Journal of Water Process Engineering 56 (2023) 104247

Contents lists available at ScienceDirect

Journal of Water Process Engineering

journal homepage: www.elsevier.com/locate/jwpe

Exploring ANFIS application based on actual data from wastewater treatment plant for predicting effluent removal quality of selected major pollutants

Liang Qiao [a,b,c], Pei Yang [a], Qi Leng [a], Liujie Xu [a], Yanxin Bi [a], Jinzhen Xu [a], Zhe Wang [a], Jianye Liu [d], Wanxin Yin [c,e], Luyan Zhang [a,c,f], Feihong Wang [a,c], Ye Yuan [a,c,*], Tianming Chen [a,c,*], Cheng Ding [a,c]

[a] School of Environmental Science & Engineering, Yancheng Institute of Technology, Yancheng, Jiangsu Province, China
[b] Jiangsu Zhen'gen Energy Conservation and Environmental Protection Technology Co., Ltd, Yancheng, Jiangsu Province, China
[c] Jiangsu Province Engineering Research Center of Intelligent Environmental Protection Equipment, Yancheng Institute of Technology, Yancheng, China
[d] Beijing Origin Water Technology Co., Ltd., Beijing, China
[e] Liaoning University, College of the Environment, Shenyang, China
[f] Jiangsu E-INTEL Environmental Technology Co., Ltd, Yancheng, Jiangsu Province, China

ARTICLE INFO

Keywords:
ANFIS
Indirect prediction
Orthogonal experiment
Wastewater treatment plant
ANFIS predictive logic

ABSTRACT

Efficient simultaneous multi-pollutant removal in wastewater treatment plants (WWTPs) is of critical importance in meeting increasingly stringent discharge standards. This study investigated the effectiveness of the Adaptive Neuro-Fuzzy Inference System (ANFIS) model for predicting the removal of major pollutants in a WWTP. The Parameters were screened using Principal Component Analysis (PCA), and Orthogonal Experiments (OE) were conducted to enhance ANFIS accuracy. The actual removal values and predicted values of major pollutants such as COD, BOD, NH_3, and SS were compared and analyzed. The study obtained satisfactory linear results within a 95 % confidence interval, with some R^2 values exceeding 0.950. Additionally, there was revealed a lack of one-to-one correspondence between predicted and actual pollutant values, for the same input value yielding different output values. This study highlighted the potential of ANFIS for pollutant removal prediction in WWTPs, yet further investigations were required to refine the model's logic and enhance accuracy for practical applications.

1. Introduction

With the advancement of sewage treatment technology and increasingly stringent sewage discharge standards, the need for highly efficient simultaneous multi-pollutant removal process in wastewater treatment plants (WWTPS) has become imperative [1]. Mere compliance with the discharge requirements is no longer sufficient [2–4]. The influent concentration of conventional pollutants, such as BOD, COD, SS, pH, and NH_4^+-N, varies, and the presence of toxic substances can significantly impact the operation of WWTPs [5–8]. Various methods have been employed for water quality and quantity analysis and prediction, encompassing aeration control of dissolved oxygen (DO), coagulant addition, carbon source addition, and other operational aspects of WWTPs [5]. Key parameters such as BOD/COD (B/C) ratio, COD/NH_4^+-N (C/N) ratio, and pH play critical roles in assessing WWTP performance and making timely adjustments [9]. However, the complex non-linear relationship between these parameters make it challenging to provide timely feedback and direct regulation of WWTP operations based on assumptions and conditional control [6,7,10,11].

To address this challenge and improve WWTP operation parameter analysis, intelligent models are utilized to enhance data processing and decision-making within an acceptable range with directness and timeliness [11,12]. The development of artificial intelligence has led to innovative approaches in sewage treatment [13]. Various intelligent models, such as Adaptive Neuro-Fuzzy Inference System (ANFIS), Artificial neural networks (ANNs), support vector machines (SVMs), fuzzy neural networks (FNN), and genetic algorithms (GA) enable independent analysis, prediction, variable optimization, parameter analysis, and output adjustment based on input data, reducing human error, and increasing productivity, shown in Table 1 [3,7,10,14]. While the

* Corresponding authors at: School of Environmental Science and Engineering, Yancheng Institute of Technology, Yancheng 224051, China.
E-mail addresses: yuanye_19840915@163.com (Y. Yuan), ycchentm@163.com (T. Chen).

https://doi.org/10.1016/j.jwpe.2023.104247
Received 29 May 2023; Received in revised form 27 July 2023; Accepted 31 August 2023
Available online 15 September 2023

Table 1
Advantages and disadvantages of machine learning models on WWTPs.

Model	Advantages	Disadvantages	Application in WWTPs	Part of the reference
ANFIS	Combines the benefits of fuzzy logic and neural nets; Self-learn and fine tune; Handle non-linear relationships and uncertainties	Complex model to design and optimize; Requires sufficient and representative training data; Decrease interpretability when ANFIS become complex	Water quality Prediction; Process optimization; Optimize aeration control, and improve treatment process efficiency.	[6,7,18,19]
ANN	Powerful for pattern recognition; Fast predictions; Handle both structured and unstructured data.	Require a large amount of labeled training data; Prone to overfitting; Computationally intensive of training process.	Water quality Prediction; Suitable for various wastewater treatment processes, especially when large datasets are available.	[6,20,21]
SVM	Effective in high-dimensional spaces; Handle non-linear decision boundaries; Strong theoretical foundations.	Sensitive to parameter tuning; Difficult to handle datasets with noisy or overlapping.	Water quality Prediction; Suitable for classification tasks in wastewater quality monitoring and prediction.	[1,10,21,22]
FNN	Capture complex relationships; Well-suited for applications with uncertain and imprecise data.	Face challenges in handling large and high-dimensional datasets; Time-consuming; Expert knowledge in selecting fuzzy membership functions and rule bases.	Real-time process control; Fault detection and anomaly detection; Water quality prediction.	[3]
Decision Trees	Easy to interpret and visualize; Handle both numerical and categorical data; Classification and regression tasks.	Prone to over fitting, especially when the tree becomes too deep and complex; Instability; Limited in handling complex relationships	Useful for initial wastewater quality screening and decision-making	[1,17]
GA	Effective global search capability; Inherently parallelizable.	Slow convergence speed of genetic algorithms.	Optimal control strategies; Process optimization.	[3,17,23]

artificial neural network was best suited for non-compliant flexible programs and black-box scenarios where determining the effect of input parameters on results is challenging. Each model has unique strengths and weaknesses, and their effectiveness depends on the specific problem at hand [3,9,12]. In contrast, ANFIS amalgamates the benefits of fuzzy logic and neural networks, allowing self-learning and parameter fine-tuning with large training data, while maintaining the reliability and simplicity of fuzzy reasoning [9]. Furthermore, before utilizing ANFIS, the applicability of data can be further strengthened by statistical and multivariate data analysis methods, such as multiple linear regression (MLR), response surface methodology (RSM), principal component analysis (PCA), and other methods [5,6,14–17]. As a result, screening major pollutants and clarifying logical relationships based on ANFIS is advantageous in improving prediction accuracy.

In ANFIS, the feedforward neural network serves to learn the mapping relationships between inputs and outputs, while the fuzzy inference system interprets and regularizes these relationships. Pollutants of WWTPs can be categorized into easy-to-treat pollutants and complex pollutants based on the length of the treatment process, with the former being predicted a white box model and the latter being predicting using a gray box model [9]. There exist fuzzy areas in the sequence, action relationship, and influence of each pollutant treatment, which cannot be directly applied to the existing model. Hence, training the model according to the data type and index composition is essential [24]. ANFIS serves as a model between the gray-box and black-box models, providing a certain degree of interpretability and comprehensibility, while also adapting to complex nonlinear systems and has high predictive power [12].

ANFIS has been successfully applied and validated in various aspects, including the stable operation of aerobic granular sludge reactors, sludge dewatering and polymer dosage optimization [25,26], anaerobic digestion methane production [27], heavy metal adsorption [19,28], sewage treatment plant influent characteristics prediction [15,24], post-epidemic era SARS-Cov-2 virus in wastewater prediction analysis [29], principal component analysis, fuzzy reasoning systems [6], drug prediction analysis [16], and others. Moreover, ANFIS can manage not only single pollutants but also multiple pollutants, considering their interdependencies, types, and concentrations during the treatment process [6,11,12]. As the treatment of pollutants forms an internal relationship through the treatment process, it is affected by type and concentration, as well as interactions among microorganisms, treatment sequences, and chemicals such as coagulants, including a certain degree of ambiguity [4,30,31]. ANFIS assists in describing changes in sewage parameters and operation through soft prediction [23,32]. Based on this information, ANFIS can facilitate the control of WWTPS based on influent and effluent conditions, encompassing the prediction of relevant pollutant removal (effluent pollutant concentration) and the necessary chemical additions during the treatment process [6].

