water

MDPI

Article

Prediction of Wastewater Treatment Plant Effluent Water Quality Using Recurrent Neural Network (RNN) Models

Praewa Wongburi [1,*] and Jae K. Park [2]

[1] Faculty of Environment and Resource Studies, Mahidol University, Nakhon Pathom 73170, Thailand
[2] Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA; jkpark@wisc.edu
* Correspondence: praewa.won@mahidol.ac.th

Abstract: Artificial Intelligence (AI) has recently emerged as a powerful tool with versatile applications spanning various domains. AI replicates human intelligence processes through machinery and computer systems, finding utility in expert systems, image and speech recognition, machine vision, and natural language processing (NLP). One notable area with limited exploration pertains to using deep learning models, specifically Recurrent Neural Networks (RNNs), for predicting water quality in wastewater treatment plants (WWTPs). RNNs are purpose-built for handling sequential data, featuring a feedback mechanism. However, standard RNNs may exhibit limitations in accommodating both short-term and long-term dependencies when addressing intricate time series problems. The solution to this challenge lies in adopting Long Short-Term Memory (LSTM) cells, known for their inherent memory management through a 'forget gate' mechanism. In general, LSTM architecture demonstrates superior performance. WWTP data represent a historical series influenced by fluctuating environmental conditions. This study employs simple RNNs and LSTM architecture to construct prediction models for effluent parameters, systematically assessing their performance through various training data scenarios and model architectures. The primary objective was to determine the most suitable WWTP dataset model. The study revealed that an epoch setting of 50 and a batch size of 100 yielded the lowest training time and root mean square error (RMSE) values for both RNN and LSTM models. Furthermore, when these models are applied to predict effluent parameters, they exhibit precise RMSE values for all parameters. The study results can be applied to detect potential upsets in WWTP operations.

Keywords: Artificial Intelligence; Recurrent Neural Networks; Long Short-Term Memory; prediction model; sequential data; wastewater treatment plants

check for updates

Citation: Wongburi, P.; Park, J.K. Prediction of Wastewater Treatment Plant Effluent Water Quality Using Recurrent Neural Network (RNN) Models. *Water* **2023**, *15*, 3325. https://doi.org/10.3390/w15193325

Academic Editor: Chaojie Zhang

Received: 17 August 2023
Revised: 13 September 2023
Accepted: 15 September 2023
Published: 22 September 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Artificial Intelligence (AI), the overarching concept of machine learning, deep learning, and artificial neural networks (ANNs), has become a powerful instrument for tackling challenging problems in various fields and solving intricate real-world applications. It is a branch of computer science focused on simulating intelligent behavior in computers, enabling them to perform tasks typically requiring human intelligence, such as visual perception, speech recognition, decision making, and language translation.

Deep learning or Deep Neural Networks (DNNs) are algorithms for setting new accuracy records in various areas like sound and image recognition, stock market prediction, recommender systems, etc. Inspired by the human neural system, these networks use multiple hidden layers to learn from extensive data, improving accuracy as data pass through these layers [1].

Wastewater introduces diverse environmental contaminants that pose significant concerns [2]. Inefficient wastewater treatment plant (WWTP) operation exacerbates these challenges, even as abundant data from WWTP sensors often remain unutilized [3]. WWTP

complexities introduce uncertainty and variability into wastewater treatment systems, necessitating innovative approaches [4]. The authors have pioneered the application of big data analytics in WWTPs, identifying data patterns and relationships among historical data parameters. This research culminated in developing novel models for monitoring, performance prediction, and control of complex nonlinear processes.

However, amid the increasing applications of AI across environmental domains, a notable research gap emerges—the comprehensive exploration of AI's transformative potential in wastewater treatment. While AI methodologies have gained prominence in environmental science, examples such as the ASEAN water quality indices for assessing spatial variations in surface water quality [5], monsoonal river classification in specific basins [6], and advanced algorithms like random forest and support vector machines for water quality index modeling exist [7], there is a discernible absence of comprehensive research that extends AI's reach into wastewater treatment processes. Therefore, it underscores the imperative for in-depth studies that explore the full spectrum of AI's applicability in wastewater treatment.

Previous Studies of the Development of Predictive Models in WWTPs

Several forecasting models have been developed so far to monitor wastewater treatment processes. However, traditional predictive models have complex structures and are involved in various amounts of parameters that must be identified [8]. The difficulty of the previous modeling technique also includes a heavy computational burden for the simulation and design process [9]. Due to the challenges in the simulation of such complex systems, numerous data, and heavy computation, many modeling techniques have been developed, such as the autoregressive (AR) model, artificial neural network (ANN), genetic algorithm, multivariate analysis, etc.

ARIMA is a method to forecast future behavior from previous historical data. It can be written as "ARIMA (p, d, q)", where parameter p is the order of the autoregressive (AR) model, parameter d is the degree of differencing, and parameter q is the order of the moving average (MA) model [10]. However, the authors recommended that nonlinear models, such as neural networks, would instead perform better than ARIMA for forecasting purposes.

A recent review paper highlighted the explosive growth in the number of publications related to machine learning in environmental science and engineering, with around 50% being in the water sector [11]. Moreover, the number of research publications on AI application to wastewater treatment was 19 times greater in 2019 than in 1995, and papers had 36 more citations on average [12].

ANNs are one of the models most applied in the simulation and prediction of the performance of biological treatment in WWTPs [11]. The models are composed of several artificial neurons, connected by links of variable weight, to form black box representations of pseudo neurological systems. The most common machine learning models used in the simulation, prediction, evaluation, and diagnosis of wastewater treatment operations are ANNs, Fuzzy Logic (FL), genetic algorithms (GA), and neural-fuzzy (NF), as well as artificial neural networks–genetic algorithm (ANN–GA) as hybrid models [11]. Table 1 shows the application of the AI model for operation management in WWTPs [13–18].

The concept of ANNs is stimulated by the human neural system, connecting neurons to develop a layered structure from inputs to outputs through several hidden layers [1]. While the capacity of ANNs is suitable for modeling nonlinear systems, architecture like feed-forward, Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) have been developed. RNNs, extensions of feed-forward neural networks, can handle variable-length sequence inputs [19]. Among RNNs, LSTM stands out for its ability to learn long patterns in sequential data and maintain information over extended periods [20]. However, RNNs and LSTM models have been limited to be applied to WWTP operations. Therefore, the exceptional learning capabilities of the models are s-suited for sequence data in wastewater treatment systems.

Table 1. The use of AI models for operation management in WWTPs.

Treatment Process	AI Model	Model Performance (RMSE)	Reference
Anaerobic digestion	ANN–GA	196.1	(Huang et al., 2016 [13])
	ANN	447.7	
Aeration diffusion	ARMA-VAR	113.56	(Nadiri et al., 2018 [14])
Aeration diffusion	BP-ANN	303.51	(Man et al., 2019 [15])
	GA-BP-ANN	232.6	
Anaerobic oxic biological	SDAE	5.94	(Shi & Xu, 2018 [16])
		1.27	
		1.26	
Activated Sludge	PFA	0.25	(Yu et al., 2019 [17])
Activated Sludge	SVM	1435.4	(Najafzadeh & Zeinolabedini, 2019 [18])
	BP-ANN	1445.9	
	ANFIS	1515.6	
	RBF-ANN	1501	

2. Materials and Methods

The digital dataset was obtained from the Nine Springs WWTP in Madison Metropolitan Sewerage District (MMSD), Madison, WI, USA, which is 155,000 m^3/day. Figure 1 represents the overview of an RNN modeling process.

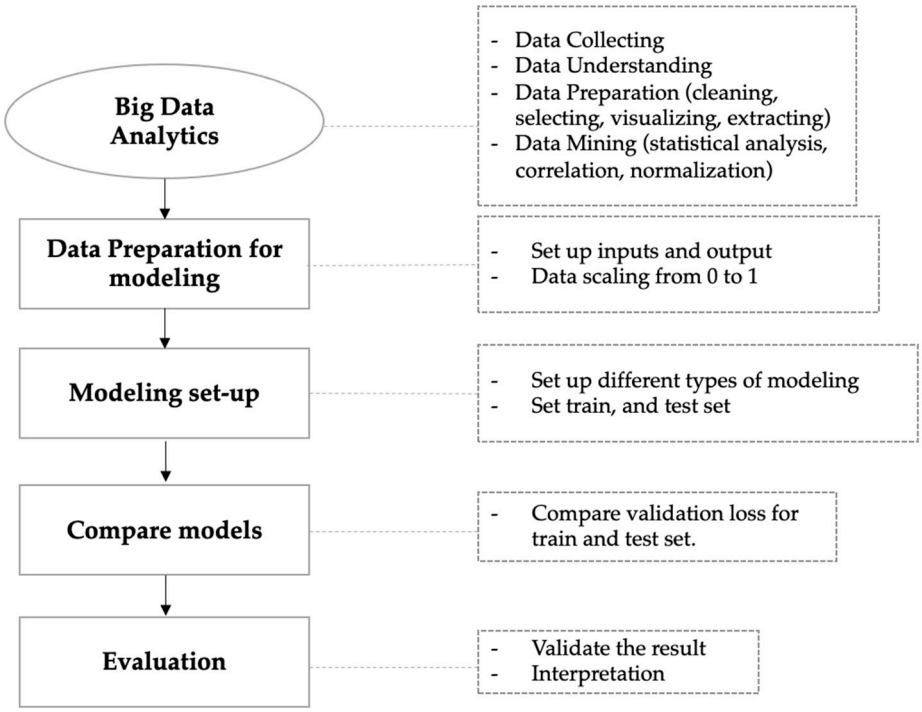

Figure 1. The overview of an RNN modeling process.

