processes

MDPI

Article

Performance Evaluation of Hydroponic Wastewater Treatment Plant Integrated with Ensemble Learning Techniques: A Feature Selection Approach

Hauwa Mohammed Mustafa [1,2,3,*], Gasim Hayder [4,5,*], S. I. Abba [6], Abeer D. Algarni [7], Mohammed Mnzool [8] and Abdurahman H. Nour [9]

1 College of Graduate Studies, Universiti Tenaga Nasional (UNITEN), Kajang 43000, Selangor Darul Ehsan, Malaysia
2 Department of Pure and Applied Chemistry, Kaduna State University (KASU), Tafawa Balewa Way, Kaduna PMB 2339, Nigeria
3 Centre for Energy and Environmental Strategy Research, Kaduna State University (KASU), Tafawa Balewa Way, Kaduna PMB 2339, Nigeria
4 Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Kajang 43000, Selangor Darul Ehsan, Malaysia
5 Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), Kajang 43000, Selangor Darul Ehsan, Malaysia
6 Interdisciplinary Research Center for Membranes and Water Security, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
7 Department of Information Technology, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
8 Department of Civil Engineering, College of Engineering, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia
9 Faculty of Chemical and Natural Resources Engineering, Universiti Malaysia Pahang (UMP), Gambang 26300, Pahang, Malaysia
* Correspondence: hauwa.mustafa@uniten.edu.my (H.M.M.); gasim@uniten.edu.my (G.H.)

check for updates

Citation: Mustafa, H.M.; Hayder, G.; Abba, S.I.; Algarni, A.D.; Mnzool, M.; Nour, A.H. Performance Evaluation of Hydroponic Wastewater Treatment Plant Integrated with Ensemble Learning Techniques: A Feature Selection Approach. *Processes* **2023**, *11*, 478. https://doi.org/10.3390/pr11020478

Academic Editors: Bipro R. Dhar and Andrea Petrella

Received: 15 November 2022
Revised: 20 January 2023
Accepted: 29 January 2023
Published: 4 February 2023

Abstract: Wastewater treatment and reuse are being regarded as the most effective strategy for combating water scarcity threats. This study examined and reported the applications of the Internet of Things (IoT) and artificial intelligence in the phytoremediation of wastewater using *Salvinia molesta* plants. Water quality (WQ) indicators (total dissolved solids (TDS), temperature, oxidation-reduction potential (ORP), and turbidity) of the *S. molesta* treatment system at a retention time of 24 h were measured using an Arduino IoT device. Finally, four machine learning tools (ML) were employed in modeling and evaluating the predicted concentration of the total dissolved solids after treatment (TDSt) of the water samples. Additionally, three nonlinear error ensemble methods were used to enhance the prediction accuracy of the TDSt models. The outcome obtained from the modeling and prediction of the TDSt depicted that the best results were observed at SVM-M1 with 0.9999, 0.0139, 1.0000, and 0.1177 for R^2, MSE, R, and RMSE, respectively, at the training stage. While at the validation stage, the R^2, MSE, R, and RMSE were recorded as 0.9986, 0.0356, 0.993, and 0.1887, respectively. Furthermore, the error ensemble techniques employed significantly outperformed the single models in terms of mean square error (MSE) and root mean square error (RMSE) for both training and validation, with 0.0014 and 0.0379, respectively.

Keywords: error ensemble methods; computational analysis; water quality forecasting; total dissolved solids; energy

1. Introduction

Potable water is a necessary component of human existence, second only to food and shelter in terms of basic survival requirements. Surface water and groundwater are the most important sources of potable water. Globally, more than 1.4 billion people subsist on

insufficient water every year [1]. Additionally, water pollution control is a critical element of environmental contamination. Water pollution is caused by the release of municipal, agricultural, and industrial waste products into bodies of water [2]. Therefore, water quality (WQ) forecasting has become essential for the management of water pollution to enhance the efficient supply of potable water. However, numerous researchers have employed and documented the benefits of phytoremediation methods for wastewater treatment [3–7]. Furthermore, the complexity and anomalies that frequently occur during data collection and the development of model structure make it challenging to create reliable and effective WQ models [8]. Thus, stakeholders and decision-makers must incorporate modern technology and new knowledge for the future upgrade of phytoremediation techniques in wastewater remediation applications.

Recently, machine learning (ML) techniques have been introduced in the forecasting of water quality (WQ) parameters due to the availability of large data and computational resources [9–14]. Additionally, several studies have used artificial intelligence (AI) techniques to forecast and model WQ parameters [5,15]. Vo et al. [16] used *Scripus validus* for the phytoremediation of hospital wastewater containing acetaminophen (ACT). In the study, the correlation between the peroxidase enzyme extruded by *S. validus* and pollutants' removal efficiency was evaluated by applying different multivariable regression models. The results showed that the concentrations of ACT in constructed wetland effluent and enzymes in *S. validus* exhibited a significant correlation ($p < 0.001$, $R^2 = 68.3\%$). Kumar et al. [17] applied two-factor multiple linear regression in the prediction of heavy metal removal using water lettuce from paper mill effluent (PME). The findings indicated that the selected input variables helped in the development of prediction models with a high model efficiency (ME), higher linear regression (h^2), and low mean average normalizing error (MANE) of 0.92–0.99, $h^2 > 0.72$, and MANE < 0.02, respectively. The authors concluded that their study demonstrated an efficient technique for simulating the absorption of heavy metals by water lettuce from PME.

Furthermore, Kumar and Deswal [18] conducted a comparative study on artificial neural networks (ANN), random forest, and M5P techniques in predicting phosphorous removal from rice mill wastewater, where 30% of the trained data was used for testing and 70% for training. The modeling findings suggest that ANN outperforms the M5P tree and random forest models. Besides, developing an accurate and reliable model is difficult in wastewater treatment systems [19]. Hence, the nature of the historical data influences the outcome of the generated models [20]. Other research on AI and wastewater can be found in [21–26]. Additionally, ensemble methods have been employed to improve the performance of trained single models by summing or averaging their individual outputs [27]. For instance, Xenochristou and Kapelan [20] proved that ensemble models could outperform single models by contrasting different bias correction methods to enhance the performance of the trained models.

However, despite the promising outcome of ensemble methods, they only just started to attract interest a few years ago [28]. Thus, this study employed an artificial neural network (ANN), support vector machines (SVM), an adaptive neuro-fuzzy inference system (ANFIS), and multilinear regression (MLR) in predicting the concentration of the total dissolved solids in the treated water samples (TDSt). Additionally, three different error ensemble learning methods were used to improve the accuracy of the trained models. Furthermore, ensemble learning methods are yet to be employed in the field of phytoremediation, either in technical proposals or published literature. As such, this research will be the first to propose the application of non-linear error ensemble methods in the phytoremediation of wastewater using *S. molesta* plants. According to the published results, AI-based models have become more and more widely known in the desalination WWT field. For example, a smart analysis research study of the widely used Scopus database (1984–January 2023) revealed that the minimum number of keyword phenomena per report is 5 out of more than 7800 documents. More than 1000 keywords were also determined to

fulfill the threshold. For each of the 1000 keywords, the total strength of the concurrency hyperlinks with their key phrases was computed (see Figure 1).

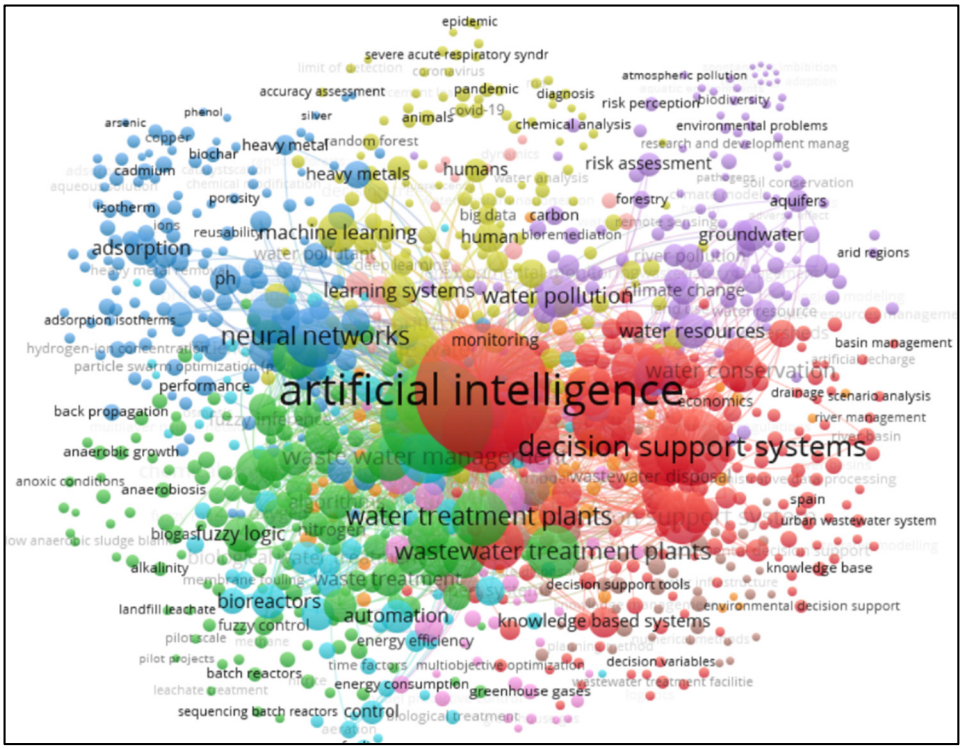

Figure 1. The major keywords used over the literature on desalination and AI-based models (1984–January 2023).