Nevertheless, ANFIS still faces some limitations in WWTPs operations. Firstly, Designing an ANFIS model with appropriate fuzzy rules and membership functions can be very complex and time-consuming, especially when dealing with large and diverse datasets [3]. Secondly, Data scarcity. Poses a challenge, as ANFIS requires abundant training data to achieve robust and accurate predictions, which may be challenging to obtain in some cases [9,22,33]. Thirdly, ANFIS involves tuning various parameters, such as learning rate and number of rules, which significantly impact performance, necessitating extensive experimentation to find optimal parameter settings [11]. Lastly, the complexity of wastewater treatment processes in WWTPs, involving numerous interacting variables, presents a challenge in accurately capturing all these relationships by ANFIS [3,23,34]. These factors introduce inherent uncertainties that may impact predictions in WWTPs.

The aim of this study is to establish and adjust the ANFIS model, directly predicting the actual data of the WWTPs and comparing the gap between the pollutant removal values and the actual values. Specific objectives include: 1) establishing the ANFIS model and analyze the relationship between parameters and treatment processes; 2) training ANFIS and adjust the prediction accuracy, and 3) indirectly predicting the effluent pollutant concentration by predicting the pollutant removal amount of WWTPs. In addition, this study discussed the limitations of the ANFIS model and introduced restrictive statements to enhance prediction accuracy. These modifications were of practical significance for the application of ANFIS in WWTPs.

2. Methods and materials

2.1. ANFIS fuzzy rules and model structure

The fuzzy rule of ANFIS of membership function Sugeno model is:

$$If\ x\ is\ A\ and\ y\ is\ B, then\ z = f(x,y) \tag{1}$$

where A and B represent the fuzzy degree of the system, and $z = f(x,y)$ represents the conclusion based on the early fuzzy numbers of A and B. $f(x,y)$ is a multinomial expression of x and y. The Sugeno model is established when $f(x,y)$ is a first-order polynomial of x and y, and its fuzzy rules are as follows [9,28,35]:

Rule 1

$$If\ x\ is\ A_1\ and\ y\ is\ B_1, then\ f_1 = p_1x + q_1y + r_1 \tag{2}$$

Rule 2

$$If\ x\ is\ A_2\ and\ y\ is\ B_2, then\ f_2 = p_2x + q_2y + r_2 \tag{3}$$

ANFIS enforces the black box problem of neural networks by making them transparent without knowing the exact mathematical model. It does this by blurring the input, converting some of the blurring rules into a map, and then making them transparent (Fig. 1) [24,36].

The first layer acts as the input layer, and the input value is the data of sewage quality parameters in the treatment process, which is expressed as $X = [X_1, X_2, X_3, \ldots, X_n]$, where n represents the number of input parameters, and $n \in N^*$, $n \geq 1$. X_1, X_2, X_3, …, X_n represent the sewage quality parameters of influents and effluents in the treatment process, namely, COD, SS, TP, TN, NH_3-N, and pH.

$$f_1(i) = X = [X_1, X_2, X_3 \cdots X_i], i = 1, 2, 3 \cdots \tag{4}$$

The second layer is the fuzzification layer. Each node of this layer corresponds to a variable value, and its function is to convert the input deterministic quantity into a fuzzy vector. The input variable is converted into the corresponding membership degree through the membership function on the fuzzy subset. The Gaussian function was selected as the membership function, and it corresponds to each input component in different fuzzy language values as:

$$f_2(i,j) = exp\left(-\frac{(f_1(i) - C_{ij})^2}{(\sigma_{ij})^2}\right) \tag{5}$$

where i = 1, 2, 3…, j = 1, 2, 3…, C_{ij} and σ_{ij} are the width and center of the membership function of the i-th input variable of the j-th fuzzy set, respectively.

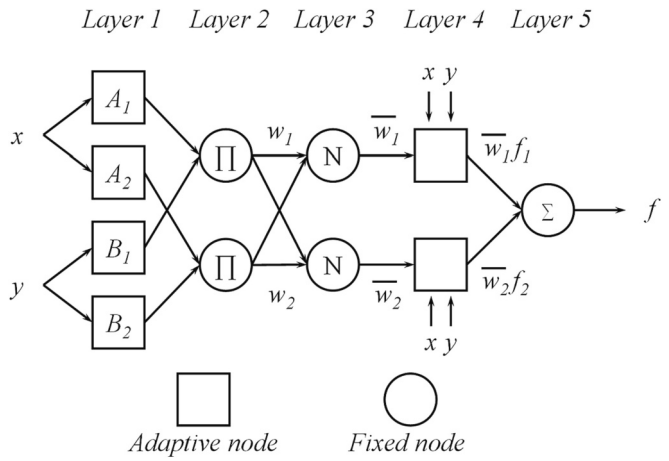

Fig. 1. ANFIS structure.

The third layer is the fuzzy rule layer. Each node in this layer represents a fuzzy rule. The matching of the fuzzy rules is completed through a connection with the fuzzification layer; the fuzzy operation between each node is realized. The output of each node is the product of all of the signals.

$$f_3(j) = \prod_{m=1}^{N} f_2(i,j) \tag{6}$$

$$N = \prod_{n=1}^{n} N_i \tag{7}$$

where N_i is the number of fuzzy segmentations of the i-th input.

The fourth layer realizes normalization calculation. The formula is defined as follows:

$$f_4(j) = \frac{f_3(j)}{\sum_{i=1}^{i} f_3(i)} \tag{8}$$

The fifth layer is the output layer. The function of this layer is to convert the blurred amount of the output into a clear amount. The formula is as follows:

$$y = \sum_{r=1}^{m_i} w_r f_4(j), r = 1, 2, \cdots, m_i \tag{9}$$

where w_r is the connection weight, $f_4(j)$ is the rule applicability, and r is an integer between 1 and m_i.

After attempting several membership functions, gaussmf, gbellmf trimf, and pimf were chosen as the membership functions of this study, which obtained relatively better fitting results.

2.2. The data source of sewage treatment plant

The operating data used in this study was collected from a sewage treatment plant in Shenzhen, Guangzhou Province, China. The plant specializes in treating domestic sewage and employs a primary treatment process consisting of biological treatment by AAO-MBR with a total designed volume of 50,000 m^3/d. The removal of contaminants was achieved through various means, including the use of composite carbon sources, composite flocculants, film washing chemicals, disinfectants, and aeration in the biochemical system. The chemicals and the discharge of the sludge were adjusted based on the needs of the sewage and the sludge concentration of the biochemical tank.

The dataset used in this study consisted of operational data collected from October 1st 2021, to May 31st 2022. The relevant parameters measured on a daily basis included Influent and Effluent (Inf and Eff) (m^3/d, daily), T (°C, daily), Inf-COD and Eff-COD (mg/L, daily), Inf-SS and Eff-SS (mg/L, daily), Inf-TP and Eff-TP (mg/L, daily), Inf-TN and Eff-TN (mg/L, daily), Inf-NH_3-N and Eff-NH_3-N (mg/L, daily), and Inf-pH and Eff-pH (\, daily), shown in Fig. 2 [3,9,24,34,37]. The maximum, minimum and average value of these parameters were presented in Table 2. The input values of the above Inf-parameters were used to predict the correlated Eff-parameters, serving as the output data in the ANFIS model [6]. The ANFIS analysis was performed using the data for the year 2022, while the remaining data were used to validate the model's output efficiency.

