Available online www.jsaer.com

Journal of Scientific and Engineering Research, 2023, 10(10):148-163

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Water Quality Prediction Using Artificial Intelligence

Siva Sathyanarayana Movva

President, Innovations for Water Corp.
Email id: sivasathya@gmail.com Contact: 0014845156305

Abstract In recent years, water quality has faced threats from various pollutants, underscoring the importance of modeling and predicting water quality to control pollution. This study focuses on utilizing advanced artificial intelligence (AI) algorithms to predict both the water quality index (WQI) and water quality classification (WQC). For WQI prediction, nonlinear autoregressive neural network (NARNET) and long short-term memory (LSTM) deep learning algorithms were developed. Additionally, three machine learning algorithms—support vector machine (SVM), K-nearest neighbor (K-NN), and Naive Bayes—were employed for WQC forecasting. The dataset utilized encompassed seven significant parameters, and the models were assessed based on various statistical parameters. Results indicate that the proposed models accurately predict WQI and demonstrate robustness in water quality classification. Specifically, the NARNET model outperformed LSTM slightly in WQI prediction, while SVM achieved the highest accuracy (97.01%) in WQC prediction. Moreover, both NARNET and LSTM models exhibited similar accuracy during testing, with a slight difference in the regression coefficient (RNARNET = 96.17% and RLSTM = 94.21%). Such promising research holds significant potential for enhancing water management practices.

Keywords water quality index, water quality classification, artificial intelligence, machine learning

1. Introduction

Water stands as the paramount lifeline, essential for sustaining the existence of myriad creatures, including humans. Vital for the continuation of life, water must maintain a certain level of quality to support living organisms. There exist thresholds for pollution that water species can endure, surpassing which jeopardizes their existence and imperils their survival. The quality of various water bodies such as rivers, lakes, and streams is gauged by specific standards, indicative of their overall health. Similarly, different applications necessitate water of distinct standards. For instance, irrigation water should not harbor excessive salinity or toxic substances detrimental to plants or soil, which could disrupt ecosystems. Industrial water quality prerequisites also vary depending on the unique demands of industrial processes. Each water application requires meticulous attention to its specific standards, ensuring the sustenance of life and the smooth operation of various activities dependent on this invaluable resource. Natural sources of freshwater, including ground and surface water, are essential resources. However, human and industrial activities, along with natural processes, can pollute these resources. Rapid industrial development has led to a concerning decline in water quality. Additionally, inadequate infrastructure and lack of public awareness contribute to the degradation of drinking water quality. The consequences of polluted drinking water are severe, affecting health, the environment, and infrastructure. According to a United Nations report, approximately 1.5 million people perish each year due to diseases caused by contaminated water. In developing nations, it is estimated that 80% of health issues stem from polluted water sources, resulting in five million deaths and 2.5 billion illnesses annually. This mortality rate surpasses that of accidents, crimes, and terrorist attacks. Some of the affordable sources of fresh water, such as ground and surface water, are natural resources. However, these resources can become contaminated due to human/industrial activities and natural processes. Consequently, rapid industrial growth has led to a concerning deterioration in water quality. Additionally, inadequate infrastructure, coupled with a lack of public awareness and poor hygiene practices, significantly impacts the quality of drinking water. The consequences of polluted drinking water pose serious risks to health, the environment, and infrastructure. According to a report by the United Nations (UN), approximately 1.5 million people die each year from waterborne diseases caused by contaminated water. In developing countries, it is estimated that 80% of health issues are linked to contaminated water. Annually, five million deaths and 2.5 billion illnesses are attributed to this issue. This mortality rate surpasses deaths resulting from accidents, crimes, and terrorist attacks.

Hence, it is crucial to propose novel methodologies for analyzing and, if feasible, predicting water quality (WQ). Incorporating the temporal dimension into forecasting WQ patterns is recommended to monitor seasonal changes effectively. However, employing a combination of specialized model variations for WQ prediction yields superior outcomes compared to relying on a single model. Numerous methodologies have been suggested for WQ prediction and modeling, encompassing statistical techniques, visual modeling, algorithmic analysis, and predictive algorithms. To ascertain correlations and relationships among various water quality parameters, multivariate statistical techniques are utilized. Geostatistical approaches, including transitional probability, multivariate interpolation, and regression analysis, have also been employed. The significant rise in population, industrial activities, and agricultural practices involving fertilizers and pesticides has exerted profound impacts on WQ environments. Therefore, having predictive models for WQ is immensely beneficial for monitoring water contamination. Presently, there are two primary types of models used for modeling and predicting water quality: mechanism-oriented and non-mechanism-oriented models. Mechanism-oriented models are relatively sophisticated, utilizing advanced system structural data to simulate water quality. Consequently, they are regarded as multifunctional models suitable for various types of water bodies. Additionally, the Streeter–Phelps (S–P) model, one of the earliest water quality simulation models, has seen widespread use. Later, some countries have developed a variety of WQ models including the QUAL model and the WASP model, which have gained wide usage in mimicking the water quality of rivers. This was followed by Warren and Bach who suggested to use MIKE21 for designing systems to model the estuaries, coastal waters, and seas. Hayes et al. have paired two models for improving the quality of downstream water, namely, quasi-static two-dimensional dissolved oxygen reservoir model (DORM-II) and a daily scale optimal dispatch model.

