Integrating Artificial Intelligence for Enhanced Water Quality Prediction and Classification: A Multi-Model Approach

[1]P. Kavitha, [2]A. Latha, [3]R. Mafaz Ahamed, [4]R. Ganesan, [5]D. Chitra, [6]G. Elumalai

[1]Dept. of Artificial Intelligence and Data Science, Panimalar Engineering College, Tamil Nadu.

[2,3,4]Dept. of Civil Engineering, Velammal College of Engineering and Technology, Madurai-09, Tamil Nadu.

[5]Dept. Management of Business Administration, Panimalar Engineering College, Tamil Nadu.

[6]Dept. of Electronics and Communication Engineering, Panimalar Engineering College, Tamil Nadu.

varshnikavitha@gmail.com[1], lathaganesan.a@gmail.com[2], mafaz.anwari@gmail.com[3], grganeshr4@gmail.com[4]

chitrambapec@gmail.com[5], elumalaig79@gmail.com[6]

Corresponding Author: P.Kavitha (varshnikavitha@gmail.com)

Abstract

In current years, the integrity of water quality has been jeopardized by a multitude of contaminants. Consequently, the modeling and forecast of water quality have assumed significant importance in the management and mitigation of water pollution. This study focuses on the development of sophisticated Artificial Intelligence (AI) algorithms for the purpose of predicting the Water Quality Index (WQI) and Water Quality Classification (WQC). Artificial Neural Network models, specifically the Nonlinear Autoregressive Neural Network (NARNN) and Long Short-Term Memory (LSTM) deep learning algorithm, have been devised for the purpose of predicting the WQI. In Water Quality Classification, we used Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM) & k-Nearest Neighbour algorithm (K-NN) has been used for the forecasting. The dataset included in this study consists of 10 parameters that are deemed to be significant. The models that were constructed were subsequently assessed using several statistical criteria. The findings indicate that the suggested models have a high level of accuracy in predicting the WQI and effectively classifying water quality with improved resilience. The findings of the study indicate that the NARNN model exhibited a slightly superior performance compared to the LSTM model in forecasting the values of the WQI. Additionally, the XGBoost algorithm attained the maximum level of accuracy, reaching 99.23%, in predicting the Water Quality Classification.

Keywords: Water Quality Index, Water Quality Classification, NARNN, XGBoost, SVM, K-NN, LSTM

1.INTRODUCTION

Water quality degradation has become a critical global concern due to increasing contamination from industrial, agricultural, and domestic activities [1]. Poor water quality directly impacts human health, ecosystems, and economic development. According to global reports, over 2 billion people consume water contaminated with fecal matter or harmful pollutants, resulting in widespread waterborne diseases and environmental damage [2]. The growing demand for clean and safe drinking water calls for advanced, scalable, and cost-effective solutions to monitor, predict, and treat water quality.

Traditional methods of water quality assessment and purification rely heavily on laboratory-based testing and basic filtration techniques [3,4]. While these approaches are effective to some extent, they are often labor-intensive, time-consuming, and lack the precision needed to address complex contamination patterns. Furthermore, conventional filtration methods are inadequate for removing emerging pollutants, while real-time monitoring capabilities remain largely underdeveloped. This gap necessitates the development of advanced tools and technologies that can deliver accurate predictions and efficient treatments for water quality improvement [5,6].

Artificial intelligence (AI) offers a transformative approach to tackling water quality challenges. By leveraging machine learning and deep learning techniques, AI models can analyze large datasets, predict water quality trends, and classify water conditions with high accuracy [7,8]. This study employs advanced AI techniques, including the Nonlinear Autoregressive Neural Network (NARNN) and Long Short-Term Memory (LSTM) models for Water Quality Index (WQI) prediction. In addition, algorithms like Extreme Gradient Boosting (XGBoost), Support Vector Machines (SVM), and k-Nearest Neighbor (KNN) are utilized for water quality classification [9,10]. In addition to predictive modeling, this study introduces a novel approach to natural filtration using Hemidesmus indicus, a cost-effective and environmentally friendly material [11].Hemidesmus indicus demonstrates promising potential for lake water purification by effectively reducing turbidity, dissolved solids, and heavy metals. This experimental approach aligns with the principles of sustainable water management by minimizing chemical use and leveraging renewable natural resources [12,13].

