Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Forecasting Dissolved Oxygen in Lakes Using Different AI Models

Zahraa Ch. Oleiwi[a]*, Karrar Khudhair Obayes [b], Nagham Kamil Hadi[c], Rahmah.Q yaseen[d], Asaad Jabbar Sahib[e]

[a,b,c]*College of Computer Science and Information Technology, University Al-Qadisiyah.*

[d,e]*College of Science , University Al-Qadisiyah.*

Email: [a] zahraa.chaffat@qu.edu.iq, [b] Karrar.khudhair@qu.edu.iq, [c] nagham.kamil@qu.edu.iq, [d] rahmahyasen@gmail.com, [e] asaad.jabbar@qu.edu.iq

ARTICLE INFO

Article history:
Received: 2 /1/2025
Rrevised form: 10 /1/2025
Accepted : 13 /1/2025
Available online: 30 /3/2025

Keywords:

Heart disease;

features selection

mutual information;

Random Forest

ABSTRACT

The freshwater ecological situation and water quality management would be impossible without the accurate prediction of the dissolved oxygen (DO) concentration in lakes. This study focuses on the use of different advanced Artificial Intelligence (AI) models in predicting values of the DO level for a given input environmental data. To make the data usable for analysis, a number of pre-processing steps are carried out. These processes include, but are not limited to, the ability to deal with the missing data and standardization of the features in the data set, so that the features are on the same level. After the pre-processing, the dataset aggregated in features and a target variable, which is the concentration of the target substance, the dissolved oxygen, is selected. Also, as a part of the building process of the model, the analysis of the features with regard to the target variable is performed in order to discern contributing features for better prediction. Additionally, in order to increase the accuracy of the machine learning models, standardization transformations to the mean of zero and to the variance of one were applied in the training set and test set using Standard Scaler. A selection of different machine learning models was therefore performed in order to identify the best predictor of the DO concentration. In the case of Linear Regression, it's R^2 has indicated a high degree of predictive accuracy at 0.9974 with a very low Mean Squared Error (MSE) of 9.6146e-05. Meagerly performing as an alternate method, Support Vector Regression (SVR) managed to attain 0.9448 of R^2 and an MSE of 0.0021, so it works but not as accurately as Linear Regression. In addition, multiple hidden layers were applied through Artificial Neural Networks (ANN) in a bid to model the data as uncertainty that appears in higher orders. Most notably, the ANN model achieved an R^2 of 0.995 and an MSE of 0.000161 which is nearly comparable to that of Linear Regression.

MSC..

https://doi.org/10.29304/jqcsm.2025.17.11959

1. Introduction

Modeling and predicting oxygen levels in lakes in a prominent factor in the management of water resources and through a better understanding of the lake ecosystem. However, in recent years, the rapid growth of artificial intelligence (AI) development has created new opportunities for modeling the oxygen level in lakes with much higher precision and less time-consuming ML algorithms can achieve. In this work, we give literature overview on

∗Corresponding author: Zahraa Ch. Oleiwi

Email addresses: *zahraa.chaffat@qu.edu.iq*

Communicated by 'sub etitor'

the studies focusing on artificial intelligence techniques that are used to predict dissolved oxygen concentrations in lakes. The value of AI technology for lake water quality monitoring and forecasting, naturally, follows from a value understanding enabling to compare the performance of various AI models [1].

There are several works that have used intelligence techniques in this field, which aims to predict the percentage of dissolved oxygen in lakes. They differ from each other in terms of sample collection methods, data processing, and general methodology steps. Portray predictive models for levels of Dissolved Oxygen in an urban lake using common water quality parameters like pH, Oxidation Reduction Potential (ORP), Conductivity, and Temperature were produced [2]. In this work, a Multiple Linear Regression model was created to predict the Dissolved Oxygen with the correlated data of water parameters. On the other hand, a model was built using the Artificial Neural Network (ANN) method with the Levenberg-Marquardt algorithm to predict the Dissolved Oxygen. The results of this study indicated that Multiple Linear Regression achieved an R^2 of 96.3%, surpassing the R^2 of 93% achieved by ANN [2]. The AdaBoost algorithm model to predict the state of water quality using data from the open artificial intelligence (AI) hub was introduced by [3]. pH, SS, water temperature, total nitrogen(TN), dissolved total phosphorus(DTP), NH3-N, chemical oxygen demand (COD), dissolved total nitrogen (DTN), and NO3-N as the input variables of the AdaBoost model were selected and dissolved oxygen (DO) was used as the target. The results of implementing this proposed predictive model showed that R2 was 0.912. The other categories of research depend on deep learning models in Forecasting Dissolved Oxygen in Lakes as in [4, 5] where the radial basis function neural network (RBFNN) and deep recurrent neural network.