A two-dimensional numerical model utilizing Environmental Fluid Dynamics Code (EFDC) was crafted to replicate the water dynamics of the Mudan River, employing distance-based points and intervals. In a separate investigation, Batur and Maktav utilized satellite image fusion, employing Principal Component Analysis (PCA), to forecast the Water Quality (WQ) of Lake Gala in Turkey. Jaloree et al. endeavored to anticipate the WQ of the Narmada River, employing a decision tree model integrating five WQ indicators. Additionally, a study recommended deploying the deep Bidirectional Stacked Simple Recurrent Unit (Bi-S-SRU) for the development of an accurate forecasting system for WQ in smart mariculture operations. Liao and Sun formulated a predictive model for forecasting the water quality of China's Chao Lake by integrating artificial neural networks (ANN) with a decision tree algorithm. Yan and Qian introduced an affinity propagation clustering model utilizing a least-squares support vector machine (AP-LSSVM), notable for its susceptibility to vacancies. Solanki et al. investigated and forecasted chemical eigenvalues of water, particularly dissolved oxygen and pH, employing a deep learning network model. Liao and Sun devised a model for forecasting the water quality of China's Chao Lake by integrating an Artificial Neural Network (ANN) with a decision tree algorithm. Yan and Qian introduced an Affinity Propagation (AP) clustering model that incorporates a Least-Squares Support Vector Machine (LSSVM), which exhibits high sensitivity to vacancies. Solanki et al. examined and predicted chemical eigenvalues of water, particularly dissolved oxygen and pH, utilizing a deep learning network model, demonstrating superior accuracy compared to supervised learning techniques. Li et al. developed an innovative hybrid model employing both neural network and Markov chain methods, facilitating dissolved oxygen prediction, a crucial water quality indicator. Khan and See incorporated dissolved oxygen, chlorophyll, conductivity, and turbidity into their water quality model utilizing an Artificial Neural Network (ANN). Yan et al. proposed employing a Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) algorithm to enhance Backpropagation (BP) neural networks for predicting oxygen demand in lakes, resulting in significantly improved prediction accuracy. Several studies have been performed to model and predict the water quality using different ANN models. These studies have approved the feasibility and effectiveness of employing ANN applications to predict the quality of drinking water. Currently, researchers mostly emphasize enhancing the applicability and reliability of water quality prediction/modelling by using a variety of new technologies such as Fuzzy logic, stochastic, ANN, and deep learning. Shafi et al. proposed four machine learning algorithms,namely, Support Vector Machines (SVM), Neural Networks (NN), Deep Neural Networks, and k-Nearest Neighbors (kNN), for the prediction of water quality. Using single feed-forward neural networks to classify water quality, 25 parameters have been included as input parameters. Ranković et al.estimated the dissolved oxygen (DO) by employing the ANN model. Gazzaz et al.estimated the WQI by using an ANN model, and the Internet of Things (IOT) technology was applied to collect the dataset from water resources. Abyaneh has applied the machine learning approaches like ANN and regression to predict the chemical oxygen demand (COD). Sakizadeh used ANN with Bayesian regularization to estimate the water quality index (WQI). However, the radial-basis-function (RBF), a type of the ANN model, was used for the prediction and classification of water quality.

In addition, it has been reported that deep learning methods showed high performance in predicting the WQ when compared to the traditional methods. Marir et al. developed a model to find out the uncommon behavior from large-scale network traffic data. While a deep learning algorithm was employed for extracting features, a multilayer ensemble support vector machine model was used for classification. Fadlullah et al. visualized a reward-based deep learning structure combining a deep convolutional neural network and a deep belief network. For the analysis and prediction of theWQ of groundwater, different algorithms including ANN, Bayesian neural networks, adaptive neuro fuzzy , decision support system (DSS), and autoregressive moving average (ARMA) have been applied. However, these mimicking models have some limitations. However, the contributions of the current study can be summarized as follows:

(i) Developing highly efficient advanced artificial intelligence models to predict the water quality index Applied Bionics and Biomechanics (WQI) based on artificial neural networks and deep learning algorithms

(ii) Applying some machine learning models, namely, support vector machine (SVM), K-nearest neighbour (K-NN), and Naive Bayes algorithms, for the prediction of water quality classification (WQC).

The highly efficient developed models can be generalized and used to forecast the water pollution process which will help the decision-makers to make the right decisions at the right time.

2. Materials and Methods

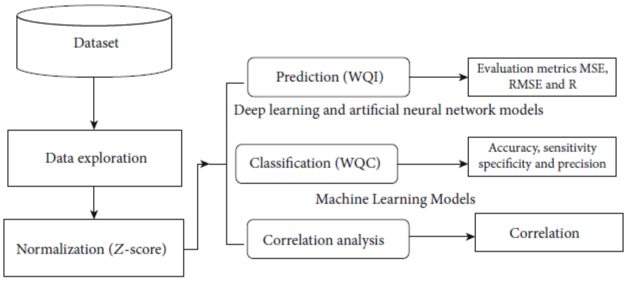


Figure 1: The proposed methodology of the present study

A. Dataset

The dataset utilized in this research was gathered from specific historical sites across India. It comprised 1679 samples taken from various Indian states spanning from 2005 to 2014. This dataset encompasses 7 important parameters, namely, dissolved oxygen (DO), pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform. The Indian government collected this data to monitor the quality of drinking water provided. The dataset was sourced from Kaggle at https://www.kaggle.com/anbarivan/indian-water-quality-data.

B. Data Preprocessing

The processing phase is very important in data analysis to improve the data quality. In this phase, the WQI has been calculated from the most significant parameters of the dataset. Then, water samples have been classified on the basis of the WQI values. For obtaining superior accuracy, the z-score method has been used as a data normalization technique.

Water Quality Index Calculation

To measure water quality, WQI is used to be calculated using various parameters that significantly affect WQ. In this study, a published dataset is considered to test the proposed model, and seven significant water quality parameters are included. The WQI has been calculated using the following formula

$$\text{WQI} = \frac{\sum_{i=1}^{N} q_i \times w_i}{\sum_{i=1}^{N} w_i}, \tag{1}$$

where: N is the total number of parameters included in the WQI calculationsqi is the quality rating scale for each parameter i calculated by equation (2) below, and wi is the unit weight for each parameter calculated by equation (3).

$$q_i = 100 \times \left(\frac{V_i - V_{\text{Ideal}}}{S_i - V_{\text{Ideal}}}\right), \tag{2}$$

where: Vi is the measured value of parameter i in the tested water samples V_{Ideal} is the ideal value of parameter i in pure water (0 for all parameters except DO = 14:6 mg/l and pH = 7:0), and Si is the recommended standard value of parameter i (as shown in Table 1).

$$w_i = \frac{K}{S_i}, \tag{3}$$

where K is the proportionality constant that can be calculated as follows:

$$K = \frac{1}{\sum_{i=1}^{N} S_i}, \tag{4}$$

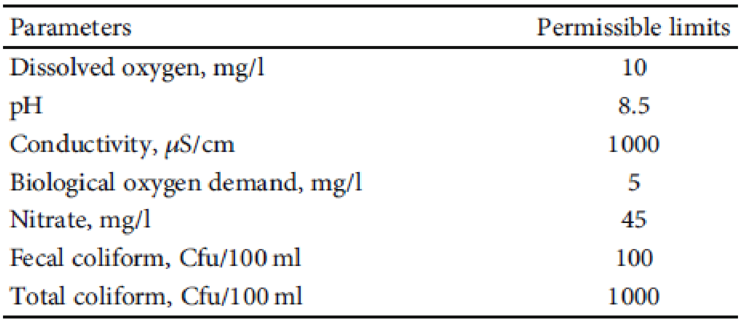
Tables 2 and 3 represent the unit weight of each parameter and the WQC, respectively.

Z-Score Normalization Method

Normalization is a way to simplify calculations. It is a dimensional expression transformed into a nondimensional expression and becomes a scalar. Z-score normalization (or normalization score) is a normalization method used to normalize parameters by using the mean (μ) and standard deviation (σ) values of the tested data. It can be calculated as follows:

$$Z\text{-score} = \frac{(x - \mu)}{\sigma}, \tag{5}$$

Table 1: Permissible limits of the parameters used in calculating WQI [43].



