https://doi.org/10.26682/csjuod.2023.26.2.17

Journal of University of Duhok., Vol. 26, No.2 (Pure and Engineering Sciences), Pp 166-178, 2023
4th International Conference on Recent Innovations in Engineering (ICRIE 2023) (Special issue)

PREDICTION OF BOD INDEX IN WASTEWATER TREATMENT PLANT USING GENETIC ALGORITHMS AND NEURAL NETWORKS

HEBA AL JADDOU*, **MOHAMAD BASHAR ALMOFTI**** and **DIALA SHEHAB*****

*College of Civil Engineering, University of Damascus, Syria

**College of Civil Engineering, University of Damascus, Syria

***College of Engineering, University of Wadi, Syria

(Accepted for Publication: November 27, 2023)

ABSTRACT

The biochemical oxygen demand (BOD is defined as the rate at which microorganisms use oxygen in water or wastewater during the fixation of biodegradable organic matter under aerobic conditions. The use of BOD as a necessary parameter for the effective control and monitoring of the wastewater treatment plants remains somewhat restricted due to the long time it takes by traditional ways, which hinders its use in real time. In this paper, a hybrid model of genetic algorithms (GA) and artificial neural networks (ANN) is used for the prediction of biochemical oxygen demand. The data used in this research was collected over ten years through the daily records of the Homs waste water treatment plant. the model was built based on the approval of each of the values of (COD, SS, Q) as inputs to predict the value of the BOD, and the performance of the model was evaluated by adopting an inverse validation error for selecting the best network structure in addition to other differential criteria. The optimal structure of the neural network was determined after a number of attempts and errors, and the results showed a high efficiency of the proposed hybrid model of (genetic algorithms and neural networks) by predicting the value of inflow BOD . As a result of this research, a neural network structure was selected to predict the value of the BOD indicator which is (2-54-1) using the Hyperbolic Tangent function in the hidden layer and the logistic function in the output layer, the Levenberg-Marquardt was used as a training algorithm for training, The value of the performance function was 0.125, and the average error value of the three groups was 1.83, while the mean maximum error of the three groups was 5.86, the value of the correlation coefficient was 0.99.

KEYWORDS: Biochemical oxygen demand, Genetic algorithm, Neuronal network, Wastewater treatment plant.

1. INTRODUCTION

Biological oxygen demand is defined as the rate at which microorganisms use oxygen in water or wastewater during the fixation of biodegradable organic matter under aerobic conditions. During the decomposition process, the organic materials present in the water form food for the bacteria, and the energy required for the bacteria is formed through the oxidation of these materials.[1]

The biological digestion of organic matter takes up to 25 days, and in order to save time in conducting the experiment, it is sufficient to determine the value of BOD after five days, and this value is called BOD_5. There is a relationship between BOD_{tot} and BOD5, and the value of BOD5 is equivalent to about 68.4% of $BOD_{tot.}$[2]

drhebajaddou@gmail.com ,
mbalmofti@gmail.com
diala89shehab@gmail.com

In any case, the use of BOD as a necessary parameter for the effective control and monitoring of the sewage treatment process remains somewhat restricted, due to the long time it takes to experiment with BOD, which impedes its immediate use, for example, the introduction of large quantities of organic pollutants into a river causes a continuous collapse of its content The dissolved oxygen in it, and the real diagnosis of the pollution problem in the river will not take place before five days, and therefore any corrective action will come late, because the traditional method for determining BOD requires incubation for a period of 5 days to determine BOD_5[3]

Accordingly, it became necessary to find a fast and accurate way to find an inferential predictive model to determine the BOD. [4]

This procedure would eliminate the delay in measuring the BOD using laboratory methods. Due to the time it takes to obtain the BOD parameter using laboratory methods, there is a tendency to dispense with this indicator for the purposes of control and monitoring in treatment plants and to replace it with the COD indicator.

However, as it is known, unlike the COD index, the BOD index refers to degradable organic pollutants, and therefore the BOD is the most important parameter in determining water's susceptibility to biological treatment [5].