This research focuses on Red Hills Lake, a critical rain-fed reservoir in Chennai, India, that serves as a primary water source for urban and rural communities. The study aims to develop an integrated system combining AI-driven predictive models with experimental filtration techniques. By

addressing the dual challenges of accurate water quality monitoring and sustainable treatment, the research aspires to provide a robust framework for improving water safety and management [14]. The findings of this study have far-reaching implications for water resource management, particularly in regions grappling with limited infrastructure and high pollution levels. By demonstrating the efficacy of multi-model AI approaches and sustainable filtration techniques, this research contributes to the broader goals of environmental sustainability and public health [15]. The proposed framework offers a scalable solution for real-time water quality monitoring and treatment, setting a benchmark for future innovations in water management.

2.MATERIALS AND METHODS

2.1 Study Area

The Red Hills Lake, commonly known as Pulhalaeri or Pulhal lake and occasionally spelt Puzhal lake, is situated in Red Hills, Chennai, India. It is located in the Tamil Nadu state's Thiruvallur district. It is one of two rain-fed reservoirs the other two being Chembarambakkam Lake and Porur Lake from which water is collected for delivery to Chennai City. 3,300 million ft^3 is the lake's total storage volume. The reservoir is equipped with two masonry weirs. Both weirs have a length of 220 meters and a depth of 15 feet, with one of them measuring 178 meters in length. The bund measures 5 meters in length and spans a distance of 7 kilometres. Over a period of time, the weirs have developed a propensity for leakage. The Ambattur residential neighbourhood and the main road connecting to Avadi junction are situated in the southwestern region of the reservoir. The eastern region of the reservoir encompasses several notable features, including a water treatment facility, Pulhal Central Prison, and a road that provides connectivity to Koyambedu and North Chennai.

The Padiyanallur village is situated along the Thiruvallur Koot Road, which is located in the northern region and serves as a direct route to Thiruvallur. There are two bund road routes that facilitate connectivity between the north and east sides of the reservoir. The first route extends from Alamaram junction to Redhills market junction, while the second route spans from Redhills bypass road junction to Surapet junction. The villages of Pammathukulam, Pothur, Attanthangal, and Naravarikuppam are situated in the northwestern region.

2.2 Proposed Method

The purpose of this study is to determine the water quality using Artificial Intelligence (AI) methods, such as XGBoost, SVM, K-NN, LSTM, and NARNN, can accurately forecast various

aspects of Water Quality Index and Water Quality classification. Therefore, in the first part of this section, the examined area is provided, and then ranges of the various components of the water quality that were assessed are discussed. Following that, the research methodology of AI models is shown in Figure 1.