Parameters	Permissible limits
Dissolved oxygen, mg/l	10
pH	8.5
Conductivity, μS/cm	1000
Biological oxygen demand, mg/l	5
Nitrate, mg/l	45
Fecal coliform, Cfu/100 ml	100
Total coliform, Cfu/100 ml	1000

Table 2: Parameter unit weights.

Parameter	Unit weight (w_i)
Dissolved oxygen	0.2213
pH	0.2604
Conductivity	0.0022
Biological oxygen demand	0.4426
Nitrate	0.0492
Fecal coliform	0.0221
Total coliform	0.0022

Table 3: Water quality classification (WQC) [42].

Water quality index range	Classification
0-25	Excellent
26-50	Good
51-75	Poor
76-100	Very poor
Above 100	Unsuitable for drinking

where x is the measured value of the parameter i in the tested sample.

Prediction of Water Quality Index

For this purpose, ANN models, namely, nonlinear autoregressive neural network (NARNET) and long short-term memory (LSTM) deep learning algorithm, were used for the prediction of water quality index.

Artificial Neural Network (ANN) Model

In general, the neural network (NN) models are used as very powerful machine learning algorithms for time-series prediction of different engineering applications. The ANN model has consisted of an input layer, a hidden layer/s, and an output layer. Each hidden layer has weight and bias parameters to manage neurons. To transfer the data from the hidden layer into the output layer, the activation function is used. The learning algorithms are used to select the weights within the NN framework. The weight selection is based on the minimum performance measures such as mean square error (MSE). The NARNET model is a very popular multilayer feedforward network. It starts with a guessed initial weight value, which is then updated using the actual data. Consequently, there is some sort of randomness in the prediction process performed by the NN model.

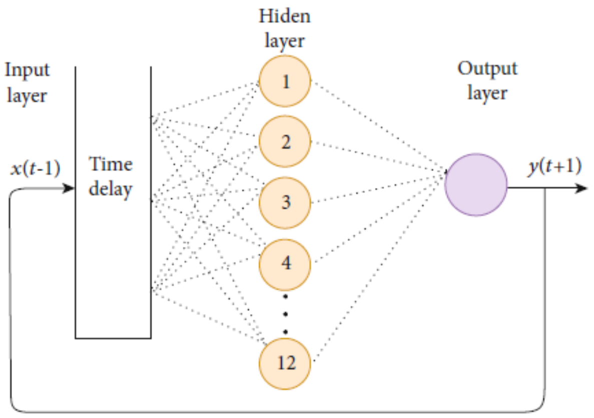


Figure 2: Computation of the NARNET model.

The network is regularly trained many times using different random values for the initialization, and the results are averaged. In the NARNET model, the number of hidden layers and nodes must be identified in advance.

Figure 2 displays the NARNET model scheme with multiple inputs and 4 hidden layers (as recommended for most of the research datasets). Equation (6) describes the NARNET time series model.

Table 4: Parameters of the developed ANN (NARNET) model.

Number of hidden layers	12
Number of delays	1 : 8
Maximum number of iterations	100
Maximum number of epochs	12
Number of gradients	1.734×10^3

$$y(t) = h(y(t-1), y(t-2), \cdots, y(t-p)) + \epsilon(t), \tag{6}$$

where y is the value of time-series data at time t and yðtÞ for employing the p observation values of the series. The function ðhÞ is used to optimize the network weights and neuron bias. Finally, the ϵðtÞ is the error obtained from the model at time t:

In this work, the NARNET model has been developed to predict the WQI. The NARNET model is a time series model that is used to predict the stationary time series compared with other ANN models like the forward neural network model. The WQI parameters seem in the form of time series; therefore, the NARNET model is proposed to predict theWQI. Table 4 shows the significant parameters of the developed model. Figure 3 represents the topology of the developed NARNET model.

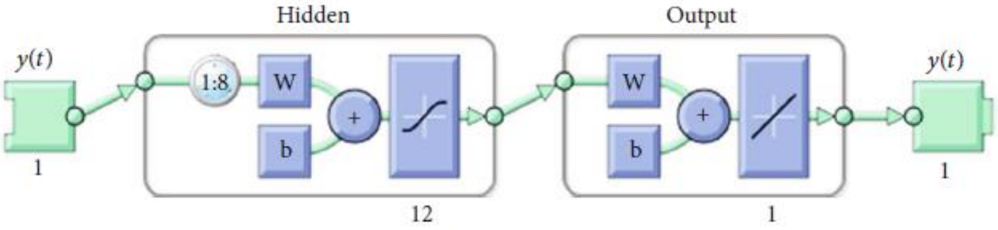


Figure 3: Architecture of the NARNET model.

Deep Neural Network (DNN) Model

The DNN model is one type of feed forward NN algorithms, which is a fundamental technique for deep learning. DNN consists of 3 levels of nodes, and each node follows a nonlinear function, except for the input node. DNN presents a technique of backpropagation supervised learning. In this work, a WQI model was developed using the DNN algorithm and the simple DNN was compared with the proposed model. This model includes the following parameters and functions: bias (b), input (x), output (y), weight (w), calculation function (α), and activation function f ðαÞ.

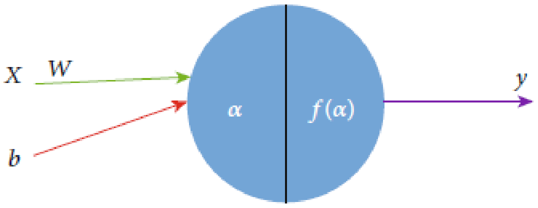


Figure 4: Architecture of the DNN model.

The neuron architecture of the DNN model is schematically shown in Figures 4 and 5. Every single neuron in the DNN employs the following equations

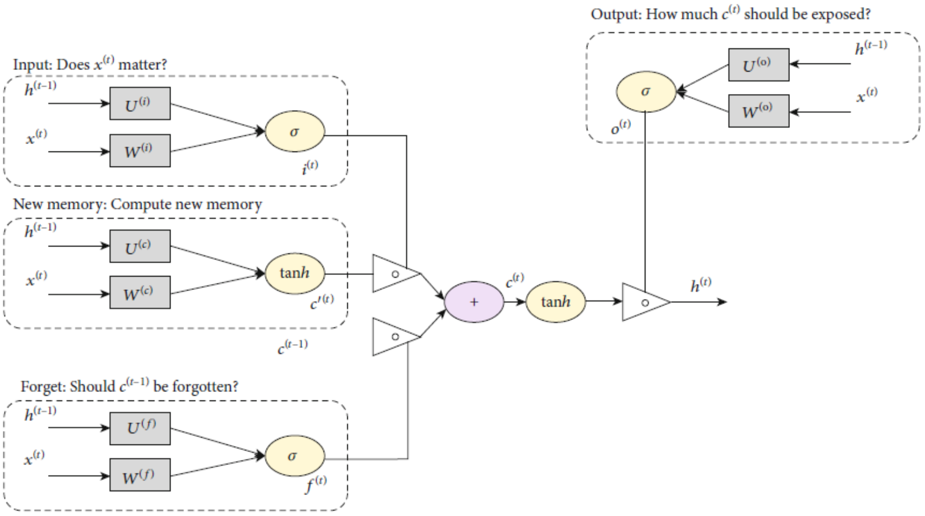


Figure 5: Architecture of the LSTM model.

$$\alpha : sum = w \bullet x + b, \tag{7}$$

$$y : f(\alpha) = f(w \bullet x + b), \tag{8}$$

Recurrent neural network (RNN) is one type of deep learning techniques used in different domains such as computer vision, natural language processing, pattern recognition, and medical image diagnosis. As compared to different feed ANNs, RNN has a directional control loop that enables the previous states to be stored, recalled, and added to the current output. One of the most powerful RNN algorithms used to predict time series data is the LSTM model.