The primary aim of this ambitious research project is to develop innovative methods that will contribute to our comprehension of the evolution of lake environments, mainly as concerns the changes of dissolved oxygen. Because oxygen is an essential element for living organisms in aquatic environments, its level must not decline otherwise it will set off a chain of calamitous consequences for the whole food web. The said consequences would include the following:

• Extinction of organisms that are dependent on waters with high oxygen content
• The degradation of water standards, resulting in heightened contaminants and the surge of disease - causing bacteria
• Stress on adjacent ecosystems which depend on the fitness of other water bodies
Four major spheres of this paper are of paramount importance:
• Environmental protection, focusing on safeguarding sensitive regions within the aquatic ecosystem
• Water resources management which is imperative to fulfilling both the human and ecological aspects
• Protection of fisheries, which rely on fish stocks and clean water resources
• Promotion of research, facilitating the collection of further information for subsequent research in the area of marine science
• Crisis management preparedness in knowing what could the environmental emergency be and taken steps in advance to avoid it.

Some of the existing issues are overfitting, not promising performance results, inability to predict oxygen concentration for different types of lakes, or insufficient consideration of spatial-temporal dependencies. More importantly, most of the above approaches have not been jointly evaluated with other methodologies to establish relationships between and select the optimal models across the range of AI methods. To this end, this research seeks to fill this gap by analyzing the ability of several machine learning algorithms in predicting DO levels in the following lakes. In particular, it should be used to compare the results obtained when applying various AI techniques. Therefore, by overcoming these limitations and filling the gaps in the current state of research in this field, this study will facilitate more precise and accurate forecasting models in order to enhance the effective management of some of the important and valuable lake ecosystems and save aquatic life missing.

This research outlines the science advancements for the prediction of dissolved oxygen levels that are crucial for aquatic life and also for management of water resources AI models such as Linear Regression, SVM, and ANN were used, these models are enhancements over the conventional methods which are many times manual and time consuming or less precise. These algorithms were compared and analyzed. The careful tuning and preprocessing steps implemented, including feature normalization using Standard Scaler, ensure that these models are optimized and demonstrate the potential for high-performance predictions.

2. Related works

The modeling and forecasting of the concentrations of oxygen plays a vital role in water quality in this case being important in the preservation of healthy water bodies. The case of water management and the preservation of water lakes, as a direction under consideration, is being reviewed through the PRISMA model that targets an agenda. For such a focus area, it is emphasized DO forecasting where more advanced methodologies, like fusion techniques, are being leveraged instead of DO methods[2-6]. It is worth mentioning that the advancements experienced in the field of deep learning have afforded the researchers improved understanding for DO levels in rivers as well as reservoirs, which in turn has enhanced the way marine environments are managed[7, 8].

As mentioned above, several algorithms and techniques have been put into use by the researchers in a bid to enhance the proper forecasting of DO levels. These comprise of diverse models such as regression models, Bayesian approaches, support vector machines (SVM) and several tree based predictive models. Prasad et al. (2011) are among those who made notable contributions in this field with their work that used multivariate regression (MLR) modelling forecasting the levels of dissolved oxygen in the Chesapeake Bay, thus demonstrating practical uses of these traditional methods[9]. In the same manner, Kisi et al. (2020) proposed a DO evaluation model based on Bayes for the DO concentration with every hour which had better results than whatever was available in the market including extreme learning machines (ELM), artificial neural networks (ANN), adaptive neuro fuzzy inference systems (ANFIS), simple classification and regression trees, MLR models etc. Their works can be regarded as the development which is in the crossroad regarding the advancement of machine learning in the area of application [10].

In the year 2020 Li and his colleagues have made a wonderful contribution by introducing an approach which they called the maximal information coefficient-support vector regression which is denoted MIC-SVR. This model was able to perform remarkably well by achieving only 3.01% measuring root mean square error which is reported as 62.36% of Nash-Sutcliffe efficiency and an R-square of 0.9 considerate enough as it indicates a remarkable predictive accuracy more than what was achieved in the SVR regression revision. It is to be noted however that the traditional SVR as well as the MIC-SVR models are fundamentally constrained by their respective decision functions which stops them from providing richer information such as the spatiotemporal relationships of time and dissolved oxygen concentration [11].

