Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Analysis of Sewage Treatment Plant's Data to evaluate Quality of Effluent using Experimental Results and Prediction based on Artificial Intelligence Models (Case Study: Tehran Wastewater Treatment Plant)

Haniyeh Malek

haniye.74.malek@gmail.com

Babol Noshirvani University of Technology https://orcid.org/0009-0001-0748-0516

Majid Ehteshami

K.N.Toosi University of Technology

Research Article

Keywords: Sewage treatment Plant, Effluent, Artificial Neural Network (ANN), Multivariate Statistical Analysis Method, Statistical Indicators

Posted Date: September 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11178321/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Analysis of Sewage Treatment Plant's Data to evaluate Quality of Effluent using Experimental Results and Prediction based on Artificial Intelligence Models (Case Study: Tehran Wastewater Treatment Plant)

Haniyeh Malek[1], Majid Ehteshami[2]

1- Ph.D. Candidate in Environmental Engineering, Department of Civil Engineering, Babol Noshirvani University of Technology, Babol, Iran
2- Associate Professor, Department of Environment, Faculty of Civil Engineering, K. N. Toosi University of Technology, Tehran, Iran

Abstract:

The use of wastewater outflowing from the municipal wastewater treatment plants for irrigation of agricultural lands is one of the important and fundamental issues for reuse of the plants' effluent. In recent years, the artificial neural networks (ANN) have received considerable attention for modeling the sewage treatment process. Accordingly, the ANN models, fuzzy logic (FL) and adaptive neuro fuzzy inference system (ANFIS) have been utilized in this paper to predict quality of the effluent running out of the treatment plant. For this purpose, first, three scenarios were chosen and their inputs were reduced using the principal component analysis (PCA) method. Finally, the process of modeling was conducted with and without this method and then, the results were compared. Evaluating the results obtained from the predictions using the statistical indicators revealed that the ANFIS model with mean error reduction value of 13.92% compared to that of the FL model and a reduction value of 8.22% in contrast to the ANN model, benefits from a higher accuracy and this trend has been held true with and without PCA. Moreover, calculating the percentage of pollutant removal efficiency indicated that the maximum removal efficiency is obtained when total suspended solids (TSS) are removed which is equal to 96.68%. It bears to mention that the rest of the pollutants had values approximately equal to that of TSS. Based on the results, as a simple, accurate, efficient and reliable approach, the ANN models can be applied to predict quality of the effluents.

Keywords: Sewage treatment Plant, Effluent, Artificial Neural Network (ANN), Multivariate Statistical Analysis Method, Statistical Indicators.

1- Introduction

Due to population growth and increasing demand for water and limited water resources, the importance of reusing wastewater from reliable sources is receiving attention to supply water and preserve the environment.

Since most processes during wastewater treatment operations are nonlinear and exhibit behaviors that are not easily interpreted by linear mathematical models, and also due to the variety of incoming wastewater characteristics, such as flow rate as well as type and amount of organic and inorganic compounds, which affect the model parameters, modeling of the wastewater treatment processes using mathematical models, is a difficult task and most of the mathematical models used in this field, provide only estimates based on hypotheses and probabilities. Due to these limitations, the use of advanced software such as neural networks in modeling the treatment processes and predicting the performance of treatment plants can provide more accurate results ([2] and [3]). Prediction is one of the most important issues for researchers in making important decisions for principled planning. Thus, to achieve this important matter, useful statistical methods as well as available models can be applied. Accordingly, the existing and available information can be used at a lower cost in using statistical methods. The structure of statistical methods is such that their use does not require much expertise and skill in the field in question. In the group of statistical methods, the artificial neural networks (ANN), adaptive neuro-fuzzy inference system (ANFIS), linear and nonlinear models can be employed. Applying these methods requires information recorded in an area whose long-term measurement is not without its drawbacks and costs. As the statistical methods have been used successfully to predict the behavior of linear and nonlinear phenomena; Researchers are trying to reduce input parameters in order to reduce costs and facilitate access to better and more accurate results and interpretation [4]. In this regard, Camdevyren et al. have emphasized the use of principal component analysis (PCA) to reduce the number of variables and better interpret information [5]. Therefore, finding a structure for the artificial intelligence network (AIN) with minimal input seems necessary because increasing the large number of variables as input, causes the size of the neural network to grow and its training would be slowed down.

Many studies have been done in this field that can be mentioned. By creating an artificial neural network (ANN), Gontarski et al. predicted the quality characteristics of effluents from industrial sludge treatment plants. By performing the sensitivity analysis on parameters that had a great impact on the quality of effluent, it was concluded that pH and inlet flow were among the parameters that had the greatest impact on treatment plant control [6].

In Egypt, Nasr et al. examined the ANN approach using a feed-forward back propagation algorithm to predict the performance of the Alexandria EL-AGAMY wastewater treatment plant. This study shows that ANN can predict the treatment plant's performance by a correlation coefficient (R) between the observed and predicted output variables up to 0.90. In addition, the ANN provides an effective analysis and diagnostic tool to understand and simulate the nonlinear behavior of the treatment plant and is used as a valuable assessment tool for operators [7]. In the research work of Abaa and Elkirana, the artificial neural network model (ANNs) has been used to develop and estimate the chemical oxygen demand effluent (COD) from the Nicosia wastewater treatment plant. The results of the ANN model were compared with multiple linear regression (MLR) analysis. Based on the results, the ANN model was more accurate than MLR in predicting the COD effluent and offered a better performance [8]. In another study, Abaa et al. Studied multiple linear regression (MLR), artificial neural network (ANN), and adaptive neuro-fuzzy inference system (ANFIS) to predict dissolved oxygen (DO) concentrations downstream of Agra city. The results of this study illustrated that both ANN and ANFIS can be used in modeling DO concentrations in Agra city and also the results showed a slight advantage of ANN model over ANFIS and a significant advantage over MLR [9]. In the study by Elkiran et al., Artificial

intelligence-based approaches to multi-station modeling of dissolved oxygen in the downstream Matura River in India were investigated. In this study, adaptive neuro-fuzzy inference system (ANFIS) and feedforward neural network (FFNN) were used as two models based on artificial intelligence with multiple linear regression model. The simulated results of dissolved oxygen showed that multiple linear regression was not efficient for predicting dissolved oxygen. In addition, both AI models had a good predictive power; the ANFIS offered a slight increase in performance compared to the feedforward neural network model in the validation stage [10]. In 2018, Djeddou et al. evaluated the performance of the Ain Beida Wastewater Treatment Plant using artificial neural networks. Their study demonstrates the capabilities of the ANN model for modeling and indicates that the ANN is an adequate tool for controlling the performance of the activated sludge process [11]. In the research work of Abbaa et al., a fuzzy neural approach was developed using two different fuzzy inference system (FIS) optimizers (i.e. hybrid and feed-forward) to predict turbidity at the Tamburawa Water Treatment Plant (TWTP). Then, a fuzzy neural group technique was used to improve the performance of the two optimizers. The results showed that both the combined FIS optimizer and the combined FIS post-release optimizer showed excellent accuracy for predicting turbidity in the treatment plant [12]. Measuring the BOD_5 levels takes five days, and using a predicting model to estimate the BOD_5 levels, saves time and allows to adopt an online control system. In the study of Abdul Rahman Al-Sulaili and Abdul Rahman Rafie, the application of artificial neural networks (ANN) in predicting BOD5 concentration and wastewater treatment plant's performance was investigated. The results showed that the newly-developed ANN model had the best performance to predict BOD concentration. The optimal models obtained (three inputs, one output) show that the input temperature and conductivity have a great effect on the performance of the treatment plant as an input in all models. In general, the ANN model offers a simple approach to predict the complex processes of wastewater treatment plants [13].

In this study, first three scenarios were selected and modeling with artificial neural network (ANN), fuzzy logic (FL) and adaptive neuro-fuzzy inference system (ANFIS) once with PCA method and again without using this method, for these three scenarios, was carried out. Finally, the results of the models were compared using common relations to measure the error of the models and according to the error determination criteria. Flowchart of this study is depicted in Figure 1.

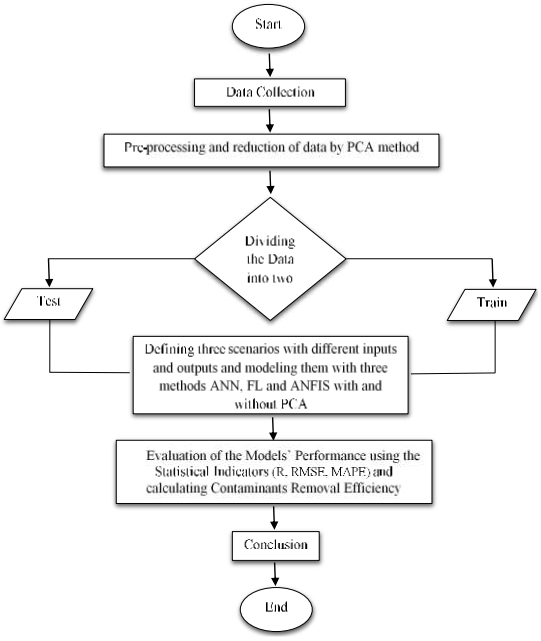


Figure 1: Flowchart of the Research Steps

2- Methodology and Model Development

2-1- Case Study: Tehran Treatment Plant

Before evaluating performance of the wastewater treatment plant by the ANN, it is necessary to obtain information about the wastewater and its treatment methods. As the industry grows, the wastewater treatment will become more and more difficult. As mentioned, population growth, water scarcity, and increasing competition between the agricultural and urban sectors for clean water make the need for treated wastewater even more apparent. International and domestic organizations and institutions, such as the World Health Organization (WHO) and the US Environmental Protection Agency (US EPA), are required to comply with wastewater reuse requirements. So far, several methods including physical, biological and chemical treatment have been developed to solve many problems of municipal wastewater pollution. Figure 2 shows the different units of municipal wastewater treatment plants.

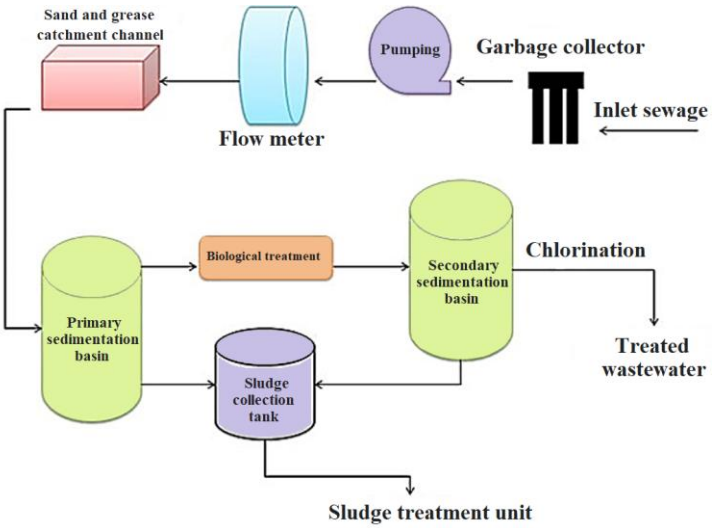


Figure 2 - Different units of municipal wastewater treatment plant

In this study, the effluent data of several treatment plants in Tehran province were investigated and among them, according to the removal efficiency of BOD, COD and TSS, Varamin wastewater treatment plant was selected for modeling. The area of the plant under study, is 7 hectares and the covered population for the project horizon in 2041, is equivalent to 160,000 people, in the first phase of which about 68,000 people were covered. This treatment plant is responsible for the collection, transfer, treatment and sanitary disposal of wastewater, which is on the agenda in two stages. The first phase of the project includes 60 km of collection network, about 1,500 branch lines, 6 km of transmission line and a phase of a wastewater treatment plant with a capacity of 11,145 m^3/day. The second phase of this project includes completion of the city collection network with a length of 80 km and construction of the second phase of the wastewater treatment plant, which will increase the total capacity of this treatment plant to 26,000 m^3/day. Since this project has been constructed in high traffic areas and with high density of underground facilities, the method of constructing the sewerage network has been mostly traditional (manual drilling). The area covered by the sewerage project is 1676 hectares. The type of sewage is domestic and industrial, about 90% of which is domestic. The place of discharge of effluent is also the surrounding agricultural lands. The layout plan of different units of the treatment plant is presented in Figure 3. Wastewater treatment method is extended aeration activated sludge (EAAS) and nitrogen

removal. This process is one of the oldest wastewater treatment processes and is routinely used today for biological treatment of municipal and industrial wastewater. In this process, the wastewater with a large amount of microorganisms is kept in a tank for several hours and enough oxygen is provided for it to create aerobic conditions. Microorganisms multiply rapidly under these conditions and provide the basis for treatment by consuming organic matter in the wastewater. Figure 4 shows the EAAS process.



