, Faculty of Engineering, Universiti Putra Malaysia, Malaysia

*Corresponding author: norashikin@uitm.edu.my

Submitted 28 August 2023, Revised 29 September 2023, Accepted 18 October 2023, Available online 31 October 2023.

Abstract: Evaluating water quality is crucial for preserving the quality of river water. However, the typical technique of getting biochemical oxygen demand (BOD) values via laboratory testing might take several days, delaying the application of real-time measurement to improve water quality. This paper suggests using machine learning to predict BOD values from eight water quality measurements. The BOD rate in the Klang River, Selangor, Malaysia, was estimated using the long short-term memory (LSTM) method. The model was trained using historical data collected from eleven water collection points along the river. The predictive test results indicated that the LSTM model with 8 water parameters as input gave the most accurate predictions compared to the models with 5 and 3 water parameters. The results of this study indicate that machine learning methods can be used to predict BOD levels in real-time. It enables water quality managers to enhance water quality and safeguard human health proactively.

Keywords: BOD prediction; Deep neural network; Klang River; LSTM; Prediction.

1. INTRODUCTION

Organic or chemical pollutants are commonly responsible for water pollution. Microorganisms such as bacteria and viruses that are generated by human and animal waste and plant residues contaminate organic substances. Chemical contamination is caused by pesticides such as nitrates and phosphates, industrial acids and hydrocarbons, household goods, and heavy metals. Water's physical, biological, and chemical components are frequently tested for pollution. These properties are referred to as physicochemical water parameters. The Biochemical Oxygen Demand (BOD) is a common criterion for water quality as stated by [1]. The BOD concentration can be calculated using an index method that assesses the environmental impact of released wastewater. According to [2], BOD is a measure of the amount of oxygen consumed by microorganisms during the oxidation of organic materials. The higher the level of BOD in the water, the lower the Dissolved Oxygen (DO), as experimented by [3]. Fish and other aquatic creatures will suffer from a shortage of DO. Furthermore, stated by Prambudy *et al*. [4], BOD is important because it offers information on water quality factors directly related to the water body's health. However, as claimed in [5], in determining water quality, the Chemical Oxygen Demand (COD) is analogous to the BOD. As agreed by [6], the COD indicates how much oxygen organic water contaminants used as they oxidize and create inorganic end products. Because of the shorter testing duration, Hasanah *et al*. [7] suggested that COD is to be used instead of BOD. However, several variables influence BOD and COD concentrations in water, as well as their temporal change. Measuring BOD value with a chemical solution is time-consuming and expensive. Existing experimental and statistical methods are incapable of resolving time limitations and have limited capability.

Alamelu *et al*. [8] stated that testing for BOD is a time-consuming technique that takes five days from data collection to analysis and requires samples to be incubated for a lengthy period. This is agreed by [9], where laboratory settings, notably changes in the microbial variety of the inoculum used, can cause results to differ by 20%. According to [10], the rate of oxygen required for the oxidation of organic wastes is addressed in laboratory BOD testing, but not the oxygen consumed by living organisms in water. Susilowati *et al*. [11] express the concern about other complicating factors including the oxygen

requirement created by algal respiration in the sample and the likelihood of ammonia oxidation. This statement is supported by [12], toxic substances, for example, may hinder microbial activity, resulting in a lower reported BOD value. The laboratory conditions used to assess BOD are not the same as those seen in aquatic systems. Furthermore, Jouanneau *et al*. [9] emphasized that significant differences in test results may exist due to the laboratories' approach to sample preservation, the grade of chemicals utilized, and the testing process used. As a result, there is an urgent need to investigate a viable secondary (indirect) technique for BOD prediction based on historical water data.

Researchers are beginning to pay attention to the application of machine learning in estimating the value of BOD. This machine learning technique has already been employed in studies by researchers [13, 14]. Ooi *et al.* [13] employed a multi-layer perceptron (MLP) to improve the accuracy of BOD readings based on physical and chemical characteristics in water. Samir *et al*. [14] conducted research on home drain water and how it might be recycled. They employed an artificial neural network (ANN) model to quickly gather BOD values so that the water quality index (WQI) could be calculated. Jiang *et al*. [15] investigated the same issue and discovered a pollution problem in the urban drainage system that is difficult to monitor and manage. To address this issue, critical water quality indicators must be identified in sewer water quality evaluation and forecasting. To predict BOD levels, they employed multiple linear regression (MLR) and multilayer perceptron (MLP).

Furthermore, its management is prone to human error and administrative challenges, resulting in the omission of critical measurements. Nevertheless, thanks to recent technological advances, this problem can be handled by employing an electronic sensor system capable of promptly and accurately monitoring the BOD concentration in water. Regrettably, this advanced electronic BOD measurement equipment is costly. Therefore, this work investigates the opportunities of using the advancement in artificial intelligence to predict the BOD level given prior information from other water parameters.

2. METHODOLOGY

A fundamental study is proposed to address these gaps and examine the relationship between BOD and other well-established characteristics. The characteristics are Chemical Oxygen Demand (COD), Dissolved Oxygen (DO), pH, temperature, turbidity, Dissolved Solids (DS), Suspended Solids (SS), and ammoniacal nitrogen (NH3-N). The parameters give a cost-effective solution, while the in-situ technique addresses the issue of time constraints via an intelligent predictive model. A long short-term memory (LSTM) recurrent neural network structure is appropriate since it can extract long-term relationships between multiple variables, resulting in accurate BOD profiling.

2.1 Data Extraction

The Klang River provided the information for this study. The Klang River passes through Selangor and Wilayah Persekutuan Kuala Lumpur. Figure 1 depicts the locations of seven tributaries, including the Batu River, Gombak River, Penchala River, Damansara River, Ampang River, Kerayong River, and Kuyuh River. Since any pollution or change in the water quality rate from any of them is likely to impact changes in the water quality or BOD parameter of the mainstream, the Klang River, studies are undertaken based on the data from these tributaries.

The water quality data from the seven tributaries used in this study were gathered by the Department of Environment (DOE) at stationary water monitoring stations (denoted as 1KXX in Figure 1) from 2012 to 2018. The data is then separated into two sections. The first portion provides data from 2012 to 2017 that will be utilized for model Training, Validation, and Testing. The latter, from 2017 to 2018, is used as prediction data. Figure 2 depicts the stations utilized to collect water information for this study. Five water data collection stations are positioned within the main river. At the same time, the remaining six are located within the six key rivers that contribute to the water changes in the main river. Based on eight water parameters used and it involves the collection of data from 11 water data collection stations, a total of 150,448 pieces of data starting from 2012 to 2017 were used to test the deep learning model. Data collection is based on daily readings from each station.

Figure 1. The main rivers that contribute to Klang River [8]