The study presented an innovative methodology called WTD-GWO-SVR, developed by Feng et al. (2024). This approach combines WTD, GWO and SVR to improve the prediction of dissolved oxygen (DO) concentration. It provides a good solution to such problems as a high level of noise and a complicated series of time. The aquaculture environment was modelled in a realistic setting using low-cost sensors to collect data. The WTD-GWO-SVR mean squared error was equal to 0.38 percent, the mean absolute error was equal to 3.81 percent, while the R squared value was 99.73 percent clearly outperforming back-propagation [12].

3. Material and Methods

All the theoretical background of methods and materials used in this research are explained in this section.

3.1 Datasets Description

The focus of this dataset is to predict the concentration of dissolved oxygen in lakes using different types of artificial intelligence (AI) models. Dissolved oxygen is an important parameter in evaluating the condition and quality of water bodies, lakes in particular. In 2019, Dr. J. Thad Scott collected water quality vertical profile data from Eagle Mountain Lake with assistance from the Tarrant Regional Water District. The data provided here consist of the final cleaned and imputed records used in the study by Durell, L., Scott, J. T., Nychka, D., and Hering, A. S. , titled "Functional Forecasting of Dissolved Oxygen in High-Frequency Vertical Lake Profiles," which is in revision since 2022. The total number of samples in this dataset is 47292. Predicting dissolved oxygen (DO) levels in lakes is crucial for assessing environmental conditions and reducing water treatment costs. High DO levels can precede toxic algal blooms, while low DO levels can lead to the precipitation of carcinogenic metals during water treatment[13]. The dataset contains information collected from multiple lakes over a period of time, including various relevant attributes:

Date.Time: Period when the DO was measured
Depth: Depth from the lake surface (0) to bottom (10)
Temp: Temperature
DO: Dissolved Oxygen (target)
DOsat: DO saturated
pH: pH of the Lake
Cond: Conductivity.

3.2 Linear Regression(LR)

Linear regression is a very well known tool to quantify the relationship between a dependent variable and its set of independent variables. With respect to the prediction of Dissolved Oxygen levels in the lakes using various AI models, a linear regression model can be used on dissolved oxygen levels by incorporating the temperature, pH, nutrient levels and even timing as factors into the model[14].
Linear regression is one of the most important algorithms in the area of supervised approaches to machine learning due to the existence of broad labeled datasets. It explains and establishes the connections among and between the data by representing the relations in the best possible linear forms. When linear regression is applied to new databases, its usefulness allows predictions to occur with a high degree of accuracy. For these reasons, linear regression is basic in the analysis of data. In essence, linear regression works out the relationship between a single dependent variable such as sales or temperature and a number of independent variables or features which are also called predictors. In case a single independent feature is being considered, it is termed as Simple Linear Regression. However, When there are several independent features being considered at the same time, it is known as Multiple Linear Regression. Among the main attributes of linear regression, it is necessary to consider the given possibility to explain the dependent and independent variables[15]. On the other hand, when we study the effect of more than one independent feature at a time, it is called Multiple Linear Regression. Linear regression has the significant benefit of easy interpretation is one of the essential strengths. The resulting model is linear equation in which each coefficient has a clear and measurable interpretation of how the variations in an independent variable affect the dependent variable. This brought about transparency not only helps in explaining the results but also offers a way of analyzing the drivers of the dataset. Nonetheless, such algorithm as linear regression is quite simple, however, it is a basis for many advanced algorithms in machine learning[16]. Those basic concepts like regularization, for checking the overfitting problem, and support vector machines are developed or derived from this theory adding more light to its multifaceted applicability in various fields. In addition to that, linear regression helps in hypothesis testing by providing the research tools that enable the testing of basic assumptions of datasets. Because R was conceived as a foundational piece of the statistical architecture, it is an invaluable tool for any analyst who intends to draw valuable conclusions out of data[17].

3.3 SVR (Support Vector Regression)

Support Vector Machine (SVM) is indeed interesting subset of tools if we specify the context to Machine Learning, quite predominantly in the classification problems area. However, SVMs are not just confined to classification since the models can be used on regression problems as well; in this case we have SVR, Support Vector Regression. Unlike mainstream regression approaches that deal with separate outputs, SVR was meant to provide a solution that deals with estimates which are the likely continuous outputs. This makes it one of the systems or approaches suitable for all predictive models. SVR is built on similar concepts as that of SVM which emphasizes on the projection of input features to high dimensional space. This projection is necessary in determining the best hyperplane which approximates the nature of these data optimally. This feature also makes it possible for SVR to perform well in the presence of both linear and non-linear relationships in data sets. One of the best features of SVR is that it can construct a model using many kernel functions, including quadratic, radial basis function, and sigmoid kernels. These kernels enable SVR to capture intricate relations in the data that the conventional techniques of regression would not be able to capture. By adapting to the intricacies of the dataset, SVR can uncover hidden patterns and deliver more accurate predictions. Moreover, Support Vector Regression is particularly versatile, finding applications across diverse fields such as financial forecasting, where it predicts stock prices and market trends, and scientific research, where it analyzes experimental data. The ability of SVR models to maintain high levels of accuracy and robustness, even when confronting complex and noisy datasets, positions them as essential tools for practitioners seeking reliable predictions in their work[18].