Figure 3: Layout Plan of Treatment Plant

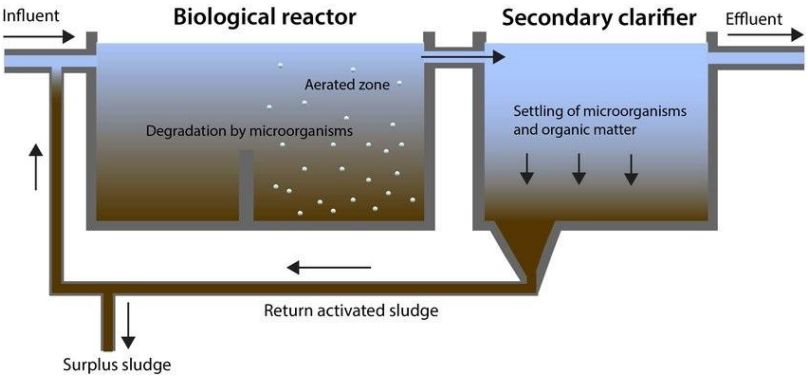


Figure 4: Extended Aerated Activated Sludge Process

2-2- Data collection and preparation

In measuring the quality of effluent and evaluating performance of the wastewater treatment plants, usually measuring and recording the parameters of temperature (T), biochemical oxygen demand (BOD), chemical oxygen demand (COD), total suspended solids (TSS) and pH at the treatment plant output and comparing them with Input values, are used. In this study, the quality parameters of inlet and outlet sewage and the parameters of aeration basin and data related to process conditions in a one-year period were analyzed. In this regard, the biochemical oxygen demand (BOD), chemical oxygen demand (COD) and total suspended solids (TSS) as qualitative parameters and mixed liquid suspended solids (MLSS), mixed liquid volatile suspended solids (MLVSS) and Sludge volume index (SVI) were used as the as the parameters of aeration basin. In addition, the temperature (T) and degree of acidity (pH) were taken as the environmental factors of the raw (untreated) sewage as input and the corresponding values of TSS, COD, BOD in the outflowing effluent were regarded as the outputs.

The data are usually gathered and recorded on a daily basis at the treatment plants, hence, first the data related to several months should be collected from the treatment plant and a review operation needs to be performed. There are often barriers to data collection from the treatment plant that cause problems and thus, before modeling process, it is necessary to review the data and identify out-of-range and insignificant data (i.e. outliers) and select the input data that has the greatest impact on the model outputs. Accordingly, this action is called data processing operation [14]. Some of the information and details about the data obtained from the wastewater treatment process is given in Table 1.

Table 1: Statistical Characteristics of the Input and Output Data

Type	Variable	Maximum	Mean	Minimum	Standard Deviation
Input	COD_{in}	520	368.84	260	62.42567
	BOD_{in}	275	220.64	178	26.51415
	TSS_{in}	283	245.76	200	20.16303
	Ph_{in}	8.35	7.29	5.6	0.69535
	T_{in}	25.5	18.94	13.5	3.58468
	$MLSS1_{in}$	4036	2720.51	1500	584.49188
	$SVI1_{in}$	472	290.69	109	72.28391
	$MLVSS1_{in}$	3527	2376.51	1420	509.95417
	$MLSS2_{in}$	3040	2373.56	1661	365.43756
	$SVI2_{in}$	544	285.10	123	109.79526
	$MLVSS2_{in}$	2600	2000.68	1426	336.02762
Output	COD_{out}	57	28.77	13.5	9.75745
	BOD_{out}	25	12.76	6.5	5.13890
	TSS_{out}	13.5	8.13	3	2.55867

2-3- Principal Component Analysis

PCA is one of the most common multivariate statistical techniques for reducing the size of large data volumes, and as its name implies, it can identify key components and help analyze a set of features that are more valuable than all other features. Past experiences have shown that conversion of variables by PCA method, in most cases increases accuracy of the learning model, and in fact, PCA calculates a matrix that indicates how our variables are all related to each other [12] and [15-16].

3- Studied Artificial Intelligence Models

3-1- Artificial Neural Network

The neural network is a mathematical structure based on the biological model of the human brain. Each neural network is made up of a set of small processed elements called neurons. Neurons connect in a special way to form monolayer or multilayer neural networks. The number of intermediate layers and neurons in each layer is usually obtained through a trial and error process. In fact, a neuron is the unit that takes inputs (P) and with their weights (W), (which show the information the network needs to solve the problem); It has the necessary effect on the inputs and is transferred to the hidden or middle layer by a fixed value called bias (b), where it produces a value according to the transfer function intended for it and sends to the network's output. This function is selected according to the problem or network of concern. Subsequently, several known functions are often used in the networks. Nevertheless, the user can utilize any kind of derivative function according to the case. One of the functions that is widely used in neural networks, is the sigmoid transfer function. Working with the ANN includes three stages of training, validation and testing. Among the various classifications of artificial neural networks, the feed-

forward neural network (FFNN) is the most common network in which data flows from the input layer through the hidden layer to the output layer. This neural network classification includes single layer perception (SLP), multilayer perception (MLP), and radial basis function (RBF). The MLP's structure is a standard sequence of inputs, linear and nonlinear neural units, and outputs. The ANN is trained to perform a practical function by adjusting the values of the weights between the elements.The purpose of network training is to achieve a situation where the network is able to provide an appropriate response to the data provided in the network training as well as similar and different data from the inputs that are used for network training. Accordingly, the back-propagation (BP) and Levenberg-Marquardt are among the most popular training algorithms. The Levenberg-Marquardt algorithm is the fastest way to train the medium-sized ANNs (up to several hundred weights). There is also an appropriate tool for this method in MATLAB software. The ANN is usually guided to the target output using real output data and the ANN is trained. The network is adjusted based on a comparison between the network output and the target output until the output matches the target. The structure of the neural network can be seen in Figure 5 [8-9], [13] and [17-30].

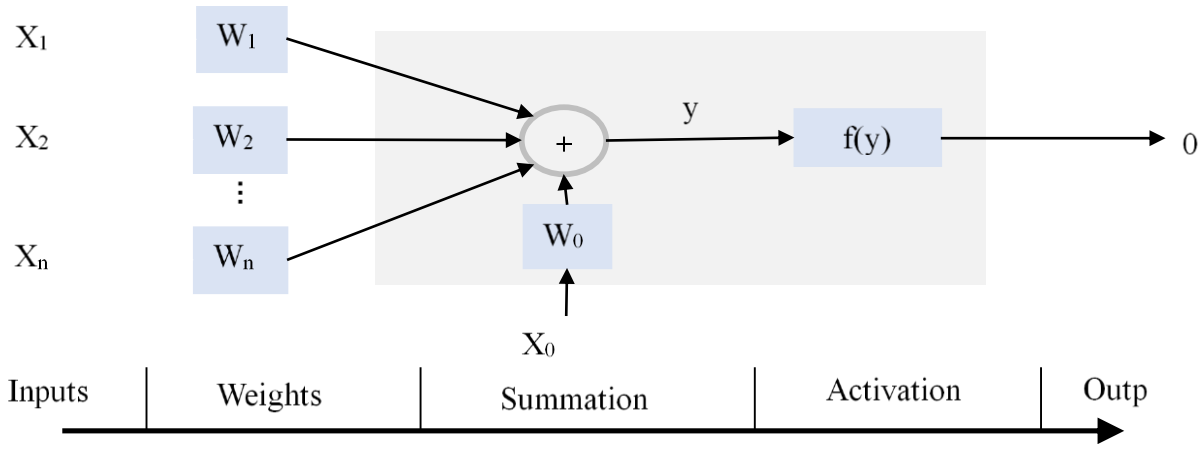


Figure 5: Structure of the Artificial Neural Network

3-2- Fuzzy Logic

The concept and study of fuzzy logic began in 1920, but the theory of fuzzy sets was first used by Professor Lotfi Ali Askarzadeh in 1965 at Berkeley University. The fuzzy logic is based on the concept of fuzzy sets that establish a soft boundary between different classes of topics, so that the membership is of degree and the degree parameter can show the degree of membership. Therefore, one of the most basic concepts in fuzzy logic is the membership function, which gives a membership value for each member of the set. There are several types of membership functions, including Gaussian, Triangular, Trapezoidal etc. Fuzzy inference systems are nonlinear models that describe the input-output relationship of a real system using if-then fuzzy rules. Two widely used types of fuzzy models include: 1. Mamdani type and 2. Sugeno type, the most important and only difference between the two is in their output, so that in Mamdani model, the output is a fuzzy set, but in Sugeno model, the output is linear or fixed. This logic is divided into four parts, which can be clearly seen in Figure 6.

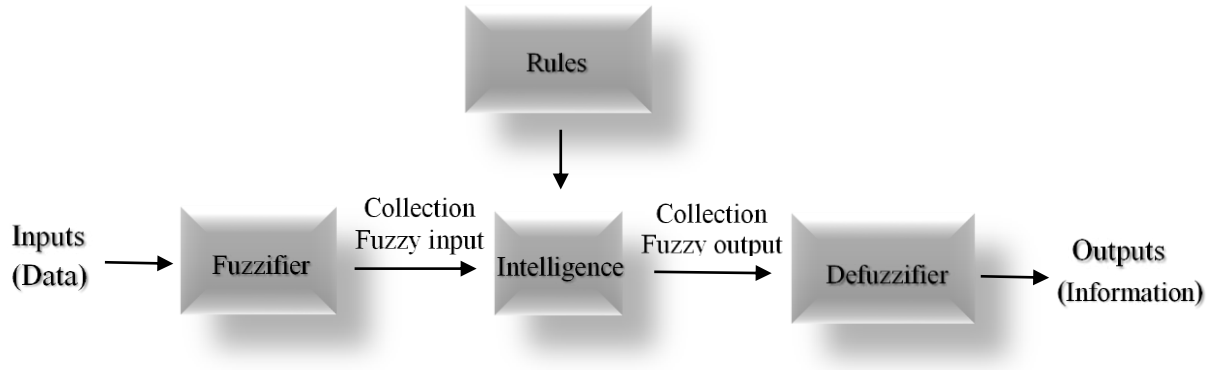


Figure 6: Structure of the Fuzzy Logic

In the fuzzyization step, the inputs are converted to the fuzzy information. The Basic Rules section contains all the rules and conditions that are specified by an expert through the "if ….then" scheme. In the inference or intelligence section, the degree to which fuzzy inputs conform to the basic rules is determined. In the last step, the results of fuzzy inference, which are fuzzy sets, are converted into quantitative data and information [9-10] and [31-32].