2.1. Big Data Analytics

As indicated by prior research titled "Big Data Analytics from a Wastewater Treatment Plant", the significance of big data analytics in constructing a wastewater treatment plant model is noteworthy. The phases of big data analytics encompass data collection, data

comprehension, data preprocessing, and data mining [21]. The details of big data analytics are as follows:

Data collection: Acquiring data from the source is crucial. For this research and prior studies, data originated from the Nine Springs Wastewater Treatment Plant operated by the Madison Metropolitan Sewerage District (MMSD) in Madison, WI, USA. Figure 2 shows the general layout of the Nine Springs Wastewater Treatment Plant and the data collection locations. According to the original dataset, the influent parameters were selected from the influent meter vault in the headworks facility where the wastewater entered the plant. The effluent parameters were chosen from the effluent building where the treated water is sent. The plants are divided into the East Complex and the West Complex. The East Complex includes Plant 1 and Plant 2, and the West Complex includes Plant 3 and Plant 4 [22]. The dataset spans from 1996 to 2019. The dataset was split into two portions following extensive big data analytics: 1997 to 2019 and 2015 to 2018.

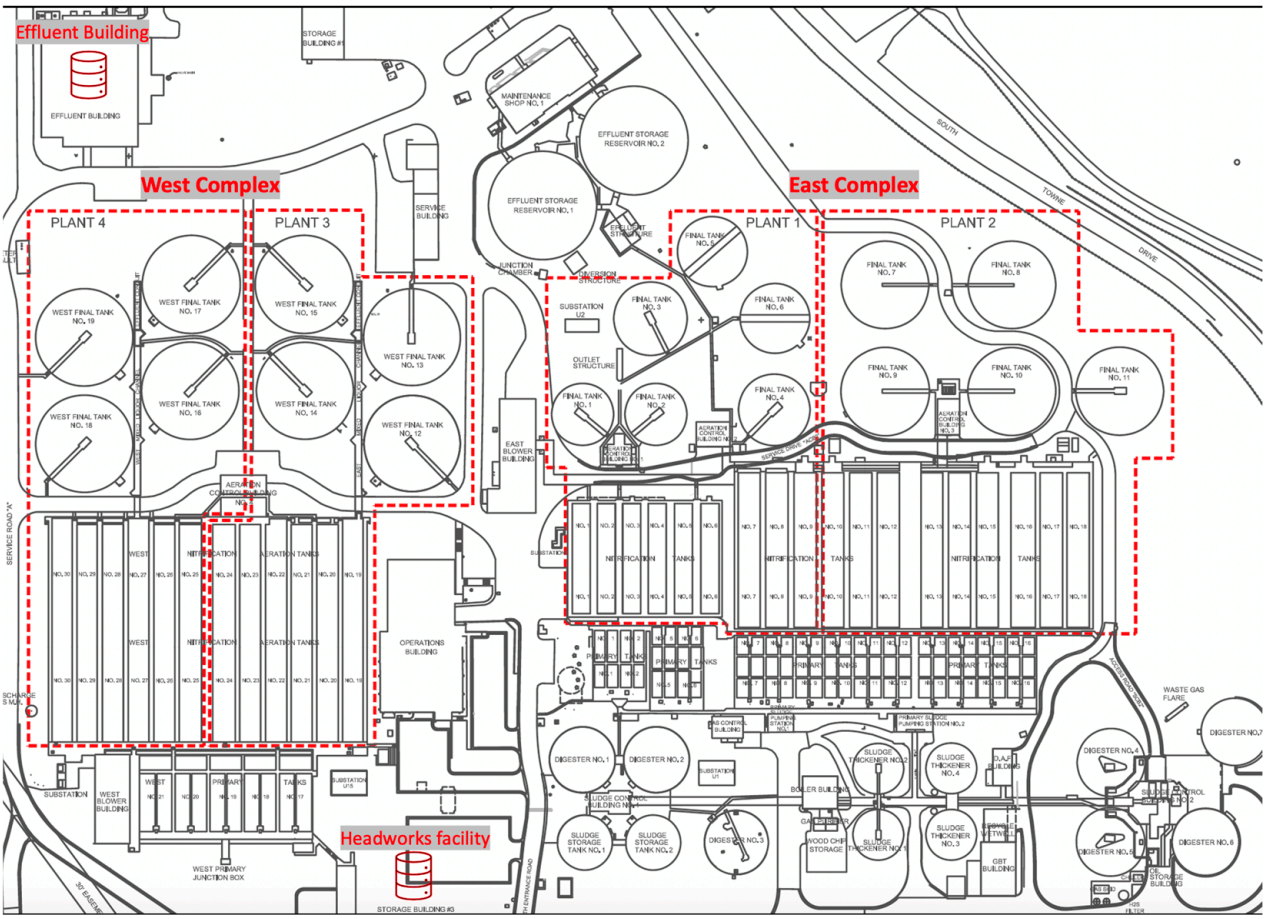

Figure 2. The general layout of the Nine Springs Wastewater Treatment Plant and the data sampling locations.

Data comprehension: A diligent examination of the collected data is essential. Given its source from the Madison Metropolitan Sewerage District (MMSD), a thorough grasp of the wastewater treatment plant's context, objectives, and liquid processes is required.

Data preprocessing: This stage transforms data into a meaningful dataset through various techniques. Data cleaning addresses outliers and missing values, data integration combines information from multiple sources, data transformation simplifies the dataset, and data reduction reduces dataset size while preserving original data integrity. The data preprocessing steps are to (1) merge date and time into one column and change to DateTime-type, (2) convert all data to numeric, (3) remove unnecessary missing values or impute some missing data using the interpolation technique, and (4) finally, apply statistical

normality tests and normal probability distribution analyses to test the data integrity. These techniques are crucial to preparing data before applying models or interpretation.

Data mining: This stage involves employing statistical methods on preprocessed data to extract knowledge to identify data patterns, conduct statistical analytics, and normalize processes.

These stages facilitate the analysis and interpretation of data. This approach aids in uncovering insights, eliminating irrelevant data, acquiring an appropriate dataset, and constructing an accurate predictive model.

2.2. Data Preparation

Continuing from the previous section, data preparation was pivotal in developing a predictive model. Within WWTPs, a critical performance metric is the biochemical oxygen demand (BOD_5). Consequently, for the construction of the model's framework, five pivotal influent quality parameters were carefully chosen: BOD_5, total suspended solids (TSSs), total phosphorus (TP), total Kjeldahl nitrogen (TKN), and NH_3-N. Furthermore, the model incorporated flow rate and sludge volume index (SVI) as inputs for the daily prediction model. The model's outcomes encompassed the effluent levels of BOD_5 and other pertinent parameters. Illustrated in Figure 3, the architecture of the Recurrent Neural Network (RNN) model incorporated these crucial influent parameters and outputs. As a stage in the previous section, the data preparation is also performed to develop a predictive model.

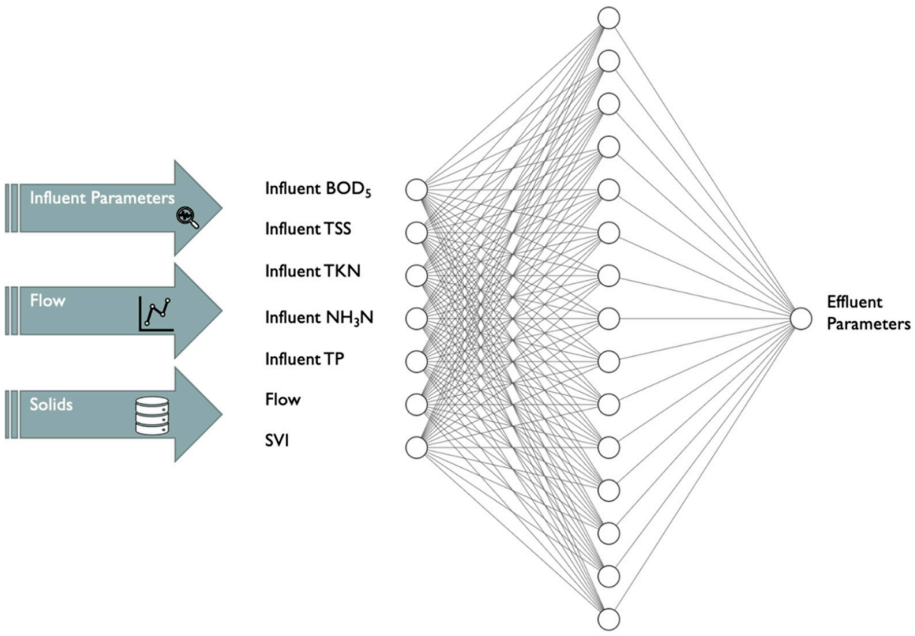

Figure 3. The architecture of the RNN model.

