Comparison Between Multiple Regression Analysis and Artificial Neural Networks for Predicting Pollution Levels in the Euphrates River

Fahad Hussein Enad[1]

Dr. Hussain Ali Abbedllah Aleagobe[2]

University of Dhi -Qar / Department of Studies and Planning Email: fahadh@utq.edu.iq

University of Thi-Qar - College of Administration and Economics - Department of Financial and Banking Sciences Email: hussain@utq.edu.iq

Abstract

Contemporary society is seeing a rising concern over environmental issues, arising as a result of continued deterioration and changes due to non-sustainable usage of natural resources; one of the biggest problems in this respect relates to the indiscriminate dumping of all kinds of wastes. These practices, therefore, have highly wounded the aquatic environment as there is a daily exposure of huge amounts of waste exposed, which results in high-level pollution that could result in damage to aquatic ecosystem services. The Euphrates River is on the top list of key bodies water used by a human for a wide variety of functions that support life; although it faces some challenges pertaining to pollution resulting from domestic and industrial and agricultural wastes.. In this context, the research aims to predict the level of pollution in the Euphrates River water over an eight-week period in 2025, using modern techniques including and multiple regression analysis and artificial neural networks. A comparison was made between the two methods based on precise statistical criteria, most notably the mean squared error. The research problem highlighted that the available studies about the Euphrates River do not sufficiently focus on predicting pollution levels, representing an ecological menace that may have unhealthy effects on humans and other parts of the environment. The results showed that artificial neural networks have more accuracy and effectiveness compared to the multiple regression model, since it handles the non-linear relationships among the variables better and makes more accurate predictions. This study represents the importance of using neural networks as a powerful and reliable tool to research and evaluate environmental pollution and hence enhance the possibility of more sustainable decision-making for the protection of water resources.

Keywords / Multiple regression , Artificial neural networks, water pollution, mean squared error.

1-Introduction:

Forecasting has become one of the essential processes in the modern era that organizations and decision-makers cannot do without, due to its importance in reducing risk factors and developing effective future strategies and plans. Among the vital areas that require forecasting future values is the field of water pollution, which is considered the backbone of modern life and a pillar of the global economy. The global water crisis has significantly worsened, especially after climate changes. Anything that alters the quality of water, whether naturally or artificially, is considered water pollution. that makes it unsuitable for various important uses in life (such as drinking, industry, agriculture, fishing, etc.) [19]. It is also defined as any change in the basic characteristics, whether chemical, physical, or biological, which leads to problems and harm to human life and other living organisms [9]. The Euphrates River has long suffered from the accumulation of waste, the leakage of sewage into it, and pollutants from fertilizers and other contaminants due to the lack of plans to address these accumulations in a scientifically advanced manner. This phenomenon needs to be studied and appropriate methods for analysis and treatment should be sought through suitable statistical means With the increasing interest in the subject of forecasting in The advent of Artificial Neural Networks in recent years technology, which is one of the branches of artificial intelligence that has proven its high efficiency in forecasting, to evaluate its effectiveness in predicting pollution levels in the Euphrates River for the next eight weeks and compare it with Multiple Regression analysis, and to determine the effectiveness of artificial neural networks technology in predicting the pollution of the Euphrates River, as conducted by the researchers (Shamshiry et al., 2014) With the aim of improving waste management, enhancing public health, and reducing costs. The results showed that artificial neural networks outperformed Multiple Regression Analysis in terms of Accuracy and efficiency in predicting waste quantities[22]. In 2015, the researcher (Al-Aidani) conducted a study aimed at predicting groundwater levels in the regions of Al-Zubair and Safwan using artificial neural networks, relying on field data collected from 13 wells over the course of a year. (2013-2014). The results showed that the optimal model contains two hidden layers, each comprising 10 neurons, reflecting the efficiency of neural networks in this field.In 2017, researchers (Nabeel et al) using regression analysis and artificial neural networks (ANN) to forecast surface roughness. The best model, according to the results, has six neurones in the hidden layer., with a coefficient of determination of 94.93% for neural networks and 93.63% for regression analysis. It was also found that high feed rates lead to greater roughness and deviation, while low feed rates result in finer roughness and improved surface finish. In 2018, researchers Elifcan and Onur predicted electricity generation in Turkey (2010-2017) Regression analysis with artificial neural networks (ANN). The results showed the superiority of neural networks in prediction accuracy

compared to regression analysis. In 2020, researchers Joseph and Martins conducted a study on predicting properties Welded joints such as notch height, Brinell hardness number, and ultimate tensile strength utilising regression analysis and artificial neural networks (ANN). The results showed that regression analysis outperformed neural networks in prediction accuracy, recording fewer errors in all studied properties[1]. In 2022, (RAMAZAN) conducted an analysis of the impact of groove opening parameters on the Ti-6Al-4V alloy using a submerged electrical discharge machine. (EDM). The results showed the impact of pulse-on time and discharge current on surface roughness, tool wear, and material removal rate. Models were developed to predict the outputs using artificial neural networks and regression analysis, where neural networks outperformed in predicting the results.8] After that, (ANN) are trained in a manner similar to human learning through examples and training. Neural networks are organized into specific applications such as discrimination and perception models or data classification through the learning process. Learning in the biological system is used to adapt synaptic points, which is the main idea behind the operation of (ANN). Since the use of (ANN) does not require assumptions about the nature of the time series being linear or nonlinear, their use is considered beneficial in addressing the prediction discussed in this study.

2- The aim of the research

The research aims to study pollution problems in the Euphrates River and predict them, presenting the (ANN) method as a computational approach to determine the accuracy of estimating model parameters. Hence, the importance of the (ANN) method emerges to obtain more accurate estimates, the goal of the research is to provide specialists with a clear and comprehensive picture of pollution predictions in the coming months.

3-Research methodology

To achieve the research objective, a simplified definition of water pollution and water quality was relied upon, along with mentioning some of its main symptoms, explaining multiple regression and its characteristics, estimating the model parameters using estimation methods and some model tests, and the artificial neural networks approach.

4-Water

4-1- Water Quality:

Human activity, the environment, or both can have an impact on the quality of water, whether it is groundwater or surface water. Natural deposition of organic materials or salts from agricultural fields, wind-driven dust and gas deposition from the atmosphere, and hydraulic forces that can change the chemical and physical characteristics of water all have an impact on water quality without human

interference. Therefore, a lot of dissolved and partially dissolved chemicals are present in natural water.Numerous dissolved minerals and salts combine to generate critical compounds that support the health and vigour of species that rely on this ecosystem's constituent parts. Regarding the primary alterations in the physical and chemical characteristics of water, they are caused by human activities, which frequently have slow, undetectable impacts on the water system [21].

4-2-Water Quality Index (WQI)

Properly cleaning water sources will encourage the existence of a comprehensive and suitable system for the living organisms that need high-quality water. According to this concept, identifying the factors through which water quality can be determined is important for diagnosing the level of water pollution. Due to the numerous changes that can indicate water quality, Finding scientific ways to evaluate the enormous amount of data pertaining to water quality that are straightforward, easy to grasp, and provide fast, reliable results without delving into the interpretation of the factors indicating water quality separately has become essential. A mathematical method for reducing vast amounts of data on water quality to a single value (number, phrase, or term) that expresses a certain degree of water quality is the water quality index [23].