Previous attempts to determine the BOD according to the COD met with limited success due to the wide changes that are usually observed in the relationship between the two water quality variables. Despite the speed of measuring the COD index, it cannot be used as a substitute for the BOD index in the control and monitoring processes of treatment plants operating with activated sludge. [5]

Mazen Hamada et al at Al-Azhar University in the Gaza Strip 2018 modeled the basic water quality parameters in the main wastewater treatment plant in Gaza using artificial neural networks (ANN) and multiple linear regression (MLR). Data which used in this study. Were used Over nine years of the station's monthly records, the following parameters were adopted as inputs to the various models (PH, T, COD, BOD, TSS), while the outputs are (BOD, COD, TSS). The model's performance was evaluated by adopting (RMSE, R), the optimal structure of the neural network was determined after several trials and errors. The results showed the superiority of neural network models (ANN) over multiple linear regression (MLR) models in modeling these parameters. The study also showed that both temperature and suspended solids have a significant impact. It predicts the studied parameters more than the rest of the parameters [6]

In Egypt, Mahmoud S. Nasr et al 2012 studied the application of an artificial neural network (ANN) to predict the performance of the Agami station in Alexandria. Neural network models were built to predict COD, BOD5, and TSS based on data collected for one year. The study concluded that the correlation coefficient reached (0.9, 0.85, and 0.77) values, respectively. [7]

2. the aim of research:

Building a model based on artificial intelligence applications to predict the BOD parameter that enters Homs wastewater treatment plant.

3. Methods and Materials:

- Genetic Algorithms (GAs):

Genetic algorithms are among the important technologies in the random search for the optimal solution, and they are a representation of the prevailing belief that human intelligence is created with humans, and is largely acquired through inheritance. It is a simulation of the mating process between organisms of the same species. [8]

3.1.1. Basic components of genetic algorithms:

- The method of coding the solution (chromosome) to suit the issue at hand.
- The target function (fitness function) is used to

evaluate solutions.

- Genetic processes (selection, crossing over, and mutation). [9]

3.2 Artificial Neural Networks (ANN):

It is a computational technique designed to simulate the way that human brain performs a specific task, and it is made up of simple processing units. These units are neurons or nodes, which have a neural property, as they store scientific knowledge and experimental information to make them available for use by adjusting weights [10].

Figure (1) shows the mechanism of action of an artificial neuron and its basic components. (ANN) processes data in parallel, which provides high speed in performance that enables it to solve complex problems that include many hypotheses and variable information, quickly and effectively.

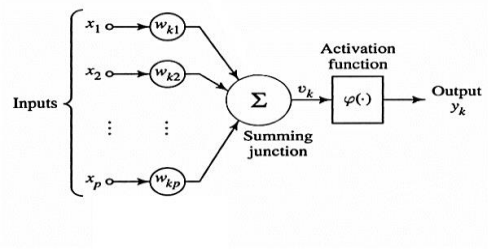

Fig.(1):- Artificial neuron and its basic components

There are many artificial neural networks used in processing patterns of data, and each of these types has a specificity in the architecture and the mechanism of information processing through the number and type of nodes in each layer in addition to the type of activation functions and the mechanism for adjusting weights. Backpropagation Feedforward which has gained a lot of interest in the field of weather prediction is chosen in this research [11]

2.2.1 *Basic components of an artificial neural network:*

Neural networks consist of the following basic components, or at least some of them

These elements are: (input layer - output layer - hidden layers – interconnections-weights) [12].

3.3 The description of Homs wastewater treatment plant:

The study was conducted at the wastewater treatment plant for the city of Homs in the village of Al-Duwair, which is about 6 km north of the city center and within the scope of the city's residential expansion, extending over an area of 240,000 m² and at an average level of 470 m above sea level.

The purpose of this treatment plant is to treat the wastewater of the city of Homs for agricultural purposes and reduce pollution from the Orontes River, where the treated water is discharged into the river. It works with traditional activated sludge technology.

The start of operation and investment for the station was on 12/16/1998, and it has been operating properly since 2000, Figure (2) shows an aerial view of the sewage treatment plant in Homs city

Table (1) shows the design parameters of the wastewater treatment plant for the city of Homs in 2022.

[13]

Table(1):- Design parameters of the wastewater treatment plant for the city of Homs in 2022

Qav	250000 m³/day
BOD	311.59 mg / l
COD	811.41 mg / l
SS	532.90 mg / l
Population	700000 cap

Fig.(2):- Aerial view of the sewage treatment plant in Homs city

3.4 Data source:

Sewage samples were collected from the sewage treatment plant of Homs city, and the study included the parameters of the Homs wastewater treatment plant influent, which are (BOD, COD, SS, Q)

This data was collected based on treatment plant records in the period from 2010 to 2020

4. Results and Discussion:

In this research, Alyuda NeuroIntelligence 2.2 (577) program was used, which is a software application for designing neural networks.