2.3. Data Scaling

Data in prediction problems need to be scaled when processing a neural network. When a network trains unscaled data with a wide range of values, it will slow the learning rate and prevent the network from effectively learning the data.

Normalization is rescaling the data within the range of 0 and 1 from the original range. In scaling, data are transformed using MinMaxScaler in the scikit-learn module. The equation below is rescaling (or min–max normalization), which is the simplest way to rescale data within the ranges [0, 1] or [−1, 1] [23]. The general formula for a min–max of [0, 1] is given as follows:

$$x' = (x - x_{min})/(x_{max} - x_{min}) \quad (1)$$

where x is an original value, and x' is the normalized value.

Data scaling or normalization is essential, especially when the data have input values with differing scales. After applying the above formula, the maximum value is 1, and the minimum values are 0. Therefore, all the values are between 0 and 1. Table 2 presents the data after rescaling.

Table 2. The result from data scaling.

	var1(t − 1)	var2(t − 1)	var3(t − 1)	var4(t − 1)	var5(t − 1)	var6(t − 1)	var1(t)
1	0.296296	0.563376	0.173804	0.517385	0.783042	0.456763	0.333333
2	0.333333	0.423265	0.154176	0.454798	0.783042	0.439024	0.251852
3	0.251852	0.364614	0.165774	0.450626	0.783042	0.331486	0.251852
4	0.251852	0.364614	0.165774	0.349096	0.783042	0.331486	0.251852
5	0.251852	0.364614	0.165774	0.479833	0.783042	0.331486	0.251852

2.4. Development of Recurrent Neural Networks (RNNs) Models

An RNN model is suitable for making predictions on sequence data. This study aimed to develop a sequence prediction model that performs the best result. There are five steps to establish a simple RNN model (Figure 4) and LSTM model (Figure 5), as follows [24]:

Figure 4. Five steps to develop a simple RNN model.

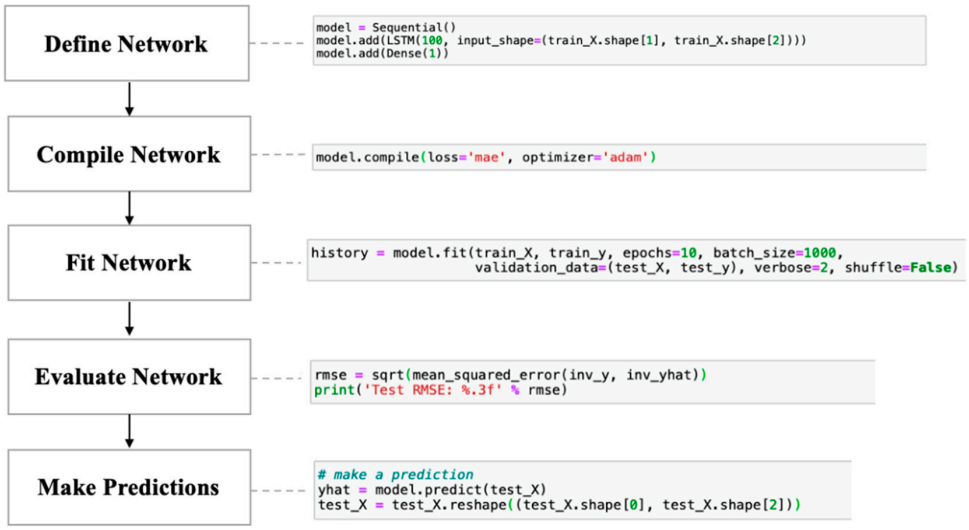

Figure 5. Five steps to develop an LSTM model.

2.4.1. Define Network

Neural networks operate as a succession of layers. The sequential class provides a framework for these layers. The procedure involves creating an instance of the sequential class, forming layers, and then adding them to establish a connection. A fully connected layer is incorporated after simple RNN or LSTM layers, and a Dense() function is added to generate predictions. The initially hidden layers must specify the number of inputs, which should be three-dimensional, encompassing samples, time steps, and features.

2.4.2. Compile Network

Complication expects to identify parameters to train a network. The optimization algorithm and loss function are required to train and evaluate the network. The standard optimization algorithms are stochastic gradient descent (sgd), Adam, and RMSprop. This model applied the Adam optimization algorithm, which is an optimization algorithm that could be used instead of the classical stochastic gradient descent (sgd) procedure to update network weights iterative based on training data [25]. The loss functions usually apply the mean squared error (MSE) and mean absolute error (MAE).

2.4.3. Fit Network

After compiling the network, it must be fitted, which means adapting the weights on training data. The model fitting process requires the training dataset to define a matrix of input patterns. The network exploits the backpropagation for training and optimizing algorithms and loss functions. The backpropagation requires a specific number of epochs, which means one passes through all data in the training set and updates the weights.

2.4.4. Evaluate Network

It needs to be evaluated after training the network. It is useful, especially for indicating the performance of a predictive model, because the network has seen all of the data before by separating data (no training). This will provide an estimate of the network's performance at making predictions for unseen data in the future.

- Root mean square error (RMSE): RMSE has been used as a standard statistical metric to measure model performance in meteorology, air quality, climate studies, and other research areas [26]. RMSE is the standard deviation of the prediction errors, which is the most popular measure of estimation accuracy to compare forecasting errors of different models, defined by

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{2}$$

  where n is the total number of samples in the model, and e is the model errors calculated as (i.e., $i = 1, 2, \ldots, n$).
- Mean absolute error (MAE): MAE measures errors between two observations. y and $\hat{y}$ include comparisons of predicted versus observed. MAE has the same unit as the initial data, and it can be evaluated between models whose errors are computed in the same units. It is typically similar in magnitude to RMSE but slightly smaller. MAE is calculated as [26].

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

  where n is the number of observations, y_i is the actual value, and $\hat{y}_i$ is the predicted value.
- Coefficient of Determination (R^2): R^2 is a statistical measure representing the proportion of the variance for a dependent variable, explained by an independent variable in a regression model. While correlation interprets the strength of the connection between an independent and dependent variable, R-squared explains to what extent one variable's variance explains the second variable's variance [27]. Thus, if the R^2

of a model is 0.50, then nearly half of the observed value can be explained by the model's inputs.

2.4.5. Make Predictions

After fitting and evaluating the model, the network will return the prediction values provided by the output layers of the network.

Table 3 displays the time steps of developing a time series model. The input parameters will be x variables to predict the output y. The model used 80% of the data for training and 20% for testing.

Table 3. Time steps to develop a time series model.

Date and Time	BOD_{5_In}	TSS_{In}	TKN_{In}	NH_3-N_{In}	TP_{In}	$Flow_{In}$	BOD_{5_Eff}
1	x_1 (t − 1)	x_2 (t − 1)	x_3 (t − 1)	x_4 (t − 1)	x_5 (t − 1)	x_6 (t − 1)	Y (t)
2	x_1 (t)	x_2 (t)	x_3 (t)	x_4 (t)	x_5 (t)	x_6 (t)	Y (t + 1)
3	x_1 (t + 1)	x_2 (t + 1)	x_3 (t + 1)	x_4 (t + 1)	x_5 (t + 1)	x_6 (t + 1)	Y (t + 2)
...	...						
n	x_1 (t + n)	x_2 (t + n)	x_3 (t + n)	x_4 t + n)	x_5 (t + n)	x_6 (t + n)	Y (t + n + 1)

3. Results

This study used the discrete historical data of the Nine Springs WWTP in developing the AI model. The parameter values in wastewater treatment from 2015 to 2018 are summarized in Table 4.

Table 4. The values of water quality parameters in wastewater treatment from 2015 to 2018.