2-Methods and techniques

2-1- Multiple Linear Regression Model

The field of statistics is concerned with studying various phenomena through data analysis of the phenomenon under study, understanding the relationships between variables, or predicting the values of a variable or variables depending on the values of a set of influencing variables using statistical analysis models. These models are classified based on The quantity of variables included in the model into a simple model, which contains two variables: the variable that is dependent and the independent variable The independent variable and the other model is called the multiple linear model, which studies several variables within the model that follow the same distribution. From this concept, we can define multivariate statistical analysis as a set of statistical methods and techniques that deal with several variables at the same time for one or more samples or experimental units. These variables are not limited to a specific field but include all types of data according to their type. Thus, the mathematical form of the multiple regression model is [2],[4],[6],[20]:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + U_i \quad (1)$$

Since

β_0 : Intercept coefficient

$\boldsymbol{\beta_j}$: The number of independent variables in the model is reflected in the regression coefficients. (j=1,2,.....,k)

$\boldsymbol{U_i}$: is the error term (i=1,2,....,n)

The Maximum Likelihood Method can be used to estimate the multiple linear regression model. The error term is distributed as follows based on the well-known fundamental assumptions of the regression model:

$$\boldsymbol{U \sim N(\sigma_u^2 . I_n)}$$

And we had the following probability density function (p.d.f).

$$\boldsymbol{f(Y_1 . Y_2 \ldots . Y_n) = \left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^n exp\left\{-\frac{\sum_{i=1}^{n}[Y_i - E(Y_i)]^2}{2\sigma^2}\right\}} \quad (2)$$

In matrix form, the above equation can be rewritten as follows:

$$\boldsymbol{L = [2\pi\sigma_u^2]^{-\frac{n}{2}} exp\{-\frac{U\acute{U}}{2\sigma_u^2}\}} \quad (3)$$

$$\boldsymbol{L = [2\pi\sigma_u^2]^{-\frac{n}{2}} exp\{-\frac{1}{2\sigma_u^2}(y - X\beta)'(y - X\beta)\}} \quad (4)$$

By taking the natural logarithm of both sides, we obtain

$$\boldsymbol{LnL = -\frac{n}{2} ln2\pi - \frac{n}{2}\sigma_u^2 - \frac{1}{2\sigma_u^2}(Y'Y - 2\beta'X'Y + \beta'X'X\beta)\}} \quad (5)$$

After setting the equation to zero and differentiating with respect to (β), we obtain:

$$\boldsymbol{\hat{\beta} = (X'X)^{-1}X'Y} \quad (6)$$

2-1-2-Multiple model assumptions

Each type of multivariate statistical analysis method has its own specific assumptions, but there are some common characteristics such as:

The dependent variable is a linear function of the independent variables, a condition known as linearity of the relationship between variables. In this context, linearity refers to the parameters.The dependent variable is a linear function of the independent variables when the relationship between the variables is linear; in this case, linearity with respect to the parameters is being referred to.

The covariance and variance matrices are homogeneous.Covariance and variance matrices are homogeneous.The absence of the multicollinearity problem.

No aggregation errors in the independent variables.No aggregation errors in the independent variables.

The independent variables must be non-random.Independent variables must be non-random.

The absence of outliers.

2-1-3- Testing the parameters of the multiple linear model

Hypothesis testing is a case of evaluating the performance efficiency of the studied model through the following tests.

2-1-3-1- t-test for the significance of regression line coefficients

The following hypothesis is used in this test to ascertain how independent variables affect the dependent variable.

$$H_o: \beta_j = 0$$

$$H_o: \beta_j \neq 0 \quad j=1,\ldots,k$$

The mathematical formula for the test is:

$$t_{\hat{\beta}_j} = \frac{\widehat{\beta_j}}{\sqrt{s^2_{\hat{\beta}_j}}} \qquad j=1,\ldots,k \qquad (7)$$

Since β_j represents the estimated parameter value to be tested and extracted according to equation (6), and $s^2_{\hat{\beta}_j}$ is extracted from the variance and covariance matrix, the extracted value is compared with the tabulated value for the t-test according to the significance level and degrees of freedom.

2-1-2-3- F-statistic test

This test aims to measure the significance of the dependent variable's linear connection with the independent factors through the following hypothesis:

$$H_o: \beta_j = 0$$

$$H_1: \beta_j \neq 0 \quad j=1,\ldots,k$$

And the mathematical formula relies on the coefficient of determination, which is:

$$F = \frac{R^2/k}{(1-R^2)/n-k-1} = \left[\frac{R^2}{1-R^2}\right]\left[\frac{n-k-1}{k}\right] \qquad (8)$$

The extracted F value is compared with its tabulated value for the degrees of freedom (n-k-1) and a specified significance level, and based on this, the null hypothesis is either rejected or accepted, and the results are analyzed.

2-2 -The concept of Artificial Neural Networks (ANN)

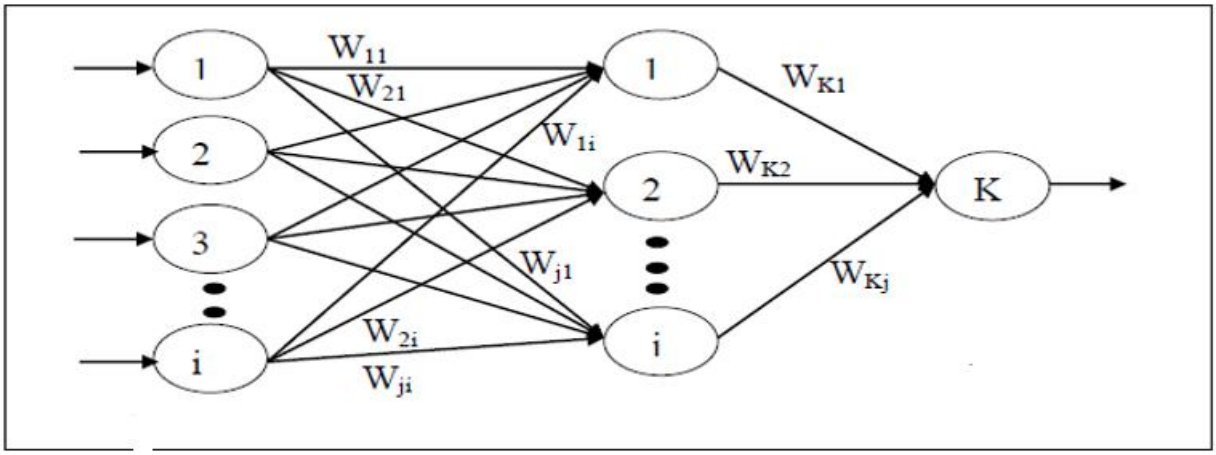
Scientific studies on the human brain serve as the inspiration for artificial neural networks. The biological neurons, the basic component of the brain, were discovered in 1836. The neuron consists of the nucleus, dendrites, and the axon that transmits signals to other neurons, while the gap between the dendrites and axon is called the synapse.