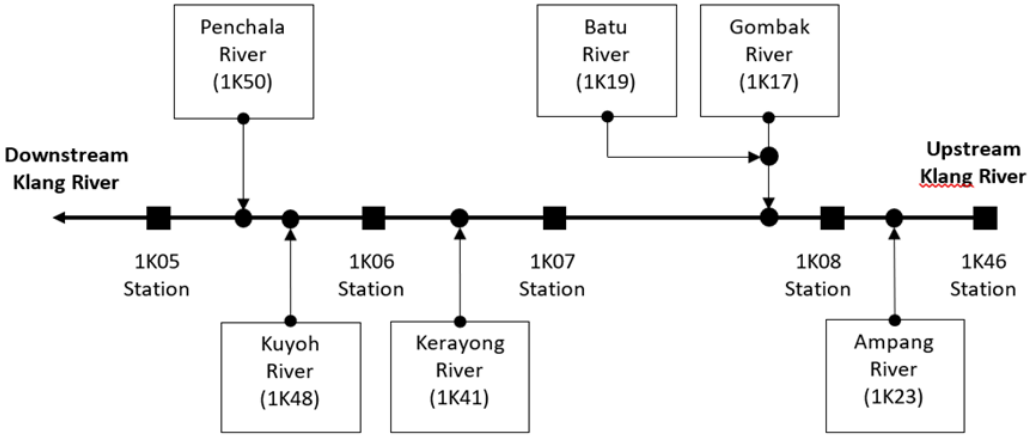


Figure 2. Water data collection station involved along the Klang River

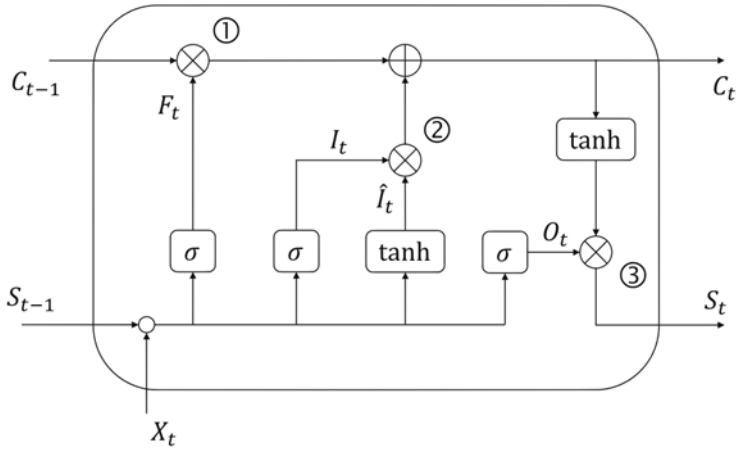


Figure 3. The internal structure of LSTM hidden layer cells [22]

2.2 Water Parameter Data Restructuring using Interpolation

The hydrological data of the Klang River collected from DOE were not evenly dispersed, and some of the data is significantly lacking due to possibilities such as sensor breakdowns, interrupting data collection schedules, and the officers' unavailability at the time. The DOE data is generally not routinely recorded by day, month, or year. Interpolation is a technique to rearrange the water parameter data needed to generate structured data. Before interpolation, the correlation coefficient test is performed on the interpolated data to see whether there has been a significant change in the test findings.

2.3 Correlation Coefficient on Non-Interpolated and Interpolated Water Parameter Data

As mentioned by Isaac *et al.* [16], the correlation coefficient is the precise measurement that indicates the degree of the linear link between two variables in correlation analysis. Meanwhile, the Pearson correlation coefficient (r) is one of the most utilized correlation metrics in practice as stated in [17]. A correlation coefficient close to 1 or -1 denotes the most significant positive or negative correlation between two variables. In contrast, Taylor [18] suggested that a number close to 0 implies no statistically significant linear link between two variables with a correlation coefficient of 0.05 and/or 0.01. There are two possibilities for probabilities: either a positive correlation or a negative correlation. There is a positive correlation when two variables respond in the same direction. In the case of a negative correlation, the two factors being compared will act in the opposite direction, with one parameter increasing while the other decreases, which has been agreed by several researchers [16],[19]. The correlation coefficient test was performed following the remodelling of the water parameter data to examine if there was a significant difference in the correlation coefficient between BOD and other water parameters before and after the interpolation process.

2.4 LSTM Neural Network Model

Hongxiang *et al*. [20] have mentioned that the LSTM structure enables the model to remember and keep a state at any time. It is accomplished using specifically engineered gates and memory cells. While on the other hand, Zhenbo *et al.* [21] discovered that the LSTM's long-term memory capabilities are appropriate for processing, classifying, and predicting time series data. LSTM is well-suited to situations involving long-term data series. This is due to the network's ability to learn long-distance temporal dependencies. The internal structure of hidden layer cells in the LSTM network is shown in Figure 3.

The first gate is known as the forget gate, the second as the input gate, and the third as the output gate. The forget gates will determine whether the information is retained. The sigmoid function will be applied to the current input data and data from the previous state, yielding a result between 0 and 1. A value near 0 indicates forgetting, and vice versa. The calculation for forget gates is shown in Equation (1).

$$F_t = \sigma\left(W_f \,.\, [S_{t-1}, X_t] + B_f\right) \tag{1}$$

where σ denotes the sigmoid function. W_f is the weight matrixes and the [.,.] represents the concatenation operation. The S_{t-1} represents the previous cell. X_t is the input sequence data and B_f is the bias vectors [23]. The transfer function is initiated by feeding the prior hidden state and current input to the sigmoid function, which determines the relevance of the value by changing it to be between 0 and 1. In this phase, 1 denotes importance and 0 denotes insignificance. The tanh function is additionally passed the current input and hidden state to regulate the value between -1 and 1. The sigmoid output will be multiplied by these numbers. The sigmoid output will determine whether the information from the tanh output is retained. The following are the input gate calculations:

$$I_t = \sigma\left(W_i \,.\, [S_{t-1}, X_t] + B_i\right) \tag{2}$$

$$\hat{I}_t = \tanh\left(W_i \,.\, [S_{t-1}, X_t] + B_i\right) \tag{3}$$

where W_i and B_i are the bias vectors. The tanh denotes the hyperbolic tangent function. The output gate plays a crucial role in determining the content of the next hidden state. At this point, the hidden state retains valuable information from prior inputs, which can also be utilized for making predictions. To update the cell state, both the current input and the previous hidden state are introduced to the sigmoid function, resulting in a freshly adjusted cell state that is then passed through the tanh function. The multiplication of the tanh and sigmoid outputs serves as the mechanism for determining which information should be retained. Subsequently, the output comprises the hidden state, the new cell state, and this new hidden state proceeds to the next time step for further processing. The calculation is as Equation (4):

$$O_t = \sigma\left(W_o \,.\, [S_{t-1}, X_t] + B_o\right) \tag{4}$$

$$S_t = O_t \,.\, \tanh\left(C_t\right) \tag{5}$$

where W_o are the weight matrices and B_o are the bias vectors. C_t is the output of the current cell, containing the cell state. For the cell state, the calculation is as in Equation (6).

$$C_t = F_t \,.\, C_{t-1} + I_t \,.\, \hat{I}_t \tag{6}$$

The cell state plays an important role in maintaining the information between each time step in the chain. Its content is modified through the forget gate and input gate every time the cell state passes through at different time-step. The primary idea behind the LSTM is the cell state and a few distinct gates. All sequencing chains receive relative data from the cell state. During the sequence's processing, the cell state will store relevant information. As a result, even data from earlier time steps will lead to later time steps, reducing the impact on short-term memory.

2.5 LSTM Model Pre-Processing

Since eight physicochemical parameters are involved as the input to the LSTM system, it is certainly not easy to get the best adjustment before the LSTM system processes all this data to get the best prediction. When working using a long time-series data will face problems mainly related to data inconsistency and outliers. The issue of outliers is due to incorrect measurements, sensor damage, incorrect procedures and natural disasters, which has been discussed by several researchers [24],[25]. In addressing this gap, pre-processing techniques are used to process the data before the LSTM model uses it.

2.5.1 Z-score Normalization

The z-score is very good at handling outliers in data problems. Cinar *et al*. [26] stated that the advantage of having the z-score in the process is that it can complement various characteristics into a single scale. Extremely obvious outliers can be reshaped so that they no longer stand out. It is accomplished by rescaling the data to have the features of a Gaussian distribution, also known as a normal distribution. The equation for the *z*-score is as in Equation (7).

$$z = \frac{x_o - \mu}{\sigma} \tag{7}$$

where x_o is the original value, μ is the mean for the x_o train, and σ is the standard deviation to measure how spread the numbers are. The *z* value represents the range between the mean of and from the standard deviation. There will be a positive and negative standard score where the value above the mean will have positive scores, and the value below the mean level will have negative scores.