The long short-term memory (LSTM) model, a deep learning algorithm, is appropriate for estimating the time series data whenever there is a randomized sized time step. The activating function used in the LSTM model is a logistic sigmoid. Providing that the forget gate is opened and the input gate is closed, the memory cell keeps reminding of the first entry and thus solving the typical RNN problems. The formulas of the RNN model are as follows:

$$h_t = \tan h(W_i \bullet h_t + w_x x_t), \tag{9}$$

$$y_t = w_y \bullet w_t, \tag{10}$$

where h_t is the hidden layer of NN for the input training data ðxt Þ. The output layer is represented by y_t . However, w_t and w_y are the weight of the neural cell and the matrix, respectively. The RNN model is used to create the LSTM model for the computing process. The LSTM consists of three significant parameters, namely, the input gate, forget gate, and output gate. The formulas used to compute the LSTM model are as follows:

$$\text{Input gate} : i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i), \tag{11}$$

$$\text{Forget gate} : f_t = \sigma\left(w_f \bullet [h_{t-1}, x_t] + b_f\right), \tag{12}$$

$$\text{Output gate} : o_t = \sigma(W_0 \bullet [h_{t-1}, x_t] + b_0), \tag{13}$$

$$\text{New memory cell} : \widetilde{c}_t = \tan h(W_c \bullet [h_{t-1}, x_t] + b_c), \tag{14}$$

$$\text{Final memory cell} : C_t = f_t \times C_{t-1} + i_t \times \widetilde{c}_t, \tag{15}$$

$$h_t = o_t \times \tan h(C_t), \tag{16}$$

where:

i_t , f_t , and o_t : input, forget, and output gates, respectively h_t : number of hidden layers

σ: the logistic sigmoid function is used to transfer the training data from a hidden layer into the output gate

w_t : the weighted neural network

$\sim c_t$: an internal memory cell is used to compute in the hidden layer

C_t : the internal memory

h_t :the output of a hidden layer state is used to derive from the new memory

i, f , and o : are subscripts that stand for input, forget, and output gates, respectively

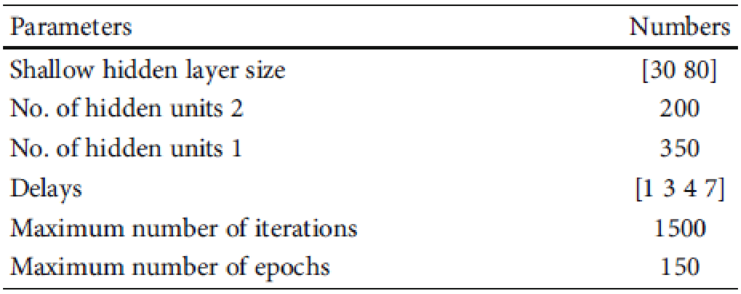
x_t : input training data

w_f , $w_o w_c$: weight vector of NN

b_f and b_o: bias vector in NN

The analysis of LSTM was performed utilizing MATLAB. Throughout the LSTM layer, 23 variables are open. We just set the units, activate the function, return the sequence, and dropout. Figure 5 illustrates the architecture of the LSTM, and the significant parameters of the LSTM model are presented in Table 5.

TABLE 5: Parameters of the LSTM model.



Parameters	Numbers
Shallow hidden layer size	[30 80]
No. of hidden units 2	200
No. of hidden units 1	350
Delays	[1 3 4 7]
Maximum number of iterations	1500
Maximum number of epochs	150

B. Prediction of Water Quality Index

In this section, some machine learning algorithms, namely, support vector machine (SVM), K-nearest neighbor (KNN), and Naive Bayes, have been used to predict the water quality classification.

Support Vector Machine (SVM) Model

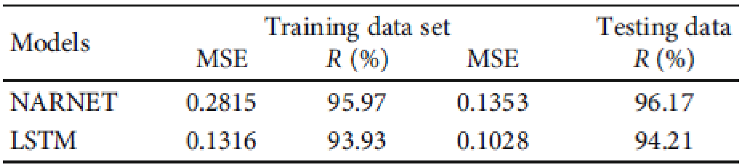
The SVM model was developed in 1995 by Corinna Cortes and Vapnik. It has several unique benefits in solving small samples, and nonlinear and high-dimensional pattern recognition. It can be extended to function in the simulation of other machine learning problems. It uses the hyperplane to separate the points of the input vectors and finds the needed coefficients.

The best hyperplane is the line with the largest margin, which is meant the distance between the hyperplane and the nearest input objects. The input points defined in the hyperplane are called support vectors. In this work, the linear SVM model along with the Gaussian radial basis function (equation (17)) is used to classify the tested water samples based on their quality.

$$K\left(X, X'\right) = \exp\left(-\frac{||X - X'||^2}{2\sigma^2}\right), \tag{17}$$

where X and X′ represent the feature vectors of the input dataset and the ll X − X′ll^2 is the squared Euclidean distance between the two feature inputs. The σ is the free parameter.

TABLE 6: Performances of the NARNET LSTM and ANN models to predict WQI.



Models	Training data set		Testing data	
	MSE	*R* (%)	MSE	*R* (%)
NARNET	0.2815	95.97	0.1353	96.17
LSTM	0.1316	93.93	0.1028	94.21

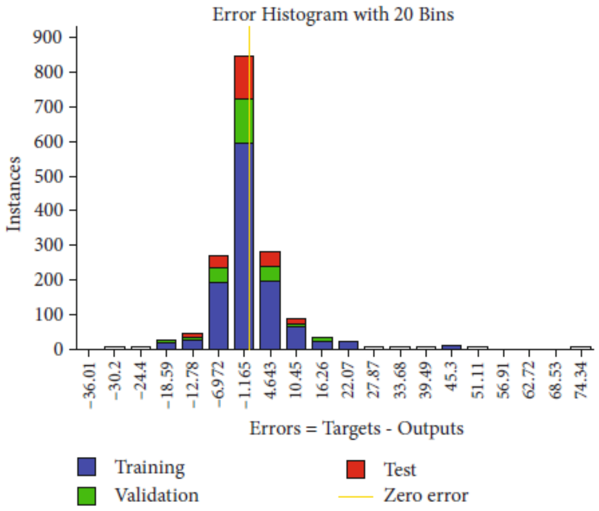


FIGURE 6: Histogram error of the NARNET model.