Parameters	Unit	Maximum	Minimum	Mean	SD
$BOD_{5,i}$	mg/L	400.00	93.10	246.53	39.44
$BOD_{5,e}$	mg/L	27.00	0.00	5.48	2.34
TSS_i	mg/L	1170.00	49.20	222.72	60.76
TSS_e	mg/L	22.30	0.00	4.74	1.67
TP_i	ppm	11.40	2.38	5.68	0.98
TP_e	ppm	1.12	0.05	0.31	0.12
TKN_i	ppm	90.00	18.10	44.80	7.56
TKN_e	ppm	9.65	0.23	1.79	0.49
NH_3-N_i	ppm	40.10	0.00	28.26	4.01
NH_3-N_e	ppm	7.22	0.00	0.26	0.36

ppm: Parts per million.

Performance Comparison

The first model performance comparison applied discrete big data from 1997 to 2019 and 2015 to 2018, separating into two scenarios. The models were performed using simple RNN and LSTM models with different numbers of epochs and batch sizes. The input parameters are TSS, TP, TKN, and NH_3N. The output is the effluent BOD_5. The performance of models includes time and root mean squared errors (RMSEs). Table 5 shows the effluent prediction models using simple RNN and LSTM models. Scenario 2, data from 2015 to 2018, achieves better timing and accuracy in both models when using many epochs and small batch sizes. The optimization model is when specifying 20~50 epochs and 100 batch sizes with a time of 1~2 s in each epoch and RMSE of 0.3~0.8.

Figure 6 shows the original dataset of effluent BOD_5 from 1 January 2015 to 31 December 2018. The data were collected hourly. The prediction result would be more clearly seen if the data were collected at the same interval.

Figure 7 shows the prediction result of effluent BOD_5 from 2015 to 2018 using the simple RNN algorithm. The original value is very close to the predicted value, which has an RMSE of 0.247 and MAE of 0.052, which are small RMSE and MAE values. It means that the model is optimal and predicts an excellent in-sample fit.

Table 5. Performances of the effluent BOD_5 prediction models.

Model	Scenario	Training Epoch	Batch Size	Time (Seconds)	RMSE
Simple RNN model	Scenario 1: Data from 1997 to 2019	10	1000	9–14	2.298
		20	100	37–84	4.688
		50	100	40–53	1.702
	Scenario 2: Data from 2015 to 2018	10	1000	0–1	1.396
		20	100	1–3	0.390
		50	100	1–4	0.361
LSTM model	Scenario 1: Data from 1997 to 2019	10	1000	10–15	1.888
		20	100	48–92	1.638
		50	100	49–62	1.529
	Scenario 2: Data from 2015 to 2018	10	1000	0–2	1.888
		20	100	2–3	0.778
		50	100	1–2	0.452

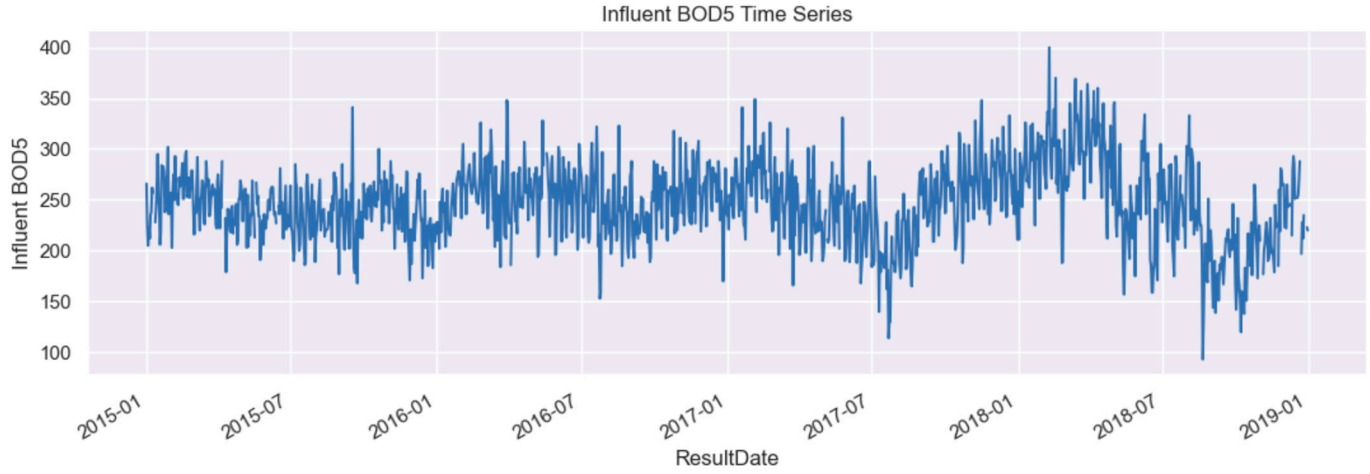

Figure 6. The original data of effluent BOD_5 from 2015 to 2018.

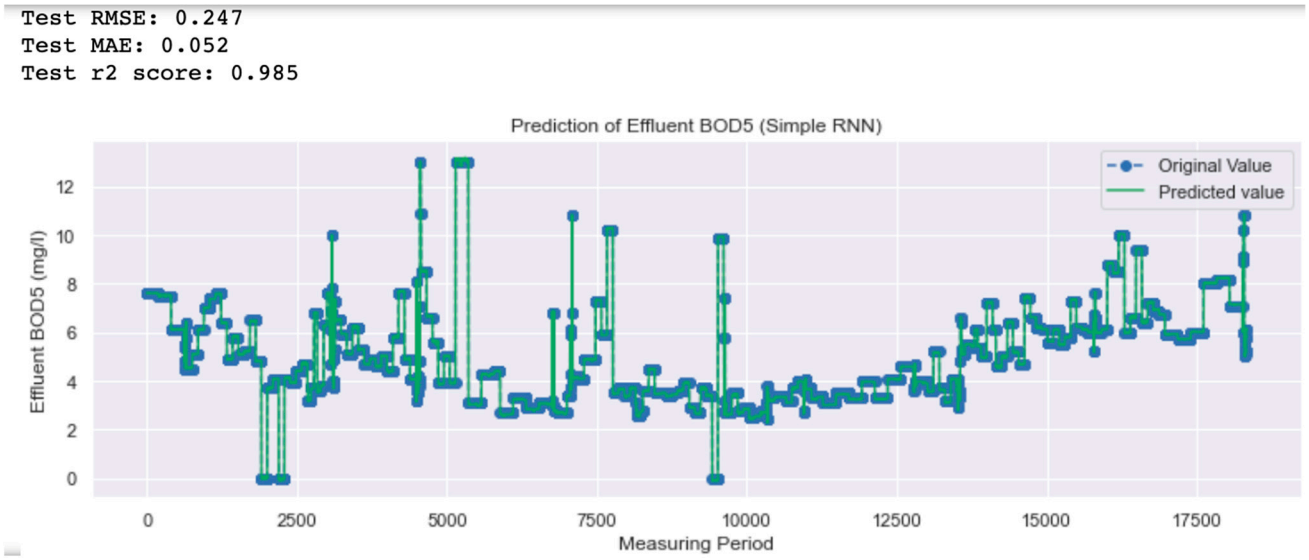

Figure 7. The prediction of effluent BOD_5 from 2015 to 2018 using the simple RNN model.

Figure 8 illustrates the predicted effluent BOD_5 from 2015 to 2018, as determined by the LSTM model. The predicted results align closely with the original data, indicating the accuracy of the prediction model. The model's root mean square error (RMSE) is 0.246, a

relatively small value, signifying that the model commits minor errors, minimizing the risk of overfitting or underfitting. Hence, the original dataset spanning from 2015 to 2018 provides a well-suited fit for the models. These models can learn and retain the pattern, enabling them to predict the effluent BOD_5 accurately.