3-3- Adaptive Neuro Fuzzy Inference System

By combining neural network and fuzzy logic, a new type of system called adaptive neural fuzzy system (ANFIS) is created, which have the advantages of both methods. In 1993, Dr. Ragarjang introduced ANFIS as a global approximation that could accurately estimate any actual performance in a compact data set. In fact, the ANFIS is a fuzzy inference system that adapts itself to the training data. As determination of the parameters of the membership functions in the first part of the rules as well as specifying the coefficients of the last part of the rules and also, determining the weights of the neural network is done by methods such back-propagation algorithms; ANFIS carries the word neural [33]. The general scheme of ANFIS is as shown in Figure 7.

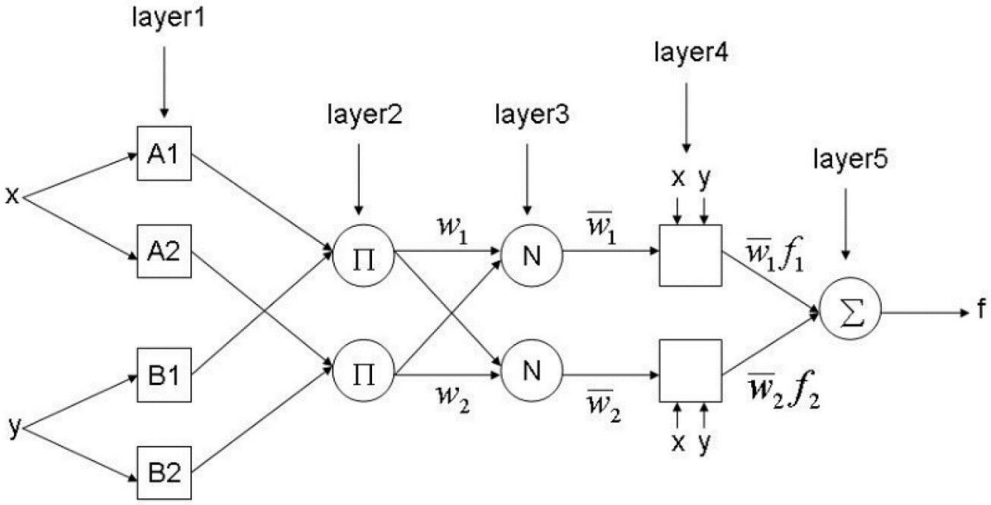


Figure 7: General Scheme of Adaptive Fuzzy Neural Network

Assuming that FIS has two inputs x and y and one output f, a first-order fuzzy Sugeno has the following rules:

$$Rule\ 1: if\ \mu(x) is\ A_1\ and\ \mu(y) is\ B_1; then\ f_1 = p_1 x + q_1 y + r_1$$

$$Rule\ 2: if\ \mu(x) is\ A_2\ and\ \mu(y) is\ B_2; then\ f_2 = p_2 x + q_2 y + r_2$$

Parameters A_1, B_1, and A_2, B_2 are the membership functions for inputs x and y. r_1, q_1, p_1, r_2, q_2, p_2 are the parameters of the output function.

ANFIS's structure is comprised of five layers:

Layer 1, Input Nodes: After the input data is entered into the system as a matrix (x, y), in this layer, fuzzy operation is performed using the membership functions. In this layer, inputs pass through membership functions.

$$Q_i^1 = \mu_{Ai}(x) \quad for\ i = 1,2 \qquad or \qquad Q_i^1 = \mu_{Bi}(x) \quad for\ i = 3,4 \tag{1}$$

Where A_i and B_i are language tags that are denoted by the membership functions μ_{Ai} and μ_{Bi}, respectively. The membership functions of each function can be a suitable parameter that in most cases Gaussian functions are selected.

Layer 2, Rule Nodes: The "and" operator is used in this layer. In fact, the output of this layer is the multiplication of the input signals, which is equivalent to the part of "if" in the rules.

$$Q_i^2 = w_i = \mu_{Ai}(x).\mu_{Bi}(y) \qquad for\ i = 1,2 \tag{2}$$

Layer 3, Intermediate Nodes: The main purpose of the third layer is to determine the ratio of each degree of activity of the first law to the total degree of activity of all laws. The output of this normalized layer is the previous layer and is standardized as follows:

$$Q_i^3 = \bar{w}_i = \frac{w_1}{w_1 + w_2} \qquad ,i = 1,2 \tag{3}$$

Layer 4, Result Nodes: Each node i in this layer is an adaptive node and executes the result of the rules. In this output layer, each node is equal to:

$$Q_i^4 = \bar{w}_i(p_1 x + q_1 y + r_1) = \bar{w}_i f_i \tag{4}$$

Where w_i is the output of the i-th node of the previous layer and $\{p_i, q_i, r_i\}$ are the coefficients of this linear combination.

Layer 5, Output Nodes: In this layer, the non-fuzzy process transforms the results of each fuzzy rule into a non-fuzzy output. The output of this layer is the total output of the system, the number of nodes being equal to the outputs:

$$Q_i^5 = \sum \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{5}$$

A network is now being developed that is equivalent to the Sugeno fuzzy inference system. To train the network, first in layer 1 all the existing rules are formed. For example, if there are 2 inputs, each with 3 membership functions, 9 rules must be formed as illustrated in Figure 8:

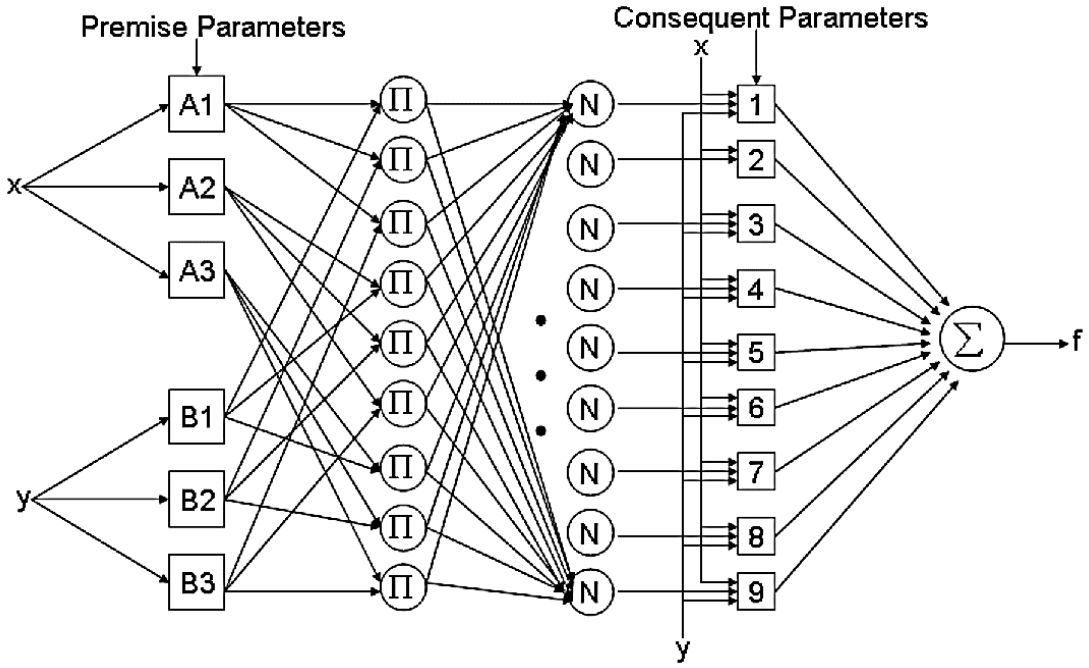


Figure 8: Illustration of Fuzzy Inference System

In the fuzzy neural model, the simulation operation is performed correctly when the following set of indicators and the set of adaptive indicators are estimated in such a way that the value of the model error function in the training and testing sections, is minimized. The number and type of inputs as well as the shape of the membership function are the factors influencing the neuro-fuzzy model [9-10], [12] and [34].

4- Model performance indicators

Various evaluation criteria can be used to determine the comparative accuracy of the predictive models. To compare performance of the models presented in the training and testing stages, statistical indicators according to relationships (6) to (8) including correlation coefficient (R), root mean square error (RMSE) and mean absolute percentage error (MAPE) were used [16] and [35].

$$R = \frac{\sum_{i=1}^{N}(q_i - \bar{q})(q_{\hat{i}} - \bar{\hat{q}})}{\sqrt{\sum_{i=1}^{N}(q_i - \bar{q})^2 \sum_{i=1}^{N}(q_{\hat{i}} - \bar{\hat{q}})^2}} \quad (6)$$

Where q_i and $q_{\hat{i}}$ are the average observational and computational values of the output of the modeling, respectively.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(q_i - \hat{q}_i)^2} \quad (7)$$

$$MAPE = \frac{100}{N}\sum_{I=1}^{N}\frac{|q_i - \hat{q}_i|}{q_i} \quad (8)$$

N is the number of data, q_i and $\hat{q}_i$ represent the observational value and the value estimated by the model, respectively.

5- Productivity Index

In this study, to compare the ability of models to remove, reduce or separate contaminants after treatment, an index with relation (9) is used. This index expresses the amount of reduction or elimination as a percentage, which the closer this number is to 100, means that the model has performed better in removing pollutants.

$$RE_X = \frac{1}{N}\sum_{i=1}^{N}\frac{X_{in}-X_{out}}{X_{in}}\times 100 \quad (9)$$

Where RE_X the productivity percentage index removes x, x_{in} is the input mass of x to the treatment plant and artificial intelligence, x_{out} is the mass output of x from the treatment plant and is estimated from artificial intelligence, and N is the number of data from each contaminant.

6- Results and Discussion

6-1- Principal Component Analysis

This method is used to determine the input parameters that have the most impact and correlation with the output parameters. Initially, 3 scenarios with 11 inputs (COD_{in}, BOD_{in}, TSS_{in}, Ph_{in}, T_{in}, $MLSS1_{in}$, $SVI1_{in}$, $MLVSS1_{in}$, $MLSS2_{in}$, $SVI2_{in}$, $MLVSS2_{in}$) and one output (BOD, COD, TSS) are modeled and then correlated according to the data. The input for each scenarios is reduced to 4 inputs to increase the accuracy of the models and improve the modeling process. The data correlation matrix is given in Table 2.

Table 2: Data Correlation Matrix

	COD_{in}	BOD_{in}	TSS_{in}	Ph_{in}	T_{in}	$MLSS1_{in}$	$SVI1_{in}$	$MLVSS1_{in}$	$MLSS2_{in}$	$SVI2_{in}$	$MLVSS2_{in}$	COD_{out}	BOD_{out}	TSS_{out}
COD_{in}	1	0.142	0.44	0.044	0.26	0.118	0.325	0.09	0.169	0.158	0.298	0.242	0.199	-0.184
BOD_{in}		1	0.143	-0.069	-0.087	-0.039	0.036	0.067	-0.074	0.028	0.089	-0.127	0.247	-0.197
TSS_{in}			1	-0.253	0.032	0.219	0.227	0.182	0.251	0.015	0.317	0.146	**0.468**	0.064
Ph_{in}				1	0.74	-0.122	0.238	-0.106	-0.16	0.691	-0.175	**-0.461**	**-0.662**	**-0.360**
T_{in}					1	-0.03	0.31	-0.085	0.138	0.868	0.122	**-0.45**	**-0.495**	**-0.232**
$MLSS1_{in}$						1	0.224	0.665	0.35	-0.159	0.413	0.039	0.23	-0.009
$SVI1_{in}$							1	0.026	-0.027	0.269	0.252	**-0.277**	-0.137	**-0.272**
$MLVSS1_{in}$								1	0.243	-0.263	0.307	-0.037	0.201	0.075
$MLSS2_{in}$									1	-0.102	0.71	0.064	0.222	0.084
$SVI2_{in}$										1	-0.075	**-0.459**	**-0.523**	**-0.234**
$MLVSS2_{in}$											1	-0.01	0.224	-0.041
COD_{out}												1	0.477	0.223
BOD_{out}													1	0.452
TSS_{out}														1

According to the correlation matrix, the four input parameters that have the highest correlation with each of the three outputs are classified in three separate models according to Table 3 and then all 3 scenarios

once with PCA method and once without PCA method and with different methods of artificial intelligence, are modeled and the results of the modeling are compared.