K-Nearest Neighbor (K-NN) Model

The K-NN algorithm is a basic classification and regression method. It is used to find the K values that are close to values in the training dataset. Most of these values belong to a certain class, and thus, tested data can be classified. The K value is used to find the closest points in the feature vectors, and the value should be unique. The following expression of the Euclidean distance function (Di) can be used.

$$D_i = \sqrt{(x_1 - x_2) + (y_1 - y_2)^2}, \tag{18}$$

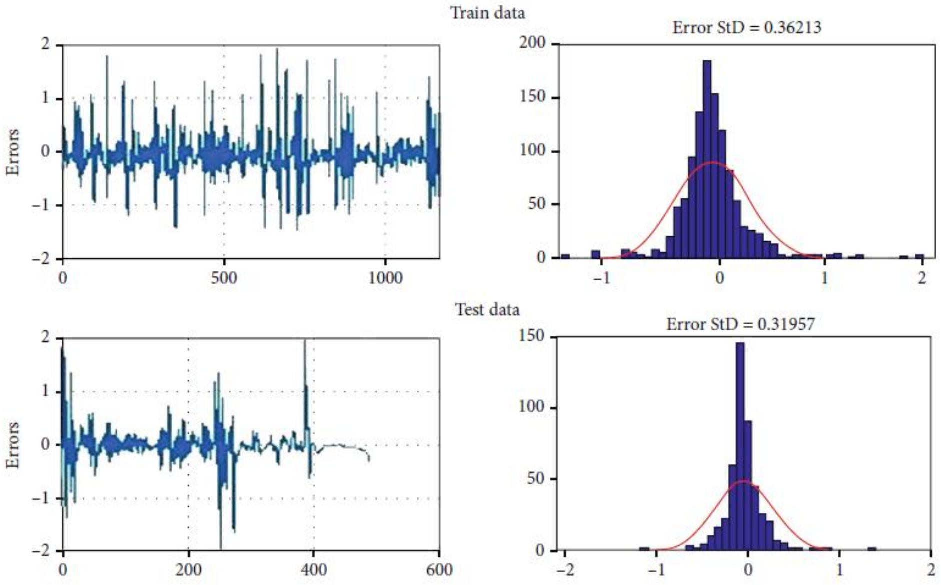


FIGURE 7: Histogram error and mean error of the LSTM model in the training and testing phases.

Naive Bayes Model

The Bayesian method uses the knowledge of probability statistics to predict and classify datasets. The Bayesian algorithm combines prior and posterior probabilities to avoid the supervisor’s bias and the overfitting phenomenon of using sample information alone. This Naive Bayes is a type of classification algorithms based on Bayes’ theorem and the assumption of the independence of characteristic conditions. Attributes are assumed

to be conditionally independent of each other when the target value is given. This method greatly simplifies the complexity of the Bayesian method.
In Bayesian analysis, the probability of an event A given an event B is not the same as the probability of B given A as in equation (18).

$$P(C \mid A) = \frac{P(C) \times P(A \mid C)}{P(A)}, \tag{20}$$

where the P(A) is a prior probability representing the feature vectors of the WQC dataset and P(A ∣ C) is the prior probability of the class of the WQC dataset.

3. Performance by Measurement

The statistical analysis, namely, mean square error (MSE), has been used to evaluate the robustness of the developed models to predict the WQI. However, the accuracy, specificity, sensitivity, precision, and F-score evaluation matrices were employed to evaluate the developed classification model to predict the WQC. The used statistical parameters were defined as follows:

(a) Mean Square Error (MSE)

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - y\wedge_i)^2, \tag{21}$$

where yi and ̂yi are the predicted and the observed responses, respectively, and N is the total number of variables.

(b) Accuracy

$$\mathrm{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} \times 100\%, \tag{22}$$