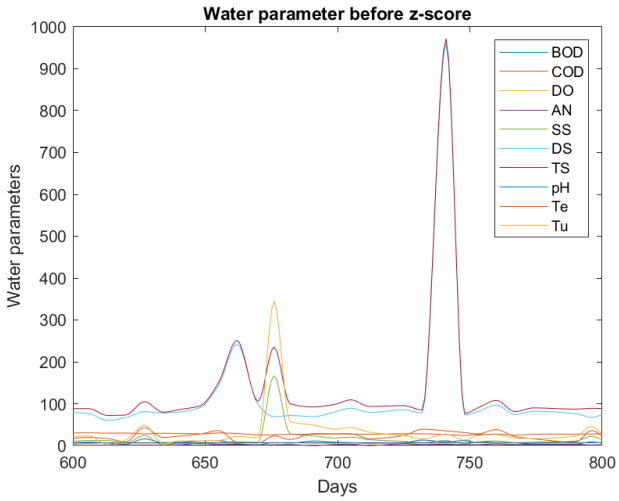


Figure 4. Water parameter before z-score process

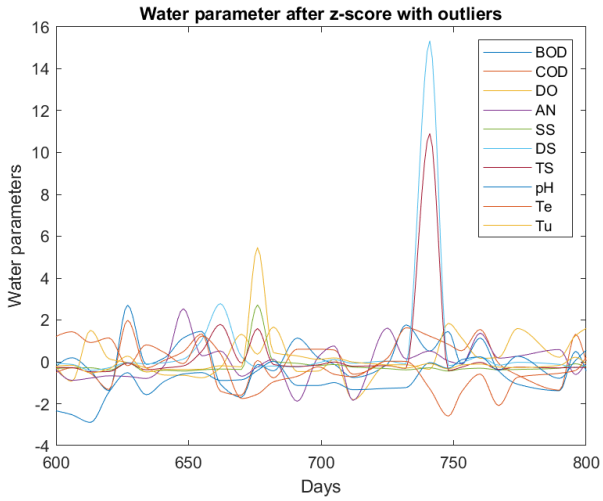


Figure 5. Water parameter after z-score process

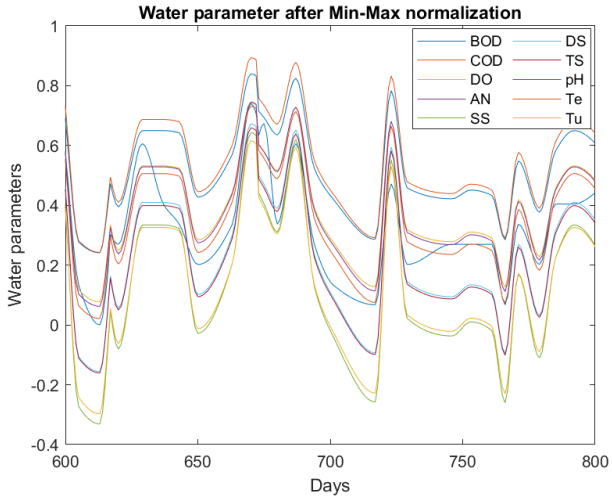


Figure 6. Water parameter after Min-Max normalization process from 1K46

2.5.2 Min-Max Normalization

The min-max technique is the primary data normalization that can normalize data as equally important. While transforming the minimum data equals 0 and the maximum data equals 1, all the other data will be transformed within the limitation between 0 and 1, as mentioned by Ghimire *et al.* [27]. The equation for the min-max calculation is shown in Equation (8).

$$x_a{}' = \frac{x_a - x_{amin}}{x_{amax} - x_{amin}} \tag{8}$$

where $x_a{}'$ is the normalized value and x_a is the original value. The x_{amax} is the maximum value in the data group and x_{amin} is the minimum value in the data group. It will prevent all the small data from being pressed smaller by the dominant big numbers. Combining the z-score and min-max normalization will cater to the pros and cons when using both techniques individually.

The initial step in preprocessing water parameter data involves applying a normalization function to the data. Each water parameter possesses unique characteristics in terms of its measurement range, as illustrated in Figure 4. As depicted in Figure 4, water parameters exhibit varying measurement ranges, which can impact the implementation of a deep learning system for analysis. Z-score normalization is commonly employed due to its ability to bring different attributes onto a consistent scale [26]. Figure 5 indicates that most water parameter data points fall within the range of -3 to 3. A z-score close to zero suggests that a data point is near the average, while values greater than or less than 3 are considered outliers, as described in reference [28]. The presence of outliers can significantly increase the standard deviation, reducing the ability to detect meaningful differences and potentially leading to errors, as noted in reference [29]. It is important to exercise caution when removing outliers because sudden spikes in pollutant levels can result in elevated readings, as discussed in reference [30].

Deep learning algorithms aim to identify trends and patterns within data points [31]. However, this can be problematic if the data points are not treated equally by the system. The min-max normalization technique is a fundamental method for ensuring that data is normalized and treated with equal importance. This is evident in Figure 6, where most of the water parameter data now falls within the range of 0 to 1, although some values remain below 0. Min-max normalization is applied to water parameter data to prevent smaller values from being unduly influenced by dominant larger values. Combining both z-

score and min-max normalization addresses the advantages and disadvantages associated with each technique when used independently.

2.6 Input Dataset of Training, Validation, and Testing for Deep Neural Network Model

For a good and effective trained model, the total data divided into a training set, validation set, and testing set plays a significant role [32]. The data presented in this study is novel and has never been assessed using the Deep Learning Neural Network model. Table 1 depicts the separation of water quality data into training, validation, testing, and final prediction. The data set is comprised of water parameter measurements taken between 2012 and 2017. This data will be divided into 3 parts: training, validation, and testing. Meanwhile, data from April 2017 to November 2018 is used as final prediction data. The distribution percentage usually consists of 80% of datasets used for training purposes and another 20% percent used for validation. This data composition is also suggested [33],[34]. However, it is not mandatory to use a specific percentage. This test aims to see the model output accuracy by testing a different number of datasets. Meanwhile, Genc *et al.* [35] suggested that building a deep learning model begins with a dataset that comprises previously collected information about the topic under investigation. Using the training dataset, the system can be constructed with different model parameter values, and it can be tested by each trained model against the validation set to see how well it performs [36]. However, as the validation set comprises samples with known provenance but unknown classifications, predictions on the validation set may be used to assess model accuracy [37]. Moreover, Choi *et al.* [38] stressed that the test dataset that is used to test the model output must be using the unknown value to the system. This value cannot be taken from the training or the validation data. The data from the test set is fed into the model, and the model's predictions are compared to the data from the test set to see if they agree. This argument has been supported by several researchers [39],[40],[41]. The tests are performed using the method in Table 2.

Table 1. Distribution of water quality data for model testing and prediction

No.	Data Purpose		Beginning Month	Ending Month	Period
1.	Training/Validation/Testing		May 2012	Mac 2017	5 years 3 months
2.	Final Prediction		April 2017	November 2018	1 year 8 months

Table 2. Various model dataset sizes and water parameter combinations

	Set 1	Set 2	Set 3
Training (%)	70	70	70
Validation (%)	15	15	15
Testing (%)	15	15	15
Number of Water Parameter	Water parameters with the highest correlation coefficient to BOD	Water parameters with the lowest correlation coefficient to BOD	Water parameters combination of highest and lowest correlation coefficient to BOD
Number of Water Parameter involved as input	3	5	8

There is no specific technique used to determine the number of dataset divisions. The division of 70% of data for training, 15% of data for validation, and another 15% for testing were used in this study. This option is made to increase the use of data for testing and validation purposes and, at the same time, reduce data for training purposes. If this happens, it will show the accuracy rate of the model output when presented with less training data. All eight water parameters can be graded from highest to lowest correlation coefficient value based on the results of the correlation coefficient test. New tests are performed to determine whether there is a difference in influence on the number of water parameters based on the water parameter with the highest correlation coefficient, the lowest correlation coefficient, or a combination of high and low correlation coefficients. These aspects are considered while determining the number of water parameters. Three experiments were carried out: (1) With the highest correlation coefficient value for the model inputs; (2) With the lowest correlation coefficient value; (3) With all high and low correlation coefficient values for all eight water parameters. The performance of each dataset is tested using Mean Absolute Error (MAE) and Root-Mean-Square-Error (RMSE).