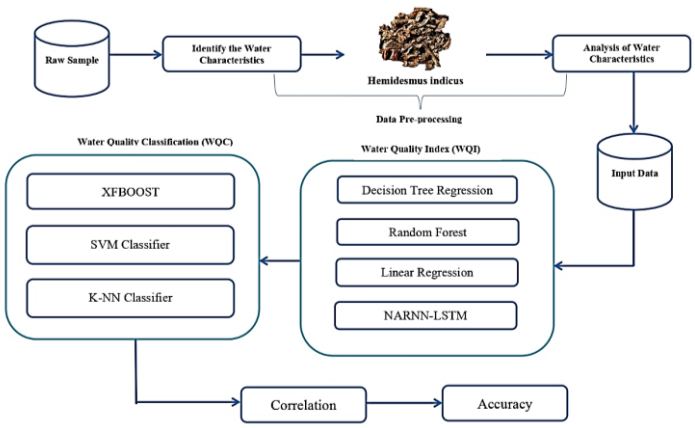

Figure 1. Research Methodology

2.3 Filtration of Hemidesmus indicus

As seen in Fig. 2, the experimental setup for filtration using species known as Indian sarsaparilla, Hemidesmus indicus, can be found in South Asia. It can be found across the majority of India, from the upper Gangetic plain in the east to Assam, as well as in a few locations in the centre, west, and south of the country. Sarsaparilla is made from the dried root of tropical Smilax species (Smilacaceae); in India, this is Smilax aspera L. and Smilax ovalifoliaRoxb. It is a thin, twining shrub that can occasionally be semi-erect or prostrate. Roots are fragrant and woody. The stem has several nodes, is thin, terete, and thickened. The opposite, short-petioled, highly varied leaves range in shape from elliptic-oblong to linear-lanceolate. The blooms grow in dense sub-sessile axillary cymes and are greenish on the outside and purple on the inside[16]. In order to compare the initial water sample findings with the treated lake water that was filtered utilising novel Hemidesmus indicus, the water sample from the Puzhal lake was taken and evaluated for the following parameters.

Figure 2: Experimental Setup for *Hemidesmus indicus* Filtration

The following procedures to filter the raw sample using Hemidesmu sindicus

A 1 litre polypropylene measuring cylinder is used to carry out the experiment for the present study. To carry out the experiment various layers with different proportions are set at first. A small hole of 10mm diameter is made in the bottom of the measuring cylinder for the treated lake water to get collected in the container kept below in the stand. In the measuring cylinder a filter paper is placed at the bottom and then a layer of fine aggregate followed by a layer of coarse aggregate of 3.5 cm depth is placed. On top of these layers, a layer of Hemidesmus indicusn root of 20 cm depth is placed. The measuring cylinder is kept on top of the stand and a container is placed on the bottom of the stand for the treated lake water to get collected. After which the initial sample of 1litre is poured on the measuring cylinder, the sample passes through various layers and the filtered water gets collected in the container. The initial and treated lake water are tested for the following parameters such as pH, turbidity, hardness, chloride, sulphate, calcium, Total Dissolved Solids (TDS), Total Suspended Solids (TSS), iron and copper.

2.4 Data Pre-processing

The preprocessing stage plays a pivotal role in ensuring the quality and reliability of data used for water quality prediction and classification. In this study, the dataset comprises measurements of ten key water quality parameters, including pH, turbidity, hardness, chloride, sulphate, calcium, Total Suspended Solids (TSS), Total Dissolved Solids (TDS), iron, and copper. The data preprocessing pipeline was designed to prepare the dataset for AI-based analysis and modelling[17,18].

2.4.1. Data Cleaning and Normalization

Initial data cleaning involved identifying and handling missing values, which were replaced using mean imputation for numerical parameters to maintain data consistency. Outliers detected through Z-score analysis were reviewed and either corrected or excluded, depending on their context within the dataset. To ensure uniformity across all features, Z-score normalization was applied, converting the data into a standard scale with a mean of 0 and a standard deviation of 1. This step was crucial for optimizing the performance of machine learning algorithms [19].

2.4.2. Data Splitting

The dataset was divided into training, validation, and testing subsets. A 70:15:15 ratio was employed to ensure adequate data for model training, hyperparameter tuning, and evaluation [20]. The training set was used to train the models, while the validation set aided in optimizing model parameters, and the testing set evaluated the models' final performance on unseen data. A stratified sampling approach was used to maintain the proportional representation of water quality classifications in all subsets.

2.4.3. Data Volume and Temporal Coverage

The dataset consists of water quality measurements collected over a specific period from Red Hills Lake. A total of [provide specific number] data points were included in the analysis, representing variations in seasonal and daily water quality [21]. Temporal dependencies in the dataset were retained, particularly for time-series models like NARNN and LSTM, to capture long-term trends and patterns in the water quality parameters.