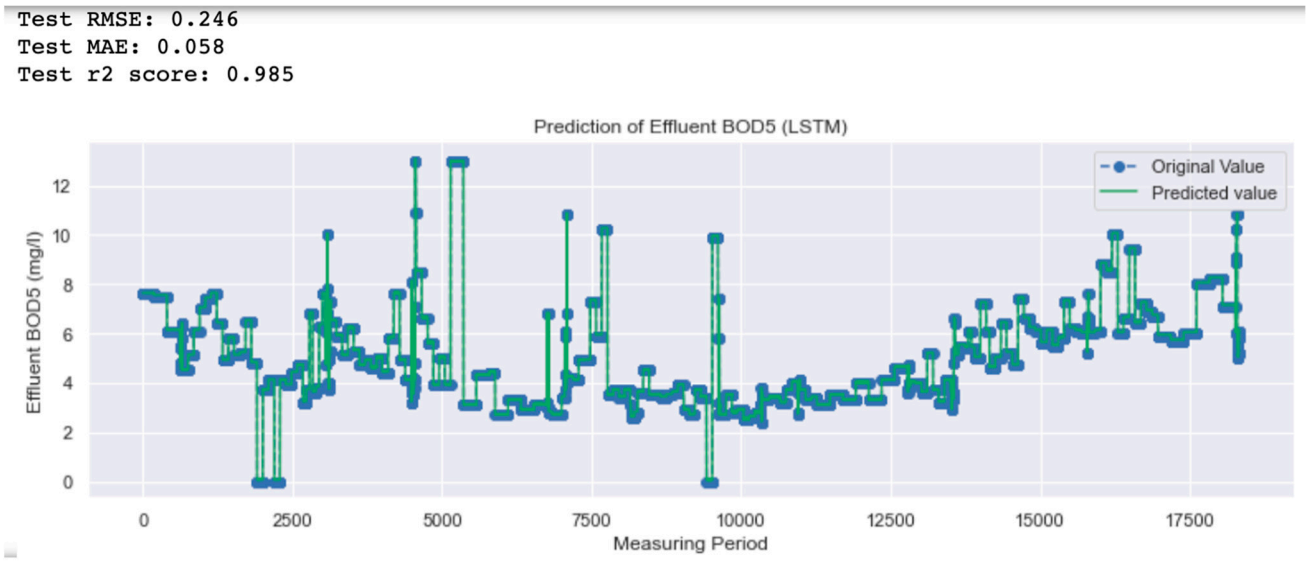

Figure 8. The prediction of effluent BOD_5 from 2015 to 2018 using the LSTM model.

Therefore, the original dataset from 2015 to 2018 is an excellent fit. The models can learn and remember the pattern to accurately predict the effluent BOD_5. The next step is to develop a model for other essential effluent parameters: TP, TKN, TSS, and NH_3-N. The model applied 100 batch sizes and 50 epochs, which created the best model performance. Figures 9–16 show the results of all effluent parameters models.

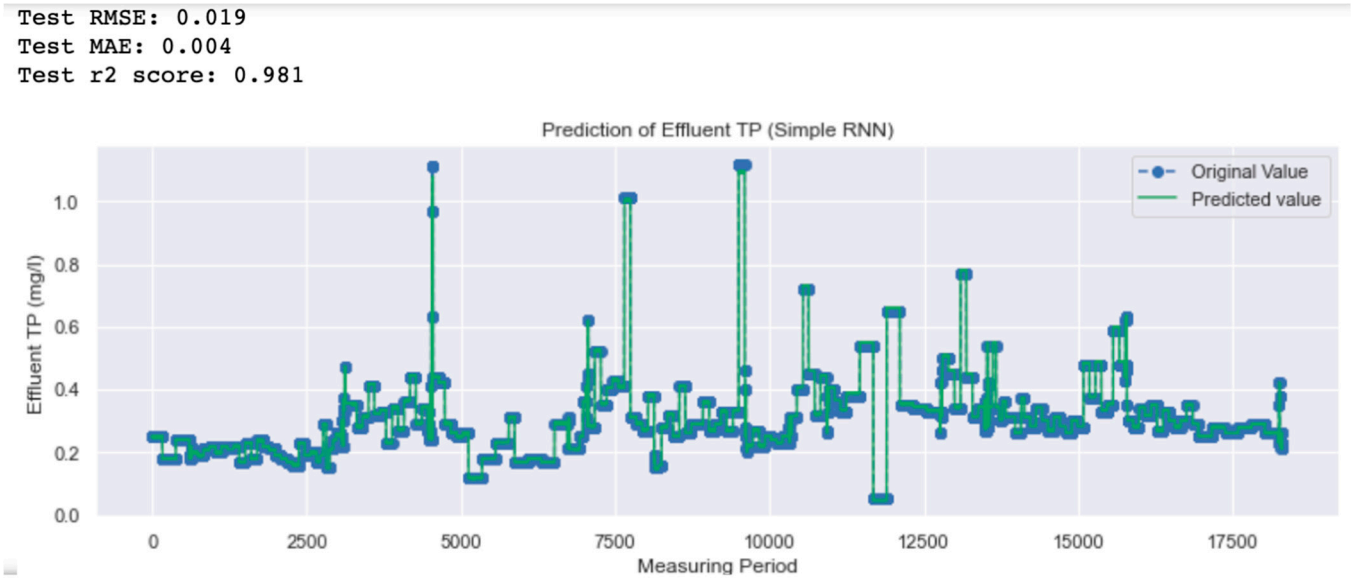

Figure 9. The prediction of effluent TP from 2015 to 2018 using the simple RNN model.

```
Test RMSE: 0.018
Test MAE: 0.003
Test r2 score: 0.981
```

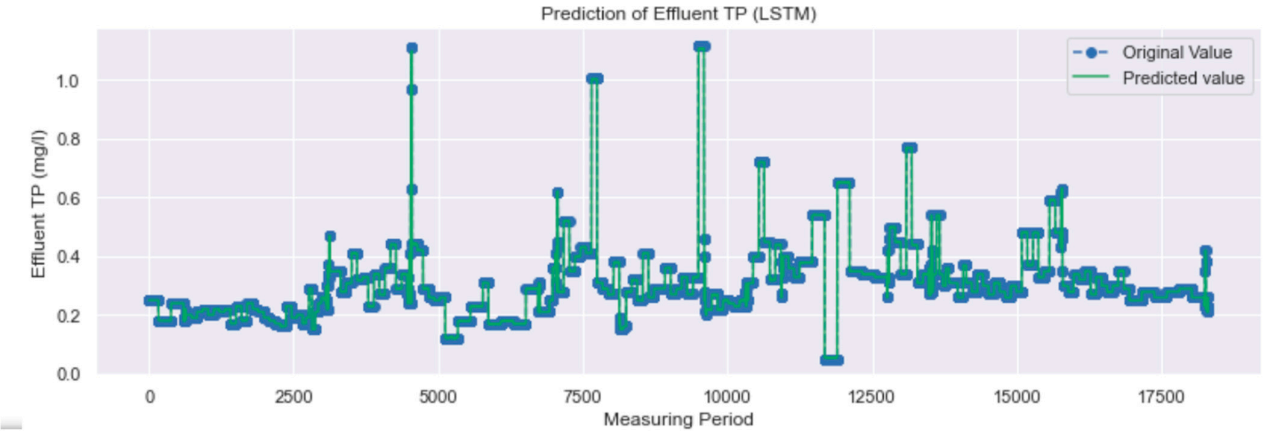

Figure 10. The prediction of effluent TP from 2015 to 2018 using the LSTM model.

```
Test RMSE: 0.091
Test MAE: 0.066
Test r2 score: 0.956
```

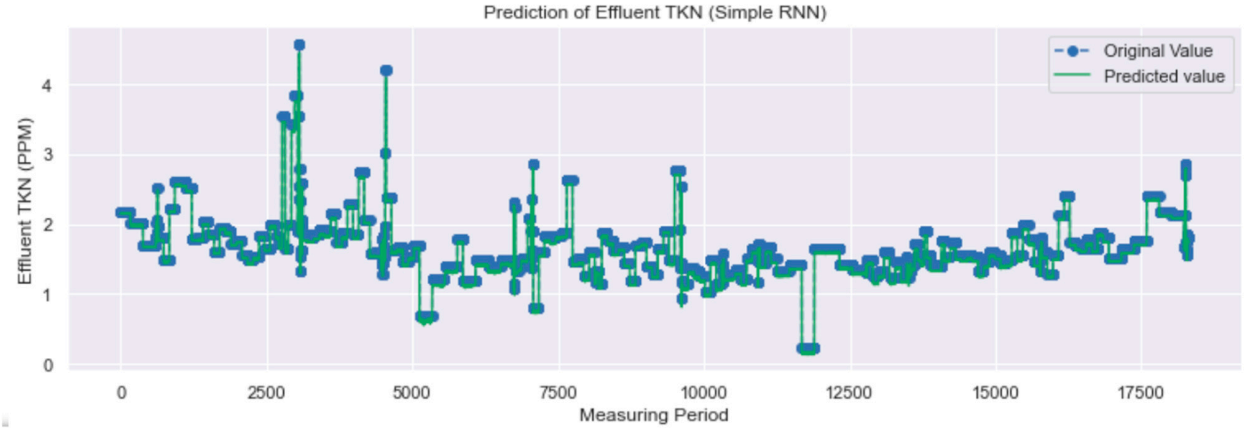

Figure 11. The prediction of effluent TKN from 2015 to 2018 using the simple RNN model.

```
Test RMSE: 0.067
Test MAE: 0.027
Test r2 score: 0.976
```