2. Methodology

2.1. Data Collection

S. molesta plants were cultivated for the phytoremediation of domestic wastewater using hydroponic systems. The Arduino Internet of Things (IoT) device was used to monitor the phytoremediation of the secondary treated wastewater samples cultivated with 280 g of the fresh *S. molesta* plants for 2 weeks. Furthermore, the data collection and WQ monitoring of the total dissolved solids (TDS), temperature, oxidation-reduction potential (ORP), and turbidity were carried out in accordance with the procedure outlined by Priyadharshini et al. [29].

2.2. Research Area

This study was carried out at a sewage treatment plant (STP) near Kajang (2°58′04″ N, 101°43′55″ E), Malaysia. The annual temperature of the environment was 27.2 °C. Additionally, the STP is separated into different chambers, where wastewater is treated at the primary and secondary stages before being released into the natural environment.

2.3. Proposed Models

WQ modeling and prediction are useful for the early detection of pollution and projections for future applications. The modeling and prediction of the TDSt parameter were proposed under three different scenarios. In the first scenario, two models were generated for the prediction of the water quality indicator. The models were built with the collected data as input and output variables. Model 1 (M1) consists of four input parameters (TDS, temperature, turbidity, and oxidation reduction potential (ORP)) for the prediction of TDSt (see, Figure 2). The details of the WQ parameter used in this study can be presented in Table 1. At the same time, Model 2 (M2) comprised three input

parameters (turbidity, ORP, and TDS) for the prediction of TDSt. An illustration of the process modeling schema and models is presented in Figure 3 (a = M1, b = M2). The second scenario explored the application of data-driven algorithms using ML techniques and one classical model (MLR) for the modeling of the phytoremediation process performance analysis based on the influent variables as the input of the model. For this purpose and approach, other possible models could be employed in the same but different manner. ANN, SVM, ANFIS, and MLR were selected due to their excellent historical performance in predicting WQ involving a large number of variables and promising abilities in several literatures of science, environment, hydrology, and hydro-informatics studies. The third scenario employed three types of ensemble learning techniques: simple averaging ensemble (SAE), weighted averaging ensemble (WAE), and nonlinear neural ensemble (NNE), using the single output of the TDS model. This was established to enhance the prediction accuracy regarding error only. Furthermore, to monitor the performance of the models, different evaluation criteria, namely the coefficient of correlation (R), the coefficient of determination (R^2), the mean square error (MSE), and the root mean square error (RMSE), were used. These parameters were used as benchmarks for determining the performance of the models for each of the water quality parameters. The flowchart of the single and error ensemble learning processes is presented in Figure 3.

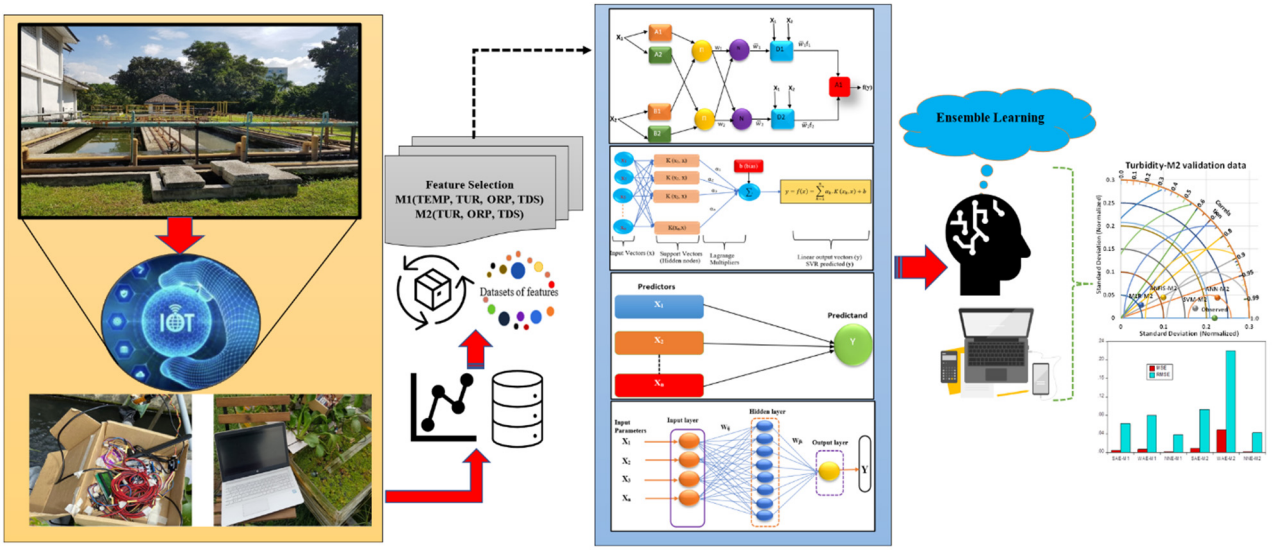

Figure 2. Proposed modeling schema and input combination.

Table 1. Parameters used (influent (raw) and effluent (treated) samples).

Parameters	Influent Parameters	Effluent Parameters
Raw Turbidity	TURBr	TDSt
Treated Turbidity	TURBt	
Raw Total Dissolve Solid	TDSt	
Treated Total Dissolve Solid	TDSr	
Raw Oxidation-Reduction Potential	ORPr	
Treated Oxidation-Reduction Potential	ORPt	
Raw Temperature	TEMPr	
Treated Temperature	TEMPt	

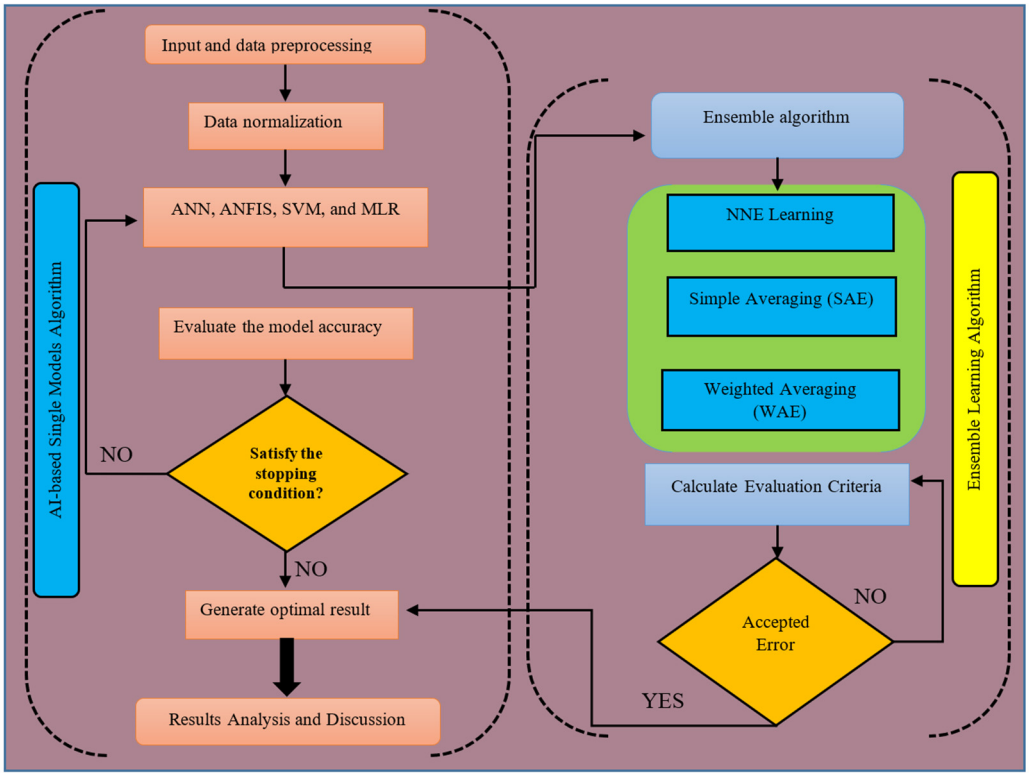

Figure 3. Flowchart of the ANN, SVM, ANFIS, MLR, and ensemble learning-based model-building process.