2.3. Parameter selection and regulation of ANFIS on WWTP

The training set size (0.5, 0.6, 0.7, and 0.8), membership functions (2, 3, 4, and 5), membership function type (gaussmf, gbellmf, trimf, and pimf), number of training times (10, 100, 1000, and 10,000), and training steps (0.05, 0.01, 0.005, and 0.001) were selected as an orthogonal experiment (OE, L16.4.5) with a total of five factors, each with four independent levels. Due to the small volume, the data was not

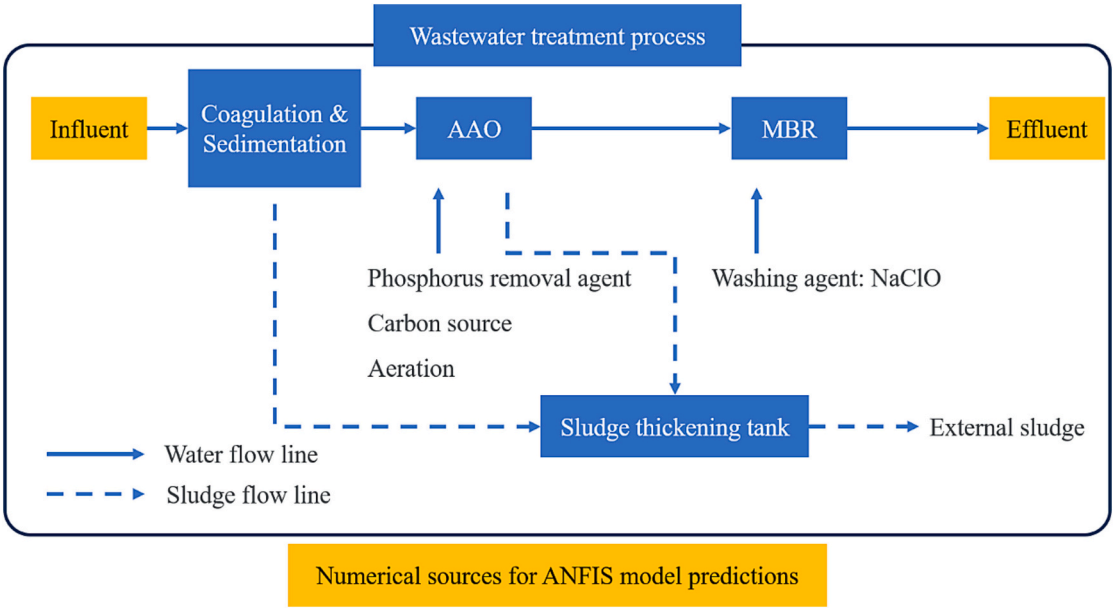

Fig. 2. Sewage treatment process and source of data.

Table 2
The maximum, minimum and average value of parameters.

	Inf	Eff	T	Inf-COD	Eff-COD	Inf-SS	Eff-SS	Inf-TP	Eff-TP	Inf-TN	Eff-TN	Inf-NH_3-N	Eff-NH_3-N	Inf-pH	Eff-pH
max	59,116	57,571	30	399	12	885	9	9.65	0.256	61.9	9.41	39.5	0.628	8.04	7.83
min	29,129	26,737	13	70	3	80	1	0.88	0.027	9.55	0.22	4.9	0.043	6.87	6.89
ave	49,361	47,338	23	264	7	249	2	4.17	0.121	41.17	6.87	29.7	0.288	7.36	7.39

divided into training (60 %), validation (20 %) and testing (20 %), or training (70 %) and independent testing (30 %) according to previous work [38].

The boxplot and normal curve of the data were drawn through the data normalized formula:

$$R_p = \frac{C_{pv} - C_{min}}{C_{max} - C_{min}} \tag{10}$$

where C_{pv} indicates the value of a parameter, C_A indicates the average value of this parameter, and C_{max} and C_{min} represent the maximum and the minimum of this parameter, respectively. R_p represents the range of this parameter [10].

Principal component analysis (PCA), construction of ANFIS, sewage quality, and quantity parameters can be found in the Supplement materials.

3. Results and discussion

3.1. Data analysis and major data processing

After performing a normalized analysis on the data range, no further analysis was conducted on pH. The effluent of COD and SS were shown as parallel data due to the accuracy of the effluent data. The design limits of the effluent of Eff-SS, Eff-NH_3-N, and Eff-TP were very small, but the range was wide after division. The relationship between the parameters was indicated within the 95 % confidence interval, and the distribution relationship of data points also determined the discreteness of the two parameters, as depicted in Fig. 3. It is worth noting that the influent index was not entirely meeting the design requirements, with ratios of the maximum values of COD, SS, TP, TN, and NH_3-N to the design requirements being 1.27, 2.86, 1.61, 1.38, and 0.99, respectively. Nevertheless, the average effluent indices were only 25.4 %, 35.1 %, 36.6 %, 67.0 % and 21.5 % lower than the design requirements. This high treatment efficiency is attributed to the addition of composite carbon source and phosphorus removal agent. However, in this WWTP, the composite phosphorus removal agent was not added daily, and there is no explicit linear relationship between its addition rate and the amount of Inf-TP. The relationship between the parameters is illustrated in the Supplementary materials (Fig. S1) within the 95 % confidence interval, and the distribution relationship of the data points determined the discreteness of the two parameters.

The number of principal components was determined by calculating the cumulative contribution rate, as shown in Table 3 and Fig. 4. The cumulative contribution rate of Influent, T, Inf-COD, and Inf-SS was well over 80 %, indicating that these four principal components contained most of the information of all indicators and were not correlated. It is worth mentioning that Inf-pH, T, Influent, Inf-COD, Inf-SS, Inf-TP, Inf-TN, and Inf-NH_3-N were vectors with different directions. The characteristic value of Influent, T, and COD was >1, indicating that the principal component explained more variation in the data than a single variable alone. To perform the analysis, it was assumed that the data needed to be classified into two categories: data indicators and hydraulic conditions. However, pH was significantly different between these two categories, suggesting that pH should be considered an independent parameter. The chosen influent wastewater indices, including pH, had significant advantages in predicting the effluent indices, whereas T and influent volume did not participate as input layer parameters in the ANFIS model [3,21].

3.2. Debugging of ANFIS parameters by OE

To perform the ANFIS analysis, an OE was set up as L16.4.4, where the four factors were the membership function type (gaussmf, gbellmf, trimf, and pimf), the number of training times (10, 100, 1000, and 10,000), the training step (0.05, 0.01, 0.005, and 0.001), and the

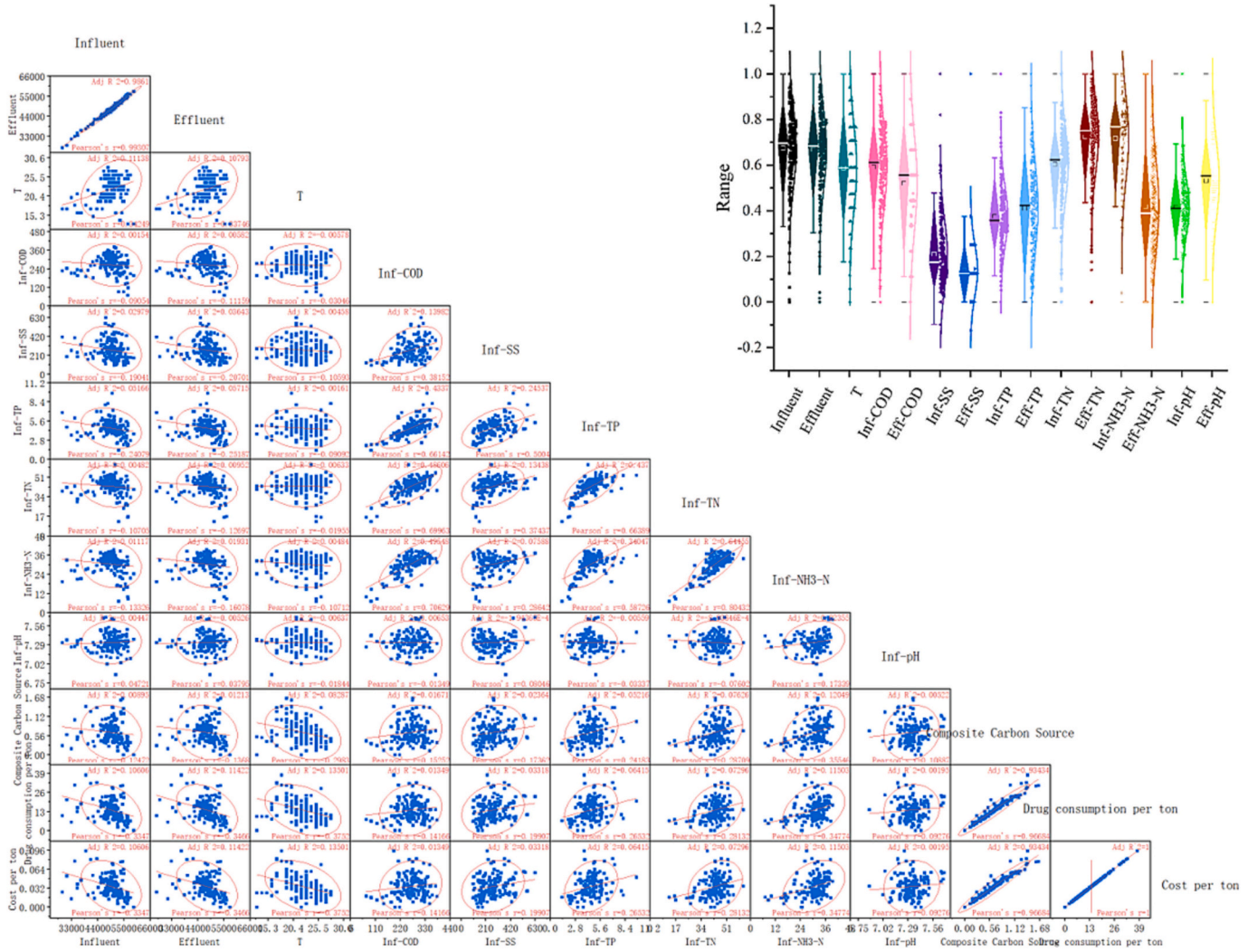

Fig. 3. The data index boxplot and the normal curve.

