Prediction of Chemical Oxygen Demand from The Chemical Composition of Wastewater by Artificial Neural Networks

Basim Ahmed Saleh¹, Hakan Kayi²
¹North Gas Company, Energy Authority, Technical Planning Section, Kirkuk, Iraq.
²Ankara University, Chemical Engineering Department, 06100 Ankara, Turkey.

Email: bsbs7791@gmail.com

ABSTRACT. In our era, many technical applications are being used. Artificial Neural Networks (ANNs) as one of the artificial intelligence tools have emerged to learn and discover a model of dynamic nonlinear. In this study, six input parameters were taken to predict the value of the Chemical Oxygen Demand (COD) in the wastewater before and after the treatment at the North Gas Company/Kirkuk, by using the standard back propagation algorithm. The network was trained with the 150 data collected from the quality indices of the untreated and treated wastewater, such as total chloride ions Cl⁻, nitrate ions NO₃⁻, phosphate ions PO₄³⁻, sulfate ions SO₄²⁻, ammonia NH₃, Biochemical Oxygen Demand (BOD₅) to predict one element, that is the COD. After properly training of the neural network, it was tested by using the test data, and the best results were selected by the consideration of the mean square error and the regression coefficient, where the best result appeared before wastewater treatment is 0.98235 and the best result after wastewater treatment is 0.99999. The findings of this study suggest that artificial neural networks are accurate and effective tools for predicting the COD values of treated wastewater.

Keywords: Artificial neural networks, industrial wastewater, chemical oxygen demand prediction, chemical composition.

1 INTRODUCTION
Chemical oxygen demand may be defined as the amount of (dissolved) oxygen required to oxidize and stabilize (organic and inorganic content of) the sample solution. It is used to measure the content of oxidizer organic as well as inorganic matter of the given sample of waters [1].

Chemical oxygen demand is an important measure of the amount of oxygen required to break down pollutants in water.

Where Chemical oxygen demand is measured using several methods, directly or indirectly. Its value is regularly determined by samples that are analyzed in the laboratory. These methods can be time consuming, the COD value is an essential parameter in the biological treatment processes[2].
The artificial neural network is a way to predict the chemical oxygen demand value by previous knowledge of the amount of each water quality parameter within the wastewater samples taken from and analyzed inside the laboratory at the North Gas Company / Kirkuk. Another helpful approach is the use of the water quality parameters in an equation and multiplying each parameter by a factor obtained from the multiple linear regression model according to the data of the experimental analysis. Hence, we can find the value of the COD without using chemical materials and analyses, and with less effort and time.

The objective of this study is to get the amount of the chemical oxygen demand (COD) without performing analyzes in the laboratory, because:

First: The wastage of the chemicals used. Chemical materials are used to analyze the amount of the chemical oxygen demand before treatment and after treatment. This is an expensive task for the company since it spends money on this analysis.

Second: This analysis takes approximately 4-5 hours to complete and to get the results, in addition to the effort exerted by the person in charge of this analysis. The analysis is taken place by adding the concentrated sulfuric acid and adding potassium dichromate and silver sulfate solution, and then heating the solution for two hours, then it will be cooled and a detector will be added. This process takes a long time, effort and cost. In the case of the absence of chemical materials for the analysis of the sample, or in the case of failure of the heating and the condensation device being used, we should dispense with this examination and the test will be failed.

Vijayan [3] used ANN models to predict wastewater treatment plant variables. This method utilizes ANNs to predict influent and effluent COD, BOD, and total suspended solids (TSS) for the effluent treatment process. Feed-forward neural network with back propagation learning algorithm, was utilized to obtain BOD, COD and TSS values in the effluent. Different network architectures were tried and the best one produced an RMSE to be 0.0984 and a regression value to be 0.99959.

Arabameri [4] used an ANN model to predict COD from landfill leachate samples which obtained from a municipal landfill site in Shahrood, Iran. All leachate samples tested by using an ultrasonic process within 2 days after taking samples from leachate lift stations. In order to calculate the MSE and R² values, Levenberg–Marquardt backpropagation algorithm is applied in training. In this study, it has been found that the predicted COD results are in line with the experimental data with the R² which equal to 0.992, and the MSE which equal to 0.000331 at epoch 31.

Abba [5] applied a feed-forward neural network model and multiple linear regression (MLR) models to predict the COD in a wastewater treatment plant in Nicosia, North Cyprus. Input parameters of ANNs are taken to be chemical oxygen demand, biochemical oxygen demand, pH, TSS, total nitrogen, total phosphates, conductivity, SS and output neuron is the COD. The best results appeared at epoch 203, the R² is equal to 0.7034 and RMSE is equal to 0.0108.