Additionally, the feature selection and data pre-processing stages are important in building ML models. This procedure has a significant impact on accuracy and prediction [6,30]. In this study, smoothing and normalization were used to describe the data trend series. Additionally, normalization was conducted to obtain uniform input and output values. The input and output values were standardized to fall within a specific range of 0 and 1 using the equation below [6,30]:

$$Y_{norm} = \frac{y - y_{min}}{y_{max} - y_{min}}$$

where y, Y_{norm}, y_{min}, and y_{max} are the observed, normalized, minimum, and maximum values of the variable, respectively.

2.4. Error Ensemble Learning Approach Development

The base learner is the most important component of ensemble learning. The general formula for ensemble learning is presented in Equation (1). Similarly, three types of ensemble learning methods were used to improve the accuracy of the TDS model.

$$P_e(x) = \sum_{i=1}^{n} p_i(x) \quad (1)$$

where $p(x)$ and $P_e(x)$ denote a single predictor and an ensemble of n predictors, respectively.

2.4.1. Simple Averaging Ensemble (SAE)

The SAE approach used in this study involved separate training and validation of ANN, ANFIS, SVM, and MLR models, followed by generating the average values of the

ANN, ANFIS, SVM, and MLR training outputs for each model. The general formula for SAE is provided in Equation (2):

$$P_{(t)} = \frac{1}{N}\sum_{i=1}^{N} P_i(ft) \quad (2)$$

where N and P_i represent the number of learners (N = 4) and the single model, $f(t)$ (ANN, ANFIS, SVM, and MLR).

2.4.2. Weighted Averaging Ensemble (WAE)

In this study, WAE was predicted by assigning different weights to individual parameters. It differs in the case of SAE because all parameters are assigned equal weights [31]. The formula for WAE is provided in Equation (3):

$$P_{(t)} = \sum_{i=1}^{N} Y_i p_i(t) \quad (3)$$

where $P_{(t)}$, Y_i, $p_i(t)$, and N are the output of SAE, the weight applied to the ith model, the output of the ith single model, and the number of single models (here, N = 4), respectively. Similarly, Y_i was calculated using Equation (4):

$$Y_i = \frac{DC_i}{\sum_{i=1}^{N} DC_i} \quad (4)$$

where DC_i is the performance efficiency of the ith single model.

2.4.3. Nonlinear Neural Ensemble (NNE)

In this study, NNE was performed through the training of another neural network. The network was trained using the backpropagation algorithm, and a tangent sigmoid was selected as the activation function of the hidden and output layers. Furthermore, the trial-and-error method was used in determining the epoch number and the best structure for the ensemble network. The obtained results are presented and discussed below.

3. Results of Single Models ANN, SVM, ANFIS, and MLR

As stated earlier, this research applied ANN, SVM, ANFIS, and MLR models in predicting the water quality parameters of the *S. molesta* treatment system. Two models were generated for the evaluation, and the model validation was conducted using four performance criteria. The obtained results are presented and discussed. TDS analysis was used to determine the number of dissolved materials in water, expressed in milligrams per liter (mg/L). TDS is important in estimating the suitability of drinking water because excess TDS in water may result in a "salty" taste [32]. The time-series and box plot for untreated and treated TDS concentrations in the phytoremediation system are presented in Figure 4.

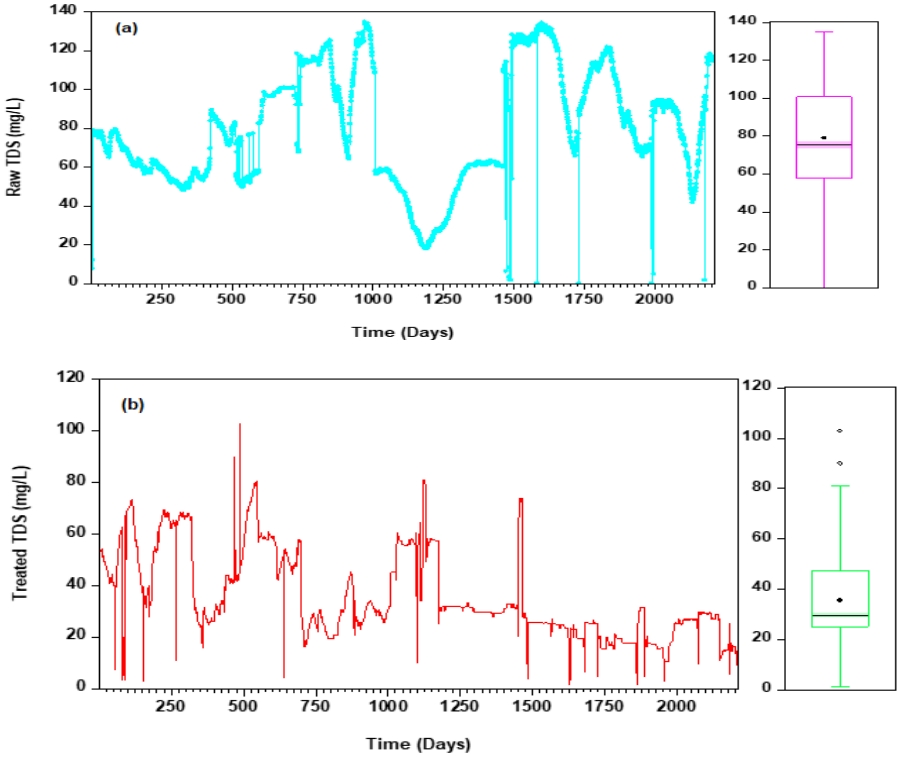

Figure 4. Plots of TDS concentration of the phytoremediation system for (**a**) influent and (**b**) effluent (treated) water samples.

3.1. Results of TDSt (ANN, SVM, ANFIS, and MLR)

The use of multiple inputs influences the performance of intelligent models. In the hydro-environmental literature, several input selection techniques, such as principal component analysis, correlation, and auto-correlation, have been published. However, these approaches are also used for linear input/output relationships. As stated above, three different ML techniques and one classical linear regression model were used in predicting the TDS of the effluent. The predicted results are presented in Table 2.

Table 2. Performance evaluation of TDSt for ANN, SVM, ANFIS, and MLR.

	Training				Validation			
Models	**R^2**	**MSE**	**R**	**RMSE**	**R^2**	**MSE**	**R**	**RMSE**
ANN-M1	0.9995	0.1176	0.9998	0.3430	0.9954	0.1140	0.9977	0.3377
ANN-M2	0.9982	0.4306	0.9991	0.6562	0.9631	0.9192	0.9814	0.9588
SVM-M1 *	0.9999	0.0139	1.0000	0.1177	0.9986	0.0356	0.9993	0.1887
SVM-M2	0.9970	0.9970	0.9985	0.9985	0.9852	0.3696	0.9925	0.6079
ANFIS-M1	0.9988	0.3024	0.9994	0.5499	0.9716	0.0011	0.9857	0.0326
ANFIS-M2	0.9926	1.8033	0.9963	1.3429	0.8309	4.2103	0.9115	2.0519
MLR-M1	0.9928	1.7588	0.9964	1.3262	0.8350	4.1066	0.9138	2.0265
MLR-M2	0.9883	2.8616	0.9941	1.6916	0.7316	6.6813	0.8553	2.5848

* Signifies the overall best model.

From Table 2, it was found that both the classical and ML models were capable of modeling TDSt, and this was proven by considering the statistical indicators (R, R^2, MSE, and RMSE). The visual investigation of the models indicated that M1 was superior to M2 for all the models. This was attributed to the fact that M1 contained additional input variables, which served as the dominant and significant variables. Concerning complexity, M1 was the best since four input variables were employed in the modeling process, while in the trial of M2, a decrease in accuracy was observed due to the disparity in the performance

between the input trials. As a result, SVM-M1 was chosen as the overall best in TDSt modeling. Figures 5 and 6 show the time-series graphs for M1 and M2 for the observed and predicted TDSt in the validation phase.