(c) Specificity

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \times 100\%, \tag{23}$$

(d) Sensitivity

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100\%, \tag{24}$$

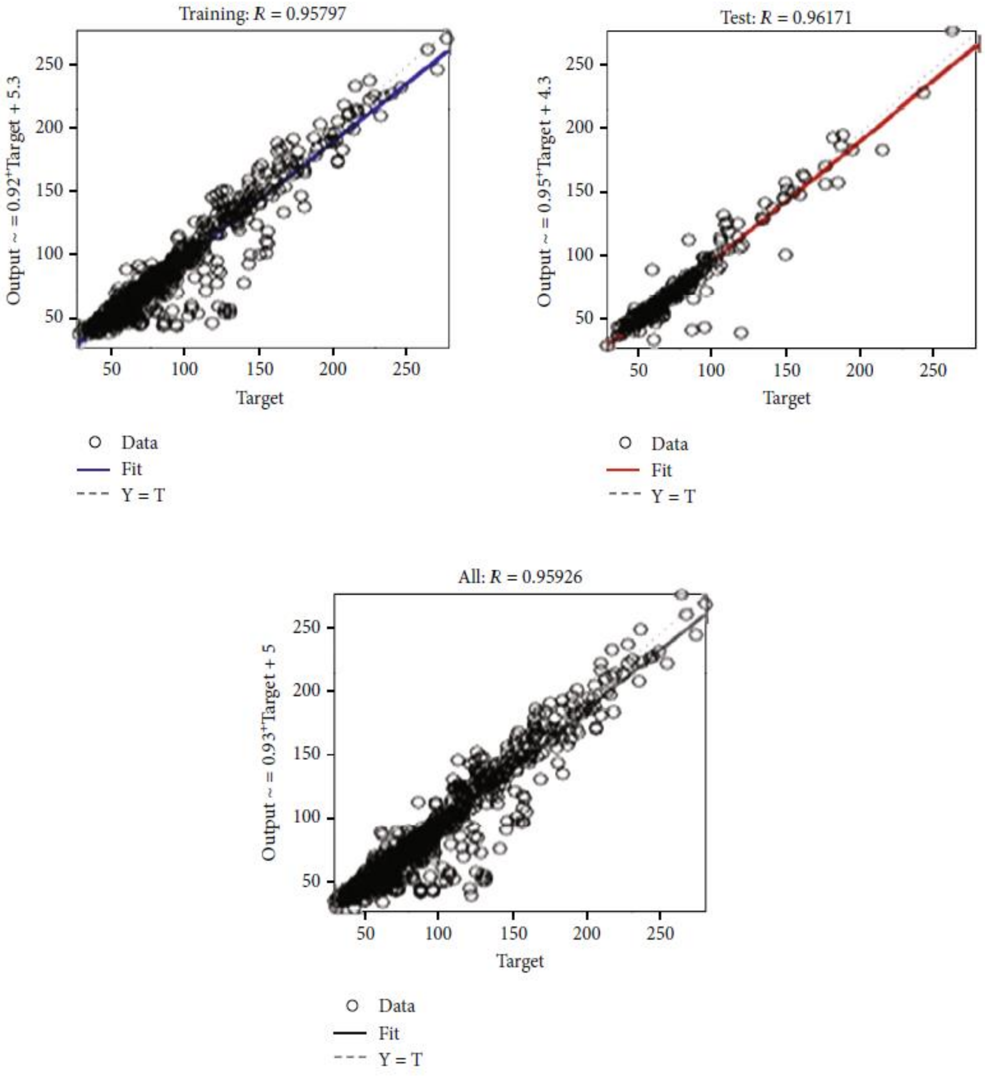


FIGURE 8: Regression plot of the NARNET model.

(e) Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%, \tag{25}$$

(f) F-score

$$F\text{-score} = \frac{2 \times \text{precision} \times \text{sensitivity}}{\textit{preision} + \text{sensitivity}} \times 100\%, \tag{26}$$

where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative, respectively.

Correlation Analysis

Pearson’s correlation coefficient approach is applied to analyze the correlation between the significant parameters of the dataset used for the prediction of the QWI values.

$$R = \frac{n\sum(x \times y) - (\Sigma x)(\Sigma y)}{[n\sum(x^2) - \sum(x^2)] \times [n\sum(y^2) - \sum(y^2)]} \times 100\%, \tag{27}$$

where: R: Pearson’s correlation coefficient approach
x: input values in the first set of the training data
y: input values of the second set of the training data
n: total number of input variables

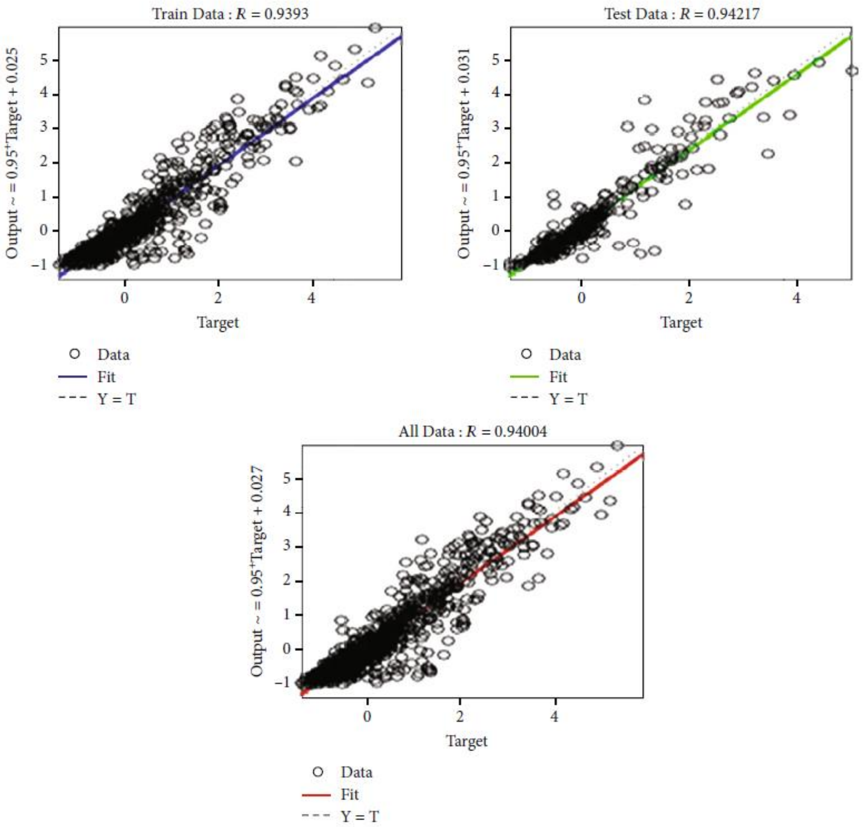


Figure 9: Regression plot of the LSTM model.

Table 7: Performance of Pearson's correlation coefficient approach.

Parameter	DO (mg/l)	pH	Conductivity (μS/cm)	BOD (mg/l)	Nitrate (mg/l)	Fecal coliform (MPN/100 ml)	Total coliform (MPN/100 ml)	WQI
DO (mg/l)	1.00	0.0466	-0.2914	-0.1819	-0.0347	0.1128	-0.1536	-0.3836
pH	0.0466	1.00	0.3268	0.2697	0.0562	-0.2082	-0.2170	0.5233
Conductivity (μS/cm)	-0.2914	0.3268	1.00	0.3288	0.1009	-0.1120	-0.0777	0.3935
BOD (mg/l)	-0.1819	0.2697	0.3288	1.00	0.2257	-0.1597	-0.1633	0.6130
Nitrate (mg/l)	-0.0347	0.0562	0.1009	0.2257	1.00	0.1408	0.0545	0.1768
Fecal coliform (MPN/100 ml)	-0.1128	-0.2082	-0.1120	-0.1597	0.1408	1.00	0.9119	0.2779
Total coliform (MPN/100 ml)	-0.1536	-0.2170	-0.0777	-0.1633	0.0545	0.9119	1.00	0.2679
WQI	-0.3836	0.5233	0.3935	0.6130	0.1768	0.2779	0.2679	1.00

Experimental Setup

The prediction experiments have been conducted in a specific environment (MATLAB 2018). The simulation has been performed using a system with i5 Processor and 4GB RAM to process all required tasks.

4. Results and Discussion

For validating the developed model, the dataset has been divided into 70% training and 30% testing subsets. While the ANN and LSTM models were used to predict the WQI,the SVM, KNN, and Naive Bayes were utilized for the water quality classification prediction.

TABLE 8: Performance of the used machine learning models to predict WQC.

Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	*F*-score (%)
SVM	97.01	99.23	97.78	94.93	98.54
KNN	83.63	84.73	94.93	87.50	85.84
Naive Bayes	75.20	77.76	91.65	78.08	81.51

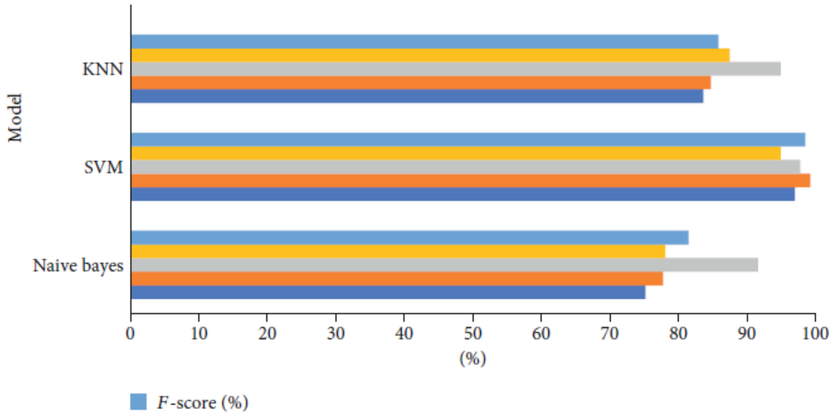


FIGURE 10: Performance of the machine learning algorithms used for the prediction of the WQC.