3. RESULTS AND DISCUSSION

3.1 Correlation Coefficient for Interpolated Water Parameter Data

The correlation coefficient test results on the interpolated data are shown in Table 3. Although, on average, the correlation coefficient test indicates a poor connection between BOD and all other water parameters except for COD, DO, and NH3-N, this does not suggest that changes in this water parameter cannot be directly related to changes in BOD. Varying water pollution levels influence these variables, and the absence of significant water pollution leads to low correlation values.

Table 3. Correlation coefficient results on interpolated water parameter data

Water Parameters	BOD										
	1K05	1K06	1K07	1K08	1K17	1K19	1K23	1K41	1K46	1K48	1K50
COD	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
DO	0.0	-0.3	-0.2	-0.1	-0.2	-0.1	-0.3	-0.3	-0.3	0.1	-0.1
NH3-N	0.4	0.3	0.1	0.2	0.4	0.4	0.3	0.5	0.3	0.4	0.1
TSS	0.1	0.0	-0.1	-0.1	-0.1	0.0	0.0	0.2	0.1	0.0	0.1
TDS	0.1	0.1	0.1	0.1	0.2	0.1	0.2	0.2	0.2	0.2	0.1
pH	-0.2	0.0	-0.1	-0.1	-0.1	-0.3	-0.1	0.2	0.0	-0.1	0.2
Temp	0.1	0.0	0.0	0.1	0.2	0.0	0.0	0.1	0.1	0.2	0.0
TUR	0.1	0.2	0.1	0.0	0.0	0.1	0.0	0.2	0.2	0.0	0.2

Table 4. Training regression coefficients results for the LSTM model utilizing 3, 5, and 8 water parameters.

No. of Parameter	Station										
	1K05	1K06	1K07	1K08	1K17	1K19	1K23	1K41	1K46	1K48	1K50
3	0.9999	0.9994	0.9988	0.9997	0.9996	0.9990	0.9996	0.9993	0.9997	0.9991	0.9994
5	0.9999	0.9997	0.9997	0.9998	0.9998	0.9998	0.9996	0.9995	0.9998	0.9993	0.9998
8	0.9999	0.9998	0.9997	0.9999	0.9998	0.9998	0.9998	0.9997	0.9999	0.9997	0.9999

Table 5. Validation regression coefficients results for the LSTM model utilizing 3, 5, and 8 water parameters.

No. of Parameter	Station										
	1K05	1K06	1K07	1K08	1K17	1K19	1K23	1K41	1K46	1K48	1K50
3	0.9933	0.9970	0.9977	0.9976	0.9987	0.9965	0.9996	0.9998	0.9996	0.9990	0.9897
5	0.9957	0.9987	0.9995	0.9986	0.9998	0.9998	0.9997	0.9999	0.9997	0.9989	0.9967
8	0.9975	0.9992	0.9997	0.9992	0.9998	0.9991	0.9999	0.9999	0.9998	0.9996	0.9978

3.2 BOD Prediction using LSTM

Referring to Table 3, three water parameters with a high correlation coefficient with BOD are COD, DO, and NH3-N. The other five water parameters with a lower correlation coefficient are TSS, TDS, pH, Temp, and TUR. This result allows the water parameters to be divided into three parts, as stated in Table 2. The three water parameters in Set 1 are COD, DO, and NH3-N. The five water parameters in Set 2 are TSS, TDS, pH, Temp, and TUR. Set 3, on the other hand, consists of a combination of all eight water parameters used.

Furthermore, each model was tested with distinct datasets: 70% training, 15% validation, and 15% testing to determine the effectiveness of training, validating, and testing the model in producing good BOD predictions. This study utilized no specific approaches, such as cross-validation, to estimate the data distribution ratio for training, validation, and testing. Therefore, the technique is agreed by [42] to be more appropriate when the available data is small and limited.

The outcomes of the tests performed using the LSTM model are organized into numerous tables. The regression coefficients test results on the LSTM model are shown in Tables 4, 5, and 6. Table 4 is the regression coefficient test for the Training process, Table 5 is for the Validation process, and Table 6 is for the Testing process. The results of these regressions are critical in determining the model's success in learning and making predictions. It also allows for a comparison with regression findings for validation to determine whether overfitting occurs between training and validation.

The results of the regression test performed by the LSTM model during the Training process may be seen in Table 4. On average, based on the readings of each water data collecting station using the 70/15/15 dataset, using three water parameters yields a regression value of 0.9994, using five parameters yields a regression reading of 0.9997, and using eight water parameters yields a regression reading of 0.9998. The average reading for this Training shows the results are encouraging and good enough for the LSTM model. A regression test is then performed on the Validation process output for the LSTM model to evaluate if there is an overshoot between the Training and Validation outputs. Table 5 displays the regression test results on the Validation model output.

According to Table 5, using three water parameters yields a regression average of 0.9971, using five water parameters yields a regression of 0.9988, and using eight water parameters yields a regression of 0.9992. When data from Table 4 are compared, the usage of three water parameters resulted in differences of 0.0023 between them, while the usage of five water parameters is 0.0009. With 8 parameters, it results in a difference of 0.0006 between them. Here, adding more water parameters minimizes the gap between the Training and Validation regressions. It shows that overfitting did not happen to the LSTM model when training was carried out due to the validation phase showing almost the same quality. Following that, regression testing is performed against the Testing procedure. Table 6 displays the regression test results for the Testing phase using three, five, and eight water parameters.

Table 6. Testing regression coefficients results for the LSTM model utilizing 3, 5, and 8 water parameters.

No. of Parameter	Station										
	1K05	1K06	1K07	1K08	1K17	1K19	1K23	1K41	1K46	1K48	1K50
3	0.9848	0.9959	0.9921	0.9972	0.9987	0.9968	0.9975	0.998	0.9958	0.9978	0.9987
5	0.9902	0.9983	0.9972	0.9972	0.9996	0.9995	0.9979	0.9984	0.9971	0.9981	0.9997
8	0.9943	0.9989	0.9979	0.9986	0.9997	0.9991	0.999	0.9989	0.9984	0.9993	0.9997

Table 7. Training MAE test results for the LSTM model utilizing 3, 5, and 8 water parameters.

No. of Parameter	Station										
	1K05	1K06	1K07	1K08	1K17	1K19	1K23	1K41	1K46	1K48	1K50
3	0.0021	0.0037	0.0027	0.0019	0.0019	0.0036	0.0015	0.0029	0.0023	0.0027	0.0045
5	0.0017	0.0026	0.0015	0.0017	0.0015	0.0014	0.0014	0.0023	0.0018	0.0024	0.0025
8	0.0014	0.0019	0.0014	0.0013	0.0015	0.0017	0.0011	0.0019	0.0016	0.0015	0.0021

Table 8. Training RMSE test results for the LSTM model with 3, 5, and 8 water parameters

No. of Parameter	Station										
	1K05	1K06	1K07	1K08	1K17	1K19	1K23	1K41	1K46	1K48	1K50
3	0.0032	0.0067	0.0051	0.0034	0.0038	0.0063	0.0026	0.0061	0.0039	0.0063	0.0069
5	0.0027	0.0046	0.0028	0.0030	0.0027	0.0026	0.0025	0.0053	0.0030	0.0058	0.0039
8	0.0023	0.0035	0.0026	0.0023	0.0025	0.0030	0.0020	0.0043	0.0027	0.0037	0.0032

Table 9. Validation MAE test results for the LSTM model employing 3, 5, and 8 water parameters

No. of Parameter	Station										
	1K05	1K06	1K07	1K08	1K17	1K19	1K23	1K41	1K46	1K48	1K50
3	0.0069	0.0038	0.0022	0.002	0.0014	0.0046	0.0010	0.002	0.0016	0.0024	0.0042
5	0.0050	0.0024	0.0010	0.0018	0.0009	0.0008	0.0009	0.0015	0.0011	0.0022	0.0020
8	0.0036	0.0017	0.0009	0.0013	0.0008	0.0020	0.0007	0.0012	0.0010	0.0013	0.0016

Table 10. Validation RMSE test results for the LSTM model utilizing 3, 5, and 8 water parameters

No. of Parameter	Station										
	1K05	1K06	1K07	1K08	1K17	1K19	1K23	1K41	1K46	1K48	1K50
3	0.0220	0.0112	0.0058	0.0068	0.0045	0.0129	0.0022	0.0036	0.0037	0.0053	0.0161
5	0.0178	0.0072	0.0027	0.0058	0.0018	0.0018	0.0019	0.0026	0.0031	0.0054	0.0091
8	0.0136	0.0058	0.0021	0.0043	0.0018	0.0066	0.0011	0.0023	0.0027	0.0035	0.0075

Referring to Table 6 for the regression test results for the LSTM model's Testing section, the average regression readings for using three water parameters are 0.9958. The result of using five water parameters is an average of 0.9976. When eight water parameters datasets are used, it gives an average of 0.9985. The number of water parameters affects the regression outcomes. MAE and RMSE tests are performed on the model according to three sections of the model, namely Training, Validation, and Testing, to understand the created LSTM model better. The outcomes of the Training tests are documented in Table 7 for Training MAE.