According to the time-series plots, it was evident that the best predictive model was M1, since the prediction trend was closer to the observed TDSt. Similarly, the trends of the observed and predicted values for ML models were found to be close to each other. The MLR models, on the other hand, revealed a discrepancy in the trends of observed and predicted values. Furthermore, the time series of the models differ from one combination to the next, implying that the vibrational pattern of the M models is determined by how they capture the relationship between the observed and target parameters. From the above figures, it can be justified that SVM-M1 captures the pattern of the time series more than the other models, with MLR-M2 being the worst model. The uniqueness of the goodness-of-fit could be indicated in the radar chart. On the other hand, it is important to involve two or more performance criteria to come up with a justifiable argument. The radar plots describing the R and R^2 values for all the models are shown in Figure 7.

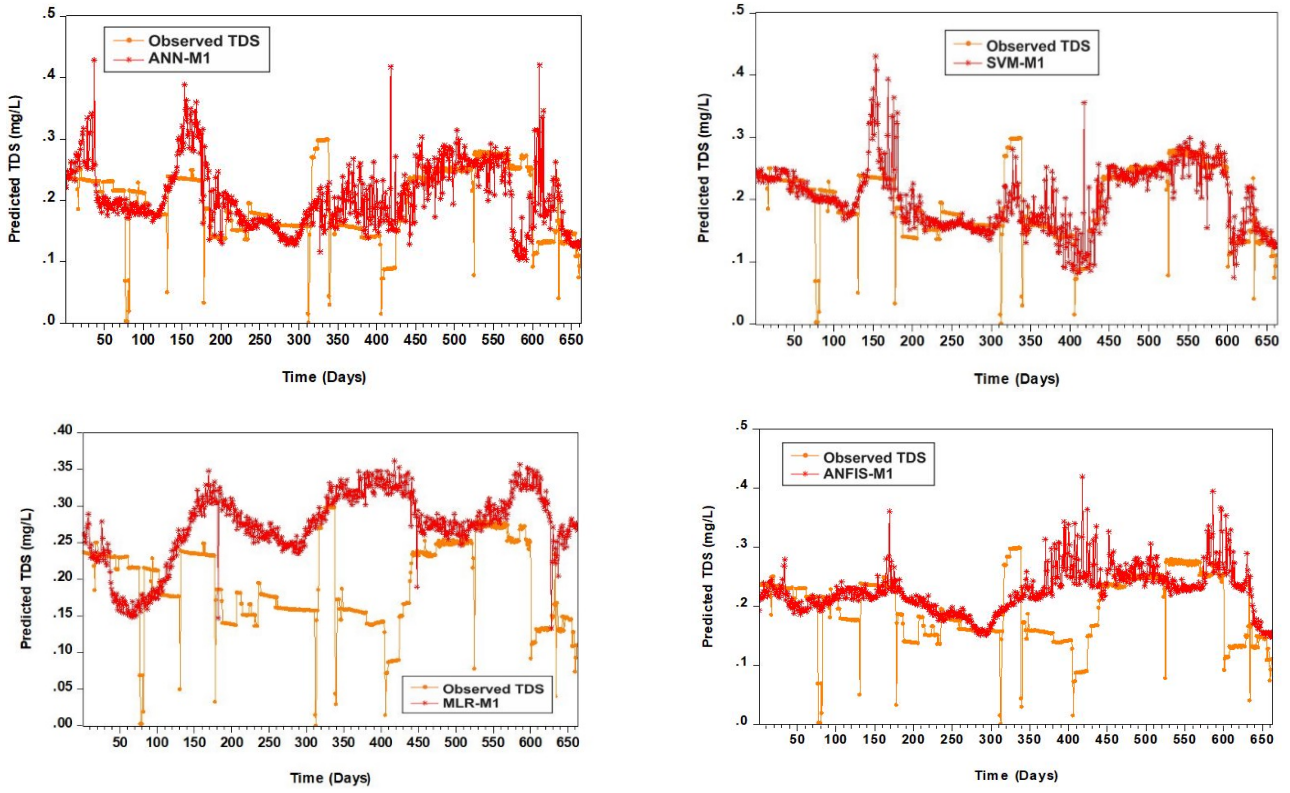

Figure 5. Time-series plot for ANN-M1, ANFIS-M1, SVM-M1, and MLR-M1.

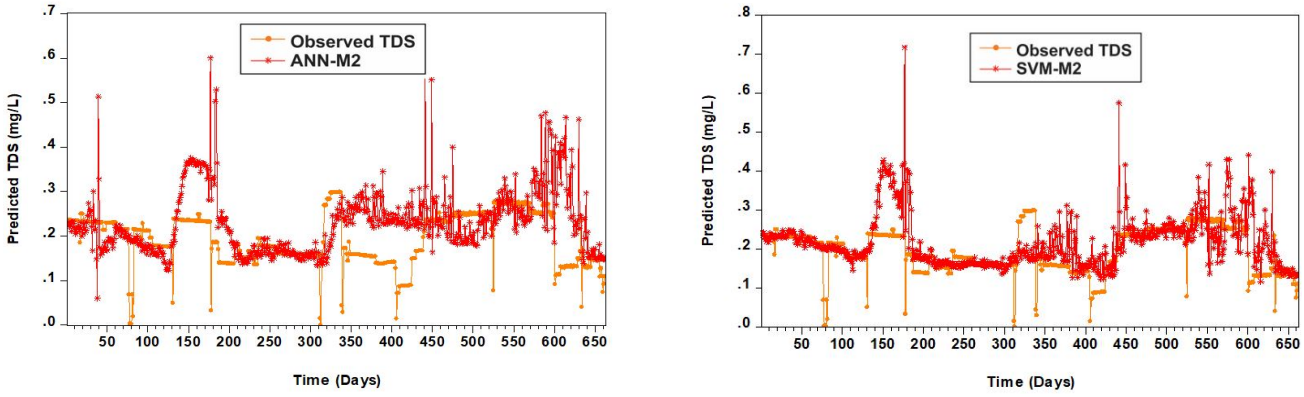

Figure 6. *Cont.*

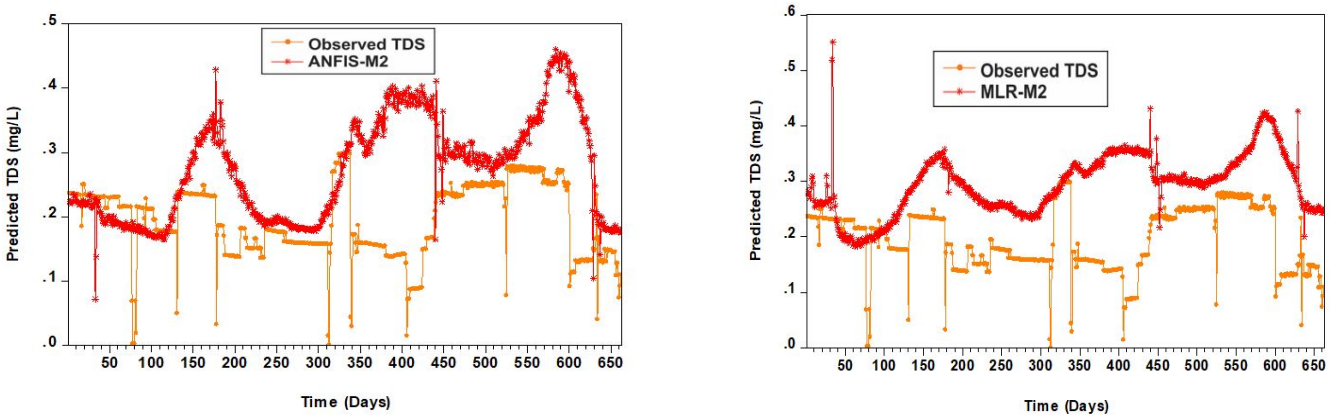

Figure 6. Time series plot for ANN-M2, ANFIS-M2, SVM-M2, and MLR-M2.

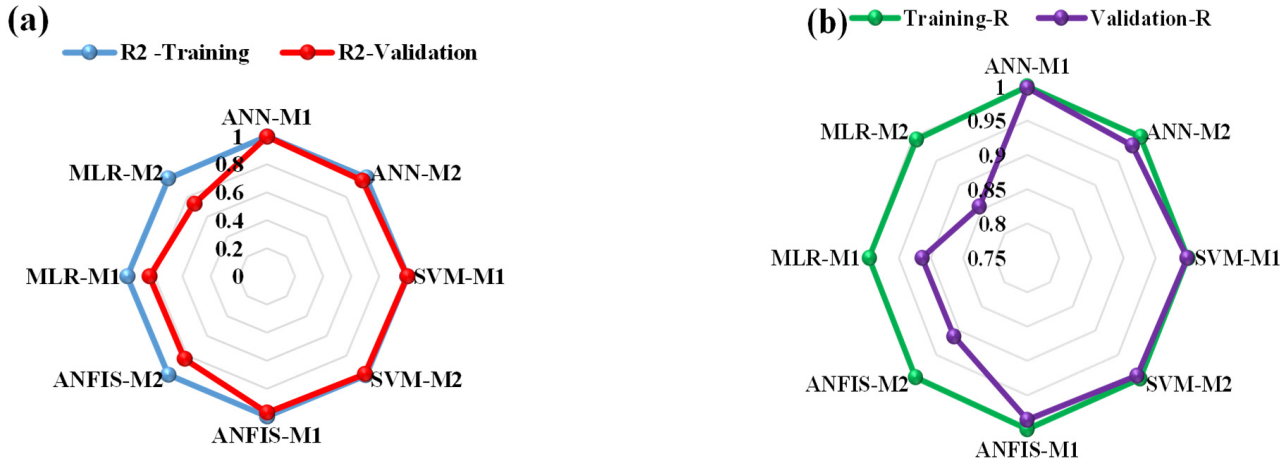

Figure 7. Radar chart for (**a**) goodness-of-fit and (**b**) correlation coefficient for TDS.