The first step in training neural networks is to enter data digitally by importing it in the form of columns from the "excel.csv" file, which included (Q; COD; SS; BOD), as shown in Figure (3).

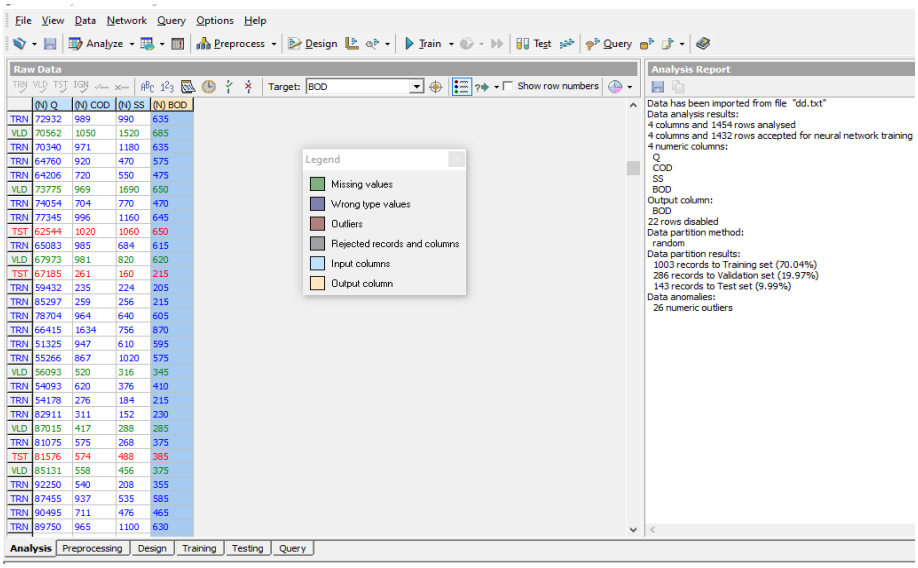

Fig.(3):- Data analyses tab in aluyda

The Alyuda NeuroIntelligence program contains six main steps to build an ANN model, which are shown in Figure 6. They include: data analysis, pre-processing, network architecture design, training networks, and validation.) and testing it as shown in the figure below(4):

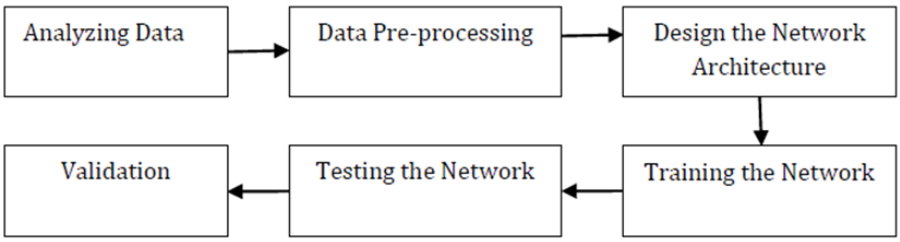

Fig.(4):- The steps of model building

4.1 -Analyzing Data:

This step allows defining the types of columns, detecting anomalies in the data, dividing the data into groups, specifying the target column for the neural network, and excluding specific rows and columns, which can be explained as follows:

 Separation of abnormal data that negatively affect network performance. Where the abnormal data is divided into two parts known as Outliers and Missing Values, and the missing value is an unknown value that is considered empty cells, while outliers are heterogeneous values that are far from the majority of the column data, and they can be just extreme cases or errors Measurement or other anomalies, which prevent proper training of the neural network and greatly degrade the performance of the neural network, so the entire log is discarded if there is an outlier or missing value in the input columns.

 Data Partition randomly into three groups:

• Training set (70%) of the data (equivalent to 1003 records).

• Validation set (20%) of the data (equivalent to 286 records).

• Test set (15%) of the data (equivalent to 143 records).

Figure (5) shows the mechanism of division by selecting the percentage or specifying the number of records so that one of them is calculated automatically through the program, knowing that the data for each group is determined randomly and automatically

drhebajaddou@gmail.com ,
mbalmofti@gmail.com
diala89shehab@gmail.com

(Automatically) or manually (Manually), and the most common random method is used An ability to search for the pairs of data that most reflect the time series of the studied data.