According to Table 7, the MAE test results demonstrate that employing 8 water parameters has the most impact and delivers the lowest average MAE value, which is 0.0016 on average, compared to 5 parameters of 0.0019 and 3 parameters of 0.0028. To further refine the results, the RMSE tests are done. Table 8 shows the results for Training RMSE. The RMSE test result in Table 8 also still shows that the number of eight water parameters gives the lowest RMSE result as in the MAE test. The trained model is tested using new data in the Validation section to validate the training data during the Training function. It is done to reduce the possibility of data overfitting. It aims to assess the Training dataset and make appropriate adjustments accurately.

The results of all MAE and RMSE tests for LSTM model Validation are shown in Table 9 and Table 10. On average, the error range values for MAE and RMSE show a downward trend compared to during the Training process. For example, if viewed from station 1K07, the lowest MAE during the Training process is 0.0014, while RMSE is 0.0026. After the Validation process is implemented, the lowest MAE is 0.0009, and RMSE is 0.0021. It shows an increase in model learning after the Validation process is implemented. One similarity in the Training process is the advantage of reading accuracy using 8 water parameters. Then, the model that has been trained and validated is tested.

The Testing phase is where the model displays the model's prediction results after the LSTM model has been trained. As with the Validation phase, where fresh data that the model does not recognize is utilized, the Testing process also employs new data that has never been used before. The MAE and RMSE testing have been carried out. The tests are the same as those performed in the Training and Validation section. Table 11 shows the MAE values, while Table 12 shows the RMSE results.

Table 11. Testing MAE test results for the LSTM model employing 3, 5, and 8 water parameters

No. of Parameter	Station										
	1K05	1K06	1K07	1K08	1K17	1K19	1K23	1K41	1K46	1K48	1K50
3	0.0154	0.0038	0.0027	0.0019	0.0022	0.0050	0.0018	0.0032	0.0024	0.0042	0.0039
5	0.0116	0.0024	0.0014	0.0016	0.0014	0.0013	0.0016	0.0025	0.0018	0.0038	0.0020
8	0.0082	0.0018	0.0012	0.0013	0.0014	0.0023	0.0013	0.0021	0.0014	0.0023	0.0017

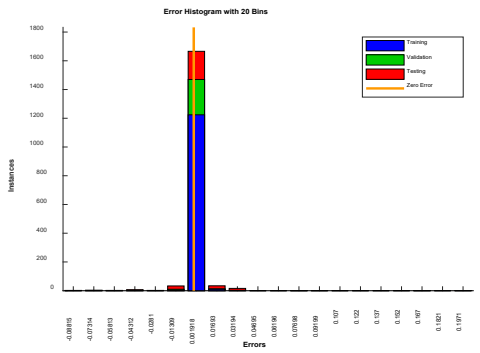
Table 12. Testing RMSE test results for the LSTM model utilizing 3, 5, and 8 water parameters

No. of Parameter	Station										
	1K05	1K06	1K07	1K08	1K17	1K19	1K23	1K41	1K46	1K48	1K50
3	0.0324	0.0114	0.0100	0.006	0.0079	0.0161	0.0063	0.0090	0.0102	0.0143	0.0106
5	0.0258	0.0075	0.0059	0.0073	0.0045	0.005	0.0057	0.0079	0.0085	0.0132	0.0056
8	0.0196	0.0060	0.0051	0.0052	0.0040	0.0083	0.0040	0.0065	0.0064	0.0081	0.0049

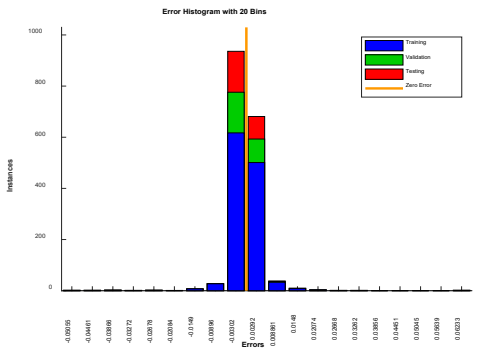
Based on Tables 11 and 12, it can be observed that using 8 water parameters as input to the model produces the best results when compared to using 5 water parameters and 3 water parameters. It is based on the results using the MAE and RMSE tests. The lowest MAE reading for all 11 water data collection stations is between 0.0012 and 0.0082, and the lowest RMSE reading is between 0.0040 and 0.0196. At this stage, it is known that using 8 water parameters gives the LSTM model acceptable accuracy. The performance of the Training, Validation, and Testing processes was then compared using 8 water parameters. The average value from the MAE test results from Tables 7, 9, and 11 and the RMSE test results from Tables 8, 10, and 12 are used for this purpose. The average MAE value for each water data collection site yields an outcome of 0.0016 for Training, 0.0015 for Validation, and 0.0023 for Testing. The average RMSE obtain for Training is 0.0029, Validation is 0.0047, and Testing is 0.0071. The average results of MAE and RMSE indicate very minor differences and are highly comparable. Overall, the results are within a reasonable range, and the model's accuracy is acceptable.

An error histogram has been plotted to see the error between the target and predicted values after Training, Validation, and Testing. Figure 7 shows the error for all water data collection stations using the LSTM model. The errors in data fitting are evenly spread around zero. It provides a strong indication of the constructed LSTM model. The error histogram depicts how near the model's predicted value is to the actual value. Positive error indicates that the outputs fell short of the goal, whereas negative error indicates that they surpassed it. It can also be noted that several stations, such as 1K06, 1K07, 1K17, 1K23, and 1K41, exhibit a very minor deviation beyond the zero area. However, it is still regarded as good due to its proximity to zero. Based on the results, it is possible to conclude that the datasets 70% Training, 15% Validation, and 15% Testing are acceptable for training the LSTM model. It was discovered that using 8 water parameters as input to the model is more optimal. According to prior research, increasing the data ratio for training can improve performance and make the model more stable, while increasing the validation ratio can improve validation performance [34]. Based on the model test findings, BOD prediction from each of the 11 tributaries was performed using 8 water parameters as input.

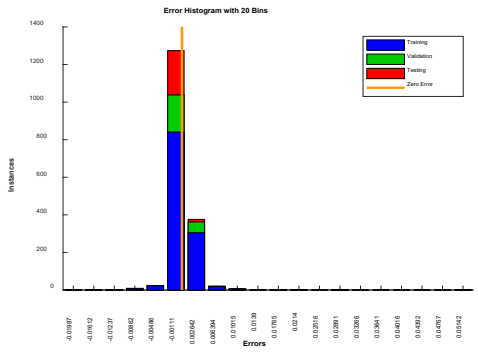
It should be stressed that the data used for Training, Validation, and Testing is utterly different from the data used in the final prediction. The final predictive data is original data that has not been subjected to all the processes applied to the data used to train the neural network model before being utilized to make predictions. Using a 70% Training, 15% Validation and 15% Testing dataset, LSTM models were built and tested using 8 water parameters: COD, DO, NH3-N, TSS, TDS, pH, Temp, and TUR to predict BOD readings. The graphs of the comparison results between the actual value of BOD and the predicted value of BOD for all water data collection stations are shown in Figure 8.