Furthermore, a radar plot was used to compare the predictive accuracy of the models. As mentioned earlier, radar plots are important in depicting the performance comparison of the models due to their transparency. In this vein, a radar chart is a type of graph used to compare three or more variables on a two-dimensional plane. Radar charts can easily be used for depicting several variables without creating a clutter and they are viewed as a better substitute for column graphs. Thus, radar plots were used regarding R and R^2 for the training and validation phases. According to Figure 8, it was observed that the R-values in the training and validation phases were ANN-M1 = 0.9977, ANN-M2 = 0.9814, SVM-M1 = 0.9993, SVM-M2 = 0.9925, ANFIS-M1 = 0.9857, ANFIS-M2 = 0.9115, MLR-M1 = 0.9138, and MLR-M2 = 0.8553. Additionally, the radar chart ranged from 0 to 1, with the best value approaching 1. It was observed that SVM-M1 was on the last spider web for both the goodness-of-fit and R. This proves the accuracy of the training and validation results presented in Table 3. Figure 8 presents a boxplot illustration of the similarities between the observed and predicted TDSt for all the predicted models.

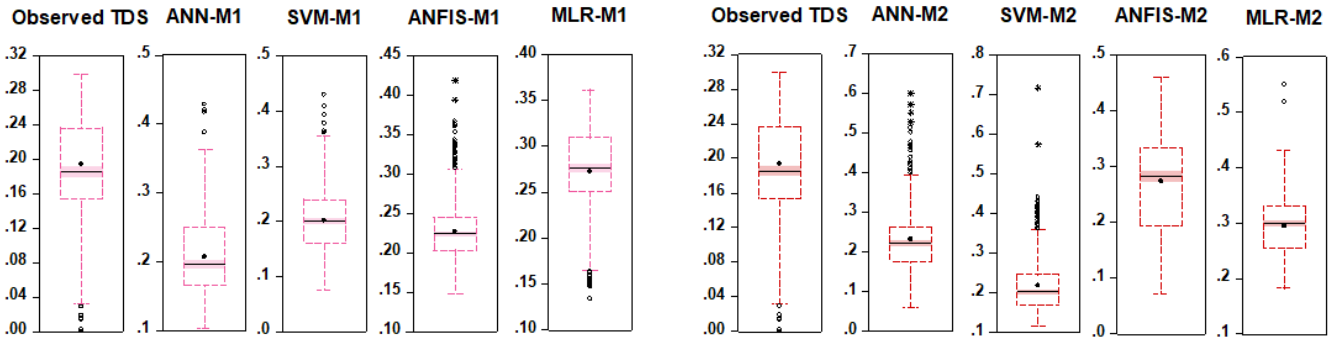

Figure 8. Boxplot for M1 and M2 in the validation phase for TDS.

Table 3. Ensemble results of SA-E, WA-E, and NN-E for TDS.

Ensemble Techniques	MSE	RMSE
SAE-M1	0.0039	0.0623
WAE-M1	0.0065	0.0806
NNE-M1 *	0.0014	0.0379
SAE-M2	0.0087	0.0933
WAE-M2	0.0486	0.2204
NNE-M2	0.0018	0.0426

* Signifies the best model.

From Figure 8, the degree of cumulative distribution in the observed and predicted values can be measured from the different quartiles and whiskers. Furthermore, the best model was selected based on its closeness to the observed values and the mean value. The degree of cumulative distribution between the models that were seen and those that were predicted showed that SVM-M1 was the best. Furthermore, there is a high likelihood of overfitting in data intelligence algorithms where the validation accuracy is greater than the training accuracy. However, the model can be improved by decreasing the variance and bias. In general, the optimal point is reached when the variance and bias are low, and the gap between training and validation accuracy is acceptable. In other words, if the class balance is considered in the training and validation data, the unbalanced data may cause the validation model to be biased towards one class. There are a lot of important points to be considered regarding regularization approaches and dropouts that induce this behavior. Furthermore, boxplot graphs take up less space, which is useful when comparing distributions across multiple datasets or groups. As a result, the variation is determined more by the spread of the data than by the quantitative accuracy of the models. Figure 9 represents the Taylor diagram for R, RMSE, and the standard deviation of TDSt.

According to Figure 9, the ML models were reliable tools for predicting TDSt because they had higher R-values and lower standard deviation values than the measured data. Additionally, the Taylor diagram revealed that SVM-M1 performed better than ANN-M1, ANFIS-M1, and MLR-M1 in terms of the R in the training and validation phases. Taylor diagrams have been used in the graphical presentation of data or information in the areas of climate, hydrological modeling, and water engineering. Interestingly, the Taylor diagram is used to compare the goodness-of-fit of different models [33,34]. In the same way, Figure 10 shows the error plot of the best model's performance indicator.

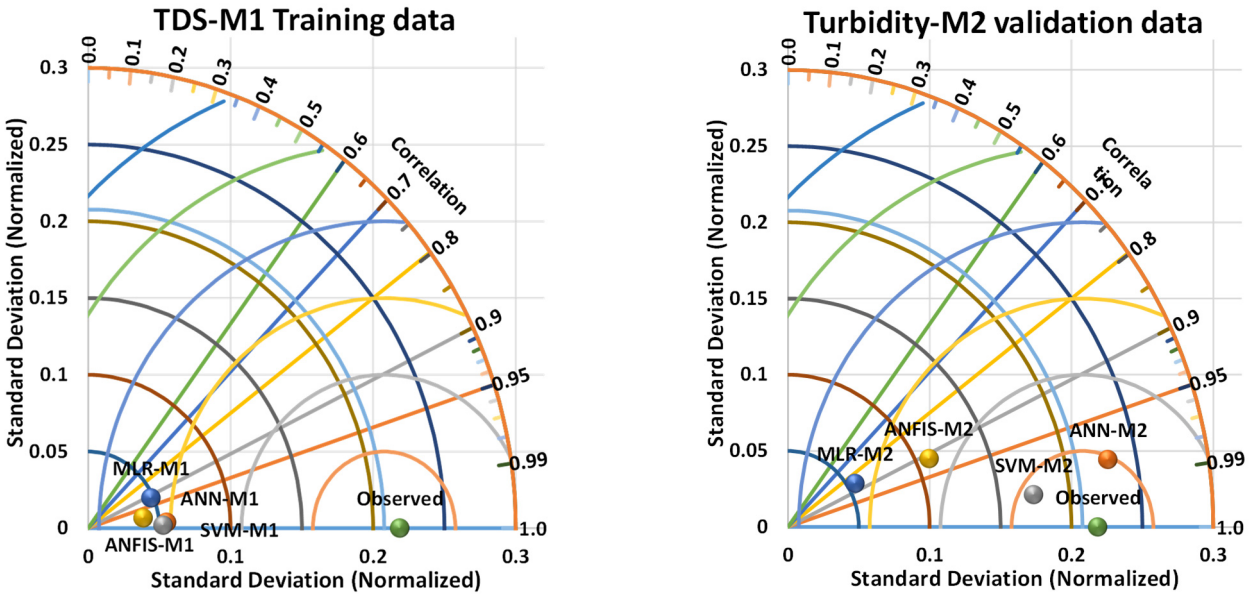

Figure 9. Two-dimensional Taylor diagram between the observed and predicted TDS in the validation phase.