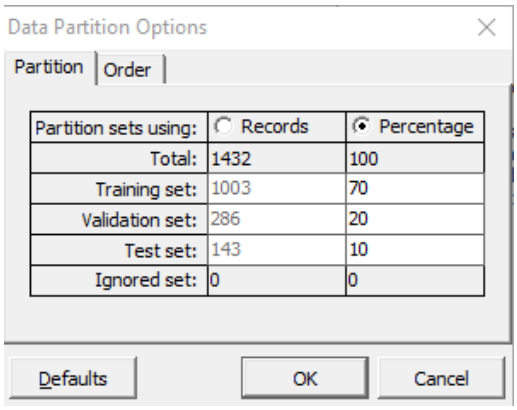

Fig.(5):- Dividing the data into groups

4.1.1 *Determine the inputs and outputs of the (ANN) model:*

This study aims to predict the values of (BOD), and accordingly we identified it as the target column (Target) and the output of the (ANN) model, and with regard to the possible income patterns of the (ANN) model, it is not enough to rely on the correlation matrix (Pearson correlation) shown in Table (2) to choose Suitable income patterns for predicting (BOD) values. Also, it is not easy to test all the possible income patterns according to the variables (Q; COD; SS). In this study there are seven patterns (3 single income, 3 binary income, 1 triple income), especially with The presence of many variables of the (ANN) model of the activation functions of the number of hidden neurons, the training algorithm, the selection of which depends on trial and error.

Table(2):- Pearson correlation between (Q; COD; SS; BOD)

	BOD	Q	cod
Q	-0.158		
COD	0.989	-0.153	
SS	0.726	-0.137	0.661

Reducing the number of explanatory variables that are used in the final model reduces effort and time as well as ensures ease of analysis and understanding. Therefore, there must be a balance between the process of reducing the explanatory variables and increasing their number to obtain accurate predictive results, so it is better to choose the model with the least number of variables. Explanatory variables so that these variables contribute and have an impact on the dependent variable

The study aimed to exclude useless input columns that do not contribute significantly to network performance through the use of genetic algorithms.

4.1.2. *Applying geniting algorithm:*

Geniting algorithm was applied to choose the best input data after adopting the following parameters figure (6):

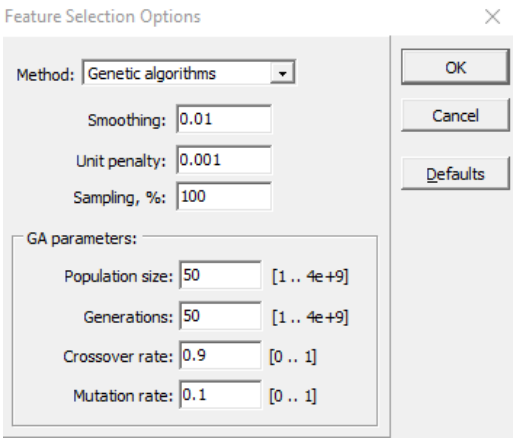

Fig.(6):- Genetic algorithm parameter

The performance function (Fitness) used in the comparison between configurations depends on calculating the network error resulting from the test set, and Figure (8) shows the most prominent proposed configurations as inputs to the model, where the number (0) indicates ignoring the input element, and the number (1) indicates the adoption of the input element as a variable affecting the output of the model.