Table 3: Inputs and Outputs of the two Samples before and after PCA

scenarios	Inputs	Outputs
1	COD_{in}, BOD_{in}, TSS_{in}, Ph_{in}, T_{in}, $MLSS_{1in}$, SVI_{1in}, $MLVSS_{1in}$, $MLSS_{2in}$, SVI_{2in}, $MLVSS_{2in}$	BOD
2	COD_{in}, BOD_{in}, TSS_{in}, Ph_{in}, T_{in}, $MLSS_{1in}$, SVI_{1in}, $MLVSS_{1in}$, $MLSS_{2in}$, SVI_{2in}, $MLVSS_{2in}$	COD
3	COD_{in}, BOD_{in}, TSS_{in}, Ph_{in}, T_{in}, $MLSS_{1in}$, SVI_{1in}, $MLVSS_{1in}$, $MLSS_{2in}$, SVI_{2in}, $MLVSS_{2in}$	TSS
1+PCA	Ph_{in}-SVI_{2in}-T_{in}-TSS_{in}	BOD
2+PCA	Ph_{in}-SVI_{2in}-T_{in}-SVI_{1in}	COD
3+PCA	Ph_{in}-SVI_{1in}-SVI_{2in}-T_{in}	TSS

6-2- Neural Network Modeling

In this study, three scenarios with different outputs were considered in order to model the input parameters, the number of hidden layers and the number of neurons in each layer, to predict the output parameter based on the best state of the model indicators. In the present study, the FFNN with the Multi Layer Perceptron (MLP) model and the Levenberg-Marquardt (LM) training algorithm along with the sigmoid transfer function in the hidden layer, have been employed. In the following, the three selected scenarios are first modeled without PCA and then with PCA, and finally the obtained results are compared. For the purpose of modeling, the input data is divided into three categories of training, testing and validation (70% data for training, 15% testing and 15% validation) and the sample is modeled several times with different neurons for the desired inputs, to achieve the minimum training error and acquire the best and most appropriate network. Figure 9 shows structure of the artificial neural network used for the treatment plant under study.

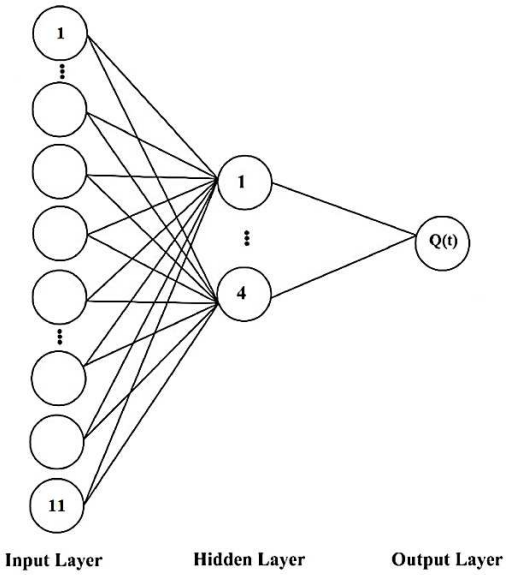


Figure 9: Structure of the ANN used for the Treatment Plant under study

Results of training the selected networks are presented in Figures 10 to 13. Based on the trial and error process, in the first scenario for the method without PCA, the network arrangement (11:8:1) and for the method with PCA, the network arrangement (4:12:1) and in the second scenario for the method without PCA, the arrangement (11:8:1) and for PCA method network arrangement (4:24:1) and in the third scenario also, the arrangement (11:16:1) and (4:24:1) for method without PCA, have been utilized. Results of the modeling are given in Table 4.

Table 4: Statistical analysis of neural network models in training and testing stages with and without PCA

Output	Without PCA					With PCA				
	Train	Test	Validation	RMSE	%MAPE	Train	Test	Validation	RMSE	%MAPE
	R					R				
BOD	0.9132	0.8748	0.941	2.17	15.78	0.9268	0.9167	0.8966	2.05	12.77
COD	0.9819	0.9434	0.8155	3.22	8.97	0.8137	0.7929	0.7246	5.84	15.37
TSS	0.9382	0.762	0.7983	1.36	12.05	0.9541	0.9039	0.8884	0.92	9.32

Figure 10 demonstrates the correlation between the data for the optimal models in the output obtained from MATLAB software. This diagram shows accuracy of the results of neural network modeling compared to the actual outputs in the experimental studies. The higher the focus of the points on the first quarter bisector, the higher the accuracy of the modeling. As can be seen, in terms of R value, there is an almost significant relationship between the output values of the ANN model and the experimental results.

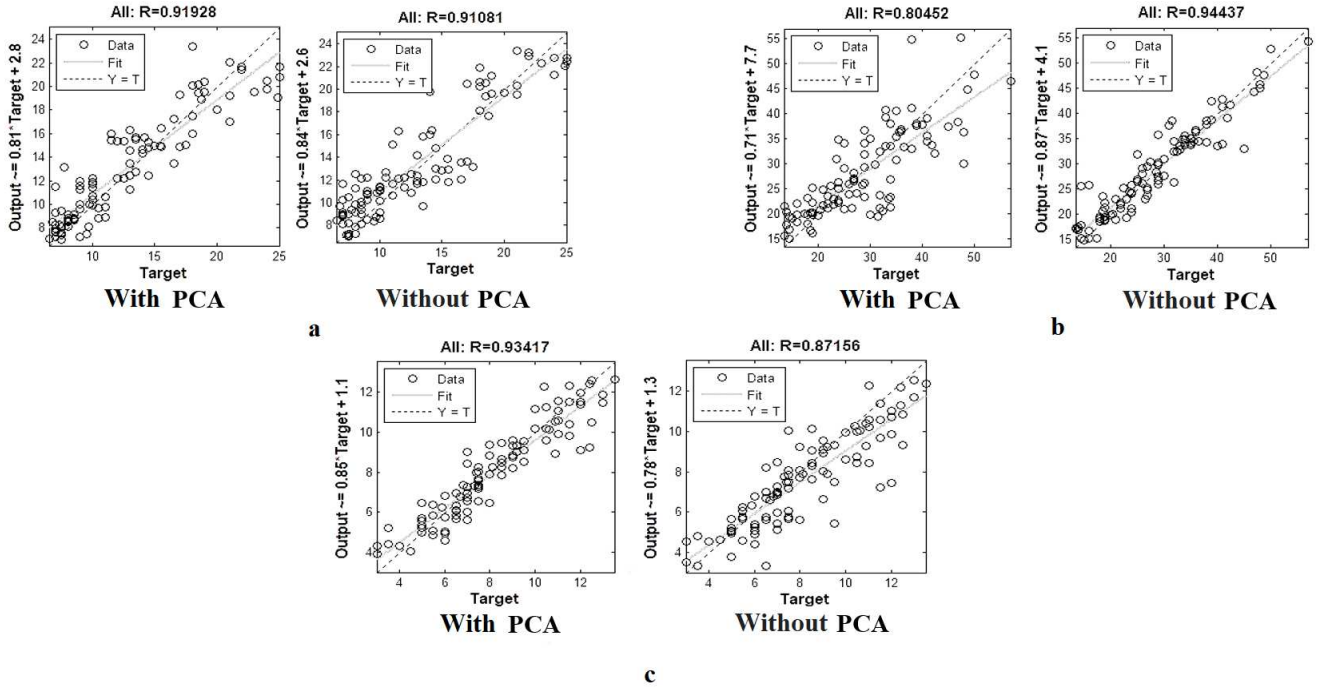


Figure 10: Correlation of neural network model data without PCA and with PCA for: a) BOD b) COD c) TSS

Figure 11 reveals the graph of changes in the mean squared error in the modeling process in terms of epoch changes that shows minimum decrease in the squared errors and if the three lines are in the same direction according to the diagram, the modeling is done correctly. As can be seen in the figure, all data (training, validation and testing data) have the highest MSE in epoch0. After multiple epochs, the MSE

value reaches a minimum in a given epoch, and the best validation performance is provided and afterwards, the model's performance remains approximately constant.

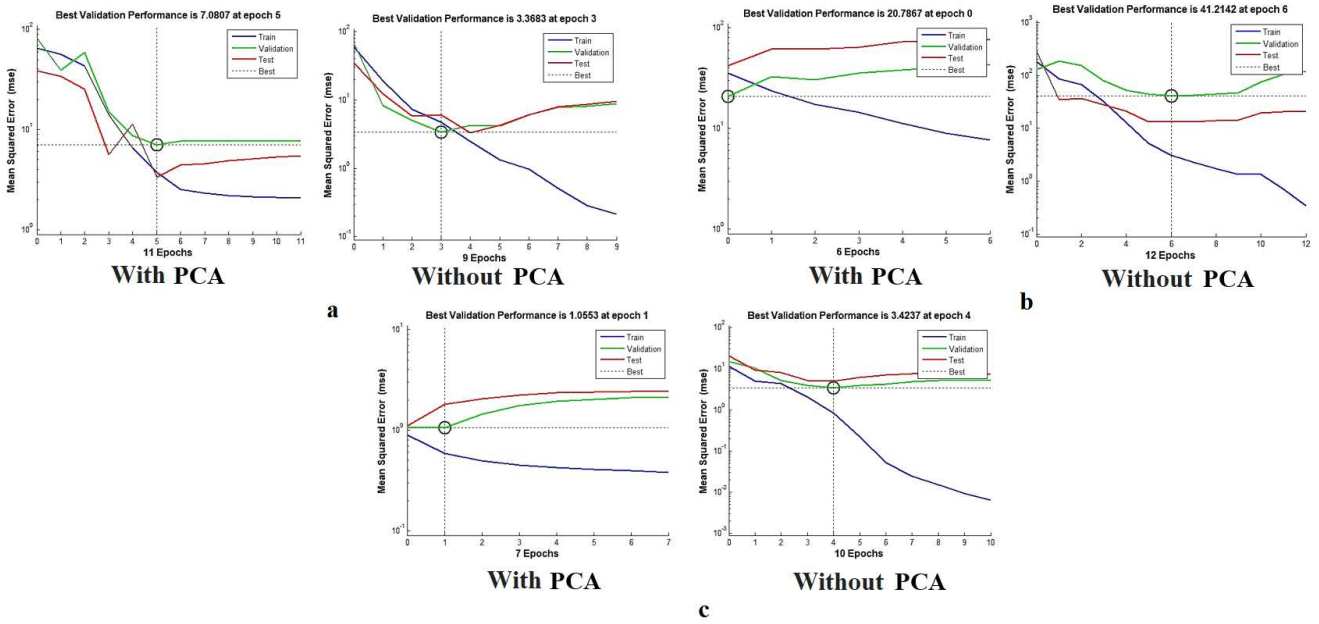


Figure 11: Diagram of MSE values without PCA and with PCA for: a) BOD b) COD c) TSS

Figure 12 demonstrates the trend of network changes in different epochs and the status of network training.