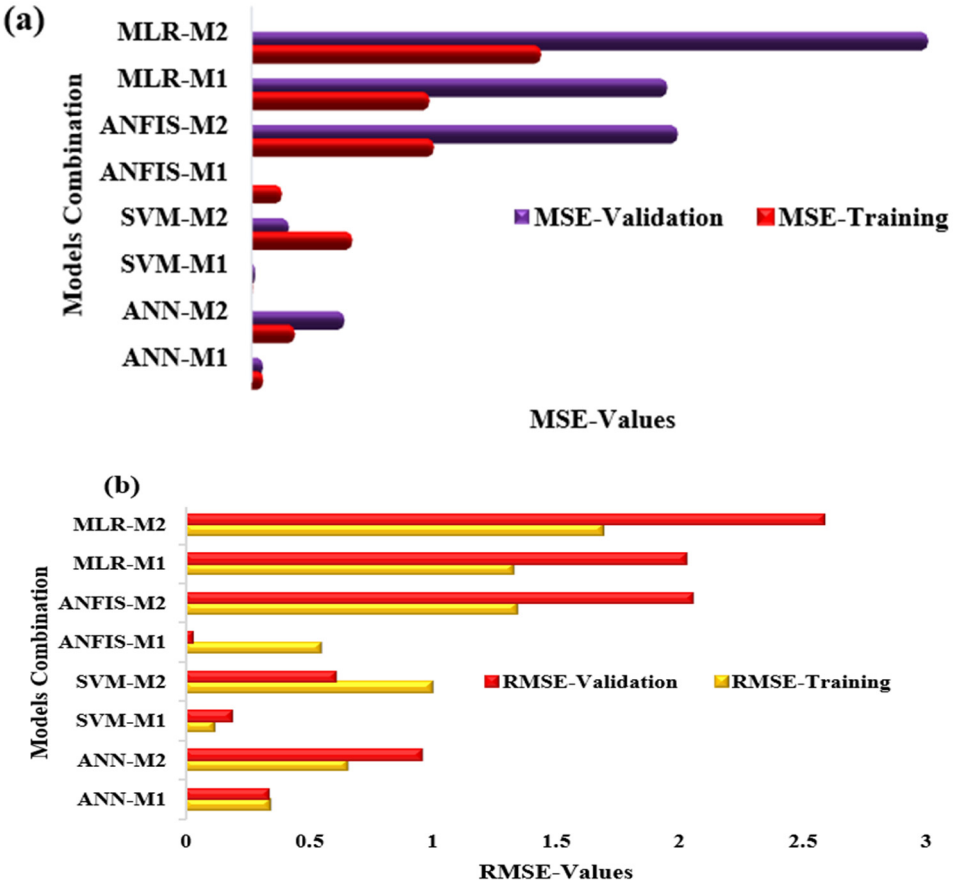

Figure 10. Error plot in both training and validation for (**a**) MSE and (**b**) RMSE.

The lower the MSE and RMSE values, the more reliable the prediction outcomes [34–38]. As a result, the nonlinear black box revealed an accurate and valuable forecasting capability in TDSt that could be seen as a valuable and accurate forecasting tool for the phytoremediation process. According to Figure 10, the outcome indicated that the model's performance

increased with the addition of input variables by approximately 1% in the testing phase, according to the RMSE value. Furthermore, the simulated TDSt recorded a better fit using ML models in the order of SVM > ANFIS > ANN > MLR for M1 and ANN > SVM > ANFIS > MLR for M2. In addition, the numerical comparison of AI-based models regarding RMSE depicts that SVM-M1 increased the prediction accuracy of ANFIS-M1 and ANN-M1 by 3% and 4%, respectively. Figure 11 represents the marginal correlation plot between the observed and predicted models in the validation phase.

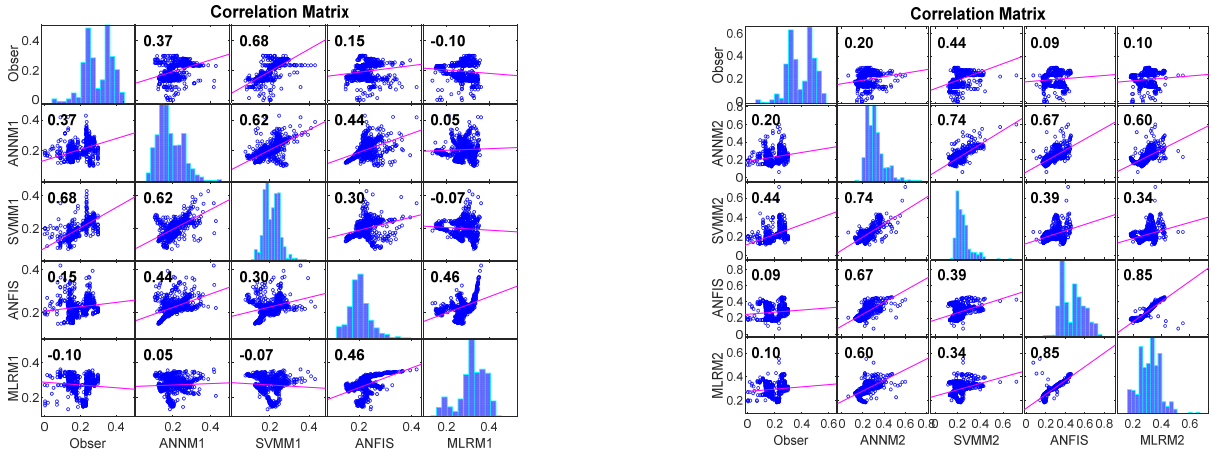

Figure 11. The marginal internal correlation plot between the observed and predicted models in the validation phase.

From Figure 11, the plots demonstrated that the marginal correlation plot corresponded with the simulated SVM and other ML models. It could be observed that the normal distribution was associated with the SVM-M1, which served as the superiority indicator among the models. More effective and accurate forecasting of the TDSt concentration in wastewater treatment and monitoring can allow effective management and protection of natural water bodies. Therefore, the ML techniques employed in this research are suitable for future application in decision-making and management practices. Furthermore, a similar result of R-values (0.993 and 0.8393) for SVM and MLR was reported by Parveen et al. [39], who used an SVM- and MLR-based model to predict Ni (II) ions from tea industry waste by taking into account the independent parameters (pH, flow rate, effluent volume, particle size adsorbent, initial Ni (II) ion, contact time, and bed depth). Additionally, Kumar et al. [40] applied ANN and MLR approaches for forecasting trace metal removal by *Agaricus bisporus* fruiting bodies. When compared to MLR models with ME (>0.96), RMSE (0.441), MNB (0.034), and R2 (0.972), the results obtained from the prediction of Zn, Mn, Cr, Fe, Cd, and Cu by ANN models demonstrated satisfactory performance in terms of model efficiency (ME > 0.99), RMSE (0.075), model normalized bias (MNB 0.009), and R2 (>0.995).

3.2. Error Ensemble Learning Results

Data characteristics such as linearity, correlation, normality, and data size have major effects on model performance [41]. In fact, there is no single model that is optimal for all datasets. It has been discovered that the application of multiple models in an ensemble technique is beneficial for a range of problems. Therefore, error ensemble learning was proposed for the TDSt models to enhance the prediction regarding error. Hence, SAE, WAE, and NNE were employed in the ensemble of ANN, SVM, ANFIS, and MLR to enhance the prediction accuracy of the TDS model. Table 3 represents the outcomes of the SAE, WAE, and NNE analyses.

From Table 3, the MSE and RMSE for both the training and validation demonstrated a noteworthy difference in efficiency performance when compared with the single models. Hence, the effectiveness of the ensemble model is contingent on the accuracy of the individ-

ual models since each model has its disadvantages in the modeling process. Additionally, the results revealed that ensemble methods were found to be superior to the application of single models for the prediction of TDS. Furthermore, Table 4 compares the best single and ensemble models. The output of the best models (SAE-M1, WAE-M1, and NNE-M1) was used as the input for the ensemble methods.

Table 4. Comparison of the best single model and ensemble for TDS.

Techniques	MSE	RMSE	Normalized % Diff MSE	Normalized % Diff RMSE
NNE-M1 *	0.0014	0.0379	3.4165	15.0820
SVM-1	0.0356	0.1887		

* Signifies the best model.

From Table 4, NNE outperformed the other model with a normalized value of 3.4165, based on the RMSE performance criterion. Additionally, MSE (0.0014) and (RMSE) 0.0379 values were obtained for NNE. Therefore, ensemble methods can be used to optimize the prediction accuracy of other water quality parameters, including turbidity and oxidation-reduction potential (ORP). Figures 12 and 13 show the frequency bars and error plots obtained for SAE-M1, WAE-M1, and NNE-M1 for TDS, respectively.

According to Figures 12 and 13, all the ensemble methods produced results that were satisfactory, but NNE was observed to be the most accurate, followed by SAE and WAE. In the validation phase, the three ensemble methods improved the error efficiency performance regarding MSE and RMSE for the prediction of TDS. This resulted in a significant improvement in the TDS error prediction, which had previously been proven to be bad using M2 models. Ensemble techniques are also used to minimize the flaws of individual models, resulting in an improved composite model that is practical, accurate, and reliable when compared to single models.