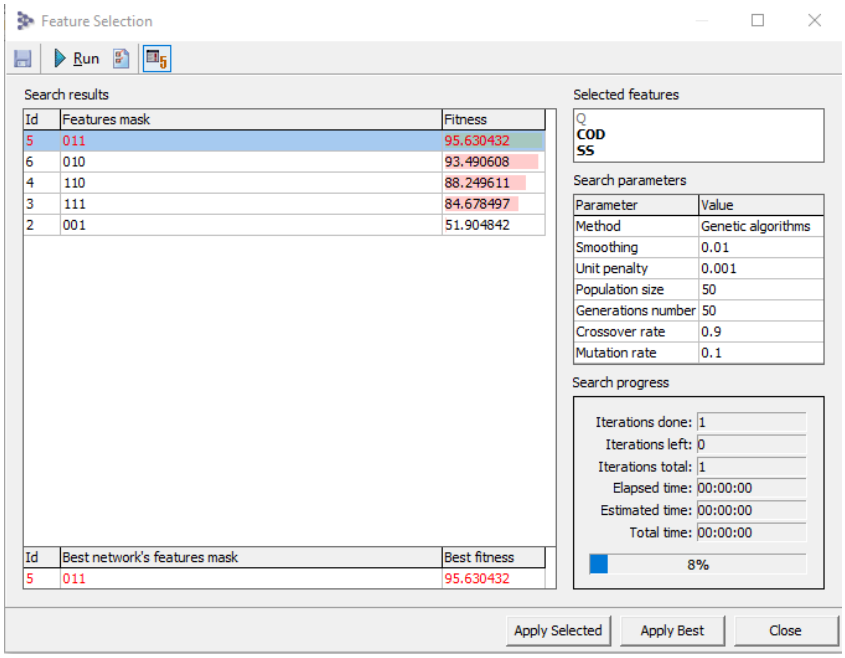

Fig.(8):- Top five proposed input configurations using the genetic algorithm.

The proposed configuration according to Figure (9) indicates the preference for using (COD; SS) as inputs for the model that aims to predict the variable (BOD) We notice a decline in the performance of the input configurations with a noticeable difference after the aforementioned model (111), and accordingly, the input configurations referred to in Table (3) will be adopted.

Table(3) :-ANN Model

Model	Feature mask	Input variables
I	011	BOD=f (COD+ss)
II	010	BOD=f (COD)
III	110	BOD=f (Q+COD)
IV	111	BOD=f (Q+COD+ss)

4.2 Preprocessing Datasets:

This stage is based on pre-processing the data before feeding the artificial neural network, and this is done through scaling or normalization of the numerical values in proportion to the artificial neurons that have a specific range of operating values. , as shown in Table (4).

Table(4) :-Scaling range for activation functions

Output layer activation function	Scaling range
Linear	**[-1.1]**
Logistic	**[0 .1]**
Hyperbolic Tangent	**[-1.1]**

• Normalization is performed in the model automatically according to equation (1)

$$S_f = \frac{SR_{max} - SR_{min}}{X_{max} - X_{min}} \ldots\ldots\ldots(1)$$

$$X_p = SR_{min} + (X - X_{min}) * S_f \ldots\ldots(2)$$

X: true value & $Xmin$: minimum true value & $Xmax$: true maximum value & $SRmin$: actual minimum of scaling range & $SRmax$: actual maximum of scaling range & Sf : scaling factor & XP: the value after processing.

According to most of the studies and research findings found in the literature, the dipole [-1..1] domain of the Hyperbolic Tangent is used for the input columns, and the domain [0..1] of the logistic function for the output column and table (5) Shows the statistical properties of the variables and the standardization coefficient for each of them.

Table(5) :-The statistical features for used data

Parameter	Q(m^3)	COD(g/m^3)	SS(g/m^3)	BOD(g/m^3)
Column type	Input	Input	Input	Output
Scaling range	[-1..1]	[-1..1]	[-1..1]	[0..1]
Min	51325	235	84	205
Max	132250	2710	2500	1295
Mean	88368.21	1062.529	852.6683	638.0831
Std. deviation	12264.31	454.3504	446.4969	205.4915
Scaling factor	0.000025	0.000808	0.000828	0.000917

4.3 Designing the network:

Designing network consists two steps :

a-designing network properties:

the network properties were as followed : A single-layer neural network with the hyperbolic tangent function in the hidden layer, the logistic function in the output layer, and the error sum of squares (MSE) function to evaluate the performance of the model according to the three groups.

b- Network Architecture:

Experimental research can be used in the absence of information about the degree of complexity of the studied problem, and to facilitate the research process, by nominating a group of designs within a specific range of the number of neurons in the hidden layer between (20) and (100) neurons, and adopting the inverse validation error criterion for the total investigation.) as a critical criterion (Fitness) for selecting network architecture in addition to other differentiating criteria; With repeated calculations (Iteration = 300) when running the model (Retrain = 1) for one time, figure (9) shows the results that was obtained for the structure of the proposed 12 best networks, where it was found that the best structure is ANN (2-50-1), ie There are two neurons in the input layer, 50 neurons in the hidden layer, and one neuron in the output layer.