Hamada [6] used ANN models which are multi-layer perceptron (MLP) model, to predict major water quality parameters in Gaza wastewater treatment plant, the input parameters were temperature, pH, COD, BOD, and total dissolved solids. The output parameters were COD, BOD, and total dissolved solids. The best results in MLP to predict COD are R equal 0.7594 and at epoch 3 the MSE equal 59.48.

Areeerachakul [7] used an ANN model to predict the COD from 11 sample sites in Bangkok, the input of the model to predict chemical oxygen demand (COD). Each record contains 13 parameters which are biochemical oxygen demand (BOD5), dissolved oxygen, temperature, pH, suspended solids (SS), hydrogen sulfide, ammonia nitrogen, nitrate nitrogen, nitrite nitrogen, total Kjeldahl nitrogen, total coliform, total phosphorous, and COD, the best results are obtained regression equal 0.89 and the root mean square error equal 15.16.
2 METHODOLOGY

The methodology is separated into two phases. The first one includes the work in the laboratory to measure the number of pollutants in industrial waters, for example, nitrate ions (NO$_3^-$) chloride ions (Cl$^-$), phosphate ions (PO$_4^{3-}$), sulfate ions (SO$_4^{2-}$), ammonia (NH$_3$), BOD$_5$, COD and to know amount of each pollutant in industrial wastewater.

In the second phase, artificial neural networks were used to predict the COD in the wastewater. The main objective of this study is to use the ANNs to predict the COD in the wastewater using the data obtained for 6 different water quality parameters. The data were obtained from an industrial wastewater treatment plant in North Gas Company, Kirkuk. In general, artificial neural networks provide highly accurate predictions in finding of both linear and nonlinear relations among the data points, hence, they are considered as the main predictive tools in this study.

2.1 Analyses in Laboratory (The First Phase) [8]

1. Nitrate ions (NO$_3^-$)
2. Chloride ions (Cl$^-$)
3. Phosphate ions (PO$_4^{3-}$)
4. Sulfate ions (SO$_4^{2-}$)
5. Ammonia (NH$_3$)
6. Biochemical oxygen demand BOD$_5$
7. Chemical oxygen demand COD

2.2 ANN Modeling (Second Phase)

ANN processes the information by working on the input data, and through the training of the network with experience, weighed signals will produce the outputs of hidden layer and the output layer.

Through continuous training, communication between the units, i.e. the input layer, the hidden layer and the output layer is improved until the error in the predictions is reduced and the network reaches the required level of accuracy.

In this study, the MATLAB 2017 program and the two sets of data are used:
The inputs ‘unnamed’ are a 150x6 matrix representing static data: 150 samples of 6 elements.
The targets ‘unnamed’ are a 150 x 1 matrix representing static data: 150 samples of 1 element.

It is selected the six inputs of the model and Table 1 indicated the input layer and Table 2 indicated the output layer for samples, below.

| No. | Input Sample                                      |
|-----|---------------------------------------------------|
| 1   | Nitrate ions (NO$_3^-$)                           |
| 2   | Chloride ions (Cl$^-$)                            |
| 3   | Phosphate ions (PO$_4^{3-}$)                      |
| 4   | Sulfate ions (SO$_4^{2-}$)                        |
| 5   | Ammonia (NH$_3$)                                  |
| 6   | Biochemical oxygen demand BOD$_5$                 |

Table 1. Input layer for samples

| No. | Output sample                                      |
|-----|---------------------------------------------------|
| 1   | Chemical oxygen demand (prediction of COD is the aim of this study) |

Table 2. Output layer for samples
3 RESULTS AND DISCUSSION

In this study, the results for before and after wastewater treatment cases are discussed.

3.1 Before Wastewater Treatment

3.1.1 The Network Performance before Treatment

After finding the number of hidden nodes, the evaluation of the network performance was achieved.

![Network performance before treatment](chart)

**Figure 1.** Network performance before treatment

The performance of the mean square error (MSE) has been used to evaluate the accuracy achieved while testing the ANN in the estimation of input for the given data. That means that a better prediction can be reached when MSE value is low or around zero, which means the training performance is without error. Figure 1 shows a plot between target values and the input values. The training performance reaches to a minimum value at the 8th iteration and the training process continues up to the iteration 14 then stops. Figure 1 also shows some problems for training, and the MSE performance at 42.2466 at 8th epoch is high.
3.1.2 The Regression Plots for Before Treatment

Regression plots are shown in Figure 2, which shows the relationship between the output-target values. If the training is a perfect fit, the output values and target values will be the same, and this means $R=1$, representing an exact linear regression between the outputs and the targets. The regression coefficient found for all data to be 0.98235; and this finding indicates that input-output mapping obtained from the ANN is not a very good fit for this case. Therefore, it can be considered that the regression among the outputs obtained from the network and the targets (desired outputs) not exact.