Table 3
PCA composition and parameters.

Principal component name	Eigenvalue	Percentage of variance (%)	Cumulative (%)	PC1 (42.2 %)	PC2 (16.5 %)	PC3 (13.5 %)
1, Influent	3.374	42.18	42.18	−1.163	3.907	−0.677
2, T	1.320	16.50	58.68	−0.733	3.923	0.424
3, Inf-COD	1.077	13.47	72.15	3.689	0.863	−0.455
4, Inf-SS	0.816	10.20	82.35	2.505	−0.642	1.050
5, Inf-TP	0.621	7.76	90.11	3.670	−0.097	−0.096
6, Inf-TN	0.359	4.49	94.6	3.826	0.886	−0.627
7, Inf-NH_3-N	0.295	3.70	98.3	3.703	0.757	0.348
8, Inf-pH	0.136	1.70	100	0.010	0.647	3.607

training set size (0.5, 0.6, 0.7, and 0.8). After comprehensive consideration of the running time and results of the model, it was found that the number of membership functions significantly affected the prediction results. The effect appeared unstable when the number of membership functions was 2, and long running times were observed when the number of membership functions was 4 and 5. As a result, the number of membership functions was fixed at 3 to ensure model stability. It is worth noting that in addition to the ANFIS prediction algorithm, the distribution of the data set and the relationship between various parameters were prominent factors that affected the prediction results [4]. Inf-COD, Inf-SS, Inf-TN, Inf-TP, and Inf-NH_3-N were all predicted and analyzed using OE. The optimal values of each factor were screened and compared based on R^2 and root mean square error (RMSE) using range analysis or multivariate analysis of variance (ANOVA) to further analyze the WWTPs parameters.

The Effluent-value for each parameter were not directly predicted, including Eff-COD, Eff-SS, Eff-TP, Eff-TN, Eff-NH_3-N, and Eff-pH. This is because the Eff-value of each parameter was small, and the Inf-values were relatively large. When a small Eff-value is predicted by a series of large input parameters, the mean absolute percentage error (MAPE) and RMSE become very large, and the R^2 value becomes relatively small, making it challenging to perform follow-up analysis, as shown in Table 4. Moreover, the difference between the treatment capacity and the effluent demand of each parameter was also substantial. For example, the Inf-COD was several hundred, while the Eff-COD was <30; similarly, the Inf-TP was <6, while the Eff-TP was <0.6. The magnitude

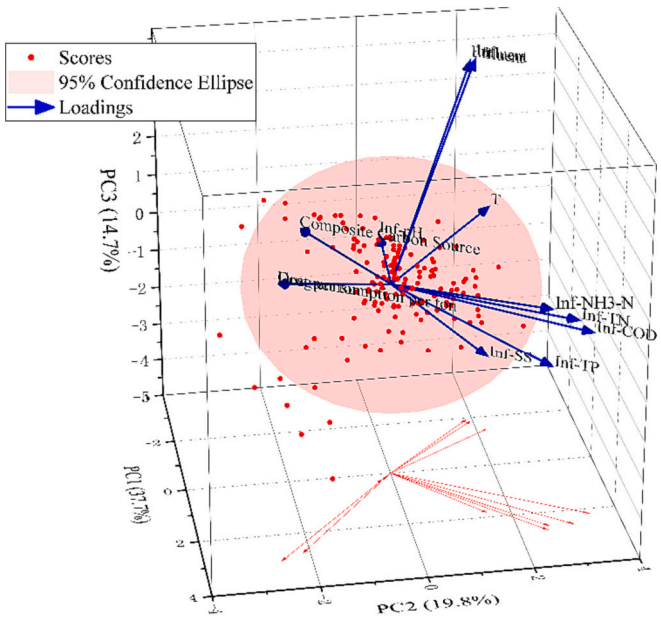

Fig. 4. Parameter distribution of PCA.

difference between these values was >100-fold, and there was no obvious magnitude relationship between Inf-pH and Eff-pH. The Effluent index values cannot be predicted by the Influent values alone, and the predicted values have no reference relationship. Therefore, it was decided to remove these parameters as output for prediction.

After considering the orthogonality and Effluent indicators, gaussmf and gbellmf were found to give better results, while pimf produced larger error values. Trimf demonstrated better predictive effect on smaller data and could serve as a reference. Therefore, a more comprehensive selection of ANFIS parameters was performed by considering multi-factor and area analysis. The analysis procedure, multivariate analysis of variance data, and the range analysis figures are provided in the Supplementary materials. Finally, the ANFIS parameters were confirmed based on the prediction of substrate processing capacity, with a training set of 0.6, membership function type of trimf, training time of 10,000, and training step of 0.005.

3.3. ANFIS prediction results

In this study, the training set was selected as 60 % of the total dataset based on previous research [18,38,39]. The linear equation between the predicted values and the pollutant removal prediction is demonstrated in Fig. 5 and Table 5. Fig. 5 displays the equation between the predicted data and the actual data. Verification of the OE is available in the Supplementary materials (Figs. S2–S5).

Upon modifying the ANFIS output to the pollutant removal, the R^2 values of COD, SS, NH_3-N, and TP remained above 0.8 with better linearity, except for TN. There was a high degree of data concentration observed between the predicted value of COD, TN, NH_3-N, and TP, but a small amount of data remained outside the 95 % prediction band, while the distribution of SS was relatively average. This suggested that relatively stable prediction results can be obtained within the existing input data range [15,39]. However, the discrete degree of larger data was higher than that of the smaller data, indicating the need for ANFIS adjustment or numerical range adjustment of input parameters [18,39].

The predicted value and actual removal value for COD, NH_3-N and TP tended to be more concentrated in a range, with COD concentrated in the 200–300 range, NH_3 concentrated in the 28–36 range and TP concentrated in the 3.5–5 range, which was acceptable by ANFIS [9]. Other data were still partially within the 95 % confidence band for COD, NH_3-N and TP, and these data at the sides or endpoint were also highly informative [11]. It is only the small amount of the data that makes this part appear to be loose [3,9]. While, TN and SS showed a different pattern. For TN, the data points were discrete. Although a part of the data was concentrated between 30 and 45, it tended to form more of a trumpet shape towards the left with an overall smaller amount of data at the 95 % confidence band. For SS was more different than TN on the majority of the data nestled within the 95 % confidence band with better linearity [10]. It is also worth noting that some of the discrete points (not in the 95 % prediction band) are relatively more represented by larger or smaller data. For COD, SS and NH_3, all the discrete points have the largest deviations from the 95 % prediction band with different R^2 values. COD, SS, NH_3 that were easier to deal with, the prediction results were often linear [11]. This portion of the data were treated as a marginal data with large values. Since most of the predicted data is less than this value, ANFIS will predict it as a borderline value and the number of predictions may be less and the accuracy is poorer [6]. As for TN and TP, there are intermediate segments of data distributed outside the 95 % prediction band, which are likely to be some extreme data. In other words, under the condition of the same TN or TP removal, this prediction will be influenced by the extreme values which will amplify the

Table 4
Comparison of Eff-COD prediction and the removal of COD prediction by OE.