As a subfield of artificial intelligence, artificial neural networks simulate how the human brain stores and processes information. This is accomplished by establishing connections between artificial neurones, which are mathematical processing units. The strength of these connections, referred to as weights, is

chosen to improve the network's capacity for learning and prediction.2-1-2-Structure of artificial neural networks

A subfield of artificial intelligence known as artificial neural networks (ANN) mimic the way the human brain processes and stores information. It consists of networked processing units called nodes or neurones, and the strength of the connections between them is determined by weights.The neurons are arranged in layers that typically include: an input layer, one or more an output layer and hidden layers. The network is instructed. by adjusting the weights based on training data that includes inputs and outputs, with repeated training to improve the accuracy of the expected results[25], as illustrated in the following figure (1):

Fig (1) illustrates the diagram of the artificial neural network model[7].



The weights between the layers and the neurons of the neural network are adjusted by comparing the network's outputs with the true values until the estimated values match the true values.

2-1-2-Characteristics of neural networks

Neural networks have many features that enable them to be a powerful and effective tool for problem-solving in various fields and applications. Most noticeably, adaptability and flexibility are among those features since the network can automatically adapt to new environments without requiring an intervention of programmers or relying on prior assumptions on the nature of the variables and their relationships. Neural networks are also characterized by their robustness, because they can still perform effectively even when there are delays in the computation processes in the neurons during training time; this reduces their susceptibility to external forces. Another remarkable ability of neural networks is that they can process a wide variety of data, including data that may be categorized as being noisy, ambiguous, probabilistic,

or inconsistent. Lastly, neural networks function through mechanisms of cooperative computation that are executed in a synchronous and distributed manner, which enhances their effectiveness and efficiency when working with various data[26].

2-2-3-Neuron Model

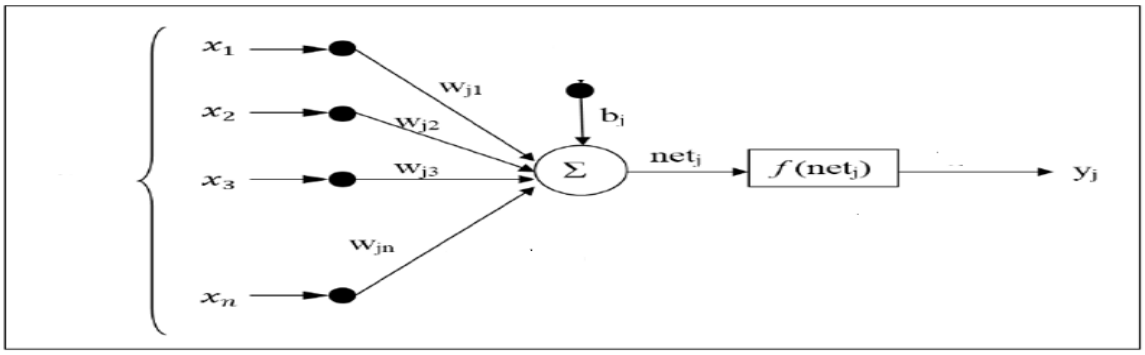
A group of information processing units known as neurones or nodes make up artificial neural networks. that serve as the fundamental building blocks in the functioning of the neural network. The design of these neurons is based on three key components, which can be described as follows[16].

1. Synaptic Weights: It is simply a set of connections or synapses, where each connection has some strength associated with it or weight, which is set randomly in an initial phase. Weighted synapses: That can be said to be a set of connections or synapses that have some strengths associated with them, generally randomly set initially.. The role of these weights is to modify the effect of incoming signals; the input signal (xn) is multiplied by the synaptic weight (wjn) associated with the neuron (j). This weight represents the strength of the connection between the connected elements, and as the network learns, these weights are continuously modified.

2. Bias:It represents another value added to the input signals after their modulation via the synaptic weights. A bias has a role to play in the regulation of neuronal outputs since it can potentiate or lessen the impact by its magnitude, irrespective of whether the bias is positive or negative.

3. Activation function: This function uses the input data to determine the neuron's output.Activation function: This function chooses the neuron's outputs according to the input values. Normalization processes are usually used to restrict the outputs within a certain range, such as [0, 1] or [-1, 1], which contributes to the stability of the network and improves its performance. The artificial neuron's structure is depicted in Figure (2).

Figure (2) Shows The Structure of a single artificial neuron [11].



To create the neural network inputs, each input signal is weighted according to its own weights, then added to the bias value and totalled together. which are referred to as the net inputs. This process can be mathematically represented as follows [11]:

$$net_j = \sum_{i=1}^{n} w_{ji}x_i + b_j = w_{j1}x_1 + w_{j2}x_2 + \cdots + w_{jn}x_n + b_j \quad (9)$$

The output of the neuron (y_j) can be expressed as follows:

$$y_j = f(net_j) = f(\sum_{i=1}^{n} w_{ji}x_i + b_j) \quad (10)$$

Since:

b:is the bias value, and f is the neurones' activation function or the transfer function of the neurons.

2-2-4-Activation Functions

Programmes for artificial neural networks use a variety of activation functions, but there are three functions that are considered the most common and widely used, which can be explained as follows[24]:

1-Log-sigmoid activation function

This function produces outputs that fall within the range [0, 1], while the inputs to the neurons are within an unlimited range between negative and positive infinity. This function is expressed in the mathematical formula:

$$a = f(n) = \frac{e^{n}-e^{-n}}{e^{n}-e^{-n}} \qquad -1 \le f(n) \le 1 \quad (11)$$

2-Linear activation function (purelin)

In this function, the function coefficient f=1, allowing the neural cell outputs to take any value based on the inputs. It can be expressed in mathematical form:

$$n = f(n) = a \qquad (12)$$

3-Activation function - Tan-sigmoid (tansig)

This function provides outputs that fall within the range [-1, +1], with inputs varying between negative and positive infinity. The mathematical formula for this function is:

$$a = f(n) = \frac{e^{n}-e^{-n}}{e^{n}+e^{-n}} \qquad -1 \leq f(n) \leq 1 \qquad (13)$$

2-2-5-Types of artificial neural networks (ANN)

The artificial neural networks can be categorized into two major types depending on the connection architecture and the way data is transmitted: feed-forward networks and feedback networks. There are no feedback links in a feed-forward network; instead, data moves in a single direction through the hidden layers, from the input layer to the output layer. The calculations are done sequentially to produce the desired outputs.

Feedback networks are distinguished by the presence of feedback connections, enabling linking of outputs with inputs within the same layer or even with previous layers. It has a bidirectional flow of data, so it will be well applied to more difficult pattern-related tasks.

Feedforward networks are the most popular choice for modeling/prediction tasks and are usually used together with the backpropagation learning algorithm. (Backpropagation). As a result, this type of network was adopted in the present study [10] [17].

2-2-6-Process of training the artificial neural network

Multilayer feedforward Neural Networks have been widely used to Solve complex and diverse problems by training the data with the backpropagation algorithm, often called the simple backpropagation algorithm. The basic algorithm relies on two passes over the network layers: one in a forward direction followed by a backward pass.

In the forward pass, it carries information to the nodes of the neural network in such a way that effects are spread gradually, layer by layer, throughout the network until the desired results are obtained, reflecting the network's answer. During this course, the inter-layer weights remain unchanged. During the backward pass, the produced outputs are compared against the target values so that the error signal can be calculated. From then on, the error signal back-propagates through the network in reverse (from outputs to inputs), while The Weights are iteratively adjusted to reduce the error and better fit the outputs to the target values.