3.4 Artificial neural network

Recently, the use of Artificial Intelligence (AI) approaches, most recently deep learning models, have been able to model some environmental paramaters such as dissolved oxygen. This paper investigates the suitability of Artificial neural network in projecting the concentration of dissolved oxygen in lakes [19]. This is also called a neural network which is a modern class of machine learning methods that is based on the working of the human neural system. In the contemporary Pangea, where the focus of the world is centered on drug discovery research, the amount of data that is at disposal also increases. This not only fosters optimism but brings with it challenges because the information is shrouded within the attributes. This warrants the employment of sophisticated procedures of analysis A lot of interests are being focused on uncovering those causal relationships that exist between a set of responses and a number of explanatory variables[20].

An artificial neural network consists of a set of units called neurons which are linked to each other and classified into k layers. Each neuron receives inputs, processes them mathematically and passes the processed information to other neurons using specific functions which are embedded in the structure. The ANN model undergoes training by making use of a particular data set and then it modifies the weights associated with the connections between different neurons. This allows the model to "learn" from data and improve the predictions it makes with the fine tuning it undergoes each time a new layer is applied or adjusted [22].

3.5 Performance measures (R^2 and MSE)

For an AI model to successfully estimate the dissolved oxygen concentration levels in lakes with a mean square error measure, it is imperative to first understand the meaning of R square the statistical coefficient of determination. It is said to be the indicator of the degree of fit of one's model to the data which is very useful information to anyone carrying out regression analysis particularly with a view to assessing the extent to which the other dependent variables included in the model account for the variations with the dependent variable which in this case is the case is oxygen levels.

In simpler terms, the R-squared tries to determine the extent of which the regression model and the data deviate or integrate with each other. A higher R-squared value indicates a greater proportion of the variability of the dependent variable that is being explained by the model. A lot of variability in the dissolved oxygen levels is therefore effectively explained through the model coefficients. The goodness of fit measure is very important in the use of statistical models since it helps to achieve the major goals of making predictions about future events and, in conjunction with hypothesis testing, it helps to prove the strength of the relationships formed in the model..Understanding R-squared is thus vital for assessing the performance of AI forecasting models in environmental studies. The mathematical formula of R^2 as in equation (1):

$$R2 = 1 - (RSS/TSS) \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where, R2 represents the required R Squared value, RSS represents the residual sum of squares, and $\boldsymbol{TSS}$ represents the total sum of squares[6].

plays a crucial role in identifying performance measure in regression analysis as well as in the domain of machine learning. This metric measures the mean of squares of the prediction-errors through the actual target values of the dataset, making it easy to compute direct measure of prediction accuracy.

The MSE makes them square out, which really magnifies large discrepancies and makes certain that the model will suffer significantly more for large forecasting errors. The main focus when computing MSE is to understand how accurate the model is to a particular set of outcomes commonly known as 'ground truth'. A lower MSE signifies better predictive performance and therefore a very important tool in the selection of better optimized models.[7]. The MSE is computed using the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \ldots\ldots\ldots (2)$$

Where MSE stands for Mean Squared Error, n represents the number of data points, Y_i denotes the observed values, and $\widehat{Y}_i$ signifies the predicted values.

3.6 General Framework of Proposed Model

The proposed regression model for forecasting dissolved oxygen levels is based on water quality vertical profile data and utilizes three AI classifiers: LR, SVR and shallow neural network. The key steps are as follows:

1. Data Preprocessing: This step consists of the data cleaning process whereby the selected dataset is examined so as to get rid of any unnecessary datums such as computes, features and records that do not assist in predicting dissolved oxygen levels like Time, Date etc. This step is critical in improving the quality of the dataset and therefore, all the subsequent analyses are based on relevant data.
2. Data Splitting: In this stage, the cleaned dataset is divided into two primary components and these are the input features and the target variable which in this case is dissolved oxygen. After this separation, the data set is split into the training set and the testing set with the aim of performing an efficient evaluation of the models developed. It is common practice to use 80 percent of the data for training the models and the remaining 20 percent to test the ability of the models to predict.
3. Data Normalization: The features should also be standardized during the training of the model so that all of them make an equal contribution The next step is normalization. This process uses StandardScaler in which the features are normalized by taking away the mean and dividing the result by SD. Thus, all features are standardized, or normalized, which is important for increasing model performance. This scaler must therefore be applied across the training as well as the testing data sets in order to prevent data leakage.
4. Exploratory Data Analysis (EDA): Initially, an exploration of data is conducted in order to get a feel of how various features and the target variable are related to each other. This step involves analysing correlation coefficients – that is, Pearson or even Spearman to determine the extent to which each of the features was associated with the dissolved oxygen levels. These correlations are presented visually, using heat maps, which help to determine the main contributors to dramatically low dissolved oxygen levels.
5. Model Creation: When the data has been preprocessed and cleaned then various machine learning models are built with the data provided. This comprises the Linear Regression and Support Vector Regression being a standard type of architectures such as Artificial Neural Networks (ANNs). While constructing a ANN two of the most important tools that can be employed are Keras or TensorFlow, into which additional layers can be incorporated to capture elaborate features of the data.
6. Model Evaluation: Once these models are trained they should be tested and evaluated using statistical measures to include the coefficient of determination- R^2 and mean squared error–MSE. All these metrics are essential in that they quantitatively compare how well each proposed model estimates dissolved oxygen.
7. Comparison and Selection: Last, it is evaluated which of the different types of models produces the best result in terms of dissolved oxygen levels. Comparing R^2 and MSE of all the models, the highest R^2 and the lowest MSE of the models are used to choose the best model. This model will be deployed into practice employing this selection or may be fine-tuned for enhanced prediction ability further.

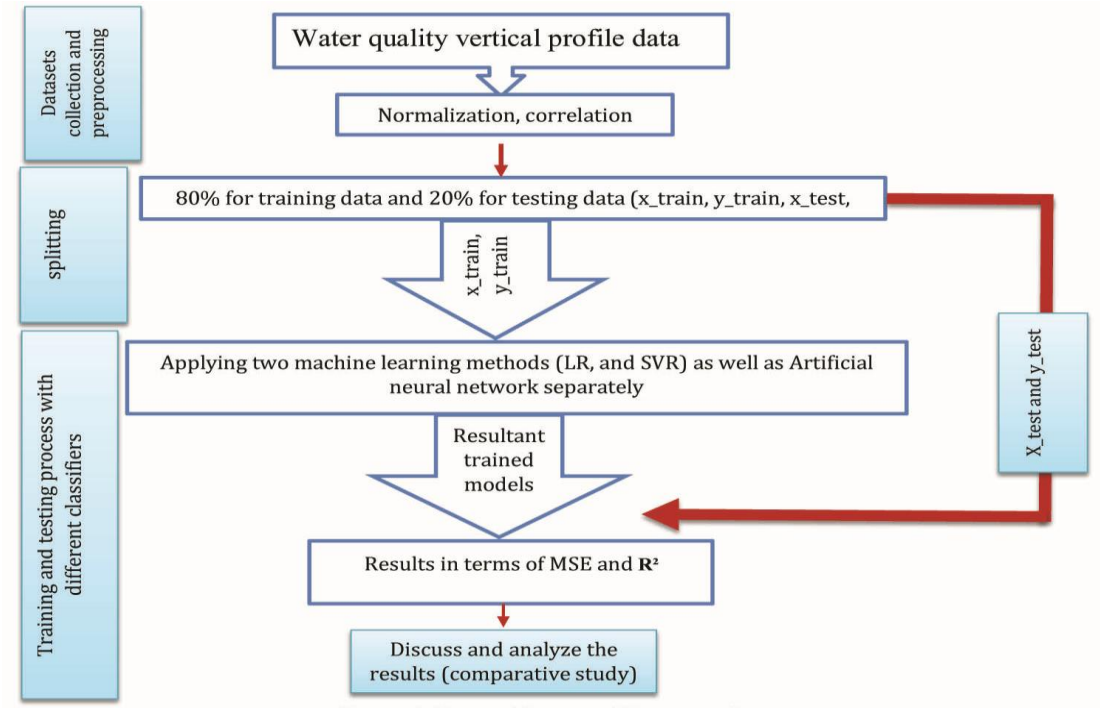

Figure 1: General Proposed Framework

4. Results and Discussions

The methods of designing a dissolved oxygen forecasting model in lakes are intricate and multiple-tiered and aimed at providing accurate and efficient results. The performance of the proposed model is evaluated using various metrics, with the two most important metrics being R^2 and mean squared error (MSE).