ID	Architecture	# of Weights	Fitness	Train Error	Validation Error	Test Error	Correlation	R-Squared
1	[2-20-1]	81	0.096	9.195	10.386	7.612	0.997	0.993
2	[2-100-1]	401	0.125	8.109	8.030	7.677	0.998	0.996
3	[2-69-1]	277	0.107	9.437	9.322	8.733	0.998	0.995
4	**[2-50-1]**	**201**	**0.131**	**7.531**	**7.643**	**7.138**	**0.998**	**0.996**
5	[2-38-1]	153	0.087	10.423	11.471	9.026	0.997	0.993
6	[2-61-1]	245	0.125	7.743	8.002	7.481	0.998	0.996
7	[2-45-1]	181	0.099	10.119	10.052	9.337	0.997	0.994
8	[2-56-1]	225	0.095	10.459	10.554	9.737	0.997	0.994
9	[2-53-1]	213	0.090	11.016	11.104	10.223	0.997	0.993
10	[2-48-1]	193	0.098	10.164	10.191	9.600	0.997	0.994
11	[2-51-1]	205	0.103	9.793	9.712	9.193	0.997	0.995
12	[2-49-1]	197	0.129	7.675	7.740	7.182	0.998	0.996

Fig.(9):- Results of an empirical search for the best neural network architecture.

The comprehensive search can now be used to accurately determine the structure of the neural network by narrowing the search scope to a number of neurons between (40) and (60) neurons and with a transition step (2) neurons, as it was found that the best structure for the neural

network is ANN (2-54-1]), as shown in Figure (10).

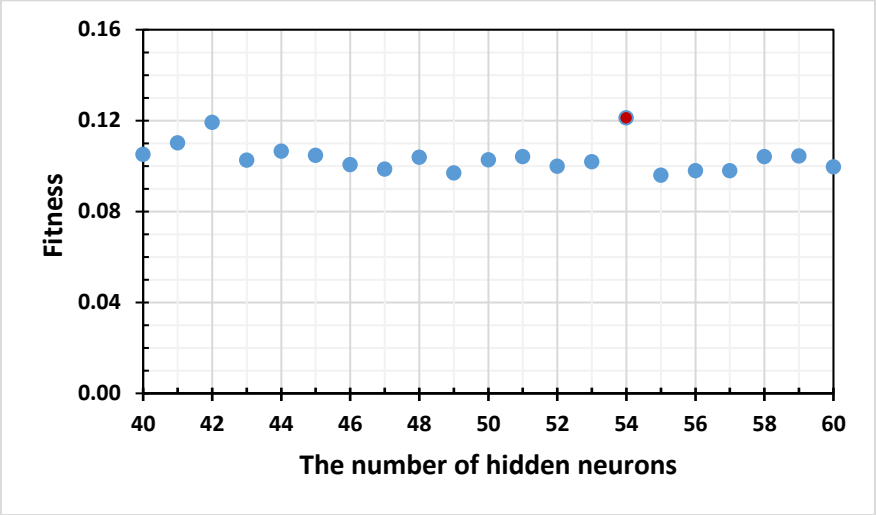

Fig.(10):- Determining the neurons in the hidden layer

Figure (11) shows the absolute error regression for the best artificial neural model ANN (2-54-1) proposed to represent the values of (BOD), where the network errors value for the training, investigation and test group reached the values (5.85; 8.24; 7.17), respectively. , with a correlation coefficient of (0.9978).

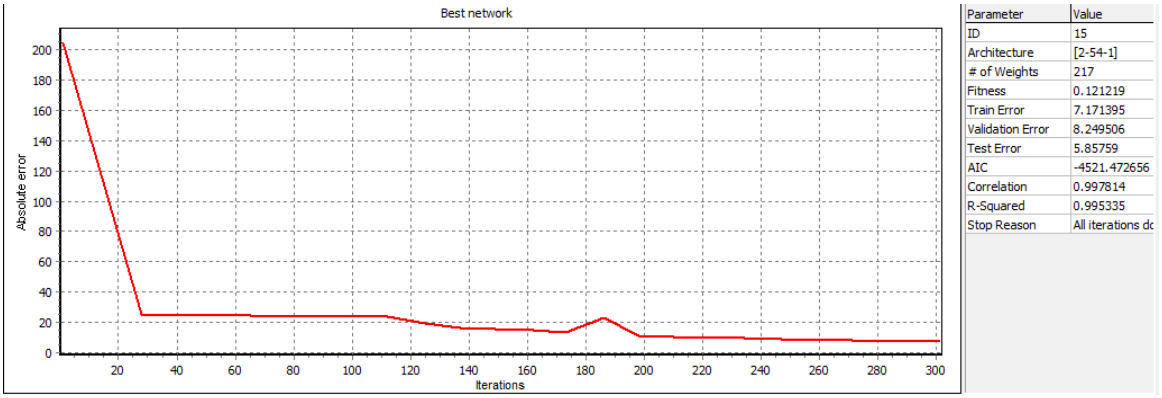

Fig.(11):- Absolute error regression of the ANN model (2-54-1)