OE	Training set	Membership function type	Training times	Training step	Eff-COD prediction[a]			The removal of COD prediction[a]		
					R^2	RMSE	MAPE	R^2	RMSE	MAPE
1	0.5	gauss	10	0.05	0.060	14.191	126.73 %	0.752	36.345	9.52 %
2	0.5	gbell	100	0.01	0.009	21.956	165.94 %	0.685	44.309	11.03 %
3	0.5	tri	1000	0.005	0.107	5.764	56.14 %	0.804	31.530	8.06 %
4	0.5	pi	10,000	0.001	0.018	238.559	1681.61 %	0.035	239.059	55.04 %
5	0.6	gauss	100	0.005	0.067	11.969	114.31 %	0.791	31.515	9.64 %
6	0.6	gbell	10	0.001	0.042	14.474	131.65 %	0.810	30.662	9.48 %
7	0.6	tri	10,000	0.05	0.025	9.749	86.45 %	0.776	32.442	9.29 %
8	0.6	pi	1000	0.01	0.056	427.028	2560.07 %	0.030	471.483	83.88 %
9	0.7	gauss	1000	0.001	0.300	11.274	102.38 %	0.771	36.262	11.02 %
10	0.7	gbell	10,000	0.005	0.311	25.057	167.02 %	0.666	45.735	11.88 %
11	0.7	tri	10	0.01	0.305	5.445	66.09 %	0.780	35.053	10.23 %
12	0.7	pi	100	0.05	0.086	231.081	1671.39 %	0.031	145.245	28.49 %
13	0.8	gauss	10,000	0.01	0.079	10.717	130.56 %	0.678	36.177	12.72 %
14	0.8	gbell	1000	0.05	0.146	18.803	225.18 %	0.671	36.572	14.68 %
15	0.8	tri	100	0.005	0.047	5.019	65.54 %	0.587	41.364	13.04 %
16	0.8	pi	10	0.001	0.035	748.003	4307.73 %	0.005	584.684	134.05 %

[a] The Eff-COD prediction and the removal of COD prediction by OE were chosen the same parameters and were taken as an example to clarify the removal pollutants as ANFIS predictions.

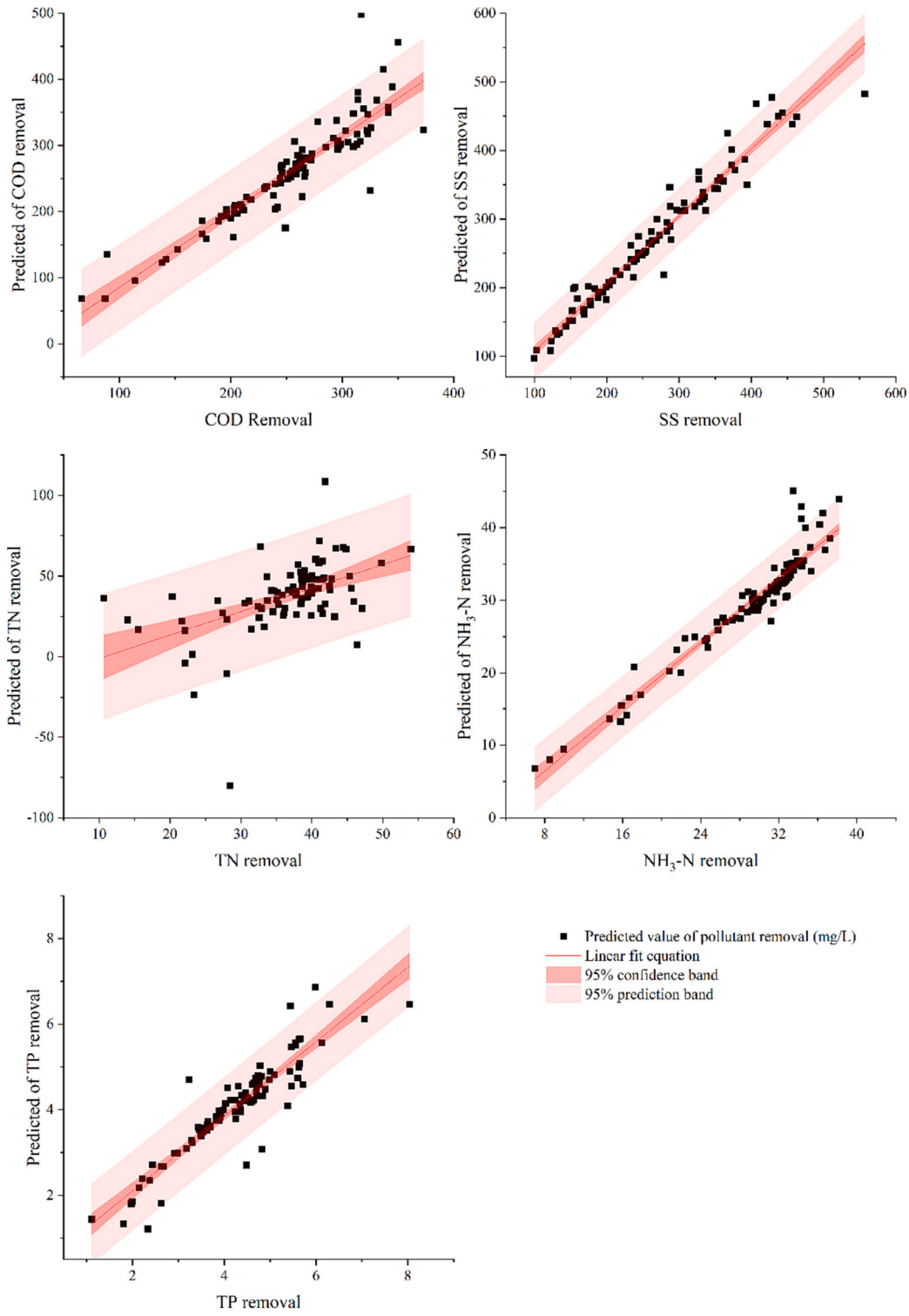

Fig. 5. Linear equation between the predicted value of pollutant removal and actual removal data.

Table 5
Linear equation of predicted values and the pollutants removal values.

Pollutants	Intercept	Slope	Pearson's r	R^2
COD	−28.924 ± 13.913	1.144 ± 0.0533	0.911	0.829
SS	9.963 ± 6.234	0.981 ± 0.0223	0.976	0.953
TN	−15.492 ± 9.588	1.452 ± 0.256	0.503	0.253
NH_3-N	−2.428 ± 1.005	1.108 ± 0.0338	0.959	0.919
TP	0.362 ± 0.174	0.871 ± 0.0395	0.915	0.837

predicted removal data [10]. This portion of the data were note eliminated from the comprehensive analysis, and it was tentatively assumed that the pollutant removal pathway is about as simple as it gets, the better linearity between the predicted value and the actual removal value. On the one hand, a pollutant like SS, which is treated primarily in the coagulation and sedimentation tank, has a single treatment unit without excessive treatment professes. On the other hand, TN is treated in the anaerobic and anoxic sections of AAO, and some of it will even be treated in the secondary sedimentation tank. TP needs multiple

pathways in the AAO, and if the treatment is not effective, additional chemical are needed. Therefore, this multi-treatment process is likely to affect the prediction of contaminants by ANFIS, resulting in some difference between the predicted value and the actual removal value [11].

The effluent cannot be controlled independently, and the influent must not exceed the influent requirements [6]. But some studies have shown that it is necessary to screen each parameter, and some parameters cannot be combined with the parameters to be predicted [3]. In this research, all the parameters were kept in ANFIS model. For COD, SS, TP, and TN parameters, the number of values exceeding the influent standard was 71, 121, 14, and 87, respectively, totaling 243 [1]. Half of the Inf-SS exceeded the Influent standard, leading to the addition of a combined phosphorus removal agent and an additional carbon source to achieve compliance with emissions regulations. Therefore, the predicted values may exceed the Influent and the actual removal amount. However, considering the actual operation of the WWTPs and the real-world scenarios, this data was not discarded.

In Fig. 6, the fitted R^2 values were relatively low because the pollutant effluents were provided in smaller integers. Most of the predicted data were concentrated in the range of [10, 20], consistent with the actual data, while only a few deviated from the expected values and exceeded the 95 % confidence interval. Meanwhile, the R^2 values of the other four pollutants had no explicit reference value, as the Effluent values were mostly whole numbers. The more decimal points, the greater the difference between actual and predicted values [6]. For example, it is difficult to trace back form Eff-COD to Inf-COD [9]. Another reason is that there can be come fluctuation in Effluent values while still meeting the requirements, causing Eff-pollutant to become a range. If the Eff-pollutant was less than this specific number, then this pollutant was considered well treated. Hence, the data fluctuation caused by volatility was generated between the predicted and actual data [11].

Predicting the removal of pollutants and then converting it to the amount of pollutants in the effluent is equivalent to amplifying the predicted value. Any predicted data <0 was removed. The relationship between the predicted data and the actual data was contrasted and explained by the color difference, as demonstrated in Fig. 7. Although most of the Eff-COD values are integers, the predicted data is not dominated by integers, as displayed in Fig. 7a. The predicted data obtained by the same Eff-COD had obvious scatter, especially the actual data of 3 and 12. The difference between the results of the predicted data was significantly large, with the absolute value much larger than the standard value [9]. However, overall, it tended to converge to the range of the Effluent. This is more apparent when the negative and positive values exceed the Eff standard.