Prediction of the WQI

A NARNET model, with 12 hidden layers, showed a good performance to predict the WQI values. As presented earlier, it has the following characteristics: 1 : 8 number of delays and 12 number of epochs. However, the developed LSTM model has a total number of 200 hidden layers,150 maximum number of epochs, and delays of [1, 3, 4, 7].

Table 6 summarizes the performance parameters of the developed models to predict WQI, although the prediction accuracy of LSTM for the testing data was slightly better than that for the training data. In addition, the LSTM model, in general, has shown a slightly better performance compared with the NARNET model according to the MSE values. However, based on the R value, the NARNET model has shown a better performance. In general, both models demonstrated an excellent prediction of the WQI values with R% > 93:93.

Figure 6 illustrate the histogram error of the NARNET model. The histogram metric is used to find errors between the target values and the predicted values of training and testing datasets. The total error range is divided into 20 smaller bins, where the y-axis refers to the number of samples located in a particular bin. Figure 7 displays the histogram metric and mean errors of the LSTM model in the training and testing phases. The mean error and histogram metric are used to find the deviation between the observation values and the predicted values of training and testing.

Figures 8 and 9 display the regression plots for the predicted values of training, testing, and whole datasets for the NARNET and LSTM models, respectively. This plot is used to find the relationship between the predicted values and actual values. The "target" values in the plot are the actual dataset, whereas the "output" is the predicted values obtained from the NARNET and LSTM models. As shown in both figures, there is a clear good agreement (R > 95:7% (NARNET) and R > 93:3% (LSMT)) between the predicted WQI values and the ones calculated from the measured parameters. This implies the highly efficient performance of both developed models.

Table 7 summarizes the Pearson's correlation coefficient approach is used to predict the WQI values. The correlation between the WQI parameters for selecting the optimal parameters has been obtained. Results revealed that all parameters have a strong relationship with WQI parameters. This indicates that these parameters are very important for predicting the quality of water.

Prediction of the Water Quality Classification

This section presents the results of the classification algorithms are used to predict the WQC. Table 8 shows the results of the used machine learning algorithms. It is noted that the performance of the SVM algorithm is very superior as compared to the KNN and Naive Bayes models. However, the Naive Bayes algorithm has shown the poorest performance. Figure 10 shows the performance of the used algorithms to predict the WQC.

5. Conclusions

Modeling and prediction of water quality are very important for the protection of the environment. Developing a model by using advanced artificial intelligence algorithms can be used to measure the future water quality. In this proposed methodology, the advanced artificial intelligence algorithms, namely, NARNET and LSTM models were used to predict the WQI. Moreover, machine learning algorithms such as SVM, KNN, and Naive Bayes were used to classify the WQI data. The proposed models were evaluated and examined by some statistical parameters. For the WQI prediction, the result has revealed that the performance of the NARNET model is slightly better than the LSTM model based on the

6. Data Availability

The dataset used in this study is collected from certain historical locations in India. It contained 1679 samples from different Indian states during the period from 2005 to 2014. The dataset has 7 significant parameters named dissolved oxygen (DO), pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform. The data was collected by the Indian government to ensure the quality of the supplied drinking water. This dataset was obtained from Kaggle https://www.kaggle.com/anbarivan/indian-waterquality-data.

References

[1]. P. Zeilhofer, L. V. A. C. Zeilhofer, E. L. Hardoim, Z. M. Lima, and C. S. Oliveira, “GIS applications for mapping and spatial modeling of urban-use water quality: a case study in District of Cuiabá, Mato Grosso, Brazil,” Cadernos de Saúde Pública, vol. 23, no. 4, pp. 875–884, 2007.

[2]. M. A. Kahlown, M. A. Tahir, and H. Rasheed, National Water Quality Monitoring Programme, Fifth Monitoring Report (2005–2006), Pakistan Council of Research in Water Resources Islamabad, Islamabad, Pakistan, 2007, http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/Water% 20Quality%20Monitoring%20Report%202005-06.pdf.

[3]. UN water, “Clean water for a healthy world,” Development, 2010, https://www.undp.org/content/undp/en/home/ presscenter/articles/2010/03/22/clean-water-for-a-healthyworld.html.

[4]. K. Farrell-Poe, W. Payne, and R. Emanuel, Water Quality & Monitoring, University of Arizona Repository, 2000, http:// hdl.handle.net/10150/146901.

[5]. T. Taskaya-Temizel and M. C. Casey, “A comparative study of autoregressive neural network hybrids,” Neural Networks, vol. 18, no. 5–6, pp. 781–789, 2005.

[6]. C. N. Babu and B. E. Reddy, “A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data,” Applied Soft Computing, vol. 23, pp. 27–38, 2014.

[7]. X. Zhang, N. Hu, Z. Cheng, and H. Zhong, “Vibration data recovery based on compressed sensing,” Acta Physica Sinica, vol. 63, no. 20, pp. 119–128, 2014.

[8]. M. M. S. Cabral Pinto, C. M. Ordens, M. T. Condesso de Melo et al., “An inter-disciplinary approach to evaluate human health risks due to long-term exposure to contaminated groundwater near a chemical complex,” Exposure and Health, vol. 12, no. 2, pp. 199–214, 2020.

[9]. M. M. S. Cabral Pinto, A. P. Marinho-Reis, A. Almeida et al., “Human predisposition to cognitive impairment and its relation with environmental exposure to potentially toxic elements,” Environmental Geochemistry and Health, vol. 40, no. 5, pp. 1767–1784, 2018.

[10]. Y. C. Lai, C. P. Yang, C. Y. Hsieh, C. Y. Wu, and C. M. Kao, “Evaluation of non-point source pollution and river water quality using a multimedia two-model system,” Journal of Hydrology, vol. 409, no. 3-4, pp. 583–595, 2011.

[11]. J. Huang, N. Liu, M. Wang, and K. Yan, “Application WASP model on validation of reservoir-drinking water source protection areas delineation,” in 2010 3rd International Conference on Biomedical Engineering and Informatics, pp. 3031–3035, Yantai, China, October 2010.

[12]. I. R. Warren and H. K. Bach, "MIKE 21: a modelling system for estuaries, coastal waters and seas," Environmental Software, vol. 7, no. 4, pp. 229–240, 1992.