In order to determine the error, the backpropagation algorithm can be summarised as calculating the difference between the network's outputs and the intended target values. The error signal is then sent backwards through connected nodes with the goal of modifying weights.. The methodology is considered one of the most important and effective learning strategies to enhance the effectiveness of multilayer Neural Networks [15].

2-2-7-Data normalization process

Normalization can be considered one of the crucial steps in enhancing the effectiveness of training for neural networks, since it ensures the input and output information is bounded within a certain interval before it is fed to the network. This method's primary objective is to avoid neuron saturation in the hidden layers, because neuron saturation can slow down the training process enormously. Log-sigmoid and Tan-sigmoid activation functions are in common use for multilayer neural networks' hidden layers due to their efficacy.

However, sigmoid functions also tend to saturate for high input values, which results in a small gradient value and thus slows down the weight adjustment during training. In the first layer of the network, the output is calculated by multiplying the inputs by the randomly assigned weights, with the addition of a bias value. To avoid saturation, the weights must be very small when the inputs have large values.

The initialization of data requires normalizing the inputs and target values within a predetermined range so that the outputs of the network are between this range while training. If the network is used for practical purposes, the outputs can easily be returned to the original values. The range of normalization depends on the transformation function: for Tan-sigmoid, the range is [−1,1]; for Log-sigmoid, it is [0,1] [14]

2-2-8-Weight initialization process

Before starting the training process of feedforward networks, it is essential to initialize the weights correctly. If the weights between the layers are too large, the network may become unstable, leading to neuron saturation. On the other hand, if the weights are too small, the weight adjustment process will be very slow, affecting the training efficiency. The selection of initial weights depends on the data normalization process and the complexity of the problem to be addressed. In general, the network performs well when the values of the random weights are within the range (-1,1) [18].

2-2-9-Performance in the school network

After training the network using the training data, it becomes necessary to evaluate its performance. The evaluation of the network's performance mainly depends on the correlation coefficient and the Mean Squared Errorwhich can be computed using the formula below. [13][14].

$$R = \frac{\sum_{k=1}^{n}(T_k-\overline{T})(O_k-\overline{O})}{(n-1)S_T S_O} \quad (14)$$

$$mse = \frac{1}{n}\sum_{k=1}^{n}(T_k - O_k)^2 \quad (15)$$

$$S_T = \sqrt{\frac{1}{n-1}}\sum_{k=1}^{n}(T_k - \overline{T}) \quad (16)$$

$$S_O = \sqrt{\frac{1}{n-1}}\sum_{k=1}^{n}(O_k - \overline{O}) \quad (17)$$

$$\overline{T} = \frac{1}{n}\sum_{k=1}^{n} T_k \quad (18)$$

$$\overline{O} = \frac{1}{n}\sum_{k=1}^{n} O_k \quad (19)$$

Where T_k represents the target data (true outputs), O_k represents the Network outputs, n is the number of data points, ($\overline{T}$) is the arithmetic mean of the target data, and ($\overline{O}$) is the arithmetic mean of the network outputs. The best case is when the correlation coefficient R equals one, and the best value for the Mean Squared Error (MSE) is zero, after completing the training process, if the MSE value and the correlation coefficient are not good, the network should be retrained by adjusting the training conditions to achieve a good convergence to the desired error value.

2-2-10-Network Testing

After the network training is completed, and the best performance is achieved on the training dataset. There has to be some checking done on the network based on new data that has not been included in training so that this new data can be said to have its range as that during the training. In a few cases, neural networks, after training, performed excellently on the used set. Nevertheless they also tend not to perform well on presented examples that are new. Therefore, testing of the network is important to ensure that it has the ability to understand the relationship between the inputs and outputs, and also to ensure that it works well when used with new data in the future. The best network is retained if it performs similarly well on both training and test datasets.

3-The practical aspect

The practical side of this research was represented by collecting data on the pollution of the Euphrates River for the years 2023-2024, with a sample size of (103) weeks distributed over (15) independent variables represented by the variables.

The independent variables are: pH, Temperature, (DO2, PO4, NO3, Ca, Mg, TH, K, Na, SO4, CL, EC, AIK, Turb) while the dependent variable represents total dissolved salts, TDS.Statistical measures such as the mean, standard deviation, and variance were calculated using statistical software (Minitab, MATLAB R2024a) for the analysis. Thus, predictions are made using multiple regression models and (ANN). To demonstrate the efficiency and accuracy of the methods used, the mean squared error criterion was employed.

3-1- Definition of variables:

X1 :represents the pH level, which measures the acidity or alkalinity of water on a scale from 0 to 14. The number 7 indicates neutrality, while values between 6.5 and 8.5 are considered ideal for drinking water according to the World Health Organization and Iraqi law. Deviation from these values may lead to pipe corrosion or environmental changes.

X2: represents the temperature (Temp) of the water that affects the dissolution of gases and salts, and it is an important environmental factor. Surface water should not exceed 25°C to ensure appropriate quality.

X3: represents dissolved oxygen (DO2) in water, which aquatic organisms need. Values above 5 mg/L are considered ideal according to Iraqi and international water quality standards.

X4: Calcium (Ca) is essential for human health and contributes to water hardness. The ideal concentration for drinking is less than 75 mg/L according to international standards.

X5: represents total hardness (TH), which expresses the concentration of calcium and magnesium together. The values should not exceed 300 mg/L to avoid the effects of excessive hardness.

X6: Potassium (K) is an essential element, but the concentration in drinking water should be less than 12 mg/L to ensure public health.

X7: Sodium (Na) is an element that contributes to the salinity of water. It should be less than 200 mg/L to avoid its effects on blood pressure.

X8:Electrical conductivity (EC) reflects the content of dissolved salts in water. Drinking water should have a conductivity of less than 1500 microsiemens/cm.

X9: alkalinity (Alk) which indicates the water's ability to neutralize acidity. The ideal values range between 20-200 mg/L as calcium carbonate.

X10:Turbidity (Turb) measures the clarity of water and indicates the presence of suspended particles. The acceptable value is less than 1 NTU according to the World Health Organization.

Y:total dissolved solids (TDS) represent the amount of salts in water, with an acceptable maximum of 1000 mg/L, and it is preferable for it to be less than 500 mg/L for high-quality water.