Many of the data points in Figs. 7b and 6e also showed a high degree of dispersion. The closer the input data was to the center, the closer it was to the actual data. This was similar to that displayed in Fig. 7c and d. Although several discrete TN data exist, most of the predicted data converged to the discharge criterion with several negative numbers. In fact, Fig. 7d depicts a relatively good prediction tendency. The predicted values of input data >6 all converge in the direction of the actual data, and only a small amount of data show errors [24]. Regardless of the prediction results, when the input data contains more quantities in a range, the convergence and concentration of that part after ANFIS prediction is better than the prediction of other ranges or index input data. This is more obvious in the Eff-COD, Eff-TN, and Eff-NH_3-N.

This study preliminarily believed that in addition to in-depth investigation on influent types and pollutants, or using mathematical methods to improve data accuracy, human factors should also be considered verify and judge each parameter. Take water temperature (T, °C) as an example, the T is higher, the functional microorganism has higher growth activity, pollutants removal activity and faster consumption of DO. However, T is entirely the opposite case in winter. This leads to the limitation of relying on PCA alone for the identification of parameters. In addition, human identification of parameters relies on practitioner experience, especially for the selection of key parameters. Therefore, increasing human identification of pollutants and environmental conditions and judging the importance of parameters subject to seasonal and environmental factors are important to improve the selection of input parameters for ANFIS and to improve prediction accuracy and precision.

It is also worth noting that although the predicted removal of pollutants achieved better results, there is no one-to-one correspondence between the calculated values and the actual Effluent values. The difference after one-to-one correspondence can be quite obvious, as some of the data have rather high dispersion, especially when compared to the actual values. To address this issue, restrictive statements were added to ANFIS to achieve the following two effects: 1) to add restrictive statements in ANFIS, such as the output data must be >0 or take the absolute value; and 2) to consider the logical relationship between the predicted data and the actual data when ANFIS was used to predict small by larger input data.

To achieve these effects, the following restrictive statements were added:

Statements 1): if p<=0 %if the predicted value p is less than 0, % p=abs(p); or return p; %p takes the absolute value or return to p to repredict until the predicted value is greater than 0. %end %end this loop. %

Statements 2): TestOutputs = evalfis(chkdata(:, 1:end-1),fismat2); % Test output set data, %if TestOutputs<=0 %if the data of testoutputs is less than 0, %TestOutputs=abs(TestOutputs); or return TestOutputs; % TestOutputs takes the absolute value or returns to TestOutputs to repredict until the value is greater than 0. %end %end this loop.%

The statement after % is the interpretive statement within the program and will not be executed. Unfortunately, neither of these restrictive statements achieved the predicted effect where each predicted value remained unchanged. Therefore, the hypothesis was that the method of setting restrictive statements outside of ANFIS does not interfere with its internal logic and is not directly executed. This approach would not affect the original prediction results. It is likely that the ANFIS logic needs to be adjusted, along with the input layer functions and output layer functions. This will ensure that the output values match the actual situation, rather than using a positive output value as a prediction criterion. This is an ongoing part of the follow-up study.

4. Conclusion

This research tentatively showed the primary parameters of treatment capacity prediction of WWTPS using multiple membership functions and OE. ANFIS programmed by MATLAB was used to predict the removal of major pollutants in WWTPS with improved results. There was a slight difference between the predicted and actual values mainly due to the volatility of the Effluent quality data. The obtained output data was difficult to retrace to the input data. Therefore, ANFIS needed further optimization. The membership function needed careful selection, as it has an important impact on the stability and repeatability of the prediction results. The logic of ANFIS needed to be adjusted to modulate the output values, rather than adding restrictive statements that cannot go deeper into the logic. Restrictive statements should go deeper into the hidden layer without altering the input layer to avoid affecting the output layer parameters. Moreover, multiple membership functions needed careful selection to predict the Effluent directly from the Influent. Notably, the prediction accuracy had to consider the dataset's dispersion and the relationship between various parameters.

For future research, collecting more long-term data from various independent WWTPS can investigate the simplicity and universality of ANFIS. With the nonlinear relationship within a single parameter and the relationship between parameters, ANFIS also needed to standardize its prediction process.

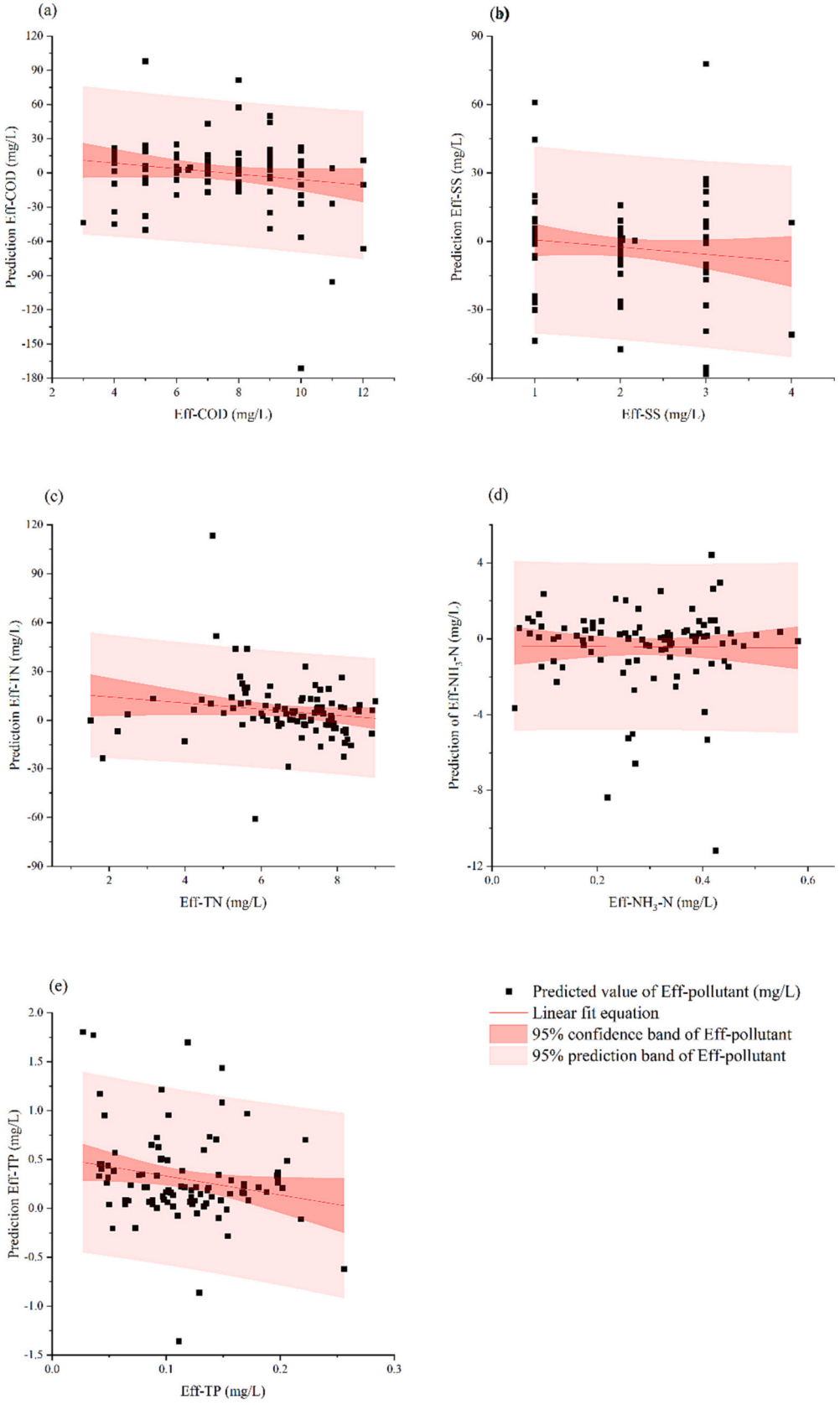

Fig. 6. The linear equation between the predicted and actual values.