Additionally, it was observed that SAE slightly outperformed WAE. This was not surprising, since weights are allocated to each parameter depending on their importance. Besides, NNE performed better than SAE and WAE in the training and validation phases due to the robustness in handling nonlinear interactions and the capability to backpropagate the generated error in the training stage until the targeted outcome is obtained.

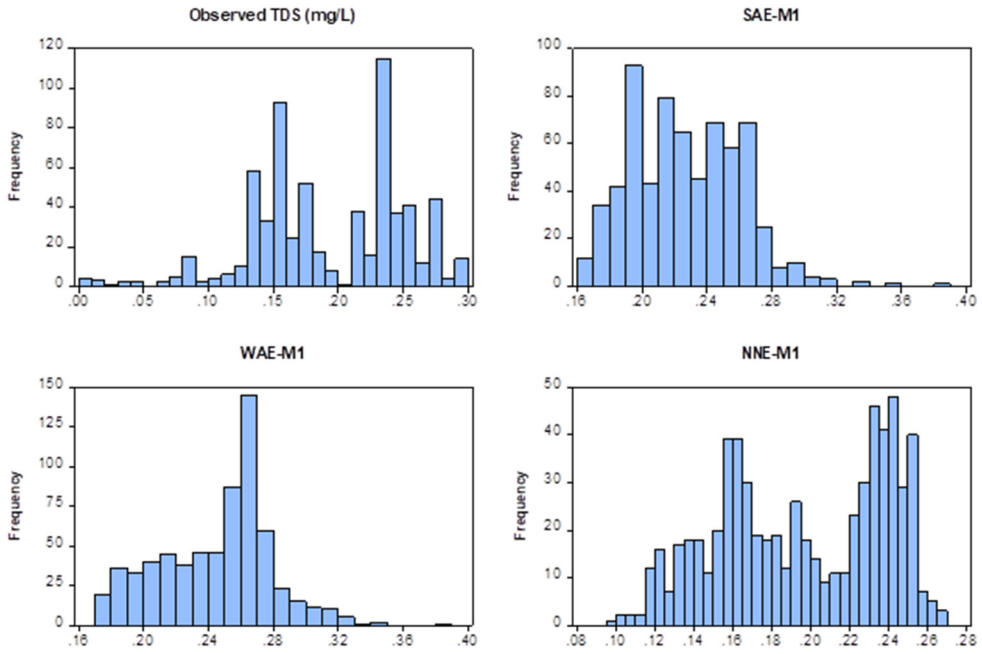

Figure 12. Frequency bar plots (SAE-M1, WAE-M1, and NNE-M1) for TDS.

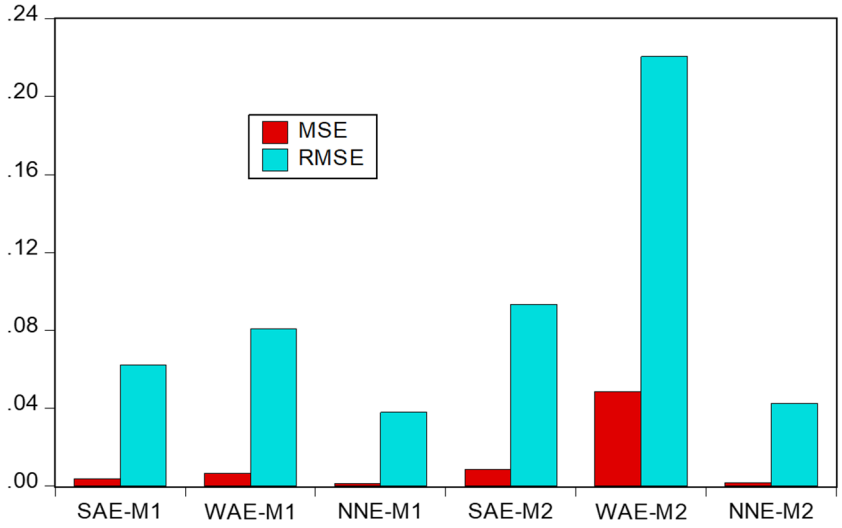

Figure 13. Error bar plots (SAE-M1, WAE-M1, and NNE-M1) for TDS.

4. Conclusions

This research employed the potential of ML tools and multilinear regression in predicting and evaluating the concentration of total dissolved solids (TDS) in the phytoremediation of wastewater using *S. molesta* plants. Subsequently, the prediction accuracy of the single models was improved using three different error ensemble techniques. The outcome showed that the ML models proved their merit with high precision in both the training and testing phases compared to the linear model method. Furthermore, the employed error ensemble techniques significantly outperformed the single models in terms of mean square error (MSE) and root mean square error (RMSE). Therefore, the incorporation of ML techniques in this study provided an ecologically friendly approach to addressing sustainable development goals (SDGs) objectives. Furthermore, this innovation can be integrated into the phytoremediation of wastewater and aquatic plant cultivation for bioenergy generation.

Recommendations for Future Work

Further studies should focus on the storage and updating of the Arduino IoT-generated data in an open-source cloud. This would allow monitoring of the WQ parameters and easy access to the data in remote areas. Applications of other single computational techniques, such as extreme learning machines, extreme gradient boosting, the Elman neural network, the emotional neural network, the kernel support vector machine, the ARIMA model (autoregressive integral moving average), logistic regression, principal component regression, and stepwise regression, should be introduced for comparison. Additionally, applications of nonlinear sensitivity analysis, such as mutual information, genetic algorithms, kernel principal component analysis, and other nonlinear input variable selection techniques, also exist.

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

Funding: This research was funded by the Universiti Tenaga Nasional (UNITEN) BOLD Refresh Fund, Malaysia. Additionally, this research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting project number (PNURSP2023R51), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

Acknowledgments: The authors would like to express their gratitude to Universiti Tenaga Nasional (UNITEN) and Princess Nourah bint Abdulrahman University for supporting the research.