[13]. D. F. Hayes, J. W. Labadie, T. G. Sanders, and J. K. Brown, "Enhancing water quality in hydropower system operations," Water Resources Research, vol. 34, no. 3, pp. 471–483, 1998.

[14]. G. Tang, J. Li, Z. Zhu, Z. Li, and F. Nerry, "Two-dimensional water environment numerical simulation research based on EFDC in Mudan River, Northeast China," in 2015 IEEE European Modelling Symposium (EMS), pp. 238–243, Madrid, Spain, October 2015.

[15]. L. Hu, C. Zhang, C. Hu, and G. Jiang, "Use of grey system for assessment of drinking water quality: a case S study of Jiaozuo city, China," in 2009 IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2009), pp. 803–808, Nanjing, China, November 2009.

[16]. E. Batur and D. Maktav, "Assessment of surface water quality by using satellite images fusion based on PCA method in the Lake Gala, Turkey," IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 5, pp. 2983–2989, 2019.

[17]. S. Jaloree, A. Rajput, and G. Sanjeev, "Decision tree approach to build a model for water quality," Binary Journal of Data Mining & Networking, vol. 4, pp. 25–28, 2014.

[18]. J. Liu, C. Yu, Z. Hu et al., "Accurate prediction scheme of water quality in smart mariculture with deep Bi-S-SRU learning network," EEE Access, vol. 8, pp. 24784–24798, 2020.

[19]. H. Liao and W. Sun, "Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method," Procedia Environmental Sciences, vol. 2, pp. 970– 979, 2010.

[20]. L. Yan and M. Qian, "AP-LSSVM modeling for water quality prediction," in Proceedings of the 31st Chinese Control Conference, pp. 6928–6932, Hefei, China, July 2012.

[21]. A. Solanki, H. Agrawal, and K. Khare, "Predictive analysis of water quality parameters using deep learning," International Journal of Computers and Applications, vol. 125, no. 9, pp. 29–34, 2015.

[22]. X. Li and J. Song, "A new ANN-Markov chain methodology for water quality prediction," in 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–6, Killarney, Ireland, July 2015.

[23]. A. A. M. Ahmed and S. M. A. Shah, "Application of adaptive euro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River," Journal of King Saud University - Engineering Sciences, vol. 29, no. 3, pp. 237–243, 2017.

[24]. Y. Khan and C. S. See, "Predicting and analyzing water quality using Machine Learning: a comprehensive model," in 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), pp. 1–6, Farmingdale, NY, USA, April 2016.

[25]. J. Yan, Z. Xu, Y. Yu, H. Xu, and K. Gao, "Application of a hybrid optimized BP network model to estimate water quality parameters of Beihai Lake in Beijing," Applied Sciences, vol. 9, no. 9, p. 1863, 2019.

[26]. H. R. Maier, A. Jain, G. C. Dandy, and K. P. Sudheer, "Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions," Environmental Modelling & Software, vol. 25, no. 8, pp. 891–909, 2010.

[27]. S. Lee and D. Lee, "Improved prediction of harmful algal blooms in four major South Korea's rivers using deep learning models," International Journal of Environmental Research and Public Health, vol. 15, no. 7, p. 1322, 2018.

[28]. U. Shafi, R. Mumtaz, H. Anwar, A. M. Qamar, and H. Khurshid, "Surface water pollution detection using internet of things," in 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), pp. 92–96, Islamabad, Pakistan, October 2018.

[29]. Z. Ahmad, N. A. Rahim, A. Bahadori, and J. Zhang, "Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks," International Journal of River Basin Management, vol. 15, no. 1, pp. 79–87, 2016.

[30]. 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.

[31]. N. M. Gazzaz, M. K. Yusoff, A. Z. Aris, H. Juahir, and M. F. Ramli, "Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors," Marine Pollution Bulletin, vol. 64, no. 11, pp. 2409–2420, 2012.

[32]. H. Z. Abyaneh, "Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters," Journal of Environmental Health Science and Engineering, vol. 12, no. 1, p. 40, 2014.

[33]. M. Sakizadeh, "Artificial intelligence for the prediction of water quality index in groundwater systems," Modeling Earth Systems and Environment, vol. 2, no. 1, p. 8, 2016.

[34]. M. I. Yesilnacar, E. Sahinkaya, M. Naz, and B. Ozkaya, "Neural network prediction of nitrate in groundwater of Harran Plain, Turkey," Environmental Earth Sciences, vol. 56, no. 1, pp. 19–25, 2008.

[35]. M. Bouamar and M. Ladjal, "A comparative study of RBF neural network and SVM classification techniques performed on real data for drinking water quality," in 2008 5th International Multi-Conference on Systems, Signals and Devices, pp. 1–5, Amman, Jordan, July 2008.

[36]. N. Marir, H. Wang, G. Feng, B. Li, and M. Jia, "Distributed abnormal behavior detection approach based on deep belief network and ensemble SVM using spark," IEEE Access, vol. 6, pp. 59657–59671, 2018.

[37]. Z. M. Fadlullah, F. Tang, B. Mao, J. Liu, and N. Kato, "On intelligent traffic control for large-scale heterogeneous networks: a value matrix-based deep learning approach," IEEE Communications Letters, vol. 22, no. 12, pp. 2479–2482, 2018.

[38]. S. Maiti and R. K. Tiwari, "A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction," Environmental Earth Sciences, vol. 71, no. 7, pp. 3147– 3160, 2014.

[39]. C. Min, "An improved recurrent support vector regression algorithm for water quality prediction," Journal of Computational Information, vol. 12, pp. 4455–4462, 2011.

[40]. R. Das Kangabam, S. D. Bhoominathan, S. Kanagaraj, and M. Govindaraju, "Development of a water quality index (WQI) for the Loktak Lake in India," Applied Water Science, vol. 7, no. 6, pp. 2907–2918, 2017.

[41]. G. Srivastava and P. Kumar, "Water quality index with missing parameters," International Journal of Research in Engineering and Technology, vol. 2, no. 4, pp. 609–614, 2013.

[42]. S. Tyagi, B. Sharma, P. Singh, and R. Dobhal, "Water quality assessment in terms of water quality index," American Journal of Water Resources, vol. 1, no. 3, pp. 34–38, 2013.

[43]. A. A. Al-Othman, "Evaluation of the suitability of surface water from Riyadh Mainstream Saudi Arabia for a variety of uses," Arabian Journal of Chemistry, vol. 12, no. 8, pp. 2104– 2110, 2019.

[44]. T. H. H. Aldhyani, M. Alrasheedi, A. A. Alqarni, M. Y. Alzahrani, and A. M. Bamhdi, "Intelligent hybrid model to enhance time series models for predicting network traffic," IEEE Access, vol. 8, pp. 130431–130451, 2020.