Figure 2. Regression plots for before treatment.

Error Histogram of Before Treatment

The blue bars indicate to training data, the green bars indicate to validation data, and the red bars indicate to testing data in Figure 3. This figure exhibits the histogram that can give us an indication of outliers and the distribution of the training. Figure 3 clearly expresses that the validation and test errors are quite high, and the error distribution results are not reasonably well, because the most of the 150 total data are in errors, and most of the errors have fall between 48.76 and -13.24 which is a quite wide range.
3.1.4 Formula for predicting the result of COD before treatment
By using POLYMATH 6.10 software, simplest formula to predict the COD to give constant factors multiplied by the variables of the laboratory data are defined as [9].

\[
COD = 0.938811 \times cl^{-} + (-0.2839987) \times NO_{5}^{-} + (-0.0592047) \times PO_{4}^{3-} + 0.0221107 \\
\times SO_{4}^{2-} + 2.061704 \times NH_{3} + 0.8517708 \times BOD_{5} \tag{1}
\]

The value of \( R^2 \) equals to 0.87507 and the value of RMSD is 1.002001.

3.2 After Wastewater Treatment

3.2.1 The Network Performance after Treatment
After finding the hidden nodes, the evaluation of the network performance was achieved.
The performance of the mean square error (MSE) has been used to evaluate the accuracy achieved while testing the ANN in the estimation of input for the given data. Figure 4 shows the relationship plot between the target and the input values. The training MSE performance value reaches to the minimum at the 8th iteration and then continues up to the 31st iteration then stops. Figure 4 shows a good performance, and the best validation performance at 25th epoch is obtained to be 0.00017594 which is close to zero.
3.2.2 The Regression Plots for After Treatment

The regression plots in Figure 5 show the relationship between the outputs of the network and the target values. As we mentioned above, if the training is perfect fit, the output from the ANN and the target will have the same value. This means R=1, which represents an exact linear regression between the outputs and the targets. The regression coefficient found to be 0.99999; this indicates that there is a perfect suitability of data (fit). So, it can be considered that the results are acceptable with reasonable errors and it’s very close to targets, and the regression between the outputs and the targets is almost exact.

Error Histogram of After Treatment

Figure 6 gives indication of outliers and exhibits the distribution of error for training, validation and test steps. Figure 6 clearly expresses that the validation and test results are very well and the results for error distribution are very reasonable. In addition, a good indication of the values of outliers is provided in this figure. However, the obtained outliers do not actually affect the performance because little errors were found out of 150 total data and most errors fall between 0.09713 and -0.2316. These errors are considered acceptable and satisfied the training distribution results.
Figure 6 Error histogram after treatment

3.2.4 Formula for predicting the result of COD after treatment
By using POLYMATH 6.10 software, simplest formula to predict the COD to give constant factors multiplied by the variables of the laboratory data are defined as [9].

\[
COD = 0.0104494 \times c_l^- + 0.2102446 \times NO_3^- + 6.406522 \times PO_4^{3-} + (-0.0044695) \times SO_4^{2-} + (-0.4085615) \times NH_3 + 0.8465617 \times BOD_5
\]

(2)
The value of \( R^2 \) equals to 0.999991 and the value of \( \text{RMSD} \) is 0.0017172.

4 CONCLUSIONS
The conclusions of the present work and some considerations relating to data analysis, model performance policy, and future recommendations are presented.
The general results of this study are:

ANN is a useful tool for predicting chemical oxygen demand. ANN units predicted chemical oxygen demand were developed with a perfect expectation based on the regression coefficient \( R \), where \( R \) equals 1 show a close relationship (that is, they are perfectly adequate fit), \( R=0 \) a random relationship (not fit). Regression value \( R \) indicates the correlation between outputs and targets.

Mean squared error, MSE, is the average squared difference between outputs and targets, as the value of MSE goes low it will be better, and the MSE= 0 means no error.