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

References

1. Mekonnen, M.M.; Hoekstra, A.Y. Four billion people facing severe water scarcity. *Sci. Adv.* **2016**, *2*, e1500323. [CrossRef]
2. Breida, M.; Younssi, S.A.; Ouammou, M.; Bouhria, M.; Hafsi, M. Pollution of Water Sources from Agricultural and Industrial Effluents: Special Attention to NO_3^-, Cr(VI), and Cu(II). In *Water Chemistry*; IntechOpen: London, UK, 2019.
3. Mustafa, H.M.; Hayder, G. Cultivation of S. molesta plants for phytoremediation of secondary treated domestic wastewater. *Ain Shams Eng. J.* **2021**, *12*, 2585–2592. [CrossRef]
4. Hayder, G.; Mustafa, H. Cultivation of Aquatic Plants for Biofiltration of Wastewater. *Lett. Appl. NanoBioScience* **2021**, *10*, 1919–1924. [CrossRef]
5. Mustafa, H.M.; Hayder, G. Performance of Salvinia molesta plants in tertiary treatment of domestic wastewater. *Heliyon* **2021**, *7*, e06040. [CrossRef]
6. Hayder, G.; Solihin, M.I.; Mustafa, H.M. Modelling of river flow using particle swarm optimized cascade-forward neural networks: A case study of kelantan river in malaysia. *Appl. Sci.* **2020**, *10*, 8670. [CrossRef]
7. Mustafa, H.M.; Hayder, G. Evaluation of water lettuce, giant salvinia and water hyacinth systems in phytoremediation of domestic wastewater. *H2Open J.* **2021**, *4*, 167–181. [CrossRef]
8. Hull, V.; Parrella, L.; Falcucci, M. Modelling dissolved oxygen dynamics in coastal lagoons. *Ecol. Modell.* **2008**, *211*, 468–480. [CrossRef]
9. Mouatadid, S.; Adamowski, J. Using extreme learning machines for short-term urban water demand forecasting. *Urban Water J.* **2017**, *14*, 630–638. [CrossRef]
10. Bata, M.H.; Carriveau, R.; Ting, D.S.-K. Short-Term Water Demand Forecasting Using Nonlinear Autoregressive Artificial Neural Networks. *J. Water Resour. Plan. Manag.* **2020**, *146*, 04020008. [CrossRef]
11. Abba, S.I.; Elkiran, G. Effluent prediction of chemical oxygen demand from the astewater treatment plant using artificial neural network application. *Procedia Comput. Sci.* **2017**, *120*, 156–163. [CrossRef]
12. Abba, S.I.; Elkiran, G.; Nourani, V. Non-linear ensemble modeling for multi-step ahead prediction of treated cod in wastewater treatment plant. *Adv. Intell. Syst. Comput.* **2020**, *1095*, 683–689. [CrossRef]
13. Abba, S.; Elkiran, G.; Nourani, V. Improving novel extreme learning machine using PCA algorithms for multi-parametric modeling of the municipal wastewater treatment plant. *Desalin. Water Treat.* **2021**, *215*, 414–426. [CrossRef]
14. Hadi, S.J.; Abba, S.I.; Sammen, S.S.; Salih, S.Q.; Al-Ansari, N.; Yaseen, Z.M. Non-Linear Input Variable Selection Approach Integrated with Non-Tuned Data Intelligence Model for Streamflow Pattern Simulation. *IEEE Access* **2019**, *7*, 141533–141548. [CrossRef]
15. Hamed, M.M.; Khalafallah, M.G.; Hassanien, E.A. Prediction of wastewater treatment plant performance using artificial neural networks. *Environ. Model. Softw.* **2004**, *19*, 919–928. [CrossRef]
16. Vo, H.N.P.; Koottatep, T.; Chapagain, S.K.; Panuvatvanich, A.; Polprasert, C.; Nguyen, T.M.H.; Chaiwong, C.; Nguyen, N.L. Removal and monitoring acetaminophen-contaminated hospital wastewater by vertical fl ow constructed wetland and peroxidase enzymes. *J. Environ. Manage.* **2019**, *250*, 109526. [CrossRef] [PubMed]
17. Kumar, V.; Singh, J.; Kumar, P. Heavy metal uptake by water lettuce (*Pistia stratiotes* L.) from paper mill effluent (PME): Experimental and prediction modeling studies. no Goheen 2018. *Environ. Sci. Pollut. Res.* **2019**, *26*, 14400–14413. [CrossRef] [PubMed]
18. Kumar, S.; Deswal, S. Estimation of Phosphorus Reduction from Wastewater by Artificial Neural Estimation of Phosphorus Reduction from Wastewater by Artificial Neural Network, Random Forest and M5P Model Tree Approaches. *Pollution* **2020**, *6*, 417–428. [CrossRef]
19. Zanfei, A.; Menapace, A.; Granata, F.; Gargano, R.; Frisinghelli, M.; Righetti, M. An Ensemble Neural Network Model to Forecast Drinking Water Consumption. *J. Water Resour. Plan. Manag.* **2022**, *148*, 04022014. [CrossRef]
20. Xenochristou, M.; Kapelan, Z. An ensemble stacked model with bias correction for improved water demand forecasting. *Urban Water J.* **2020**, *17*, 212–223. [CrossRef]
21. Khayet, M.; Cojocaru, C.; Essalhi, M. Artificial neural network modeling and response surface methodology of desalination by reverse osmosis. *J. Memb. Sci.* **2011**, *368*, 202–214. [CrossRef]
22. Salami, E.S.; Ehetshami, M.; Karimi-Jashni, A.; Salari, M.; Sheibani, S.N.; Ehteshami, A. A mathematical method and artificial neural network modeling to simulate osmosis membrane's performance. *Model. Earth Syst. Environ.* **2016**, *2*, 1–11. [CrossRef]

23. Abba, S.I.; Pham, Q.B.; Saini, G.; Linh, N.T.T.; Ahmed, A.N.; Mohajane, M.; Khaledian, M.; Abdulkadir, R.A.; Bach, Q.-V. Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index. *Environ. Sci. Pollut. Res.* **2020**, *27*, 41524–41539. [CrossRef] [PubMed]
24. Jeong, K.; Son, M.; Yoon, N.; Park, S.; Shim, J.; Kim, J.; Lim, J.-L.; Cho, K.H. Modeling and evaluating performance of full-scale reverse osmosis system in industrial water treatment plant. *Desalination* **2021**, *518*, 115289. [CrossRef]
25. Castilla-herná, P. Water Quality of a Reservoir and Its Major Tributary Located in East-Central Mexic. *Int. J. Environ. Res. Public Heal.* **2014**, *6*, 6119–6135. [CrossRef] [PubMed]
26. Nourani, V.; Elkiran, G.; Abba, S.I. Wastewater treatment plant performance analysis using artificial intelligence—An ensemble approach. *Water Sci. Technol.* **2018**, *78*, 2064–2076. [CrossRef] [PubMed]
27. Agarap, A.F.; Azcarraga, A.P. k-Winners-Take-All Ensemble Neural Network. In Proceedings of the 28th International Conference (ICONIP 2021), Bali, Indonesia, 8–12 December 2021; pp. 250–261.
28. Ghalehkhondabi, I.; Ardjmand, E.; Young, W.A.; Weckman, G.R. Water demand forecasting: Review of soft computing methods. *Environ. Monit. Assess.* **2017**, *189*, 313. [CrossRef]
29. Priyadharshini, N.R.; Vanishree, R.; Sebasteenav, P.R. Smart water quality management system. In Proceedings of the Global Research and Development Journal for Engineering | National Conference on Advancement in Emerging Technologies (NCAET'18), Chennai, India, 27–28 April 2018; pp. 25–29.
30. Hsu, H.-H.; Hsieh, C.-W.; Lu, M.-D. Hybrid feature selection by combining filters and wrappers. *Expert Syst. Appl.* **2011**, *38*, 8144–8150. [CrossRef]
31. Frazão, X.; Alexandre, L. Weighted Convolutional Neural Network Ensemble. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*; Springer International Publishing: Cham, Swizterland, 2014. [CrossRef]
32. Kitt, F.-P.; Will, P.; Robert, E. Arizona Watershed Stewardship Guide: Water Quality & Monitoring. *Coll. Agric. Life Sci. Univ. Ariz.* **2005**, 18.
33. Taylor, K.E. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.* **2001**, *106*, 7183–7192. [CrossRef]
34. Usman, A.G.; Işik, S.; Abba, S.I. A Novel Multi-model Data-Driven Ensemble Technique for the Prediction of Retention Factor in HPLC Method Development. *Chromatographia* **2020**, *83*, 933–945. [CrossRef]
35. Gaya, M.S.; Abba, S.I.; Abdu, A.M.; Tukur, A.I.; Saleh, M.A.; Esmaili, P.; Wahab, N.A. Estimation of water quality index using artificial intelligence approaches and multi-linear regression. *IAES Int. J. Artif. Intell.* **2020**, *9*, 126–134. [CrossRef]
36. Mubarak, A.; Esmaili, P.; Ameen, Z.; Abdulkadir, R.; Gaya, M.; Ozsoz, M.; Saini, G.; Abba, S. Metro-environmental data approach for the prediction of chemical oxygen demand in new nicosia wastewater treatment plant. *Desalin. Water Treat.* **2021**, *221*, 31–40. [CrossRef]
37. Pham, Q.B.; Gaya, M.; Abba, S.; Abdulkadir, R.; Esmaili, P.; Linh, N.T.T.; Sharma, C.; Malik, A.; Khoi, D.N.; Dung, T.D.; et al. Modeling of bunus regional sewage treatment plant using machine learning approaches. *Desalin. Water Treat.* **2020**, *203*, 80–90. [CrossRef]
38. Abba, S.I.; Gaya, M.S.; Yakubu, M.L.; Zango, M.U.; Abdulkadir, R.A.; Saleh, M.A.; Hamza, A.N.; Abubakar, U.; Tukur, A.I.; Wahab, N.A. Modelling of Uncertain System: A comparison study of Linear and Non-Linear Approaches. In Proceedings of the 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS 2019—Proc.), Selangor, Malaysia, 29 June 2019; pp. 1–6. [CrossRef]
39. Parveen, N.; Zaidi, S.; Danish, M. Support vector regression (SVR)-based adsorption model for Ni (II) ions removal. *Groundw. Sustain. Dev.* **2019**, *9*, 100232. [CrossRef]
40. Kumar, V.; Kumar, P.; Singh, J.; Kumar, P. Use of sugar mill wastewater for Agaricus bisporus cultivation: Prediction models for trace metal uptake and health risk assessment. *Environ. Sci. Pollut. Res.* **2021**, *28*, 26923–26934. [CrossRef]
41. Yassin, M.A.; Tawabini, B.; Al-Shaibani, A.; Adetoro, J.A.; Benaafi, M.; Al-Areeq, A.M.; Usman, A.G.; Abba, S.I. Geochemical and Spatial Distribution of Topsoil HMs Coupled with Modeling of Cr Using Chemometrics Intelligent Techniques: Case Study from Dammam Area, Saudi Arabia. *Molecules* **2022**, *27*, 4220. [CrossRef]