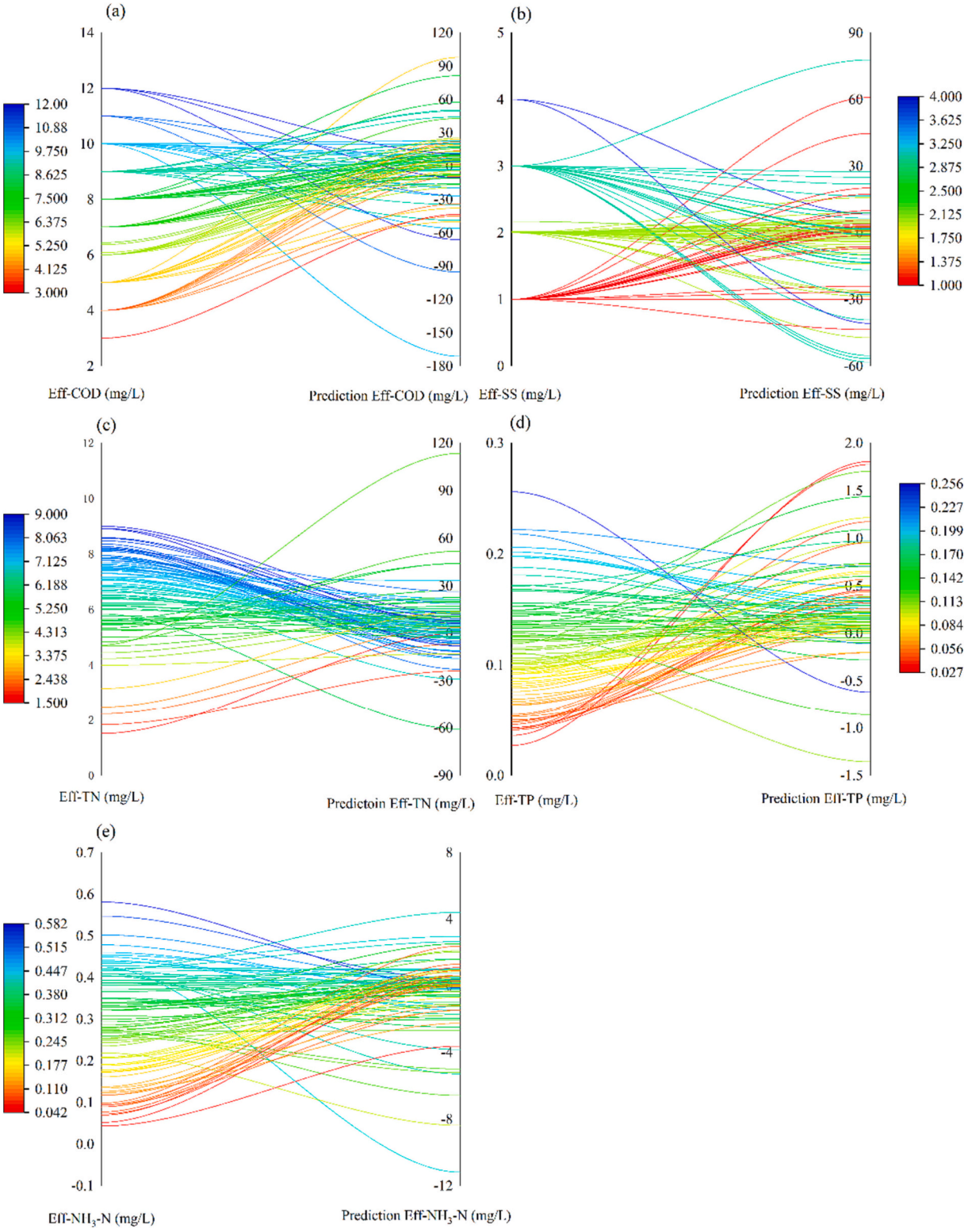

Fig. 7. Parallel coordinate plot of predicted data (right axis) and actual data (left axis).

CRediT authorship contribution statement

Liang Qiao: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Resources, Visualization, Funding acquisition. **Pei Yang:** Writing – review & editing, Visualization. **Qi Leng:** Writing – review & editing, Visualization. **Liujie Xu:** Writing – review & editing, Visualization. **Yanxin Bi:** Visualization, Software, Formal analysis. **Jinzhen Xu:** Visualization, Software, Formal analysis. **Zhe Wang:** Writing – review & editing, Visualization, Software, Formal analysis. **Jianye Liu:** Data curation, Resources, Investigation, Formal analysis. **Wanxin Yin:** Supervision, Resources. **Luyan Zhang:** Resources, Investigation, Formal analysis, Writing – review & editing, Funding acquisition. **Feihong Wang:** Writing – review & editing, Visualization. **Ye Yuan:** Supervision, Resources, Methodology,

Funding acquisition. **Tianming Chen:** Resources, Supervision, Project administration. **Cheng Ding:** Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This project was supported by the National Natural Science Foundation of China (NSFC, Grant No. 51608467 and No. 52170054), by the Funding for school-level research projects of Yancheng Institute of Technology (Grant No. xjr2020020), by the Natural Science Foundation of Jiangsu Province (Grants No. BK20210946), and by 'Qing Lan Project' of Colleges and Universities in Jiangsu Province.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jwpe.2023.104247.

References

[1] M. Bahramian, R.K. Dereli, W. Zhao, M. Giberti, E. Casey, Data to intelligence: the role of data-driven models in wastewater treatment, in: Expert Systems With Applications Vol. 217, Elsevier Ltd., 2023, https://doi.org/10.1016/j.eswa.2022.119453.

[2] M. Huang, Y. Ma, J. Wan, H. Zhang, Y. Wang, Modeling a paper-making wastewater treatment process by means of an adaptive network-based fuzzy inference system and principal component analysis, Ind. Eng. Chem. Res. 51 (17) (2012) 6166–6174, https://doi.org/10.1021/ie203049r.

[3] M. Kamali, L. Appels, X. Yu, T.M. Aminabhavi, R. Dewil, Artificial intelligence as a sustainable tool in wastewater treatment using membrane bioreactors, Chem. Eng. J. 417 (2021), https://doi.org/10.1016/j.cej.2020.128070.

[4] Q.V. Ly, V.H. Truong, B. Ji, X.C. Nguyen, K.H. Cho, H.H. Ngo, Z. Zhang, Exploring potential machine learning application based on big data for prediction of wastewater quality from different full-scale wastewater treatment plants, Sci. Total Environ. 832 (March) (2022), 154930, https://doi.org/10.1016/j.scitotenv.2022.154930.

[5] M. Andreides, P. Dolejš, J. Bartáček, The prediction of WWTP influent characteristics: good practices and challenges, in: Journal of Water Process Engineering Vol. 49, Elsevier Ltd., 2022, https://doi.org/10.1016/j.jwpe.2022.103009.

[6] S. Safeer, R.P. Pandey, B. Rehman, T. Safdar, I. Ahmad, S.W. Hasan, A. Ullah, A review of artificial intelligence in water purification and wastewater treatment: recent advancements, J. Water Process Eng. 49 (2022), 102974, https://doi.org/10.1016/j.jwpe.2022.102974.

[7] J. Zhao, H. Dai, Z. Wang, C. Chen, X. Cai, M. Song, Z. Guo, S. Zhang, X. Wang, H. Geng, Self-organizing modeling and control of activated sludge process based on fuzzy neural network, J. Water Process Eng. 53 (2023), https://doi.org/10.1016/j.jwpe.2023.103641.

[8] M. Zhu, J. Wang, X. Yang, Y. Zhang, L. Zhang, H. Ren, B. Wu, L. Ye, A review of the application of machine learning in water quality evaluation, Eco-Environ. Health 1 (2) (2022) 107–116, https://doi.org/10.1016/j.eehl.2022.06.001.

[9] V. Nourani, P. Asghari, E. Sharghi, Artificial intelligence based ensemble modeling of wastewater treatment plant using jittered data, J. Clean. Prod. 291 (2021), https://doi.org/10.1016/j.jclepro.2020.125772.

[10] S.I. Abba, H.C. Kilinc, M.L. Tan, V. Demir, I. Ahmadianfar, B. Halder, S. Heddam, A.H. Jawad, A.M. Al-Areeq, Z.M. Yaseen, Bio-communal wastewater treatment plant real-time modeling using an intelligent meta-heuristic approach: a sustainable and green ecosystem, J. Water Process Eng. 53 (2023), https://doi.org/10.1016/j.jwpe.2023.103731.

[11] Y. Yu, R. Wang, S. Huang, F. Wang, H. Zeng, L. Wang, H. Zhou, Z. Tan, Y. Chen, Simultaneous optimal prediction of various influent indexes based on a model fusion algorithm in wastewater treatment plant, Biochem. Eng. J. 198 (2023), https://doi.org/10.1016/j.bej.2023.109009.

[12] L. Zhao, T. Dai, Z. Qiao, P. Sun, J. Hao, Y. Yang, Application of artificial intelligence to wastewater treatment: a bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse, Process. Saf. Environ. Prot. 133 (92) (2020) 169–182, https://doi.org/10.1016/j.psep.2019.11.014.