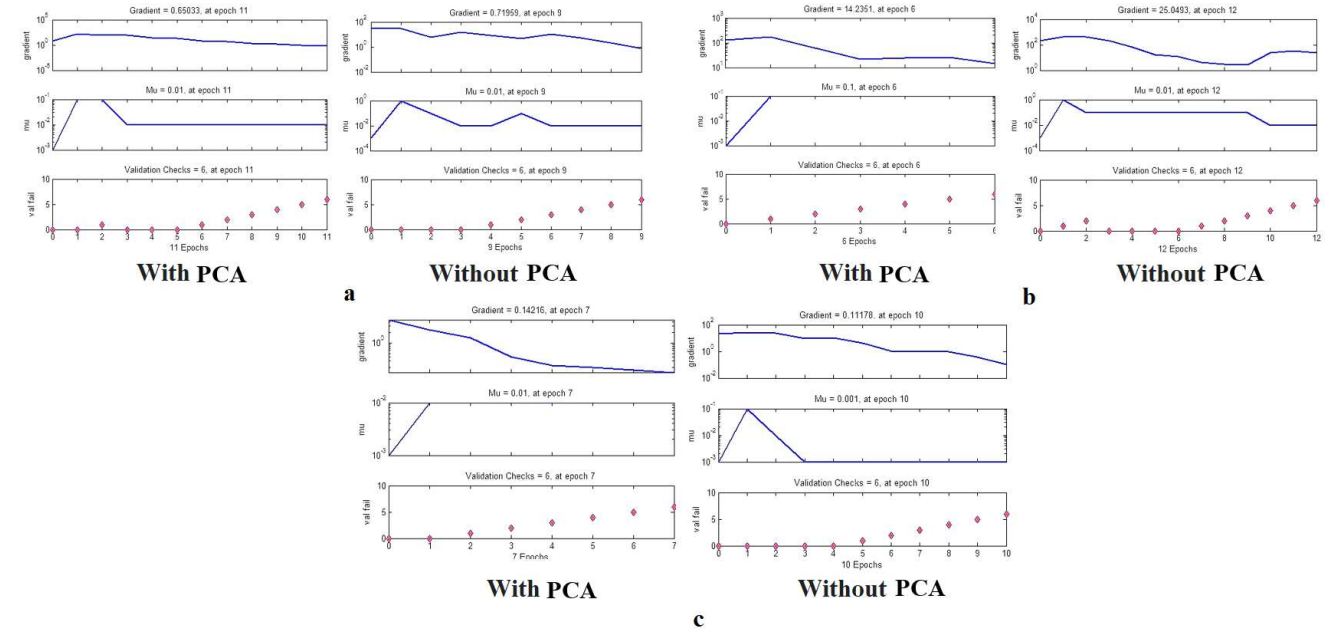
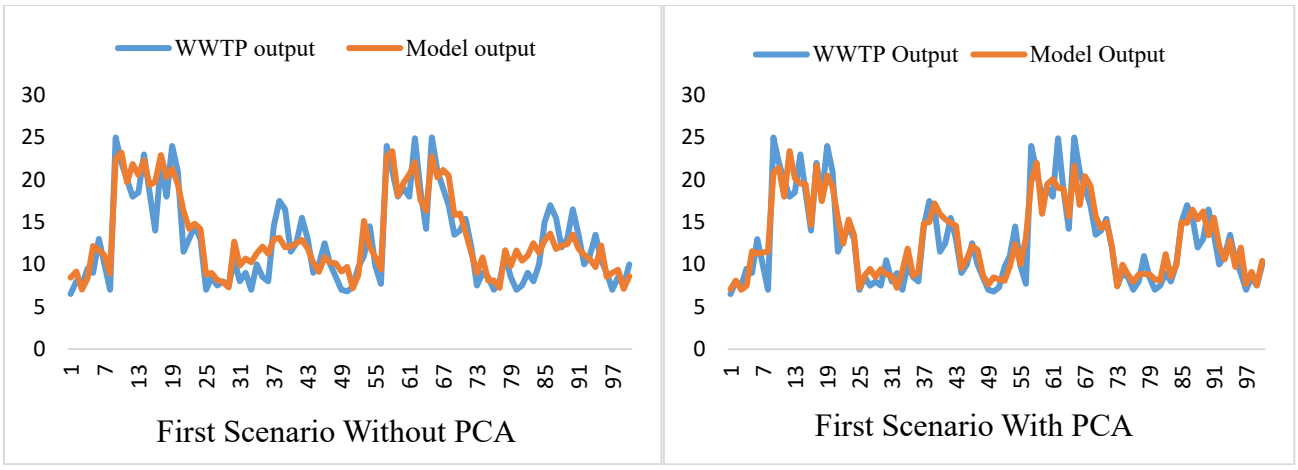
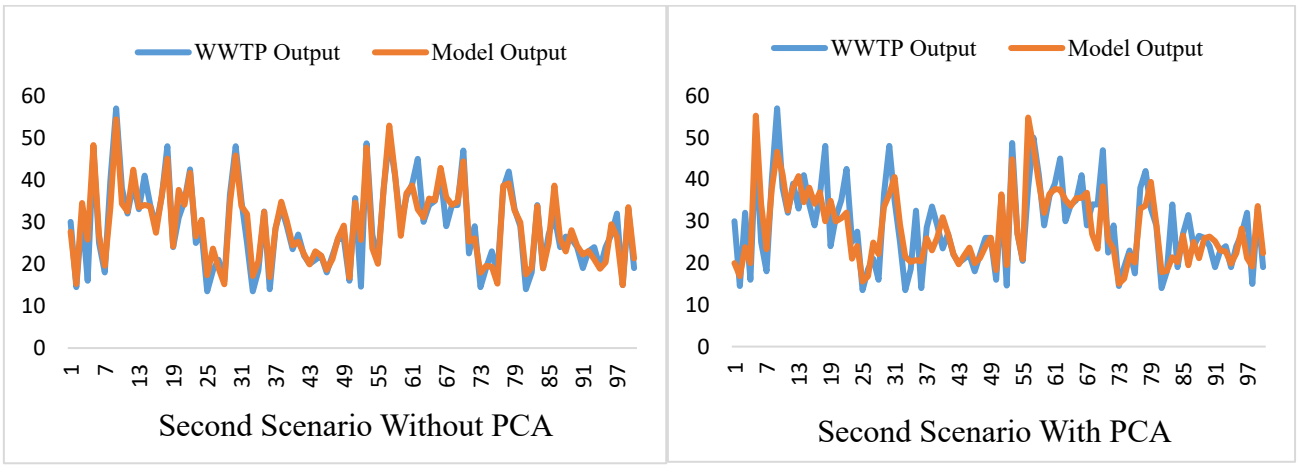


Figure 12: Trend of Network changes without PCA and with PCA for: a) BOD b) COD c) TSS

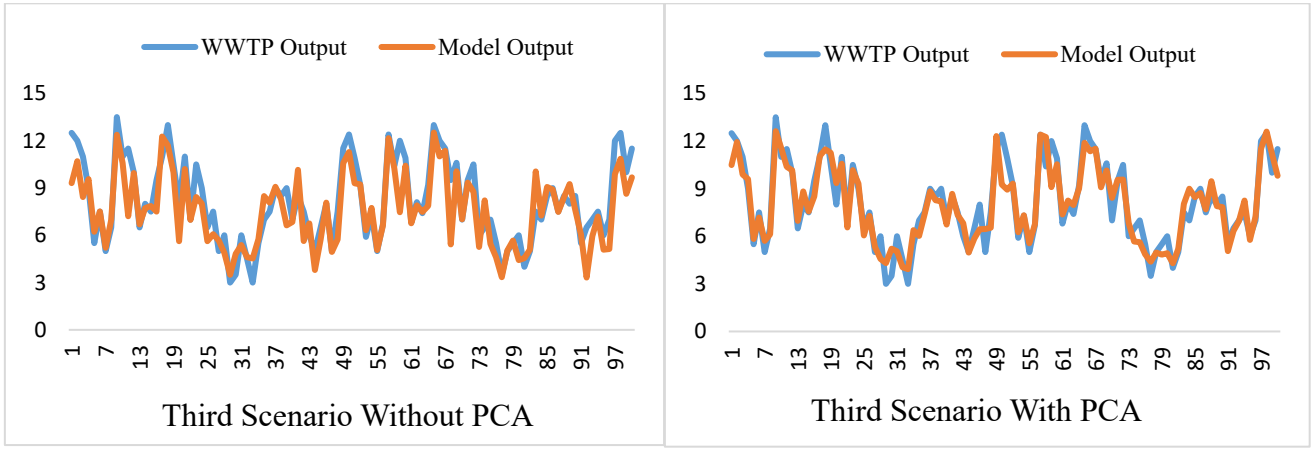
Finally, Figure 13 shows the comparison diagram of the experimental results with prediction of the neural network model with and without PCA. As can be seen, the distance between the points is insignificant, which indicates that there is a slight difference between the output of the model and the treatment plant.



a



b



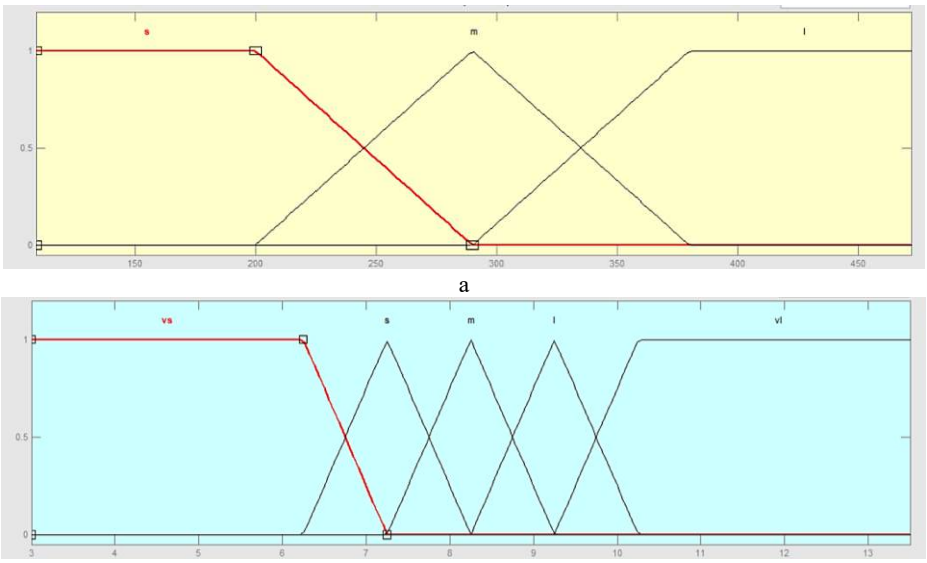
c

Figure 13: Comparison diagram of experimental values with the values predicted by ANN for: a) BOD b) COD c) TSS

In general, it can be concluded that the neural network has the ability to predict changes in wastewater treatment efficiency and shows good performance, and the use of data reduction method or PCA in the first and third scenarios leads to increased accuracy of models.