4.4 Networks Training:

The network is trained by providing the training data prepared in the first step and by selecting the best ANN (2-54-1) structure that was determined in the previous step. Training starts from initializing weights, and changes with repeated calculations towards improving network performance and achieving the lowest possible error, with determining the learning rates and momentum, and choosing the training stop criteria to avoid excessive training. Training will stop when the MSE reaches a value of (0.01) or training is completed at 1000 iterations, whichever occurs first. Also, the average absolute error (Average Absolute Error) and the network error (Sum-of-Squares) will be calculated according to the chosen training algorithm, and compared between them in terms of accuracy and speed in performance (number of iterations per second), as shown in figure (12), which It shows the superiority of the (LM) algorithm over other algorithms in terms of

(MSE) less for the three groups (training, investigation, test), and the correlation coefficient close to one.

Training algorithm	Avg training error	Avg validation error	Avg test error	Correlation
Quick Propagation	7.314	7.157	7.007	0.99857
Conjugate Gradient Descent	1.609	1.804	1.549	0.99994
Quasi-Newton	1.385	1.534	1.411	0.99997
Limited Memory Quasi-Newton	1.880	1.920	1.706	0.99992
Levenberg-Marquardt	**1.365**	**1.497**	**1.393**	**0.99997**
Online Back Propagation	5.238	5.534	4.825	0.99926
Batch Back Propagation	1.880	1.920	1.706	0.99992

Fig.(12):- Correlation coefficient, mean squared errors (MSE) for training algorithms

In general, the statistical criteria used are appropriate to some extent, but not sufficient due to the large sample size and the variation in the values, which in turn affects the result. To ensure the previous comparison, we made a comparison between the greatest errors contained in each group, as shown in figure (13), which also indicated This indicates the superiority of the (LM) algorithm despite the large data volume, which leads to a decrease in the learning speed of about (2.5 repetitions / sec), which is a slow rate compared to other algorithms. Network training in the event that there is no improvement in the performance of the network after a certain number of iterations.

Training algorithm	Iterations	Training speed	Max training error	Max validation error	Max test error
Quick Propagation	2006	56.67	60.227	55.439	46.443
Conjugate Gradient Descent	2002	4.25	19.715	20.107	5.913
Quasi-Newton	2002	6.10	10.227	12.422	4.100
Limited Memory Quasi-Newton	2002	21.88	21.164	6.499	9.120
Levenberg-Marquardt	**345**	**2.50**	**4.417**	**9.146**	**4.096**
Online Back Propagation	2001	44.26	46.365	44.657	34.196
Batch Back Propagation	2002.000	52.50	134.182	126.902	112.962

Fig.(13):- Comparison of the accuracy (Max Error) and performance speed of the algorithms.

4.5 Testing the neural network:

Observing the error distribution during network training helps in understanding the change in the performance of the neural network, which is supposed to improve the accuracy of the model by reducing the number of large errors and increasing the number of small errors. Through Figure (12), the frequency distribution of the series of errors can be observed The proposed model, which refers to the amount of agreement between the real values of (BOD) and the predicted values using the model, which can be illustrated by the test group graph in Figure (14).