[13] L. He, L. Bai, D.D. Dionysiou, Z. Wei, R. Spinney, C. Chu, Z. Lin, R. Xiao, Applications of computational chemistry, artificial intelligence, and machine learning in aquatic chemistry research, Chem. Eng. J. 426 (May) (2021), 131810, https://doi.org/10.1016/j.cej.2021.131810.

[14] Y. Azimi, M. Talaeian, H. Sarkheil, R. Hashemi, R. Shirdam, Developing an evolving multi-layer perceptron network by genetic algorithm to predict full-scale municipal wastewater treatment plant effluent, J. Environ. Chem. Eng. 10 (5) (2022), https://doi.org/10.1016/j.jece.2022.108398.

[15] M. Ansari, F. Othman, A. El-Shafie, Optimized fuzzy inference system to enhance prediction accuracy for influent characteristics of a sewage treatment plant, Sci. Total Environ. 722 (2020), https://doi.org/10.1016/j.scitotenv.2020.137878.

[16] K. Keyvan, M.R. Sohrabi, F. Motiee, An intelligent approach based on hybrid of principal component analysis-fuzzy inference system-adaptive neuro-fuzzy inference system for the simultaneous spectrophotometric determination of sofosbuvir and velpatasvir as antiviral drugs in pharmaceutical formulation and urine sample, Chemom. Intell. Lab. Syst. 220 (2022), https://doi.org/10.1016/j.chemolab.2021.104473.

[17] P. Rip Jeon, J.-H. Moon, O. Nafiu Olanrewaju, S. Hoon Lee, J. Lih Jie Ling, S. You, Y.-K. Park, Recent advances and future prospects of thermochemical biofuel conversion processes with machine learning, Chem. Eng. J. 144503 (2023), https://doi.org/10.1016/j.cej.2023.144503.

[18] J. Jawad, A.H. Hawari, S. Javaid Zaidi, Artificial neural network modeling of wastewater treatment and desalination using membrane processes: a review, in: Chemical Engineering Journal Vol. 419, Elsevier B.V, 2021, https://doi.org/10.1016/j.cej.2021.129540.

[19] M.S. Netto, J.S. Oliveira, N.P.G. Salau, G.L. Dotto, Analysis of adsorption isotherms of Ag+, Co+2, and Cu+2onto zeolites using computational intelligence models, J. Environ. Chem. Eng. 9 (1) (2021), 104960, https://doi.org/10.1016/j.jece.2020.104960.

[20] R. Abdi, G. Shahgholi, V.R. Sharabiani, A.R. Fanaei, M. Szymanek, Prediction compost criteria of organic wastes with biochar additive in in-vessel composting machine using ANFIS and ANN methods, Energy Rep. 9 (2023) 1684–1695, https://doi.org/10.1016/j.egyr.2023.01.001.

[21] P. Mathur, S. Singh, Analyze mathematical model for optimization of anaerobic digestion for treatment of waste water, Mater. Today: Proc. 62 (2022) 5575–5582, https://doi.org/10.1016/j.matpr.2022.04.606.

[22] P.P. Mondal, A. Galodha, V.K. Verma, V. Singh, P.L. Show, M.K. Awasthi, B. Lall, S. Anees, K. Pollmann, R. Jain, Review on machine learning-based bioprocess optimization, monitoring, and control systems, in: Bioresource Technology Vol. 370, Elsevier Ltd., 2023, https://doi.org/10.1016/j.biortech.2022.128523.

[23] M. Faisal, K.M. Muttaqi, D. Sutanto, A.Q. Al-Shetwi, P.J. Ker, M.A. Hannan, Control technologies of wastewater treatment plants: the state-of-the-art, current challenges, and future directions, in: Renewable and Sustainable Energy Reviews Vol. 181, Elsevier Ltd., 2023, https://doi.org/10.1016/j.rser.2023.113324.

[24] M. El-Rawy, M.K. Abd-Ellah, H. Fathi, A.K.A. Ahmed, Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques, J. Water Process Eng. 44 (July) (2021), 102380, https://doi.org/10.1016/j.jwpe.2021.102380.

[25] E. Hong, A.M. Yeneneh, T.K. Sen, H.M. Ang, A. Kayaalp, ANFIS based modelling of dewatering performance and polymer dose optimization in a wastewater treatment plant, J. Environ. Chem. Eng. 6 (2) (2018) 1957–1968, https://doi.org/10.1016/j.jece.2018.02.041.

[26] M.S. Zaghloul, R.A. Hamza, O.T. Iorhemen, J.H. Tay, Comparison of adaptive neuro-fuzzy inference systems (ANFIS) and support vector regression (SVR) for data-driven modelling of aerobic granular sludge reactors, J. Environ. Chem. Eng. 8 (3) (2020), 103742, https://doi.org/10.1016/j.jece.2020.103742.

[27] B. Najafi, S. Faizollahzadeh Ardabili, Application of ANFIS, ANN, and logistic methods in estimating biogas production from spent mushroom compost (SMC), Resour. Conserv. Recycl. 133 (February) (2018) 169–178, https://doi.org/10.1016/j.resconrec.2018.02.025.

[28] Z. Ye, J. Yang, N. Zhong, X. Tu, J. Jia, J. Wang, Tackling environmental challenges in pollution controls using artificial intelligence: a review, in: Science of the Total Environment Vol. 699, Elsevier B.V, 2020, https://doi.org/10.1016/j.scitotenv.2019.134279.

[29] I.D. Amoah, T. Abunama, O.O. Awolusi, L. Pillay, K. Pillay, S. Kumari, F. Bux, Effect of selected wastewater characteristics on estimation of SARS-CoV-2 viral load in wastewater, Environ. Res. 203 (April 2021) (2022), 111877, https://doi.org/10.1016/j.envres.2021.111877.

[30] G. Badalians Gholikandi, B.I. Beklar, M. Amouamouha, Performance prediction and upgrading of electroanaerobic baffled reactor using neural-fuzzy method, J. Environ. Chem. Eng. 9 (5) (2021), 106029, https://doi.org/10.1016/j.jece.2021.106029.

[31] K.J. Wang, P.S. Wang, H.P. Nguyen, A data-driven optimization model for coagulant dosage decision in industrial wastewater treatment, Comput. Chem. Eng. 152 (2021), 107383, https://doi.org/10.1016/j.compchemeng.2021.107383.

[32] A. Nawaz, A.S. Arora, C.M. Yun, H. Cho, S. You, M. Lee, Data authorization and forecasting by a proactive soft sensing tool-anammox based process, Ind. Eng. Chem. Res. 58 (22) (2019) 9552–9563, https://doi.org/10.1021/acs.iecr.9b00722.

[33] A.G. Kravets, P.P. Groumpos, M. Shcherbakov, M. Kultsova, Creativity in Intelligent Technologies and Data Science, 2019.

[34] H. Guo, K. Jeong, J. Lim, J. Jo, Y.M. Kim, J. pyo Park, J.H. Kim, K.H. Cho, Prediction of effluent concentration in a wastewater treatment plant using machine learning models, J. Environ. Sci. (China) 32 (2015) 90–101, https://doi.org/10.1016/j.jes.2015.01.007.

[35] W. Asnake Metekia, A. Garba Usman, B. Hatice Ulusoy, S. Isah Abba, K. Chirkena Bali, Artificial intelligence-based approaches for modeling the effects of spirulina

growth mediums on total phenolic compounds, Saudi J. Biol. Sci. 29 (2) (2022) 1111–1117, https://doi.org/10.1016/j.sjbs.2021.09.055.
[36] D. Dutta, S.R. Upreti, Artificial intelligence-based process control in chemical, biochemical, and biomedical engineering, Can. J. Chem. Eng. (2021), https://doi.org/10.1002/cjce.24246.
[37] K. Elmaadawy, M.A. Elaziz, A.H. Elsheikh, A. Moawad, B. Liu, S. Lu, Utilization of random vector functional link integrated with manta ray foraging optimization for effluent prediction of wastewater treatment plant, J. Environ. Manag. 298 (2021), https://doi.org/10.1016/j.jenvman.2021.113520.
[38] Y. Zhang, X. Gao, K. Smith, G. Inial, S. Liu, L.B. Conil, B. Pan, Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network, Water Res. 164 (2019), 114888, https://doi.org/10.1016/j.watres.2019.114888.
[39] B. Heydari, E. Abdollahzadeh Sharghi, S. Rafiee, S.S. Mohtasebi, Use of artificial neural network and adaptive neuro-fuzzy inference system for prediction of biogas production from spearmint essential oil wastewater treatment in up-flow anaerobic sludge blanket reactor, Fuel 306 (2021), https://doi.org/10.1016/j.fuel.2021.121734.