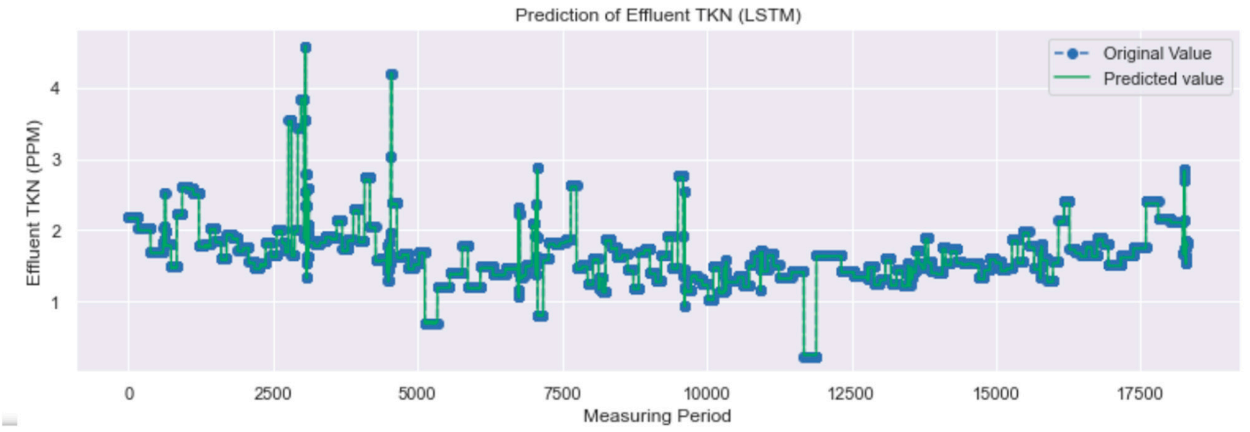

Figure 12. The prediction of effluent TKN from 2015 to 2018 using the LSTM model.

```
Test RMSE: 0.319
Test MAE: 0.051
Test r2 score: 0.980
```

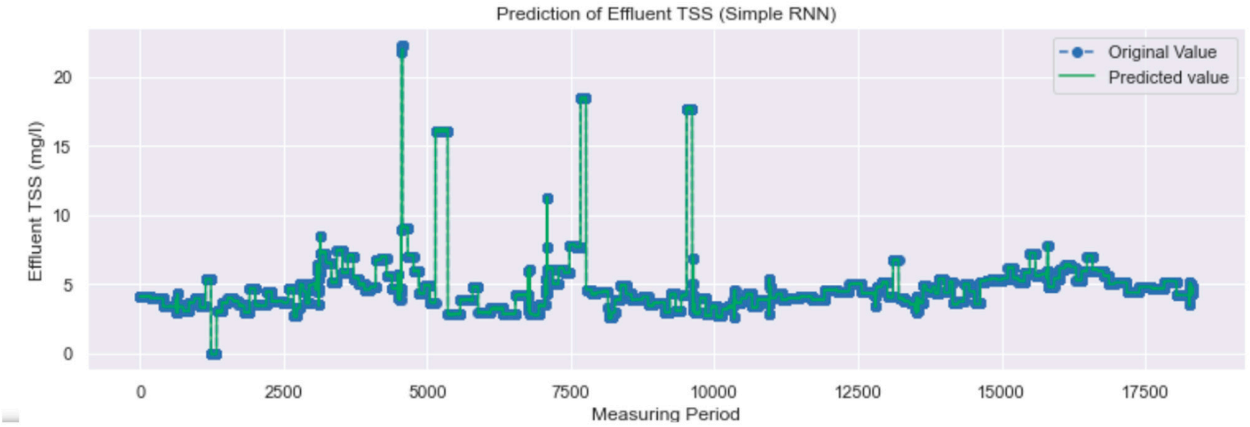

Figure 13. The prediction of effluent TSS from 2015 to 2018 using the simple RNN model.

```
Test RMSE: 0.320
Test MAE: 0.074
Test r2 score: 0.980
```

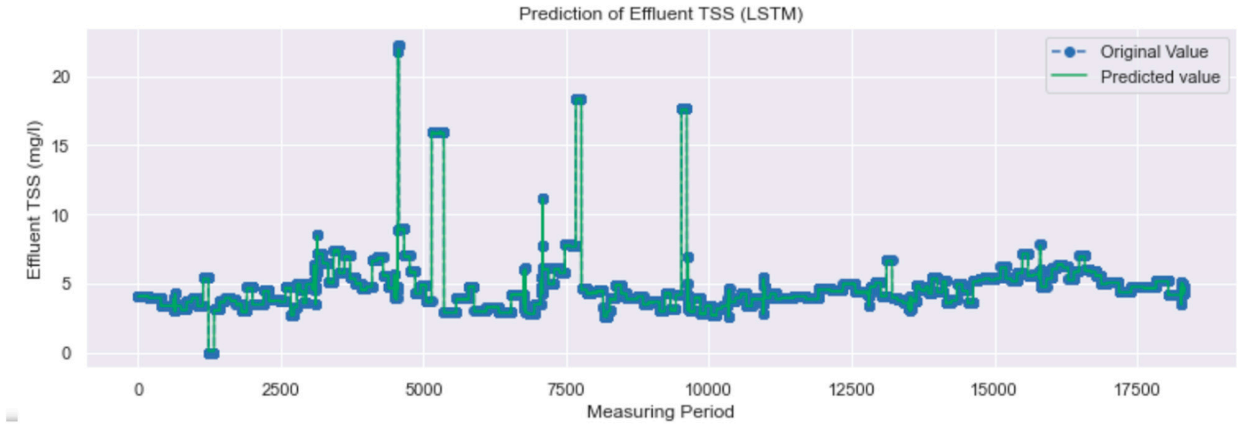

Figure 14. The prediction of effluent TSS from 2015 to 2018 using the LSTM model.

```
Test RMSE: 0.054
Test MAE: 0.025
Test r2 score: 0.955
```

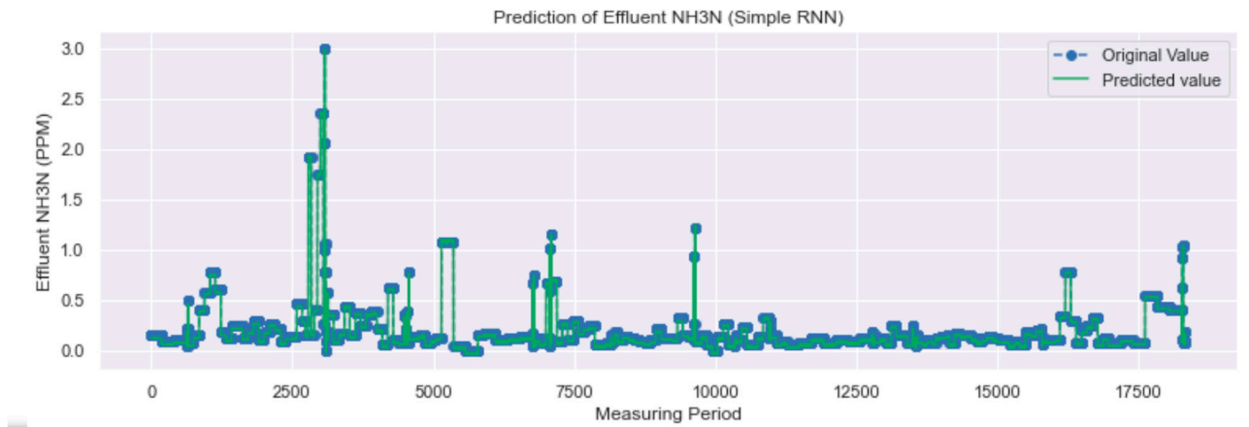

Figure 15. The prediction of effluent NH_3-N from 2015 to 2018 using the simple RNN model.

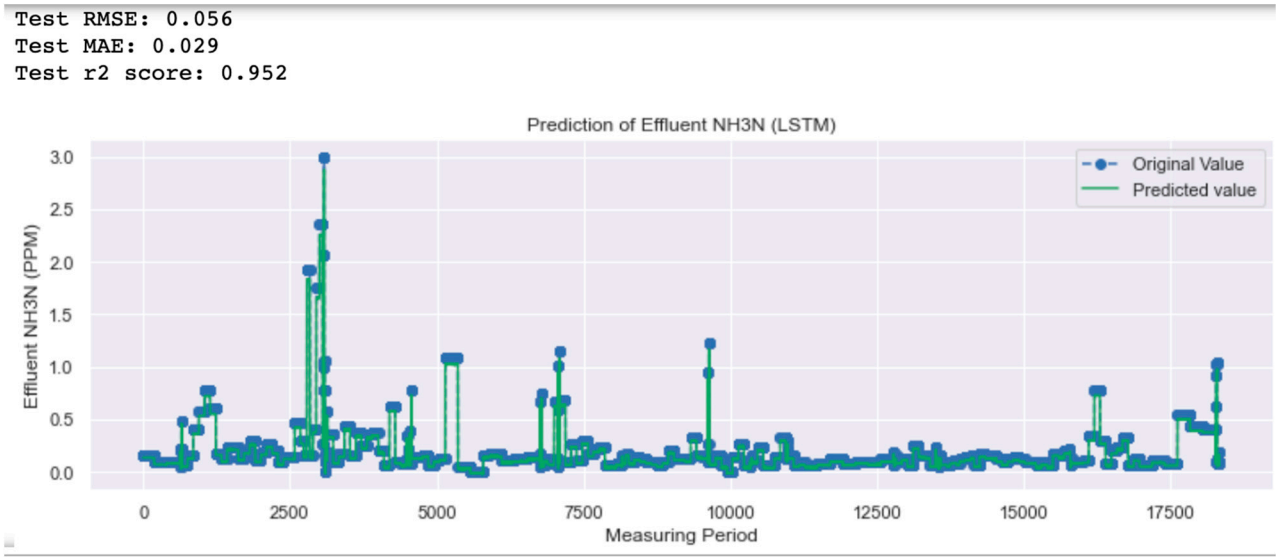

Figure 16. The prediction of effluent NH_3-N from 2015 to 2018 using the LSTM model.