3-2-Descriptive statistics

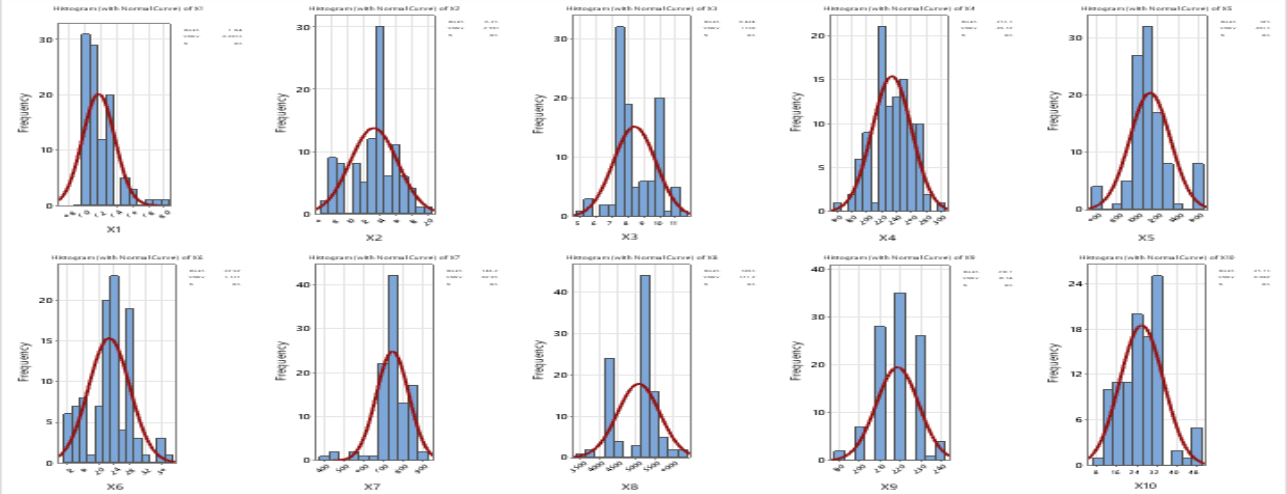
The results were calculated using the ready-made Minitab program, and the following results were obtained.

Table (1) represents the descriptive statistics.

Variable	Mean	SE Mean	Variance
X1	7.1640	0.0200	0.0414
X2	13.225	0.295	8.948
X3	8.424	0.133	1.822
X4	233.52	2.63	713.85
X5	1125.1	19.8	40535.2
X6	22.623	0.528	28.703
X7	746.25	8.16	6861.52
X8	5063.4	56.9	333154.9
X9	218.69	1.04	111.07
X10	25.748	0.875	78.896
Y	3022.1	26.9	74499.2

The statistical values of the variables X1 to X10 and the dependent variable Y indicate a clear variation in The Nature of the data, as the means vary significantly Between The variables. For example, X1 has a low mean (7.1640) with a very small variance (0.0414), compared to X8, which has a very high mean (5063.4) and a large variance (333154.9), reflecting a Wide Range of values. the standard error Of the mean (SE Mean) shows variation in the accuracy of mean estimates, with very precise estimates for some variables like X1 (SEMean = 0.0200), while other variables like X8 (SEMean = 56.9) showed greater dispersion, indicating significant variation in individual values. The dependent variable Y, with a mean of 3022.1 and a variance of 74499.2, reflects data that is widely distributed, necessitating a focus on analyzing the factors influencing it. These results indicate the importance of studying the relationship between variables using tools such as regression analysis to comprehend how each independent variable affects the dependent one.Figures (3) below show that the distribution of the data is normal.

Figure (3) illustrates the distribution of the data.



3-3 - Application of the Multiple Linear Regression Model

It is employed to ascertain how the dependent variable and the available independent variables relate to one another. which is the total dissolved solids.Here,we try to understand the impact of total dissolved solids on water quality in order to create a multiple regression model for predicting future periods. It is necessary to know whether the model is significant or not, and this can be determined through the following hypotheses:

H_0 : There is no effect of total dissolved solids on water quality.

H_1 : There is an effect of total dissolved salts on water quality.

The multiple linear regression's findings are displayed in Table (2).

Table (2) Shows The Results Of Multiple Linear Regression.

Term	Coef	SE. Coef	T_Value	P_Value	VIF
Constant	1100	453	2.43	0.017	
X1	-49.5	77.4	-0.64	0.024	2.41
X2	-9.36	5.81	-1.61	0.011	2.93
X3	-22.4	13.9	-1.61	0.010	3.42
X4	-1.394	0.744	-1.87	0.064	3.83
X5	0.2838	0.0805	3.52	0.001	2.55
X6	-4.11	2.80	-1.47	0.046	2.18
X7	3.028	0.234	12.96	0.000	3.63

X8	0.0006	0.0382	0.02	0.038	4.70
X9	1.80	1.34	1.35	0.012	1.93
X10	1.27	1.64	0.77	0.041	2.06

These regression results show that the impacts of independent variables X1 through X10 on the dependent variable differ. However, a good number of them have P < 0.05, indicating statistical significance. The Constant factor gives the expected value of the dependent variable at which zero is the value of all independent variables (1100),with its significance at P = 0.017.The variables with a positive and significant impact include X5 (Coef = 0.2838, P = 0.001) and X7 (Coef = 3.02, P = 0.000), which means that an increase of one unit in X5 or X7 leads to a clear increase in the dependent variable.The variables with a negative and significant impact include X6 (Coef=−4.11, P=0.046) and X1 (Coef=−49.5, P=0.024), indicating that their increase reduces the dependent variable, with a clear significance of their effect.On the other hand, the variables X2 (Coef=−9.36, P=0.011) and X3 (Coef=−22.4, P=0.010) show a negative and significant impact, while X8 (Coef=0.0006, P=0.038), X9 (Coef=1.80, P=0.012), and X10 (Coef=1.27, P=0.041) show positive but relatively weak effects.Although X4 (Coef=−1.394, P=0.064) is close to statistical significance, its effect cannot be confirmed based on this analysis. Every variable's VIF value is less than 5., indicating that there is no multicollinearity problem affecting the model.Based on these results, variables X5, X7, X6, and X1 can be considered the most important in elucidating the dependent variable's fluctuation.The following format can be used to express the multiple regression model:

$$y = 1100 - 49.56X_1 - 9.363X_2 - 22.45X_3 - 1.384X_4 + 0.284X_5 - 4.111X_6 + 3.038X_7 + 0.006X_8 + 1.801X_9 + 1.276X_{10}$$

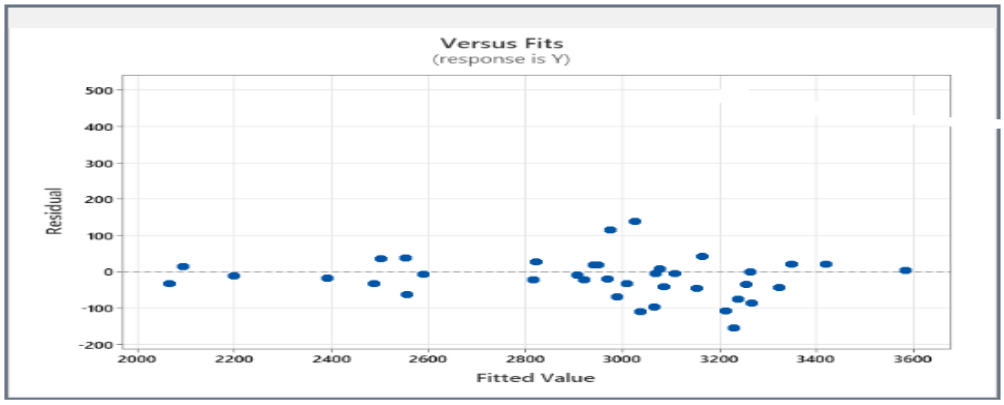
Table (3) Shows The results of The statistical values for the model.