6-3- Fuzzy Logic Modeling

In this section, the Mamdani model has been used to build each of the models. First, for each data, three small, medium and large intervals (S, M, L) for inputs and five very small, small, medium, large and very large intervals (VS, S, M, L, VL) for output are considered. After this step, the desired rules are extracted from the data that the number of rules can be guessed by trial and error. Again, with the trial and error method and giving different membership functions, the optimal state with trapezoidal and triangular membership function is obtained for inputs and outputs. In the last step, the output of the model, which is fuzzy, is numbered by one of the non-fuzzyization methods. An example of a membership function for a sample of the inputs and outputs is given in Figure 14. The modeling results are presented in Table 5.



a

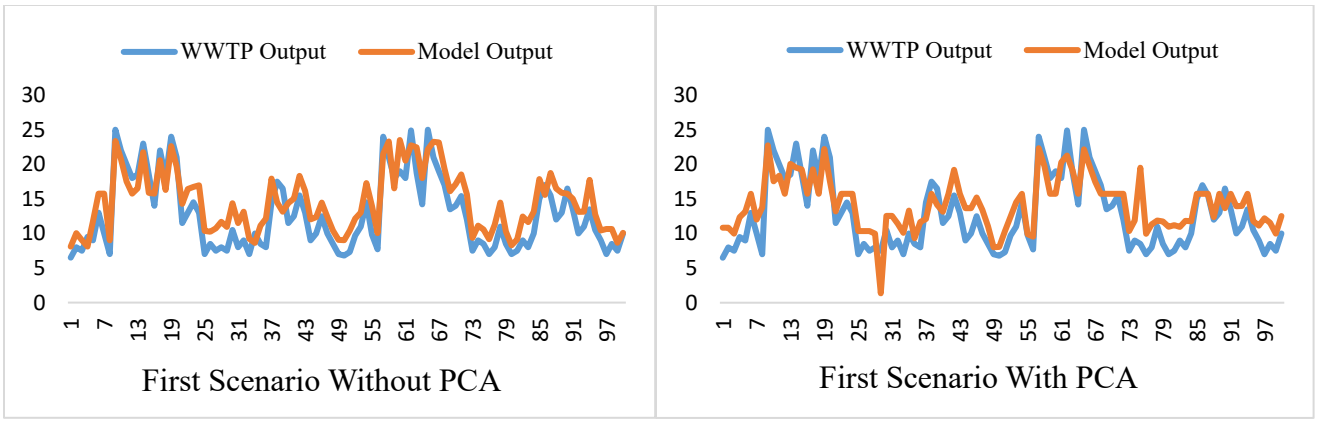
b

Figure 14: Membership function for: a) a sample of inputs, b) a sample of outputs

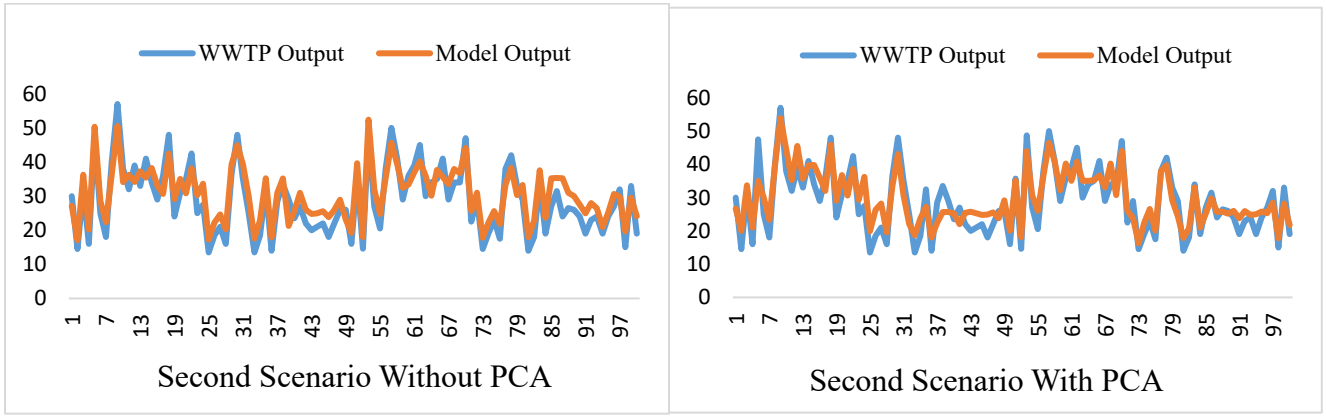
Table 5: Statistical analysis of fuzzy logic in training and testing with and without PCA

Output	Without PCA					With PCA				
	Train	Test	Train	Test	%MAPE	Train	Test	Train	Test	%MAPE
	R		RMSE			R		RMSE		
BOD	0.9114	0.9040	2.68	2.79	23.58	0.8926	0.6605	2.81	3.38	25.37
COD	0.9328	0.8756	4.14	4.48	15.57	0.9202	0.8981	3.06	4.76	15.22
TSS	0.9325	0.8933	1.13	1.26	14.56	0.9517	0.9161	0.83	0.91	10.19

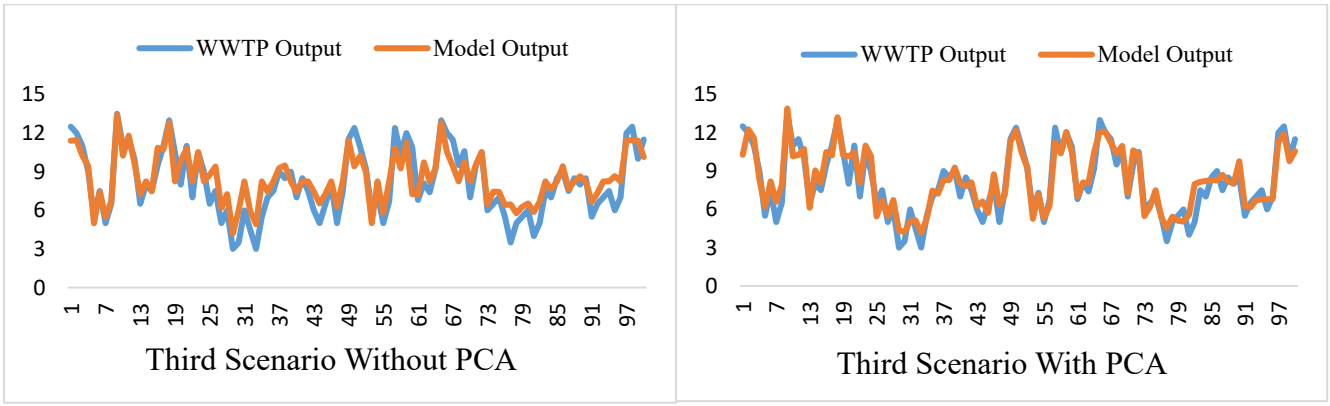
A comparison diagram of fuzzy logic output data (with and without PCA) with wastewater treatment plant's outputs can be seen in Figure 15.



a



b



c

Figure 15: Comparison diagram of experimental values with those predicted by FL for: a) BOD b) COD c) TSS

In the first and second scenarios, the results of modeling without using PCA method are better, but in the third scenario, the use of PCA has improved performance of the model. In general, it can be concluded that fuzzy logic has the ability to predict changes in wastewater treatment efficiency and also, shows a decent performance.

6-4- Modeling of adaptive neuro-fuzzy inference system (ANFIS)

In this section, as in the previous two sections, three scenarios with different outputs are considered to compare the results after modeling the input parameters. For the purpose of modeling, the input data is

divided into three categories of training, testing and validation (70% data for training, 15% testing and 15% validation) and various types of membership functions with different number of iterations are used by trial and error to determine the most appropriate model with its membership functions. An example of the adaptive neuro-fuzzy inference system's (ANFIS) structure used in this treatment plant, is depicted in Figure 16. In addition, some of the rules of the ANFIS model are presented in Figure 17.

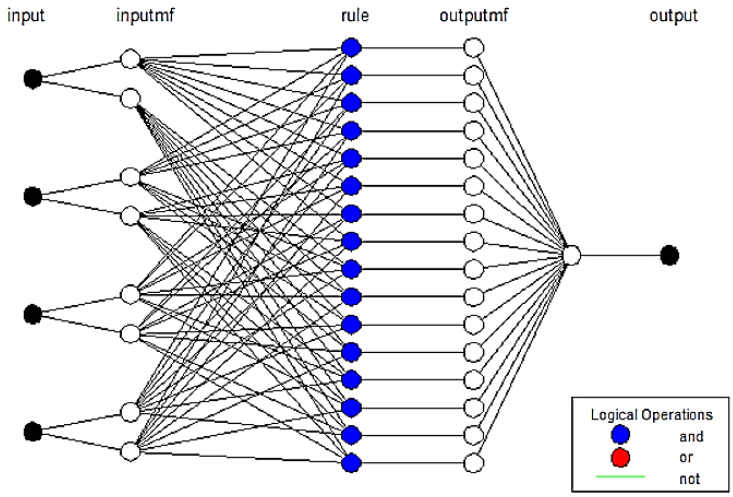


Figure 16: Structure of the adaptive neuro-fuzzy inference system (ANFIS) used in the treatment plant

5. If (input1 is in1mf1) and (input2 is in2mf1) and (input3 is in3mf2) and (input4 is in4mf2) then (output is out1mf5) (1)
6. If (input1 is in1mf1) and (input2 is in2mf1) and (input3 is in3mf2) and (input4 is in4mf3) then (output is out1mf6) (1)
7. If (input1 is in1mf1) and (input2 is in2mf1) and (input3 is in3mf3) and (input4 is in4mf1) then (output is out1mf7) (1)
8. If (input1 is in1mf1) and (input2 is in2mf1) and (input3 is in3mf3) and (input4 is in4mf2) then (output is out1mf8) (1)
9. If (input1 is in1mf1) and (input2 is in2mf1) and (input3 is in3mf3) and (input4 is in4mf3) then (output is out1mf9) (1)
10. If (input1 is in1mf1) and (input2 is in2mf2) and (input3 is in3mf1) and (input4 is in4mf1) then (output is out1mf10) (1)
11. If (input1 is in1mf1) and (input2 is in2mf2) and (input3 is in3mf1) and (input4 is in4mf2) then (output is out1mf11) (1)
12. If (input1 is in1mf1) and (input2 is in2mf2) and (input3 is in3mf1) and (input4 is in4mf3) then (output is out1mf12) (1)
13. If (input1 is in1mf1) and (input2 is in2mf2) and (input3 is in3mf2) and (input4 is in4mf1) then (output is out1mf13) (1)
14. If (input1 is in1mf1) and (input2 is in2mf2) and (input3 is in3mf2) and (input4 is in4mf2) then (output is out1mf14) (1)
15. If (input1 is in1mf1) and (input2 is in2mf2) and (input3 is in3mf2) and (input4 is in4mf3) then (output is out1mf15) (1)
16. If (input1 is in1mf1) and (input2 is in2mf2) and (input3 is in3mf3) and (input4 is in4mf1) then (output is out1mf16) (1)
17. If (input1 is in1mf1) and (input2 is in2mf2) and (input3 is in3mf3) and (input4 is in4mf2) then (output is out1mf17) (1)

Figure 17: Some Rules of the ANFIS' Model

In the ANFIS modeling, for scenarios 1, 2, and 3 without PCA, two triangular membership functions are assigned to each input by the hybrid training method so that the fuzzy rules are connected to a membership function and the membership function forms the output. For the scenarios with PCA, 3 triangular membership functions are also used. The results of each scenario are given in Table 6 and Figure 18.

Table 6: Statistical analysis of adaptive neuro-fuzzy inference system (ANFIS) with and without PCA method

Output	**Without PCA**					**With PCA**				
	Train	Test	Validation	RMSE	%MAPE	Train	Test	Validation	RMSE	%MAPE
	R					R				
BOD	0.9997	0.8924	0.9258	0.68	3.12	0.973	0.9621	0.9803	1.16	6.91
COD	0.9953	0.9787	0.7113	1.73	3.22	0.9762	0.9697	0.9464	2.09	5.7

TSS	0.9947	0.8269	0.8234	0.83	5.79	0.9919	0.9148	0.9512	0.49	3.93

Figure 18 shows a diagram comparing the output data of an adaptive neural fuzzy system with and without PCA with the output of a wastewater treatment plant.