4. Discussion

The result of the study shows the model performance comparison between two data scenarios by applying a simple RNN model and an LSTM model with different numbers of epochs and batch sizes. The input parameters are influent TSS, TP, TKN, and NH_3-N. The output is the effluent BOD_5. Scenario 2, the dataset from 2015 to 2028, with the epoch of 50 and batch size of 100, results in the most optimal model with a training time of 1~2 s in each epoch and RMSE of 0.3–0.8.

Figures 17 and 18 represent the train and test loss plot over the epochs to evaluate the best two predictive models, simple RNN and LSTM, for BOD_5 prediction with a training epoch of 50 and batch size of 100. The result showed that both LSTM and RNN models fit the sequence data because training loss decreased over time, achieving low error values. A good fit is determined by a training and validation loss that decreases to the point of constancy with a minimal gap between the two ending loss values [24]. Moreover, the loss of the model should be lower on the training set than on the test set. Thus, the LSTM model tends to be better than the RNN model because the loss for training and testing decreased consistently, and the final test loss was above the training loss. The BOD_5 prediction model from both simple RNN and LSTM models achieved similar RMSEs of 0.247 and 0.246, respectively.

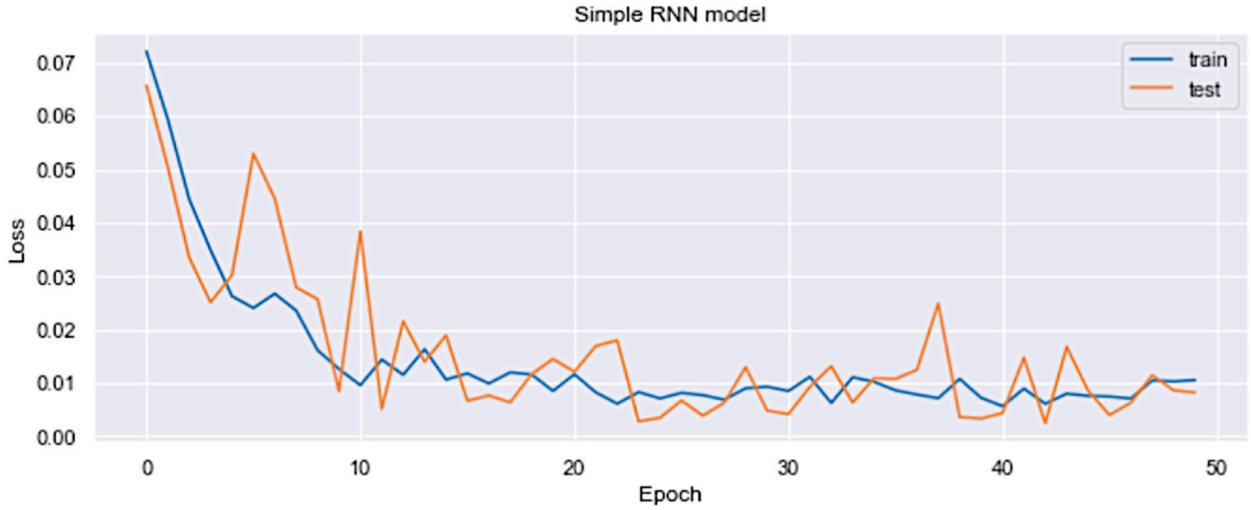

Figure 17. Train and test loss over epoch for effluent BOD_5 prediction of simple RNN model.

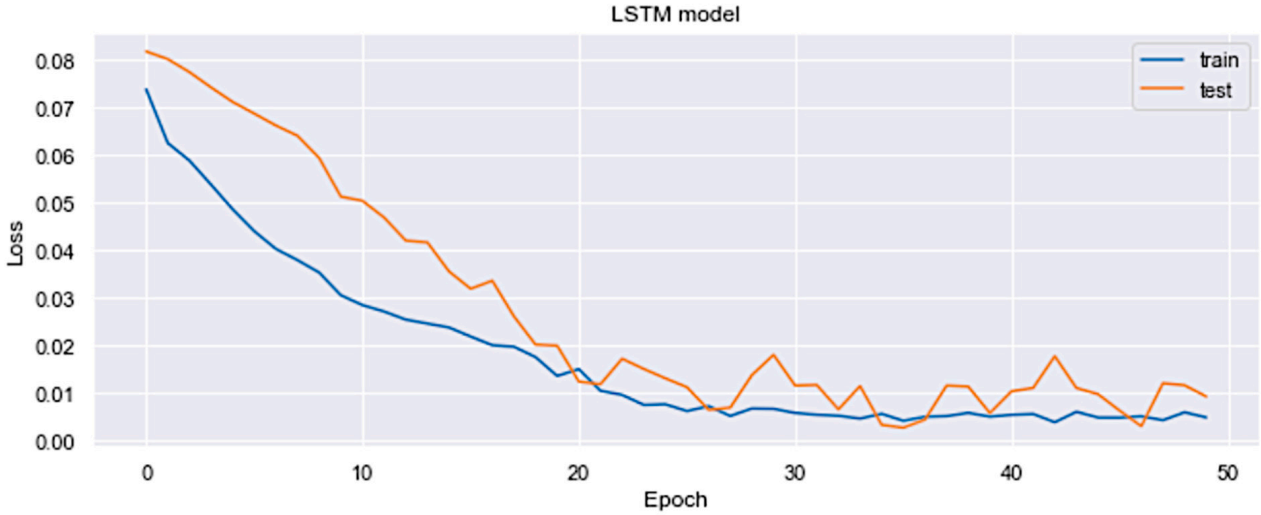

Figure 18. Train and test loss over epoch for effluent BOD_5 prediction of LSTM model.

Then, the same models with 100 batch sizes and 50 epochs for other important effluent parameters in wastewater treatment plants, including TP, TKN, TSS, and NH_3N, created the best model performance. The prediction model of TP from both the simple RNN and LSTM models achieved an RMSE of ~0.02 and an MAE of 0.004, which are very low. It means the models can predict with an error of less than 0.02. Other prediction models can also achieve shallow errors. The prediction of TKN for both models achieved an RMSE of 0.09 and an MAE of 0.07. The RMSE and MAE for the prediction of TSS were 0.032 and 0.05, respectively. Lastly, the prediction of NH_3N had an RMSE of 0.06 and an MAE of 0.03.

Therefore, the simple RNN and LSTM models displayed high prediction accuracy for effluent parameters, as illustrated in Table 6. These models exhibited low RMSE and MAE values, indicating high precision. Furthermore, their R^2 scores were close to one, confirming the model's accuracy. However, the R^2 score may only partially be suitable for time series model prediction as it is more tailored to linear regression [28]. In the context of a neural network model, RMSE was a more suitable metric to measure model accuracy because it operates in the same unit as the model. As such, the RMSE for BOD_5 prediction in the simple RNN and LSTM models showed an error of 0.246–0.247, between the predicted and original values. Regardless, TSS and BOD_5 had a more extensive value range than TP, TKN, and NH_3-N. Thus, the RMSE needs to be considered from the original data scaling. The more extensive range of the data will create a higher RMSE value. Notably, for TKN prediction, the LSTM model displayed considerably lower RMSE and MAE values than the simple RNN model. This result indicates that the LSTM model performs well with the high deviation parameters, as shown in Table 5, compared to its original data. Still, the other parameters show that simple RNN and LSTM models have very high accuracy that could apply to effluent wastewater prediction.

Table 6. The summary of the accuracy of effluent parameters models.