Following the division of the dataset into training and testing sets, the sizes of each set are illustrated in Figure 2. The training set is designed for model development, while the testing set is reserved solely for evaluating the model's performance. The figure provides a clear representation of the distribution of data between the two sets, highlighting their respective sizes.

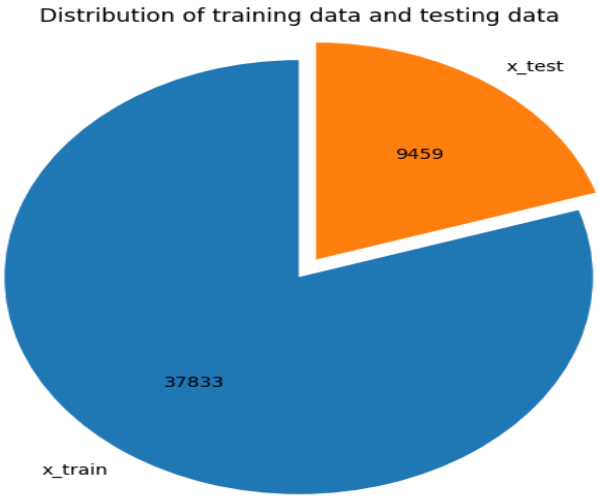

FIGURE 1: THE DISTRIBUTION OF SPLITTING DATASET TO TRAINING AND TESTING

The parameter values for the Linear Regression (LR) and Support Vector Regression (SVR) models were configured to the standard default settings commonly utilized in Python programming, where we used Colab notebook. The shallow neural network component of this paper consists of one hidden layer with leaky ReLU activation function, containing sixteen neurons, along with an input layer and an output layer with one neuron. A linear activation function is employed in the output layer to perform the regression task. Table 1 illustrates the architecture of the Artificial neural network used. This network trained with 20 epochs.

TABLE 1: THE ARCHITECTURE OF THE SHALLOW ARTIFICIAL NEURAL NETWORK

Layer(type)	Activation function	Output shape	# parameters
Input layer	-	(None, 5) =numper of features in dataset=x_trian size	0
Dense hidden layer (16 units)	Leaky-ReLU	(None, 16)	96
Dropout layer with (0.2) rate	-	(None, 16)	0
Output dense layer (one unit)	linear	(None, 1)	17

The history of training a proposed shallow neural network shows the training and validation MSE across 20 epochs, as illustrated in the figure (3). The data presented in Figure (3) illustrates that the Mean Squared Error (MSE) steadily converges over the course of 20 epochs. This indicates that as the training progresses, the model's

predictions become increasingly accurate, demonstrating effective learning and improvement in performance throughout the training process.

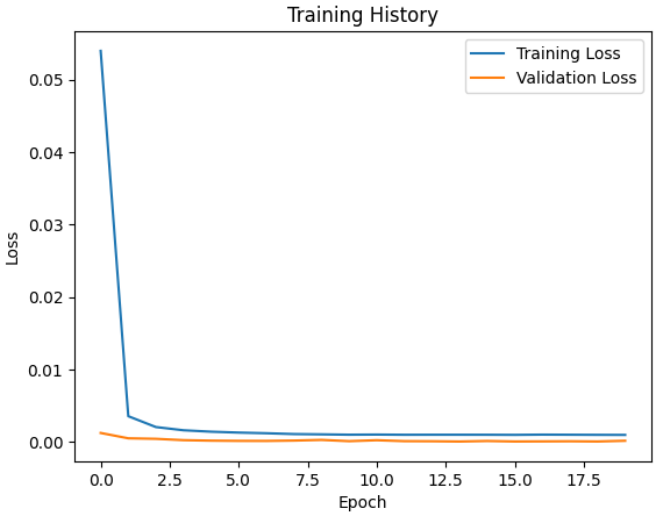

FIGURE 2: TRAINING HISTORY OF SHALLOW ANN IN TERM OF MSE

After training the three proposed classifiers, we conducted a thorough evaluation using the testing dataset. The performance metrics assessed included the coefficient of determination (R^2) and the mean squared error (MSE). The results of this evaluation are detailed in Table 2, which presents a comprehensive comparison of each classifier's accuracy and reliability.