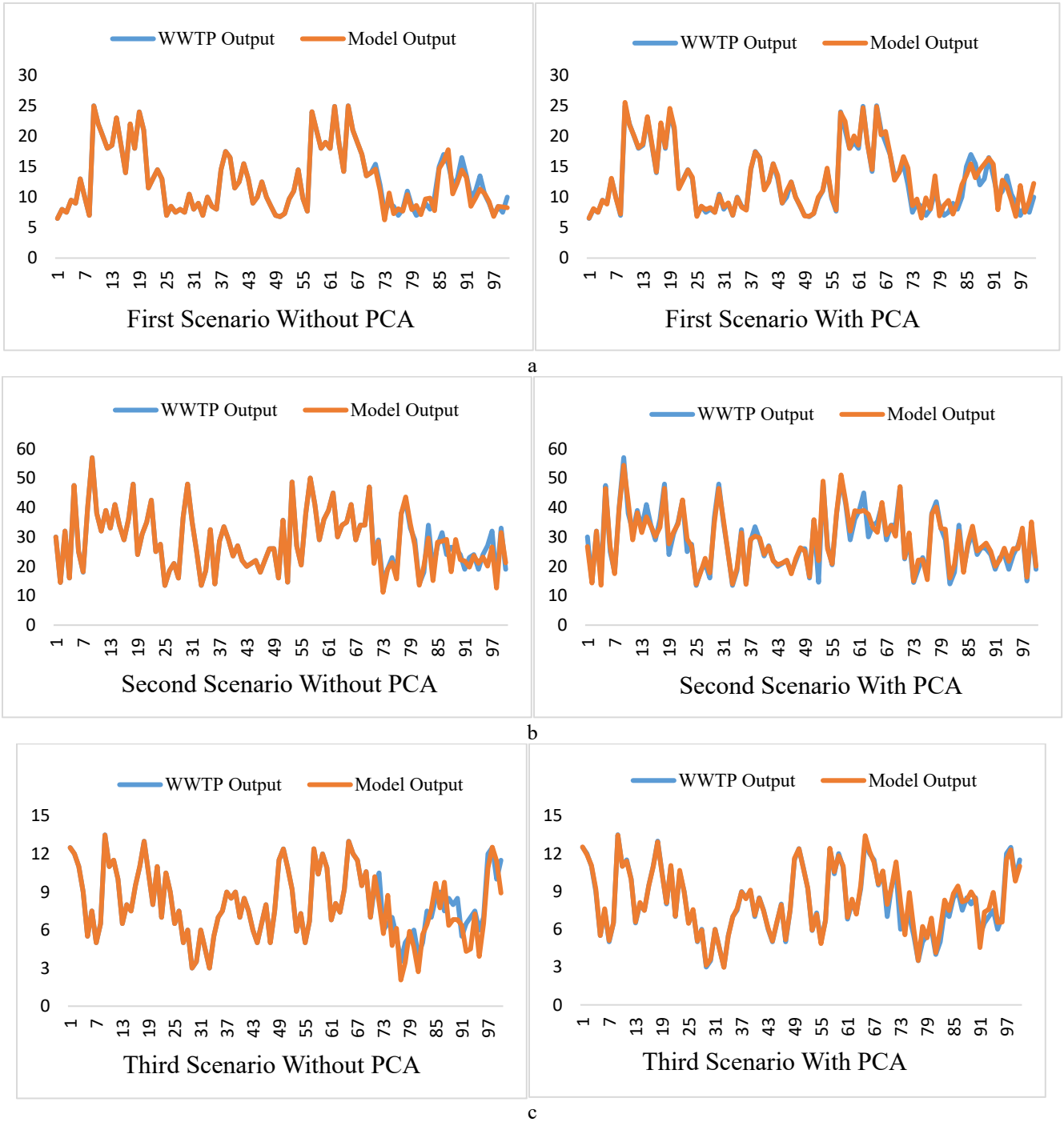
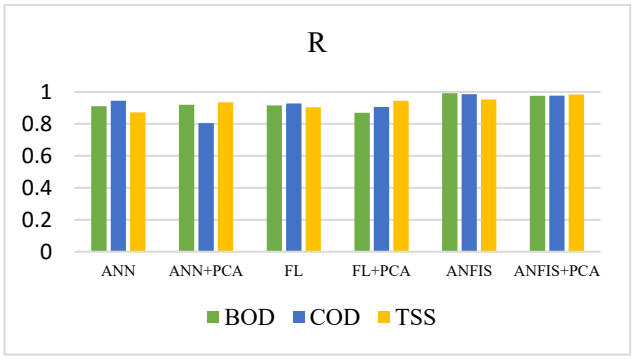


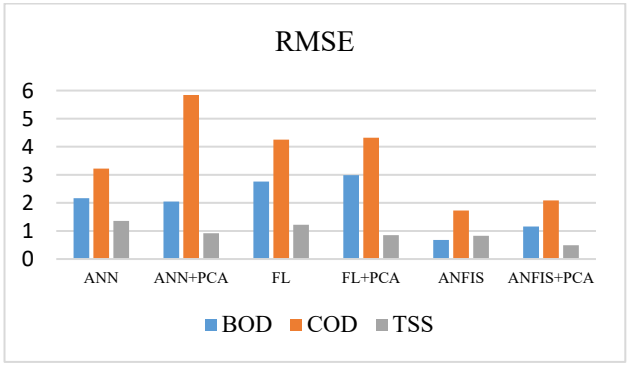
Figure 18: Comparison diagram of experimental values with the values predicted by ANFIS for: a) BOD b) COD c) TSS

In general, the results obtained from both methods with and without PCA were acceptable; although, in the first and second scenarios, the results obtained from the modeling of ANFIS without PCA method,

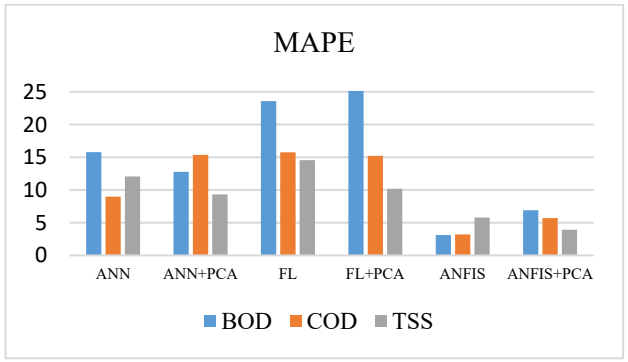
with some differences, were better results. It can be generally concluded that the ANFIS has the ability to predict changes in wastewater treatment efficiency and offers a promising performance. According to Figure 19, it can be seen that ANFIS has a better performance and yeilds better results compared to the other models.



a



b



c

Figure 19: Prediction ability of Models: a) R (b) RMSE (c) MAPE

As can be seen in Figure 20, the values of the output data from the modeling are very close to the output of the treatment plant in all three diagrams, which indicates that artificial intelligence models can be used as a suitable tool for predicting output for the existing inputs.

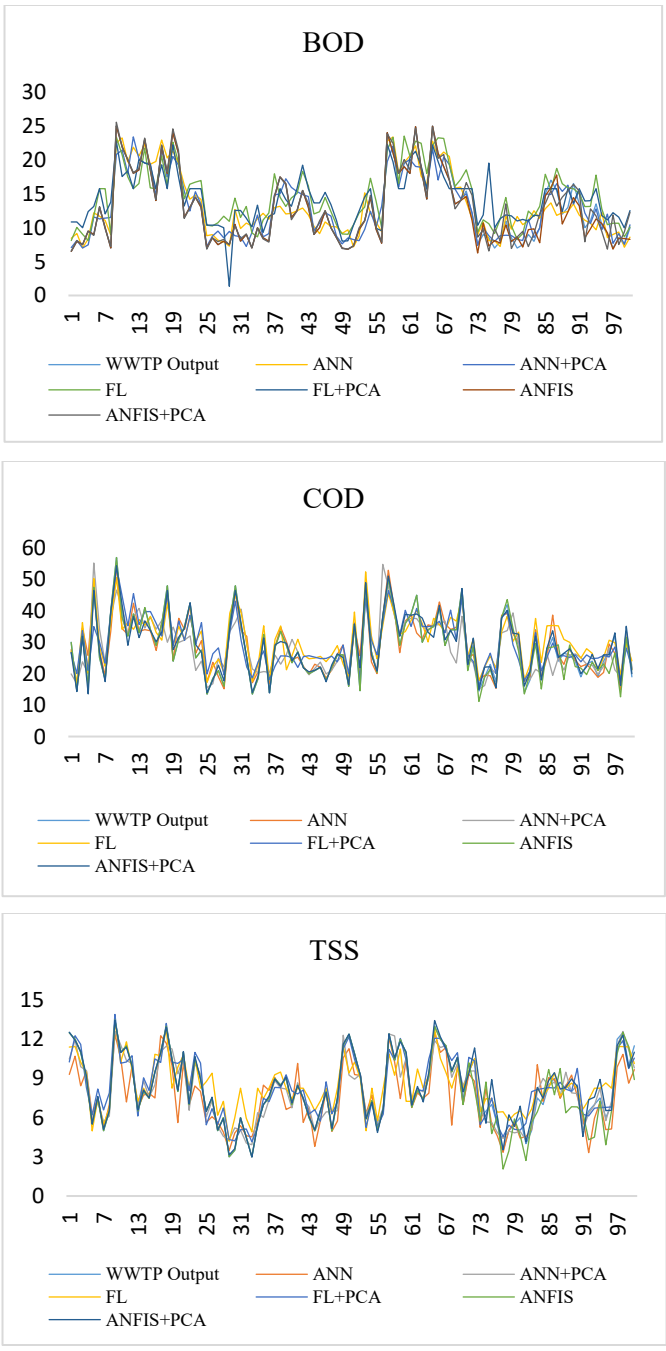


Figure 20: Comparison Diagram for the output data of the models with the treatment plant

7- Comparing the percentage of Contaminants Removal Efficiency

Finally, the removal efficiency index was used to evaluate the output of the treatment plant and the results of artificial intelligence models in separating and reducing the studied features. This index denotes the percentage of separation of each of the quality features of wastewater after treatment.

This relationship was applied in two cases, one based on the data and measured information of each contaminant at the input and output of the treatment plant and the other one, based on the data and information measured at the input and estimated from the artificial neural network. Finally, the

contaminant reduction efficiency was calculated and compared separately for each contaminant in all cases. According to Table 7, the highest removal efficiency is related to TSS, with a return of 96.68%

Table 7: Comparison of the removal efficiency of each of the studied quality features in the treatment plant output with the artificial neural network model

Unit of measurement		BOD	COD	TSS
WWTP	Input	220.64	368.84	245.76
	Observed	12.76	28.77	8.13
	removal efficiency	94.20	92.09	**96.68**
ANN	Input	220.64	368.84	245.76
	Estimated	13.27	29.12	7.60
	removal efficiency	93.94	91.98	96.90
ANN+PCA	Input	220.64	368.84	245.76
	Estimated	13.06	28.25	8.01
	removal efficiency	94.03	92.16	96.73
FL	Input	220.64	368.84	245.76
	Estimated	14.56	30.62	8.48
	removal efficiency	93.36	91.56	96.53
FL+PCA	Input	220.64	368.84	245.76
	Estimated	14.17	29.75	8.30
	removal efficiency	93.54	91.76	96.61
ANFIS	Input	220.64	368.84	245.76
	Estimated	12.66	28.45	7.95
	removal efficiency	94.25	92.19	96.75
ANFIS+PCA	Input	220.64	368.84	245.76
	Estimated	13.02	29.02	8.28
	removal efficiency	94.08	92.01	96.61

According to Figure 21, the selected models are proper for predicting performance of the treatment plant, and due to the proximity of the contaminant removal efficiency in the models with that of the wastewater treatment plant, it can be concluded that they also have an acceptable performance.