Effluent Parameters	Simple RNN			LSTM		
	RMSE	MAE	R^2	RMSE	MAE	R^2
BOD_5	0.247	0.052	0.985	0.246	0.058	0.985
TP	0.019	0.004	0.981	0.018	0.003	0.981
TKN	0.091	0.066	0.956	0.067	0.027	0.976
TSS	0.319	0.051	0.980	0.320	0.074	0.980
NH_3-N	0.054	0.025	0.955	0.056	0.029	0.952

5. Conclusions

Recurrent Neural Networks (RNNs) are excellent for developing a predictive model for sequential big data in WWTPs. Simple RNN and LSTM models were used to evaluate the model performance in predicting effluent parameters with different data sizes, epochs, and batch sizes. According to the dataset, the result shows that the historical data from 2015 to 2018 with epochs of 50 and batch sizes of 100 are the optimum model architecture, which results in the lowest RMSE value between 0.3 and 0.8 and less training time. The prediction model was developed in five steps: define, compile, fit, and evaluate the network, and then make a prediction. The predicted results were precise from low RMSE and MAE values compared to the original data range, while R^2 was close to one. The most suitable evaluation method calculated an RMSE value, determining the error between the predicted and original values. Both RMSE and MAE values were similar for both simple RNN and LSTM models in BOD_5, TP, TSS, and NH_3-N predictions. For TKN prediction, LSTM achieved a higher accuracy of 0.067, while the simple RNN model had an RMSE of 0.091 due to the high deviation of data. Thus, the LSTM model achieved better results for variation values. Finally, both simple RNN and LSTM models proposed in this study are robust and can be used to predict effluent water quality in WWTPs. From the trained model, predicting the effluent quality for a few days should be possible with a real-time optimization model that could automatically select the most suitable model architecture, including epoch, batch size, and even the dataset. This model development will aid WWTPs in operation and process optimization.

Author Contributions: Conceptualization, J.K.P.; data curation, P.W.; formal analysis, P.W.; methodology, P.W.; software, P.W.; supervision, J.K.P.; validation, P.W.; visualization, P.W.; writing—original draft, P.W.; writing—review and editing, J.K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the organizational privacy.

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

References

1. Haider, S.A.; Naqvi, S.R.; Akram, T.; Umar, G.A.; Shahzad, A.; Sial, M.R.; Khaliq, S.; Kamran, M. LSTM Neural Network Based Forecasting Model for Wheat Production in Pakistan. *Agronomy* **2019**, *9*, 72. [CrossRef]
2. Grant, S.B.; Saphores, J.-D.; Feldman, D.L.; Hamilton, A.J.; Fletcher, T.D.; Cook, P.L.M.; Stewardson, M.; Sanders, B.F.; Levin, L.A.; Ambrose, R.F.; et al. Taking the "Waste" out of "Wastewater" for Human Water Security and Ecosystem Sustainability. *Science* **2012**, *337*, 681–686. [CrossRef] [PubMed]
3. Durrenmatt, D.J. Data Mining and Data-Driven Modeling Approaches to Support Wastewater Treatment Plant Operation 2011. Available online: https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/42293/eth-4649-02.pdf?sequence=2&isAllowed=y (accessed on 16 August 2023).
4. Romero, J.M.P.; Hallett, S.H.; Jude, S. Leveraging Big Data Tools and Technologies: Addressing the Challenges of the Water Quality Sector. *Sustainability* **2017**, *9*, 2160. [CrossRef]
5. Wong, Y.J.; Shimizu, Y.; He, K.; Nik Sulaiman, N.M. Comparison among Different ASEAN Water Quality Indices for the Assessment of the Spatial Variation of Surface Water Quality in the Selangor River Basin, Malaysia. *Environ. Monit. Assess.* **2020**, *192*, 644. [CrossRef]
6. Wong, Y.J.; Shimizu, Y.; Kamiya, A.; Maneechot, L.; Bharambe, K.; Chng Saun, F.; Nik Sulaiman, N.M. Application of Artificial Intelligence Methods for Monsoonal River Classification in Selangor River Basin, Malaysia. *Environ. Monit. Assess.* **2021**, *193*, 438. [CrossRef]
7. Sakaa, B.; Elbeltagi, A.; Boudibi, S.; Chaffaï, H.; Islam, A.R.M.T.; Kulimushi, L.C.; Choudhari, P.; Hani, A.; Brouziyne, Y.; Wong, Y.J. Water Quality Index Modeling Using Random Forest and Improved SMO Algorithm for Support Vector Machine in Saf-Saf River Basin. *Environ. Sci. Pollut. Res.* **2022**, *29*, 48491–48508. [CrossRef]

8. Harrou, F.; Dairi, A.; Sun, Y.; Senouci, M. Statistical Monitoring of a Wastewater Treatment Plant: A Case Study. *J. Environ. Manag.* **2018**, *223*, 807–814. [CrossRef] [PubMed]
9. Martin, C.; Vanrolleghem, P.A. Analysing, Completing, and Generating Influent Data for WWTP Modelling: A Critical Review. *Environ. Model. Softw.* **2014**, *60*, 188–201. [CrossRef]
10. Boyd, G.; Na, D.; Li, Z.; Snowling, S.; Zhang, Q.; Zhou, P. Influent Forecasting for Wastewater Treatment Plants in North America. *Sustainability* **2019**, *11*, 1764. [CrossRef]
11. Duarte, M.S.; Martins, G.; Oliveira, P.; Fernandes, B.; Ferreira, E.C.; Alves, M.M.; Lopes, F.; Pereira, M.A.; Novais, P. A Review of Computational Modeling in Wastewater Treatment Processes. *ACS EST Water* **2023**. [CrossRef]
12. Zhao, L.; Dai, T.; Qiao, Z.; Sun, P.; Hao, J.; Yang, Y. Application of Artificial Intelligence to Wastewater Treatment: A Bibliometric Analysis and Systematic Review of Technology, Economy, Management, and Wastewater Reuse. *Process Saf. Environ. Prot.* **2020**, *133*, 169–182. [CrossRef]
13. Huang, M.; Han, W.; Wan, J.; Ma, Y.; Chen, X. Multi-Objective Optimisation for Design and Operation of Anaerobic Digestion Using GA-ANN and NSGA-II. *J. Chem. Technol. Biotechnol.* **2016**, *91*, 226–233. [CrossRef]
14. Nadiri, A.A.; Shokri, S.; Tsai, F.T.-C.; Moghaddam, A.A. Prediction of Effluent Quality Parameters of a Wastewater Treatment Plant Using a Supervised Committee Fuzzy Logic Model. *J. Clean. Prod.* **2018**, *180*, 539–549. [CrossRef]
15. Man, Y.; Hu, Y.; Ren, J. Forecasting COD Load in Municipal Sewage Based on ARMA and VAR Algorithms. *Resour. Conserv. Recycl.* **2019**, *144*, 56–64. [CrossRef]
16. Shi, S.; Xu, G. Novel Performance Prediction Model of a Biofilm System Treating Domestic Wastewater Based on Stacked Denoising Auto-Encoders Deep Learning Network. *Chem. Eng. J.* **2018**, *347*, 280–290. [CrossRef]
17. Yu, P.; Cao, J.; Jegatheesan, V.; Shu, L. Activated Sludge Process Faults Diagnosis Based on an Improved Particle Filter Algorithm. *Process Saf. Environ. Prot.* **2019**, *127*, 66–72. [CrossRef]
18. Najafzadeh, M.; Zeinolabedini, M. Prognostication of Waste Water Treatment Plant Performance Using Efficient Soft Computing Models: An Environmental Evaluation. *Measurement* **2019**, *138*, 690–701. [CrossRef]
19. Qiao, J.F.; Yang, W.W.; Yuan, M.Z. Recurrent High Order Neural Network Modeling for Wastewater Treatment Process. *J. Comput.* **2011**, *6*, 1570–1577. [CrossRef]
20. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. *arXiv* **2014**, arXiv:1409.3215.
21. Wongburi, P.; Park, J.K. Big Data Analytics from a Wastewater Treatment Plant. *Sustainability* **2021**, *13*, 12383. [CrossRef]
22. McGowan, S.; Wang, E. *50-Year Master Plan Review of Existing Treatment Facilities*; Malcolm Pirnie: New York, NY, USA, 2008.
23. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 3rd ed.; Morgan Kaufmann Publishers: Waltham, MA, USA, 2011; ISBN 978-0-12-381479-1.
24. Brownlee, J. *Long Short-Term Memory Networks with Python, v 1.0*; Jason Brownlee: Melbourne, Australia, 2017.
25. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980.
26. Chai, T.; Draxler, R.R. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?—Arguments against Avoiding RMSE in the Literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
27. Figueiredo, D.; Júnior, S.; Rocha, E. What Is R2 All About? *Leviathan-Cad. Pesqui. Polútica* **2011**, *3*, 60–68. [CrossRef]
28. Hagquist, C.; Stenbeck, M. Goodness of Fit in Regression Analysis—R2 and G2 Reconsidered. *Qual. Quant.* **1998**, *32*, 229–245. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.