TABLE 2: THE PERFORMANCE RESULT OF THREE REGRESSION MODELS IN TERMS OF R^2 AND MSE

AI model	R^2	MSE
LR	99.7%	0.000096
SVR	94.4%	0.002068
ANN	99.5%	0.000161

Table 2 clearly illustrates that the Linear Regression (LR) model outperforms the others, boasting the highest R^2 value and the lowest Mean Squared Error (MSE). This superiority can be attributed to the inherent strengths of the LR model in handling regression tasks, where it excels due to its straightforward design and focus on linear relationships.

In the course of this research work, the artificial neural network (ANN) was trained using only one hidden layer, and it turned out that this particular configuration gave an R^2 value that was higher than that even recorded when using ANN with multiple hidden layers configuration. The use of four layers gave an R^2 of 92% while much better results of R^2 of 95% was attained when three layers were utilized. In relation to the results, it indicates that expansion on the hidden layer number can cause serious overfitting which is very detrimental to the model actualization because it captures the noise instead of the real information in the training data. In case the practice of overfitting is prevalent in a design, reducing the number of hidden layers can help the opposite of overfitting, and in such a case the design becomes simpler, and the ability of the model to generalize improves.

While comparing our outcomes with the prior research analyses, it is crucial to state that adding new variables to the model and expanding the sample increased the R^2 value for this investigation to only 0.977, which is considered breakthrough in the field. For instance, the study conducted in [2] achieved R^2 values of 96.3 % by using linear

regression (LR) analysis and 93 % by artificial neural network (ANN) analysis demonstrations good performance but lower than our pioneer studies. Similarly, in [3] a 91.2% R^2 was observed using AdaBoost; this together with the MIC-SVR technique having a 90% R^2, found in the study outlined in [11], shows that RMSE differs in efficiency for differing techniques.

Curiously, only the study discussed in [12] succeeded in achieving the same level of R^2 as high as our presented performance, which equals 99.7%. However, as it will be remembered, this research used the WTD-GWO-SVR method which has more computational complication when compared with the simple linear regression method adopted in this study. It also provides a foundation to emphasize the affordability and superiority of our approach while also pointing out the opportunity for improvement in computer resource usage in this subfield in the future.

5. Conclusion

This research addresses a key problem in environmental science, which is the forecasting of dissolved oxygen levels that are essential for supporting aquatic life and water quality. Artificial intelligence models were used for the study which bolsters the age old methods which are often inefficient and inaccurate thus providing immediate gains towards the monitoring of water and conserving the lake. Various AI techniques were applied, including Linear Regression, Support Vector Regression (SVR), and Artificial Neural Networks (ANN). Each of the models was specifically tuned and pre processed, including standardization of features with StandardScaler, to achieve superior performance. The analysis showed that R^2 value for Linear Regression was 0.9974 whereas for SVR and DNN It was 0.9448 and 0.9875 respectively. This support the strength of the model as well as the ability of it to predict in this form of regression with high accuracy.

The research provides an impressive platform for subsequent studies that can seek to investigate more sophisticated prediction or further use multi-modal data. This encompasses local weather effects, human activities, geographical features etc., to improve forecasting accuracy further. By showcasing the ability to adapt and expand methodologies to various ecosystems and environmental parameters, this study strengthens its contribution to the field of AI in environmental science. Overall, the findings illuminate the transformative potential of AI-driven approaches in environmental modeling, setting the stage for innovative tools that can effectively forecast and manage water quality within lake ecosystems. Future avenues of research may include the exploration of ensemble models, additional feature integration, and real-time deployment strategies, all aimed at enriching our understanding of environmental dynamics. At this point, this study is limited to one dataset only which may hamper the generalization of the results of this research. To improve the reliability of the results and widening the application for this methodology, it is advised to use it with other data sets in the future. This will give an opportunity for a broader look at the success of the approach to other areas of practice.

Acknowledgements

This study was completely supported by the authors and has not received any financial support from any organization

References

[1] D. Pan, Y. Zhang, Y. Deng, J. Van Griensven Thé, S. X. Yang, and B. Gharabaghi, "Dissolved Oxygen Forecasting for Lake Erie's Central Basin Using Hybrid Long Short-Term Memory and Gated Recurrent Unit Networks," Water, vol. 16, no. 5, p. 707, 2024.

[2] N. Wu, J. Huang, B. Schmalz, and N. Fohrer, "Modeling daily chlorophyll a dynamics in a German lowland river using artificial neural networks and multiple linear regression approaches," Limnology, vol. 15, pp. 47-56, 2014.