4.1 Before wastewater treatment:
Some large errors had obtained from artificial neural network after the entering of input and output parameters, biochemical oxygen demand, ammonia \( NH_3 \), phosphates ions \( PO_4^{3-} \), nitrates ions \( NO_3^- \), sulfates ions \( SO_4^{2-} \) and chloride ions \( Cl^- \) and chemical oxygen demand, because these parameters had taken from the samples of the wastewater before the treatment procedures (oil separation, addition of chemicals such as aluminum sulfate and floatation aid 5165, and filtration) applied.
Regression coefficient R value was 0.98235, and the MSE value was 43.02203 when all data is considered, which is not an acceptable good performance indeed.

In addition, the validation performance obtained at epoch 8 was obtained to be 42.2466 which is a quite high value.

In the error histogram for before treatment case, most errors fall between 48.76 and -13.24. These errors are considered to be very high.

This is why the results of the equation of COD-result prediction that had developed basing on previous laboratory results that applied in polymath 6.10 program had showed errors. And the results of sampling before treatment were obtained for $R^2$ to be 0.87507 and for RMSD to be 1.002001.

In overall, performances of both artificial neural network modeling and multiple linear regression approach are a bit poor. The high errors in this part may originate from the unpredictable nature of the untreated wastewater since there can be some other dominating factors and parameters that affect chemical oxygen demand significantly, other than the six parameters considered here.

4.2 After wastewater treatment:

Some positive changes were found in the artificial neural networks modeling when the parameters, chemical oxygen demand, biochemical oxygen demand, ammonia $\text{NH}_3$, phosphate ions $\text{PO}_4^{3-}$, nitrate ions $\text{NO}_3^-$, sulfate ions $\text{SO}_4^{2-}$ and chloride ions $\text{Cl}^-$ had used from a sample of the wastewater after the application of treatment processes such as oil removal and passing the wastewater through the filters and other several processes of treatment mentioned before.

The straight line shows a direct relationship between the target and the output.

The regression coefficient, R, for all data is found to be 0.99999 and the mean square error, MSE, is to be 0.000441. These findings indicated that prediction results are very close to the actual values obtained from the experimental analyses.

In addition, the best performance for validation is obtained to be 0.0017594 at 25th epoch, this result is very well.

The error histogram obtained for after treatment case indicated that the most of the errors fall between 0.09713 and -0.2316. These errors are considered very low and the performance of the artificial neural network is very satisfactory.

Our multiple linear regression approach yielded the coefficients of the COD equation as a function of six variables based on the laboratory analyses, and after the evaluation of the performance of this equation it was found out that $R^2$ is 0.999991 and RMSD is 0.0017172. Hence, even MLR equation yields highly accurate results for the prediction of COD for after wastewater treatment case.

In summary, findings of this study suggest that artificial neural networks, as one of the artificial intelligence tools, and multiple linear regression model equation may contribute to laboratories or operators in similar industrial plants in the prediction of the value of COD by saving time, effort and money and without waste of chemicals.
References

[1] A. Tasnim, “Estimation of Chemical Oxygen Demand in WasteWater using UV-VIS Spectroscopy,” Sch. Mechatron. Syst. Eng., pp. 10–18, 2015.

[2] V. Kimstach, “Selection of Water Quality Variables,” Water Qual. Assessments, vol. 5, no. 1989, 1996.

[3] A. Vijayan and G. S. Mohan, “Prediction of Effluent Treatment Plant Performance in a Diary Industry Using Artificial Neural Network Technique,” J. Civ. Environ. Eng., vol. 6, no. 6, 2016.

[4] M. Arabameri, A. Javid, and A. Roudbari, “Artificial neural network (ANN) modeling of cod reduction from landfill leachate by the ultrasonic process,” Environ. Prot. Eng., vol. 43, no. 1, pp. 59–73, 2017.

[5] S. I. Abba and G. Elkiran, “Effluent prediction of chemical oxygen demand from the astewater treatment plant using artificial neural network application,” Procedia Comput. Sci., vol. 120, pp. 156–163, 2017.

[6] M. Hamada, H. A. Zaqoot, and A. A. Jreiban, “Application of artificial neural networks for the prediction of Gaza wastewater treatment plant performance-Gaza strip,” J. Appl. Res. Water Wastewater, vol. 9, no. 1, pp. 399–406, 2018.

[7] S. Areerachakul, “The Using Artificial Neural Network to Estimate of Chemical Oxygen Demand,” Int. J. Chem. Mol. Nucl. Mater. Metall. Eng., vol. 7, no. 7, pp. 578–581, 2013.

[8] R. Koch, Annual Book of ASTM Standards,” part 31, American Society for Testing and Materials. Philadelphia, 1982.

[9] I. Rivals and L. Personnaz, “A statistical procedure for determining the optimal number of hidden neurons of a neural model,” Second Int. Symp. Neural Comput., pp. 23–26, 2000.