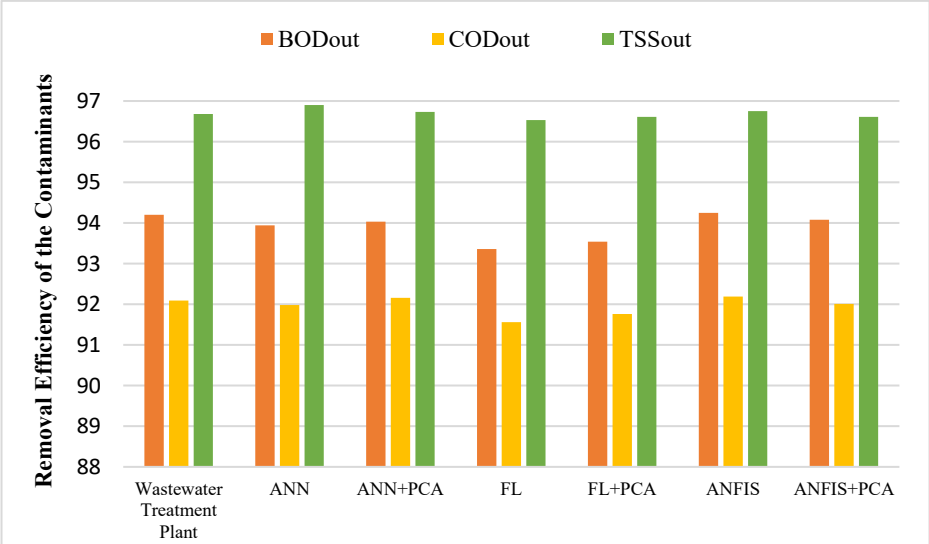


Figure 21: Comparison of Removal Efficiency of the Contaminants

8- Conclusions

As mentioned, in this study, artificial intelligence network (AIN) models have been used. The artificial neural network, fuzzy logic and adaptive fuzzy neural system were included in the methods utilized in this study.

Initially, after data collection, in order to increase accuracy, the data were reviewed and processed, which includes obtaining out-of-range data (outliers) and principal component analysis (PCA). In the next step, it was time to select three scenarios with corresponding inputs and outputs. Finally, modeling was performed on the selected scenarios and the final performance of each of them was measured using performance evaluation and predicting criteria and the results were compared. The most notable conclusions are as follows:

1- The neural network system is highly accurate and is a simple and accurate method but the adaptive fuzzy neural system is more appropriate in terms of training speed and accuracy of results. However, the fuzzy logic model can make the process of modeling difficult due to its time consuming training, while adaptive fuzzy neural system does not have this problem, but it can limit the number of inputs in ANFIS model, which increases the number of rules and complicates the model structure. Consequently, this can be deemed as a problem. In short, each one of the models is selected with respect to the desired objective.

2- Artificial intelligence models can be used as a convenient tool for predicting output over existing inputs.

3- The used models have high accuracy, which is valid with and without PCA. In addition, the use of multivariate PCA statistical method for data processing, although has negatively affected the prediction performance of some models with an average error of 3.615% , the ability to predict all models, even with PCA based on training and validation performance, remained very high.

4- Evaluating the results of the predictions using statistical indicators (R, RMSE, MAPE) showed that the ANFIS model was more accurate than the other two models.

5- The ANFIS model was more accurate than the FL model with an average error reduction of 13.92% and an ANN reduction of 8.22% compared to the ANN model, which was valid with and without PCA.

6- By calculating the removal efficiency in the treatment process, it is observed that all three contaminants have high removal capability, which indicates the efficiency of the system used in this treatment plant.

7- It was observed that by calculating the percentage of contaminant removal efficiency in the treatment plant's output, it is determined that the maximum removal efficiency is related to the TSS and is equal to 96.68%. Other contaminants also had values close to the TSS.

8- The selected models were suitable for predicting performance of the treatment plant and due to the proximity of the contaminant removal efficiency in the models with the wastewater treatment plant, it can be concluded that the models used in this study have a great performance.

9- Finally, the results obtained in this study reveal that the use of artificial intelligence network models to predict the quality of effluent from municipal wastewater treatment plants is possible and is a simple, accurate, efficient and reliable method.

References

[1] Row, D. Abdelmagid, I. (1995). *Handbook of Waste Water Reclamation and Reuse*. Lewis Publisher, by CRC press INC.

[2] Hamed, M. Khalafallah, M.G. Hassanein, E.A. (2004). Prediction of wastewater treatment plant performance using artificial neural network. *Environmental Modeling and Software*. Vol 19, pp: 919–928.
[3] Hong, Y.T. Rosen, M.R. Bhamidimarri, R. (2003). Analysis of a municipal wastewater treatment plant using a neural network-based pattern analysis.*Water Research*, Vol 37, pp: 1608–1618.
[4] Ouyang, Y. Nkedi-Kizza, P. Wu, Q.T. Shined, D. Huang, C.H. (2006). Assessment of seasonal variations in surface water quality. *Water Res*, 40, pp: 3800-3810.
[5] Camdevyren, H. Demyr, N. Kanik, A. Keskyn, S. (2005). Use of principal component scores in multiple linear regression models for prediction of Chlorophyll- in reservoirs. *Ecol. Modell.* 181, pp: 581-589.
[6] Gontarski, C.A. Rodrigues, P.R. Mori, M. Prenem, L.F. (2000). Simulation of an industrial wastewater treatment plant using artificial neural network. *Computer & Chemical Engineering*, 24, pp: 1719-1723.
[7] Mahmoud, S. Nasr, Medhat, A.E. Moustafa, Hamdy, A.E. Seif, Galal El Kobrosy. (2012). Application of Artificial Neural Network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-EGYPT. *Alexandria Engineering Journal,* 51, PP: 37–43.
[8] Abba, S.I. Elkirana, G. (2017). Effluent prediction of chemical oxygen demand from the astewater treatment plant using artificial neural network application. *Procedia computer science,* 120, pp: 156-163.
[9] Abbaa, S.I. Hadia, S.J. Abdullahia, J. (2017). River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques. *Procedia Computer Science*, 120, pp: 75–82.
[10] Elkiran, G. Nourani, V. Abba, S.I. Abdullahi, J. (2018*).* Artificial intelligence-based approaches for multi-station modelling of dissolve oxygen in river. *Global J. Environ. Sci. Manage*., 4(4), pp: 439-450.
[11] Djeddou, M. Aouatef, H. Loukam, I. (2018). Wastewater Treatment Plant Performances Modelling Using Artificial Neural Networks. *The international water association ecotechnologies for wastewater treatment*. Western University, London, Ontario, Canada, pp: 25-27.
[12] Abba, S.I. Abdulkadir, R.A. Gaya, M.S. Saleh, M.A. Esmaili, P. Jibril, M.B. (2019). Neuro-fuzzy ensemble techniques for the prediction of turbidity in water treatment plant. *2nd International Conference of the IEEE Nigeria Computer Chapter*, pp: 1-6.
[13] Alsulaili, A. Refaie, A. (2021). Artificial neural network modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance. *Water Supply,* 21(5), pp: 1861–1877.
[14] Tarun Gupta, (2019). Data Preprocessing in Data Mining & Machine Learning. *Towards Data Science.*
[15] Abba, S.I. Elkiranb, G. Nouranic, V. (2021). Improving novel extreme learning machine using PCA algorithms for multi-parametric modeling of the municipal wastewater treatment plant. *Desalination and Water Treatment*, 215, pp: 414-426.
[16] Chandak, A. Lee, W. Stamp, M. (2021). A Comparison of Word2Vec, HMM2Vec, and PCA2Vec for Malware Classification. *Publisher: Springer International Publishing*.
[17] Real-Olvera, J. Morales-Rivera, J. Patricia, González-López, A. Sulbarán-Rangel, B. Zúñiga-Grajeda, V. (2020). Adsorption of Organic Pollutants from Cold Meat Industry Wastewater by Electrochemical Coagulation: Application of Artificial Neural Networks. *Water*, Vol. 12(11).
[18] Qasaimeh, A. Al-Ghazawi, Z. (2020). Regression modeling for rapid prediction of wastewater BOD_5. *Desalination and Water Treatment*, 201, pp: 165-172.
[19] Gaya, M.S. Abba, S.I. Abdu, A.M. Tukur, A.I. Saleh, M.A. Esmaili, P. Wahab, N.A. (2020). Estimation of water quality index using artificial intelligence approaches and multi-linear regression. *IAES International Journal of Artificial Intelligence (IJ-AI)*, Vol. 9, No. 1, pp: 126-134.
[20] El-Nasr, M.S. Nguyen Dinh, T. H. Canossa, A. Drachen, A. (2021). *Game Data Science*. Published to Oxford Scholarship Online.
[21] Mariani, M.C. Tweneboah, O.K. Beccar-Varela, M.P. (Edition November 2021). *Data Science in Theory and Practice*, p: 400.
[22] Hanteh, M. Malek, H. Akbari, M. (2022). Development of the integrated MARS-PSO and ELM-PSO models for estimating the compressive strength of concrete in circular concrete columns confined with FRP polymer fibers. *Scientific and Research Journal of Civil Engineering Sharif*.

[23] Hanteh, R. Hanteh, M. Kheyroddin, A. & Rezaifar, O. (2020). Determination of Strength Parameters in Roller Compacted Concrete (RCC) Dams using Laboratory Results and Forecasting based on Artificial Neural Networks. *Modares Civil Engineering journal*, 20(2), pp: 55-70.
[24] Akbari, M. & Henteh, M. (2019). Comparison of Genetic Algorithm (GA) and Particle Swarm Optimization Algorithm (PSO) for Discrete and Continuous Size Optimization of 2D Truss Structures. *Journal of Soft Computing in Civil Engineering*, 3(2), pp: 76-97.
[25] Naderpour, H. Rafiean, A.H. Fakharian, P. (2018). Compressive Strength Prediction of Environmentally Friendly Concrete Using Artificial Neural Networks. *Journal of Building Engineering*, Vol 16, pp: 213-219.
[26] Naderpour, H. Sharei, M. Fakharian, P. Heravi, M.A. (2022). Shear Strength Prediction of Reinforced Concrete Shear Wall Using ANN, GMDH-NN and GEP. *Journal of Soft Computing in Civil Engineering* 6 (1), pp: 66-87.
[27] Ghanizadeh, A.R. Ziaee, A. Khatami, S.M.H. Fakharian, P. (2022). Predicting Resilient Modulus of Clayey Subgrade Soils by Means of Cone Penetration Test Results and Back-Propagation Artificial Neural Network. *Journal of Rehabilitation in Civil Engineering* 10 (4), pp: 146-162.
[28] Naderpour, H. Fakharian, P. (2018). Predicting the Torsional Strength of Reinforced Concrete Beams Strengthened with FRP Sheets in terms of Artificial Neural Networks. *Journal of Structural and Construction Engineering* 5 (1), pp: 20-35.
[29] Naderpour, H. Rezazadeh Eidgahee, D. Fakharian. P. Rafiean A. H. Kalantari. S.M. (2020). A New Proposed Approach for Moment Capacity Estimation of Ferrocement Members Using Group Method of Data Handling. *Engineering Science and Technology, an International Journal*, Volume 23, Issue 2, pp: 382-391.
[30] Naderpour, H. Nagai, K. Fakharian, P. Haji, M. (2019) .Innovative Models for Prediction of Compressive Strength of FRP-Confined Circular Reinforced Concrete Columns Using Soft Computing Methods. *Composite Structures*, Vol 215, pp: 69-84.
[31] El Alaoui, M. (2021). *Fuzzy TOPSIS*. Boca Raton, 1st Edition, p: 216.
[32] Vilela, M. Oluyemi, G. (2022). *Value of Information and Flexibility*. Springer International Publishing.
[33] Mansouri, I. Kisi, O. (2015). Prediction of debonding strength for masonry elements retrofitted with frp composites using neuro fuzzy and neural network approaches. *ELSAVIER, Composites Part B: Engineering*, Vol. 70, pp: 247-255.
[34] Manu, D. S. Thalla, A.K. (2017). Artificial intelligence models for predicting the performance of biological wastewater treatment plant in the removal of Kjeldahl Nitrogen from wastewater. *Appl Water Sci* 7, pp: 3783–3791.
[35] Zhu, S.L. Heddam, S. Nyarko, E.K. Hadzima-Nyarko, M. Piccolroaz, S. Wu, S.Q. (2019). Modeling daily water temperature for rivers: comparison between adaptive neuro-fuzzy inference systems and artificial neural networks models. *Environ. Sci. Pollut. Res*, 26, pp: 402–420.