MSE	R^2	R – sq (adj)	R - sq (pred)
10522.04092	88.26%	86.88%	83.08%

The statistical values of the model's performance indicate its high quality in explaining the variation in the data. The mean squared error (MSE = 10522.04) reflects the model's prediction accuracy, since this value is The amount of unexplained variance in the variable that is dependant. The coefficient of determination R2 = 88.26% of the model is indeed a strong indicator of its goodness of fit since this means that 87.26% of variation of the dependent variable in this model is explained by the independent variables. The R-Sq(adj) value, R-sq(adj) = 86.88, shows a moderate effect of the removal of bias due to irrelevant variables added into the model, hence not being affected by overfitting. The predictive value of the coefficient of determination (R-

sq(pred)=83.08%) indicates The capacity of the model to forecast fresh data not used in the training process, reflecting good performance in generalization. Based on these values, the model is strong and effective in interpreting and predicting the data, and The residuals are shown in Fig(4) below.

Figure (4) Shows The residuals.



The plot "Residuals vs. Fitted" indicates the spread of residuals. Residuals vs. Fitted (Actual - Predicted Values) Versus expected values from the model: The points indicate that variation around the zero horizontal line is completely random, reflecting proper change depiction by the model with no pattern existing in the residuals. This random scattering of the residuals indicates that there is a normal distribution of the error, as would be expected, supporting independence and symmetry in residuals. Additionally, the plot depicts constant variance across the expected values, hence not indicating the problem of heteroscedasticity. This plot contributes to the quality assessment The model's ability to predict new data error.

3-4- Application of the artificial neural network model:

When applying feedforward artificial neural networks and using the training function (Trainlm), we obtained the results as shown in table (4) below:

Table (4) Shows The Results of Artificial Neural Networks.

Overall Correlation Coefficient	Correlation Coefficient for Testing	Correlation Coefficient For training	mean squared error (mse)	number of nodes in the second hidden layer	number of nodes in the First hidden layer
0.9293	0.8931	0.9541	0.02699	5	10
0.9778	0.9768	0.9801	0.00194	10	15
0.9421	0.8236	0.9993	0.03606	15	20
0.9644	0.97471	0.9606	0.00768	20	30
0.9382	0.9084	0.94739	0.012487	40	50
0.6919	0.6326	0.7100	0.044958	3	5

Table (4) refers to the process of assessing how well an artificial neural network predicts outcomes using various hidden layer node configurations. Three main criteria were relied upon to evaluate performance: Mean Squared Error (MSE) and correlation coefficients (R) for both training data, test data, and the overall total. The goal is to determine The optimal network configuration That achieves The highest prediction accuracy and the best generalization capability.

3-5-Performance analysis based on the number of nodes:

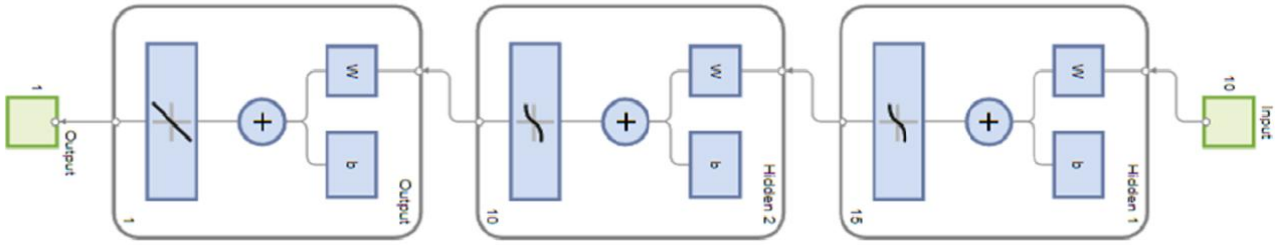
Configuration [15, 10]: The best performance was realized for this configuration with the lowest MSE of 0.00194, which indicates high accuracy in the prediction. It also recorded the highest correlation coefficients for training of (0.9801) and testing of (0.9768), respectively, with an overall value of 0.9778. This further affirms that the network has learned the pattern in the data with good generalization to the test data without overfitting or underfitting.

Setup [20, 15]: Although a very high correlation coefficient was achieved on the training data, R = 0.9993, the correlation coefficient for the test data was significantly reduced, R = 0.8236.This significant discrepancy reflects an overfitting problem, where the model has become overly specialized in the training data and has lost its ability to predict effectively using new data.

Setup [5, 3]: This configuration showed the weakest performance, recording the highest mean squared error (MSE = 0.044958) and the lowest correlation coefficients for training (R = 0.7100) and testing (R = 0.6326This would imply that the network, with the few nodes in it, was not able to learn The complex Patterns in the data, thus underfitting and failing to predict.

The more complex settings ([30, 20] and [50, 40]): These settings showed an improvement in performance with relatively good correlation coefficients, such as R = 0.9606 and R = 0.97471 for the setting of [30, 20], but they did not outperform the setting of [15, 10] in overall efficiency.Increasing the number of nodes in these cases led to greater consumption of computational resources with only slight improvements, which may be impractical from an efficiency perspective. The structure of the artificial neural network for 10 and 15 nodes can be illustrated in the following figure (5):

Figure (5) illustrates the structure of Artificial Neural Networks.



Fig(5) Shows The Architecture of the (ANN) designed for prediction, with an input layer that contains 10 inputs and two hidden layers composed of 15 and 10 nodes, respectively. The last layer is the output layer, which includes one nodeThe network's hidden layers employ non-linear activation functions to enable it to learn complex patterns in data. For the output layer, Utilising a linear activation function, order to make predictions. The architecture is suitable for prediction tasks since it generalizes and provides relatively accurate prediction results. The following figure 6 shows the correlation coefficient for the selected network.

Figure (6) illustrates the correlation coefficient.