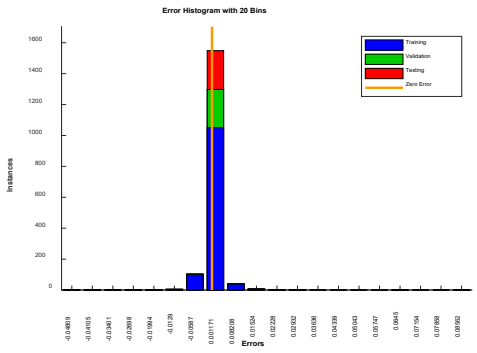
(a) LSTM 1K05 station



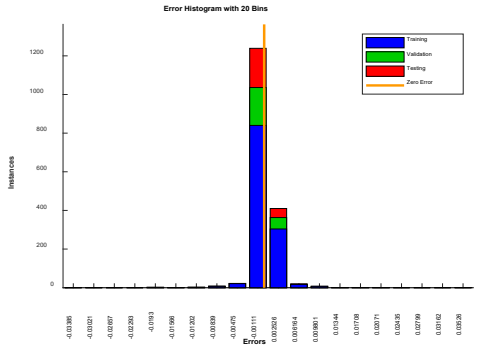
(b) LSTM 1K06 station



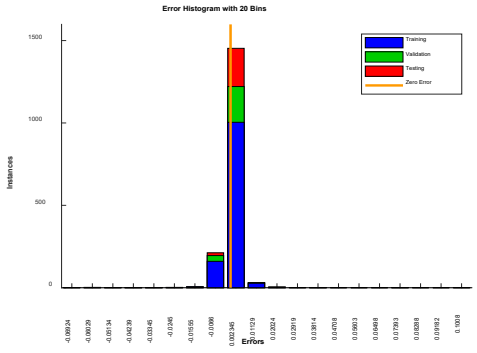
(c) LSTM 1K07 station



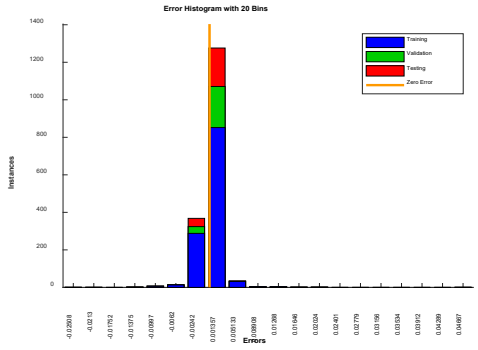
(d) LSTM 1K08 station



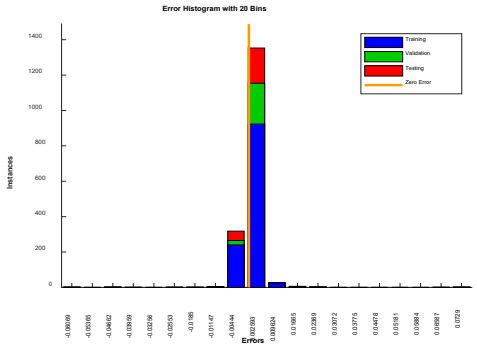
(e) LSTM 1K17 station



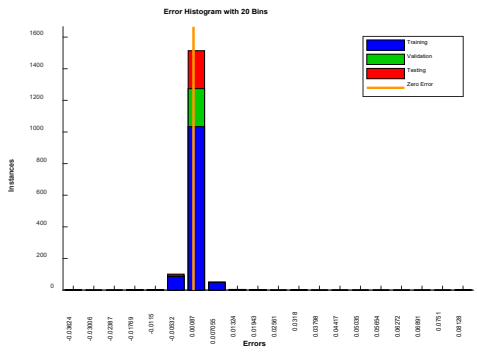
(f) LSTM 1K19 station



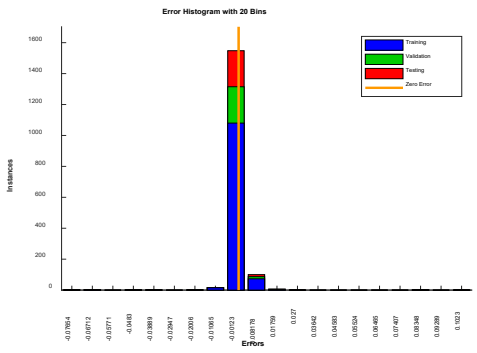
(g) LSTM 1K23 station



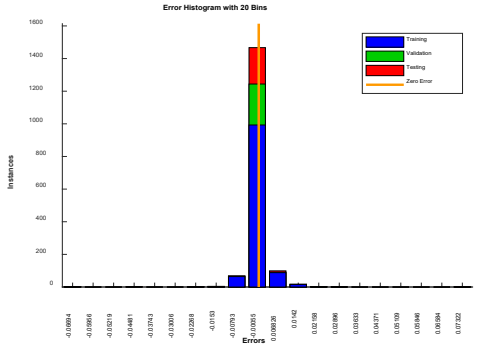
(h) LSTM 1K41 station



(i) LSTM 1K46 station

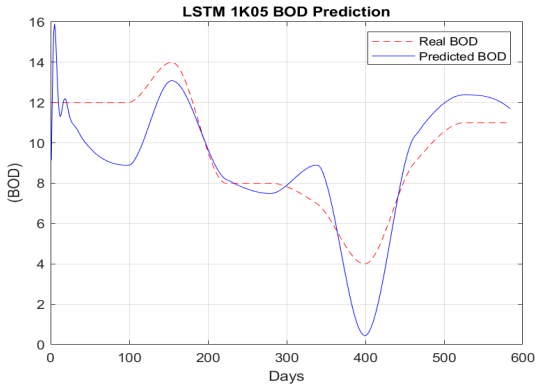


(j) LSTM 1K48 station

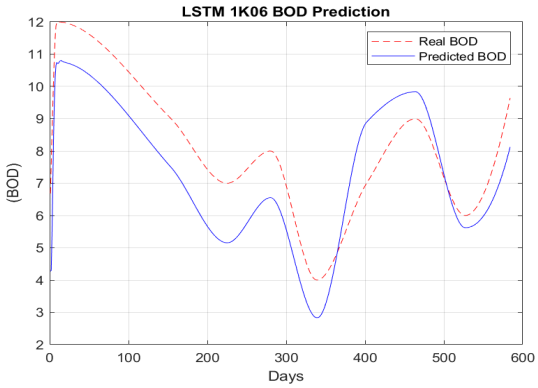


(k) LSTM 1K50 station

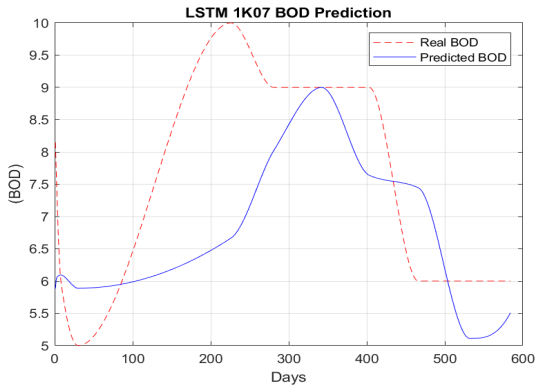
Figure 7. Error histograms for each water data collection station to view the Training, Validation and Testing errors for the LSTM model