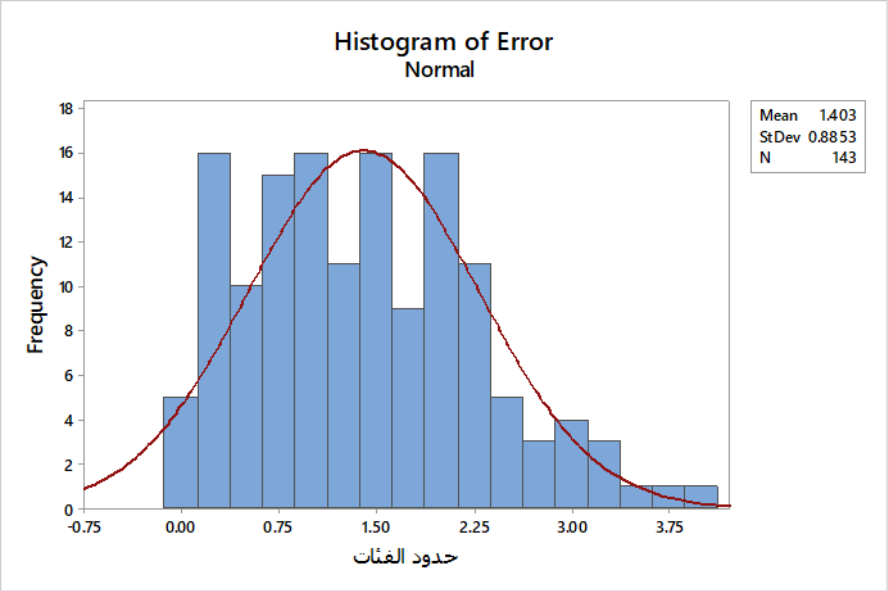

Fig.(14):- Frequency distribution of ANN errors (2-54-1) using the (LM) algorithm

5. CONCLUSIONS

Applying the previously mentioned algorithms gave highly reliable results, as the R^2 value reached 99%, while the value of R^2 was equal to 17% when the Multiple linear regression was applied [14]

The results obtained when building the model demonstrate the ability of smart models to overcome the nonlinear nature of relationships within the activated sludge system, in addition to the ability of these systems to discover relationships between the inputs and outputs of each model.

Modeling using neural networks is characterized by simplicity, generalization and efficiency, which encourages its wider use in activated sludge processes. However, the models built using artificial intelligence need to be re-adjusted and calibrated when the case study changes.

REFERENCES

Alireza Bahdori Scott.T. Smith Dictionary of Environmental and wastewater Treatment,Springer 2016 pp 485.

Yan Zheng Liu; Zhiyuan Chen Prediction of biochemical oxygen demand with genetic algorithm-based support vector regression water quality reaserch journal ,May 2023

Mohanty, S.; Scholz, M.; Slater, M. J. ,2002. Neural Network Simulation of the Chemical Oxygen Demand Reduction in a Biological Activated Carbon Filter, J.Ch. Inst. Wat. Environm. Managem.

[4]Sandeep K Sunori; P Bhakuni Negi; A Rana; A Mittal; P Juneja Prediction of Biological Oxygen Demand using Artificial Intelligence & Machine Learning , 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) 24-26 Nov. 2022

Rumana Riffat and Taqsim , Fundamentals of wastewater Treatment and Engineering ,Second Edition,USA,2022,PP 88 ,89.

Mazen Hamada, Hossam Adel Zaqoot,*, Ahmed Abu Jreiban, Application of artificial neural networks for the prediction of Gaza wastewater treatment plant performance-Gaza strip, Journal of Applied Research in Water and Wastewater 9(2018) 399-406.

Mahmoud S. Nasr, Medhat A.E. Moustafa, Hamdy A.E. Seif, Galal El Kobrosy . Application of Artificial Neural Network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-EGYPT . Alexandria Engineering Journal, In Press, Corrected Proof, Available online 29 August 2012.

غبن، زينب كريم؛ محمد، دعاء جاسم، حلِ الكينماتيكا العكسية للإنسان الالي باستعمال الخوارزمية الوراثية, 2009 , 1-24.

حيدر، بادية، -2010 البحث عن الشكل الأمثل للمقاطع العرضية في الجيزان رقيقة الجدران، رسالة دكتوراه، جامعة تشرين، سورية, 186.

Kim, M. N.; Kwon, H. S., 1999. Biochemical oxygen demand sensor using Serratia marcescens LSY 4. Biosen. Bioelectr., 14 (1), 1-7.

Scholz, M., 2006. Wetland Systems to Control Urban Runoff; Elsevier:

Amsterdam, The Netherlands.

Mazen Hamada, Hossam Adel Zaqoot,*, Ahmed Abu Jreiban, Application of artificial neural networks for the prediction of Gaza wastewater treatment plant performance-Gaza strip, Journal of Applied Research in Water and Wastewater 9(2018)

Azzar Tony, Studying the effect of fine screens in improved the efficiency of treatment in Homs wastewater treatment plant,Master thesis ,ALbaath university,2013.

jaddou,Heba ,The modeling for operation of activated sludge treatment plant using artificial intelligence techniques

"case study –Homs wastewater treatment plant" phd thesis, Damascus university, Syria ,2022