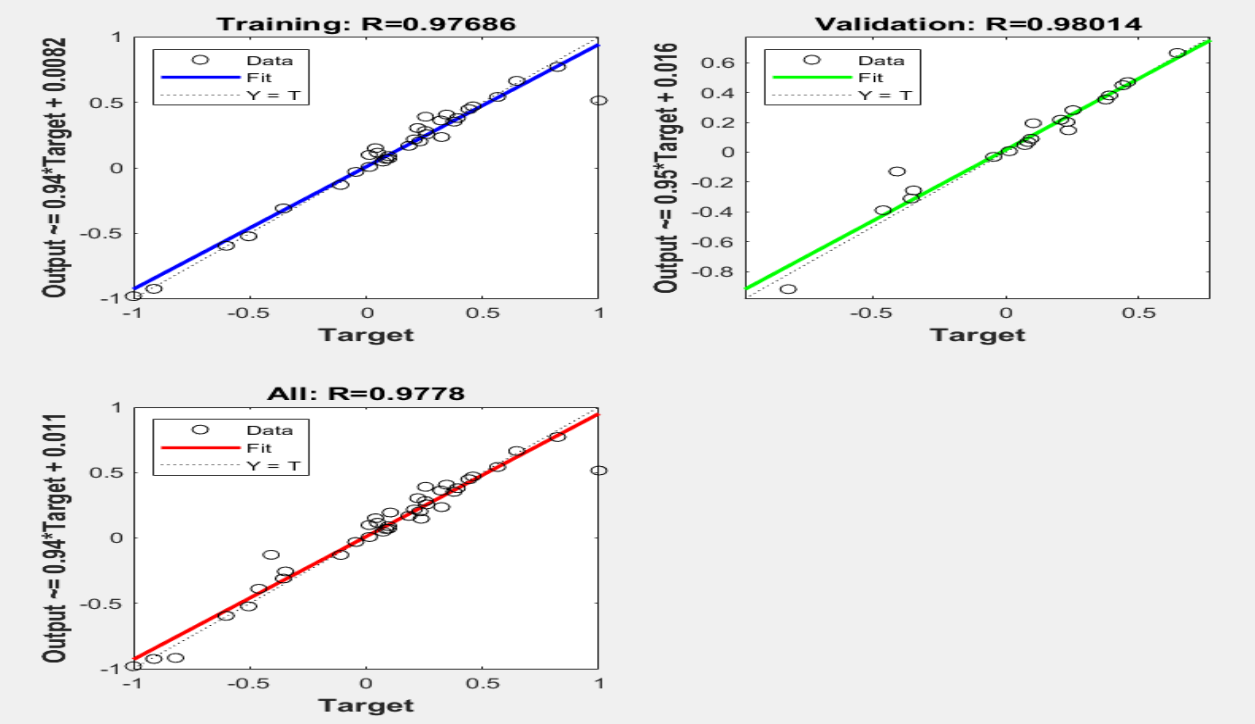
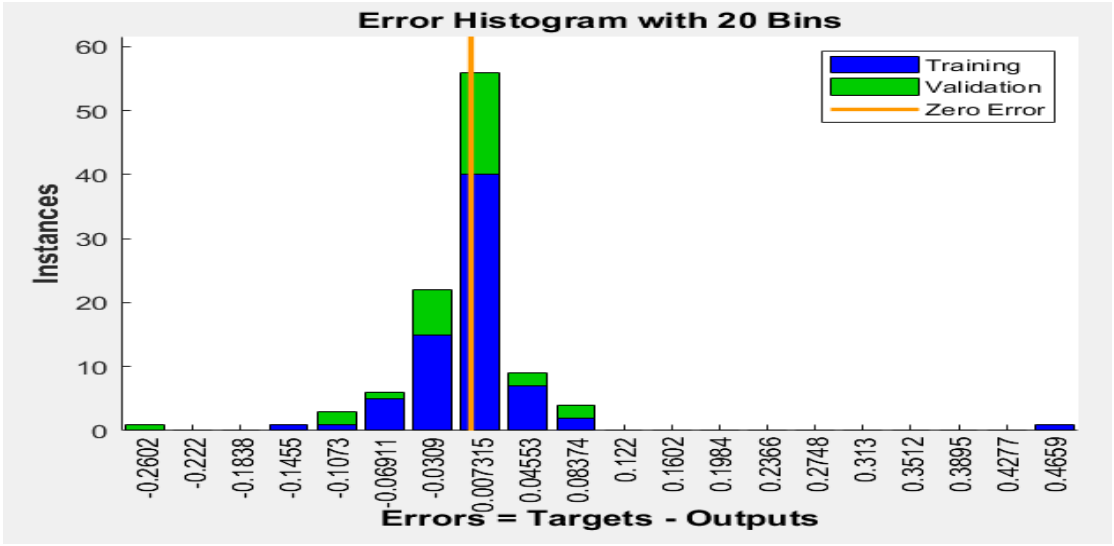


Figure 6 shows the scatter plots of the target values against the predicted outputs, which present the performance of the neural network in three different stages: training, validation, and overall. The graph in the top left of the figure presents the results of the training phase, where the correlation coefficient R=0.97686, indicating high agreement between the desired values and the outcomes. The top right graph

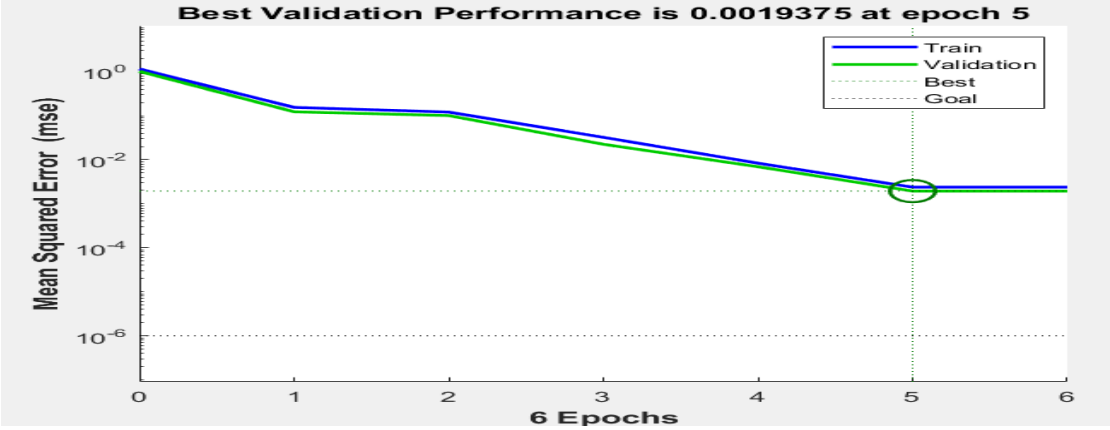
represents the performance of the network during the verification data, where (R=0.98014), showing the ability of the model to generalize on untrained data with stable performance, the bottom graph represents the overall results for all data, where the overall correlation coefficient reached R=0.9778, enhancing the model's reliability and prediction accuracy. The quality of the predictions is reflected in how close the regression lines are to the ideal line Y=T, which testifies to the effectiveness of the neural network in processing and performing predictions with high accuracy. Such results allow for the possibility of using the model in practical **applications related to forecasting. Figure (7) depicts the distribution of errors**

Figure (7) shows the distribution of errors.



The graph illustrates the distribution of errors between The target values (targets) and the predicted (Outputs) of the Neural Network, divided into 20 categories. (Bins). The horizontal axis represents the error values (Errors=Targets−Outputs) - while the vertical The axis shows how many instances there are. that fall within each category of error values.The graph shows that most errors are concentrated around zero, indicating high accuracy neural network's performance in predicting target values. The blue bands represent the training data, while the green bands represent the validation data. We notice that the errors are almost evenly spaced throughout zero with a limited number of outliers, indicating that the model is capable of generalizing well.Moreover, the orange line at zero represents the ideal no-error line, where the data is closely distributed around it, reflecting a high convergence between the predictions and the target values. This distribution reflects the model's reliability in prediction and confirms its effectiveness in various prediction applications.And Figure (8) shows the Mean Squared Error.

Figure (8) Illustrates The Mean Squared Error.



3-6- Comparison Between the Multiple Regression Method and Artificial Neural Networks

A comparison between the MSE of neural networks and the MSE of multiple regression indicates significant differences between the two methods in terms of performance. In the case of neural networks, the Mean Squared Error (MSE) value was (0.00194), indicating very accurate performance in predicting outcomes. In contrast, the MSE for multiple regression was very high, reaching (10522.04092), which indicates that multiple regression did not provide accurate results,and the prediction results of multiple regression and Artificial Neural Networks:

Table (5) Shows The Prediction Results for (ANN) and the multiple Regression Model.