[3] Y. Seo, S. Kim, O. Kisi, and V. P. Singh, "Daily water level forecasting using wavelet decomposition and artificial intelligence techniques," Journal of Hydrology, vol. 520, pp. 224-243, 2015.

[4] E. Olyaie, H. Z. Abyaneh, and A. D. Mehr, "A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in Delaware River," Geoscience Frontiers, vol. 8, no. 3, pp. 517-527, 2017.

[5] M. H. Ahmed and L.-S. Lin, "Dissolved oxygen concentration predictions for running waters with different land use land cover using a quantile regression forest machine learning technique," Journal of Hydrology, vol. 597, p. 126213, 2021.

[6] H. G. Kim, S. Hong, K.-S. Jeong, D.-K. Kim, and G.-J. Joo, "Determination of sensitive variables regardless of hydrological alteration in artificial neural network model of chlorophyll a: Case study of Nakdong River," Ecological modelling, vol. 398, pp. 67-76, 2019.

[7] D. Antanasijević, V. Pocajt, A. Perić-Grujić, and M. Ristić, "Modelling of dissolved oxygen in the Danube River using artificial neural networks and Monte Carlo Simulation uncertainty analysis," Journal of Hydrology, vol. 519, pp. 1895-1907, 2014.

[8] V. Ranković, J. Radulović, I. Radojević, A. Ostojić, and L. Čomić, "Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia," Ecological Modelling, vol. 221, no. 8, pp. 1239-1244, 2010.

[9] M. B. K. Prasad, W. Long, X. Zhang, R. J. Wood, and R. Murtugudde, "Predicting dissolved oxygen in the Chesapeake Bay: applications and implications," Aquatic sciences, vol. 73, pp. 437-451, 2011.

[10] O. Kisi, M. Alizamir, and A. Docheshmeh Gorgij, "Dissolved oxygen prediction using a new ensemble method," Environmental Science and Pollution Research, vol. 27, no. 9, pp. 9589-9603, 2020.

[11] W. Li et al., "Concentration estimation of dissolved oxygen in Pearl River Basin using input variable selection and machine learning techniques," Science of The Total Environment, vol. 731, p. 139099, 2020.

[12] D. Feng, Q. Han, L. Xu, F. Sohel, S. G. Hassan, and S. Liu, "An ensembled method for predicting dissolved oxygen level in aquaculture environment," Ecological Informatics, vol. 80, p. 102501, 2024.

[13] L. Durell, J. T. Scott, D. Nychka, and A. S. Hering, "Functional forecasting of dissolved oxygen in high-frequency vertical lake profiles," Environmetrics, vol. 34, no. 4, p. e2765, 2023.

[14] I. Suaza Sierra, "Predictive Understanding of Lake Water Temperature and Dissolved Oxygen Profiles Across the Red River Basin Through Interpretable Machine Learning," 2024.

[15] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Linear regression," in An introduction to statistical learning: With applications in python: Springer, 2023, pp. 69-134.

[16] T. M. Hope, "Linear regression," in Machine learning: Elsevier, 2020, pp. 67-81.

[17] D. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning," Journal of Applied Science and Technology Trends, vol. 1, no. 2, pp. 140-147, 2020.

[18] F. Zhang and L. J. O'Donnell, "Support vector regression," in Machine learning: Elsevier, 2020, pp. 123-140.

[19] M. Ibrahim et al., "Artificial neural network modeling for the prediction, estimation, and treatment of diverse wastewaters: a comprehensive review and future perspective," Chemosphere, p. 142860, 2024.

[20] M. Al-Shawwa and S. S. Abu-Naser, "Predicting birth weight using artificial neural network," in International conference on multidisciplinary science, 2024, vol. 2, no. 3, pp. 5-10.

[21] P. Thakkar, S. Khatri, D. Dobariya, D. Patel, B. Dey, and A. K. Singh, "Advances in materials and machine learning techniques for energy storage devices: A comprehensive review," Journal of Energy Storage, vol. 81, p. 110452, 2024.

[22] M. A. Chowdhury et al., "Recent machine learning guided material research - A review," Computational Condensed Matter, vol. 29, p. e00597, 2021/12/01/ 2021, doi: https://doi.org/10.1016/j.cocom.2021.e00597.

[23] J. Gao, "R-Squared (R2)–How much variation is explained?," Research Methods in Medicine & Health Sciences, vol. 5, no. 4, pp. 104-109, 2024.

[24] T. O. Hodson, T. M. Over, and S. S. Foks, "Mean squared error, deconstructed," Journal of Advances in Modeling Earth Systems, vol. 13, no. 12, p. e2021MS002681, 2021.