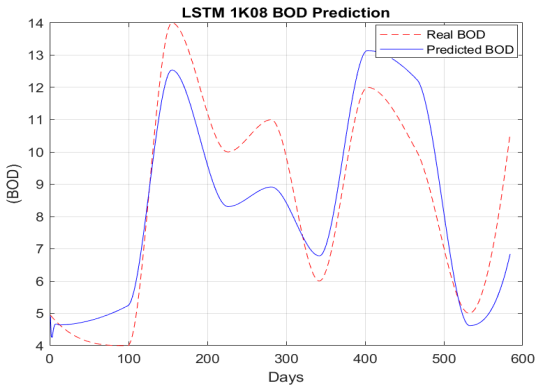
(a) BOD Prediction for 1K05 station



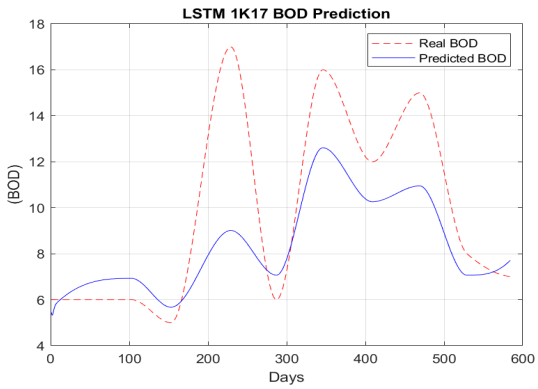
(b) BOD Prediction for 1K06 station



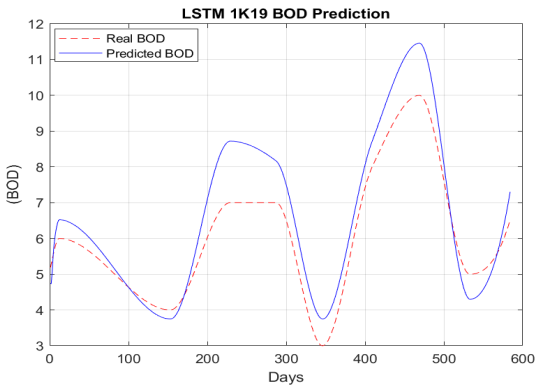
(c) BOD Prediction for 1K07 station



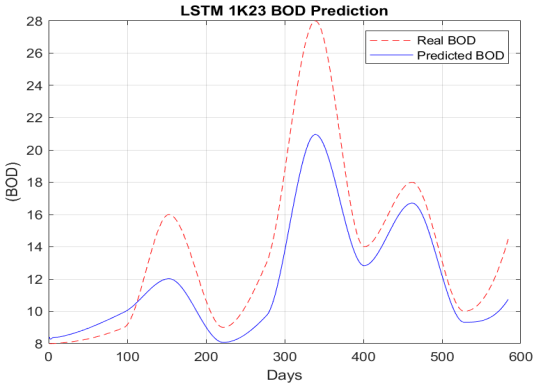
(d) BOD Prediction for 1K08 station



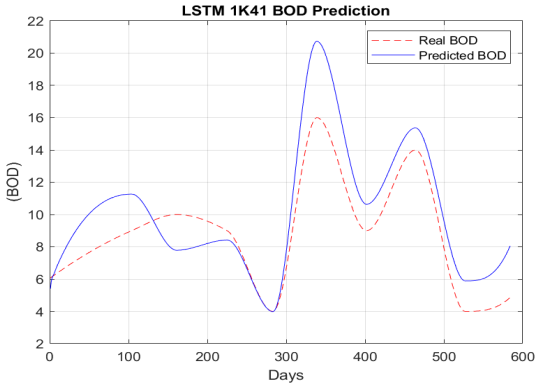
(e) BOD Prediction for 1K17 station



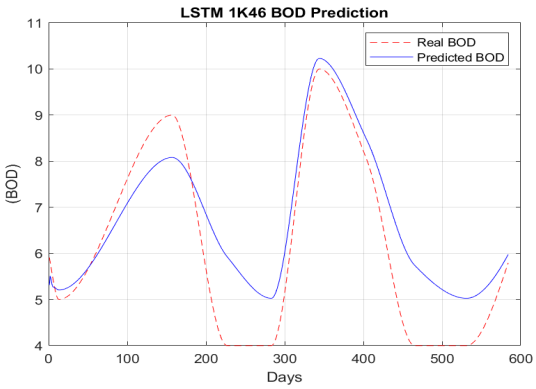
(f) BOD Prediction for 1K19 station



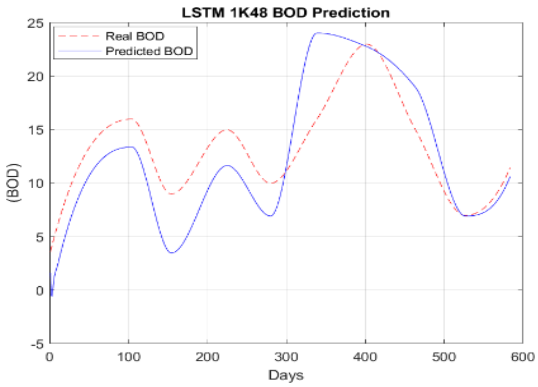
(g) BOD Prediction for 1K23 station



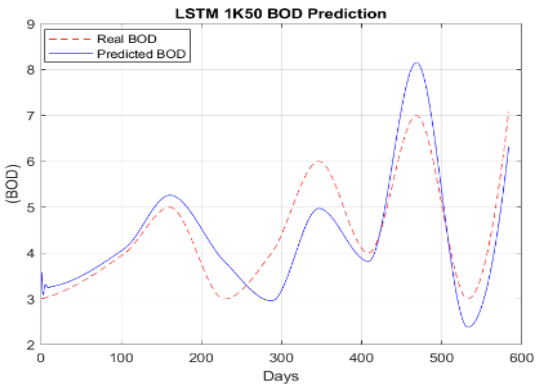
(h) BOD Prediction for 1K41 station



(i) BOD Prediction for 1K46 station



(j) BOD Prediction for 1K48 station



(k) BOD Prediction for 1K50 station

Figure 8. Comparison graph between the actual value and the predicted BOD for each water data collection station using the LSTM model

4. CONCLUSION

This study has successfully demonstrated that utilizing various water parameters and LSTM neural network makes it possible to predict the BOD water parameters needed to measure river water quality even though specific water parameters have a substantial correlation coefficient relationship with BOD. In contrast, others have a more modest relationship with BOD. This study also discovered that minor changes to water parameters with a weak correlation coefficient with BOD can also affect changes in BOD readings. The RMSE test results show that the LSTM model has a value of 0.4020 and produces nearly identical prediction results to the actual value based on the average computation against the actual value. The usage of eleven water data collection stations throughout the Klang River is intended to ensure that the deep learning model technique may very well be deployed everywhere along the Klang River. It is significant in the context of the Klang River since the river has several tributary branches. After testing the data from each water data collecting station, it was discovered that the deep learning model's BOD prediction results are consistent and suggest that this model can be deployed anywhere along the Klang River.

ACKNOWLEDGEMENT AND FUNDING

The authors would like to thank the Department of Environment (DOE) Malaysia for providing the hydrological data of the Klang River. This work is also supported by the School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA for the knowledge, facilities and funded with the (600-RMC/GPK 5/3 (137/2020)) grant by the Research Management Centre (RMC), UiTM.

DECLARATION OF CONFLICTING INTERESTS

The authors declare no potential conflicts of interest with respect to the research and publication of this article.

REFERENCES

[1] N. Jiao, J. Liu, B. Edwards, et al., Correcting a major error in assessing organic carbon pollution in natural waters, *Science Advance Journal*, 7(16), 2021, 1-11.

[2] K. S. Ooi, Z. Y. Chen, P. E. Poh and J. Cui, BOD5 prediction using machine learning methods, *Water Supply Journal*, 22(1), 2022, 1168-1183.

[3] W. Li and J. Zhang, Prediction of BOD concentration in wastewater treatment process using a modular neural network in combination with the weather condition, *Applied Sciences Journal*, 10(21), 2020, 7477.

[4] H. Prambudy, T. Supriyatin and F. Setiawan, The testing of Chemical Oxygen Demand (COD) and Biological Oxygen Demand (BOD) of river water in Cipager Cirebon, *Journal of Physics: Conference Series*, 1360, 2019, 012010.