Prediction Using Artificial Neural Networks	Prediction Using Multiple Regression Model	Weeks
2860.23	2823.19	1
2870.43	2950.97	2
3105.12	3109.00	3
2585.87	2591.65	4
3110.45	3154.89	5
3065.78	3069.40	6
2950.18	2970.45	7
3210.10	3166.67	8

Table (5) presents a comparison of prediction results using a multiple regression model and artificial neural networks over an eight-week period. It's apparent from this table that the artificial neural network stands out with respect to its prediction accuracy and closeness to real values. It is clear that the predictions yielded by the neural networks were more constant and stable, whereas those provided by the multiple regression model presented larger deviations from the reference values. The performance of neural networks underlines a clear superiority in catching nonlinear variable relations and achieving accurate responses over various time periods. Conversely, it expresses the limitation of the multiple regression model in handling nonlinear-pattern data since it depends on a linear assumption, which again confines its predictions to high levels of accuracy. Based on these results, neural networks can be considered a powerful and reliable tool in predicting complex multidimensional data and are optimum for application in the future when high levels of accuracy are required.

4-Conclusions & Recommendations

These results indicated that the network configuration [15, 10] is most suitable for artificial neural network predictions due to an ideal compromise between accuracy and the generalization capability of the model. By contrast, [5, 3]-sized configurations had poor performance due to the inability of the model to learn complex patterns, while larger configurations like [50, 40] became too complex with no significant improvements, pointing toward a risk of overfitting. Therefore, based on the obtained results, it is advisable to use the setting of [15, 10] as the best option by considering techniques such as early stopping or increasing the size of the training data to avoid overfitting when settings are complex. The optimal setting also needs to be tested on diverse data for verification of generalizability and accuracy in different applications. It is also recommended to conduct further research that analyzes the impact of changing. The Number of Nodes in the Hidden Layers on overall performance. A comparison of the outcome with other prediction techniques using neural networks, such as multiple regression, will determine the efficiency according to the nature of data and application conditions. We also recommend that there be an increase in environmental awareness and enhancement of the struggle to conserve water resources through the implementation of comprehensive awareness programs to encourage responsible behaviors towards the environment. We suggest the study of the impact of other variables on living organisms in rivers and effective strategies to mitigate the damage to such ecosystems. Moreover, we recommend the use of artificial intelligence techniques in conducting analyses on

environmental data through comprehensive comparisons with traditional models and methods for the most valid and efficient way to address environmental issues.

References

1. Achebo, J. I., & Eki, M. U. (2020). "prediction of mild steel weld properties using artificial neural network and regression analysis". *Tropical Journal of Science and Technology*, *1*(2), 37-49.
2. Ahmed, H. A.,(2017)," Development models of artificial neural network and multiple linear Regression for predicting compression index and compression ratio for soil compressibility of eamadi city",Al-Nahrain Journal for Engineering Sciences (NJES) Vol.20 No.4, pp.924-936
3. Al –Aidani & Ali S. Q., (2015)," prediction of groundwater level in safwan-zubair Area using artificial neural networks", Master thesis, college of engineering, University of Basrah.
4. Al-Hanoon, O. B. S., & Yahya, M. M. M. (2019). "Using the MLR-RNN method to predict air pollution data". IRAQI Journal of Statistical Sciences, 16(2).
5. Alharthi , N. H., Bingol , S., Abbas, A. T., Ragab , A. E., El-Danaf , E. A., & Alharbi , H. F. (2017). "optimizing cutting conditions and prediction of surface roughness in face milling of AZ61 using regression analysis and Artificial Neural Network". *Advances in Materials Science and Engineering*, *2017*(1), 7560468.
6. Al-Sabah, S. A., & Al-Kufayshi, S. M. (2020)." Estimating parameters of the multiple regression model under the problem of multicollinearity". Journal of College of administration and economics for economic, administrative, and financial studies, 12 (4) , 1–28.
7. Begum, S.A., Fujail, A.K., Barbhuiya, A.K., (2012), "artificial neural network to predict equilibrium local Scour depth around semicircular bridge abutments", 6th International symposium on advances in science and technology, kuala lumpur, malaysia
8. ÇAKIROĞLU, R. (2022). "analysis of EDM Machining Parameters for keyway on ti-6Al-4V alloy and modelling by Artificial Neural Network and Regression Analysis Methods". *Sādhanā*, *47* (3) , 150.
9. Crompton,I.R. (1997) ."Toxicants in the aqueous ecosystem,John wiley and Sons LTD".england.
10. Dawood, A.S. and Yilian Li, (2014), " Taguchi and ANN modeling of turbidity removal using hybrid flocculant ", Research Journal of applied Science, Engineering and Technology, Vol. 7, No.18, pp. 3691 3698.
11. Gadoue , S ., "artificial neural networks ", industrial automation lecture notes, school of electrical, electronic and computer engineering, Newcastle University, England.
12. Göçmen, E., & Derse, O. (2018). "Forecasting of alectricity generation shares by fossil fuels using artificial neural network and regression analysis in turkey". *International Scientific and Vocational Studies Journal*, *2*(2), 20-30.

13. Hagan, M., Beale, M. and Demuth, H., (2009), " Neural Network toolboxTM user's guide ", The Math Works, 6th edition, The mathWorks, Inc.
14. Hagan, M.T., Demuth, H.B., Beale, M.H. and Jesus, O.D., (2002), " Neural network design ", 2nd edition.
15. Hajek , M., (2005) , " Neural networks ", University of Kwazulu-Natal, 1st Edition, South Africa, Neural networks.doc.
16. Haykin, S., (1999), " neural networks a comprehensive foundation ", Pearson Education, 2nd Edition, Mc Master University, Canada.
17. Jha, G.K., (2007) , " artificial neural networks and Its applications ", Journal of Indian Agricultural Research Institute, pp. (41-49).
18. Massie, D. and Curtiss, P.S., (2001)," neural networks fundamentals for scientists and Engineers", In Proceeding of the International Conference on Efficiency, Costs, Optimization, Simulations and Environmental Impact of Energy Systems (ECOS-01), Turkey.
19. Mousa, H. (2006). Environmental Pollution (2nd ed.). Dar Al-Fikr Al-Muasir, Beirut, Lebanon.
20. Saad, K. S, Abbas, F. I, (2008) ," Prediction of surface roughness in end-milling with multiple regression model",Eng.&Tech.Vol.26,No.3,pp.165-187
21. Scheffer, M.; Carpenter, S.; Foley, J.A.; Folke, C.and Walker, B. (2001). "catastrophic shifts in ecosystems". Nature 413:591-596.
22. Shamshiry, E., Mazlin. M. & Abdul. A, (2014) , " Comparison of artificial neural network (ANN) and multiple regression analysis for predicting the amount of solid waste generation in a tourist and tropical area—langkawi Island", International Conference on Biological, Civil and Environmental Engineering (BCEE-2014) Dubai (UAE), pp. 161-166.
23. Štambuk-Gilijanovic, N. (1999) ."Water Quality evaluation by index in dalmatia". water Rese., 33 (16): 3423-3440.
24. Stark,.R. Hanson,.E. Goldstein, M. Fallon, D. Fong.L. Lee, K.E. Kroening, S.E. and Andrews, W.J. (2000). "Water Quality in the upper mississippi river basin", Minnesota, Wisconsin, South dakota, Iowa, and north dakota, 1995-98. United States Geological Survey, Circular 1211, Reston, Virginia, 36pp. available on-line at.
25. Taylor,B., (2006), " Methods and procedures for verification and validation of artificial neural networks ", Springer, United States of America.
26. Yegnanarayana, B., (2005) , "Artificial Neural Networks", prentice-Hall of India Private limited.