[5] M. A. Mottalib, S. Roy, M. S. Ahmed, M. Khan and A. N. M. Al-Razee, Comparative study of water quality of Buriganga and Balu River, *International Journal of Current Research*, 9(10), 2017, 59132-59137.

[6] A. Fernandes, H. Chaves, R. Lima, J. Neves and H. Vicente, Draw on artificial neural networks to assess and predict water quality, *IOP Conference Series: Earth and Environmental Science*, 612, 2020, 012028.

[7] U. Hasanah, A. H. Mulyati, Sutanto, D. Widiastuti, S. Warnasih, Y. Syahputri and Tri Panji, Development of COD (Chemical Oxygen Demand) analysis method in waste water using Uv-Vis spectrophotometer, *Journal of Science Innovare*, 3(2), 2020, 35-38.

[8] J. A. Mangai and Bharat B. Gulyani, Induction of model trees for predicting BOD in River Water: A data mining perspective, *Lecture Notes in Computer Science*, 2016, 9728, 1-13.

[9] S. Jouanneau, L. Recoules, M. J. Durand, A. Boukabache, V. Picot, Y. Primault, A. Lakel, M. Sengelin, B. Barillon and G. Thouand, Methods for assessing biochemical oxygen demand (BOD): A review, *Water Research Journal*, 49, 2014, 62-82.

[10] F. Dara, A. Devolli and A. Kodra, An artificial neural networks modell for predicting BOD of Ishëm River, *International Agricultural, Biological & Life Science Conference*, Edirne, Turkey, 2018.

[11] S. Susilowati, J. Sutrisno, M. Masykuri and M. Maridi, Dynamics and factors that affects DO-BOD concentrations of Madiun River, *AIP Conference Proceedings*, 2049(1), 2018, 020052.

[12] O. Thomas, J. Causse and M. -F. Thomas, Aggregate organic constituents, *UV-Visible Spectrophotometry of Waters and Soils (Third Edition)*, 2022, 161-192.

[13] K. S. Ooi, Z. Y. Chen, P. E. Poh and J. Cui, BOD5 prediction using machine learning methods, *Water Supply Journal*, 22(1), 2022, 1168-1183.

[14] S. S. Shaikh and R. Shahapurkar, Predicting BOD of greywater using artificial neural networks, *International Journal of Engineering Trends and Technology*, 70(3), 2020, 195-200.

[15] Y. Jiang, C. Li, L. Sun, D. Guo, Y. Zhang and W. Wang, A deep learning algorithm for multi-source data fusion to predict water quality of urban sewer networks, *Journal of Cleaner Production*, 318, 2021, 128533.

[16] E. I. Obilor and E. C. Amadi, Test for significance of Pearson's correlation coefficient (r), *International Journal of Innovative Mathematics, Statistics & Energy Policies*, 6(1), 2018, 11-23.

[17] E. Saccenti, M. H. W. B. Hendriks and A. K. Smilde, Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models, *Science Reports*, 10(1), 2020, 1-9.

[18] R. Taylor, Interpretation of the correlation coefficient: A basic review, *Journal of Diagnostic Medical Sonography*, 6(1), 1990, 35-39.

[19] A. M. Alsaqr, Remarks on the use of Pearson's and Spearman's correlation coefficients in assessing relationships in ophthalmic data, *African Vision, and Eye Health*, 80(1), 2021, 1-10.

[20] H. Fan, M. Jiang, L. Xu, H. Zhu, J. Cheng and Jiahu Jiang, Comparison of long, short term memory networks and the hydrological model in runoff simulation, *Water*, 12(1), 2020, 175.

[21] Z. Li, F. Peng, B. Niu, G. Li, J. Wu, and Z. Miao, Water quality prediction model combining sparse auto-encoder and LSTM network, *IFAC PapersOnLine*, 51(17), 2018, 831-836.

[22] Q. Zou, Q. Xiong, Q. Li, H. Yi, Y. Yu and C. Wu, A water quality prediction method based on the multi-time scale bidirectional long short-term memory network, *Environmental Science and Pollution Research*, 27, 2020, 16853-16864.

[23] M. A. Mustafa Azizi, M. N. Mohd Noh, I. Pasya, A. I. Mohd Yassin and M. S. A. Megat Ali, Pedestrian detection using doppler radar and LSTM neural network, *IAES International Journal of Artificial Intelligence*, 9(3), 2020, 394-401.

[24] T. Su, Y. Shi, J. Yu, C. Yue and F. Zhou, Nonlinear compensation algorithm for multidimensional temporal data: A missing value imputation for the power grid applications, *Knowledge-Based System*, 215, 2021, 106743.

[25] Y. G. Cinar, H. Mirisaee, P. Goswami, E. Gaussier and A. Aıt-Bachir, Period-aware content attention RNNs for time series forecasting with missing values, *Neurocomputing*, 312, 2018, 177-186.

[26] S. Urolagin, N. Sharma and T. K. Datta, A combined architecture of multivariate LSTM with Mahalanobis and Z-score transformations for oil price forecasting, *Energy*, 231, 2021, 120963.

[27] S. Ghimire, Z. M. Yaseen, A. A. Farooque, R. C. Deo, J. Zhang and X. Tao, Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks, *Scientific Reports*, 11(1), 2021, 1-26.

[28] P. V. Anusha, C. Anuradha, P. S. R. Chandra Murty and C. S. Kiran, Detecting outliers in high dimensional data sets using Z-score methodology, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(1), 2019, 48-53.
[29] D. Cousineau and S. Chartier, Outliers detection and treatment: a review, *International Journal of Psychological Research*, 3(1), 2010, 58-67.
[30] J. Benhadi-Marín, A conceptual framework to deal with outliers in ecology, *Biodiversity and Conservation*, 27(12), 2018, 3295-3300.
[31] L. Alzubaidi, J. Zhang, A. J. Humaidi, et al., Review of deep learning: concepts, CNN architectures, challenges, applications, future directions*, Journal of Big Data*, 8(1), 2021.
[32] T. Boulmaiz, M. Guermoui and B. Hamouda, Impact of training data size on the LSTM performances for rainfall–runoff modeling, *Modeling Earth Systems and Environment*, 6(4), 2020, 2153-2164.
[33] V. R. Joseph and A. Vakayil, SPlit: An optimal method for data splitting, *Technometrics*, 64(2), 2020, 166-176.
[34] Q. H. Nguyen, H. -B. Ly, L. S. Ho, et al., Influence of data splitting on performance of machine learning models in prediction of shear strength of soil, *Mathematical Problems in Engineering*, 2021, 4832864.
[35] G. Burkay and T. Hüseyin, Optimal training and test sets design for machine learning, *The Turkish Journal of Electrical Engineering & Computer Sciences*, 27(2), 2019, 1534-1545.
[36] Y. Xu and R. Goodacre, On splitting training and validation set: A comparative study of cross-validation, *Journal of Analysis and Testing*, 2(3), 2020, 249-262.
[37] T. P. Quinn, V. Le and A. P. A. Cardilini, Test set verification is an essential step in model building, *Methods in Ecology and Evolution*, 12(1), 2021, 127-129.
[38] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang and J. P. Campbell, Introduction to machine learning, neural networks, and deep learning, *Translational Vision Science & Technology*, 9(2), 2020, 1-12.
[39] H. Yoon, The adequacy assessment of test sets in machine learning using mutation testing, *International Journal of Engineering and Advanced Technology (IJEAT)*, 9(1), 2019, 4390-4395.
[40] J. Sadowski, When data is capital: Datafication, accumulation, and extraction, *Big Data & Society*, 6(1), 2019, 1-12.
[41] J. T. Saura, B. R. Herráez, and A. Reyes-Menendez, Comparing a traditional approach for financial brand communication analysis with a big data analytics technique, *IEEE Access*, 7, 2019, 37100-37108.
[42] Y. Chen, L. Song, Y. Liu, L. Yang and D. Li, A review of the artificial neural network models for water quality prediction, *Applied Science*, 10(17), 2020, 5776.