2.4.4. Feature Engineering and Transformation

Correlation analysis was performed to examine relationships between the parameters, aiding in feature selection and dimensionality reduction where necessary. Parameters with strong correlations were retained for prediction and classification to enhance model interpretability and reduce computational complexity. Additionally, categorical labels were encoded using one-hot encoding for the classification tasks [22,23].

2.4.5. Cross-Validation

K-fold cross-validation with 5 folds was employed during model training to mitigate overfitting and ensure robust evaluation. Each fold allowed the model to train on different subsets of data while validating on unseen portions, providing a comprehensive understanding of its generalization capabilities.

2.5 Water Quality Index Calculation

The estimation of WQI involves the consideration of several elements that significantly influence the quality of water. In this work, a set of ten significant water quality metrics is employed to evaluate the efficacy of the proposed model in comparison to a dataset that has been previously published. In water quality index, to assign the constant value proportionality calculated as follows:

$$\kappa = {}^{1}/_{\sum_{i=1}^{n} \mathcal{S}_i} \quad \text{---------- (1)}$$

Where,

κ -> Constant value

$\mathcal{S}$ -> Standard value of parameter i

In Eq. (1) were using to be calculated k for each parameter i. The Unit weight value for each parameter was calculated by Eq.2

$$\text{Unit Weight } (\mathcal{W}_i) = \frac{\kappa}{\mathcal{S}_i} \quad \text{---------- (2)}$$

In Eq.3 were using to be calculated the quality rate for each parameter i. To find the quality rate of each parameter as follows:

$$\text{Quality Rate } (\mathcal{Q}_i) = \frac{\mathcal{V}_i - \mathcal{V}_{ideal}}{\mathcal{S}_i - \mathcal{V}_{ideal}} \quad \text{---------- (3)}$$

The following formula was used to determine the WQI by using Eq.4

$$WQI = \frac{\sum_i^N \mathcal{Q}_i * \mathcal{W}_i}{\sum_{i=1}^N \mathcal{W}_i} \quad \text{---------- (4)}$$

In the given context, N denotes the total count of parameters included in the calculations of the water quality index. Q_i represents the quality rating of each individual parameter i, which is determined using equation (3), while the unit weight for each parameter i is computed using equation (2).

Table 1. The acceptable thresholds for the parameters included in the calculation of WQI

S.NO	PARAMETER	LIMITS
1	pH (range)	6.5-8.5
2	Turbidity (NTU)	1.0-5.0
3	Hardness(mg/l)	200-600
4	Chloride (mg/l)	250-1000

5	Sulphate(mg/l)	200-400
6	Calcium(mg/l)	75-200
7	TDS (mg/l)	500-2000
8	TSS (mg/l)	500-2000
9	Iron(mg/l)	1.0-2.0
10	Copper(mg/l)	0.05-1.5

Table 1 shows the list of the parameter to analyses the water quality prediction in lake water and the permissible limits of each parameter. Table 2 shows the Water Quality Index Classification.

Table 2: Water Quality Index Classification

WOI Range	**Classification**
0-24	Excellent
25-50	Good
51-75	Poor
76-100	Very poor
Above 100	Unfitting for Drinking Water

2.6 Z- Score Calculation

The process of normalisation might facilitate the simplification of calculations. The given statement undergoes a conversion process from a dimensional form to a scalar form. Z-score normalization, sometimes referred to as normalization score, is a commonly employed technique for data normalization [24,25]. It involves utilising the mean and standard deviation values of the data being tested in order to normalise the parameters. The computation can be expressed in Eq. (5)

$$\text{Z-Score Calculation} = \frac{(X-\mu)}{\sigma} \text{ ---------- (5)}$$

In this context, x represents the numerical value assigned to the parameter i in the samples that were subjected to testing.

2.7 Prediction of WQI

The construction of the water quality prediction model involves analyzing patterns within the dataset of lake water quality using a range of techniques, including decision tree regression, random forest, linear regression and support vector regression. To accomplish this goal, the ANN model of

NARNN-LSTM (nonlinear autoregressive neural network- long short-term memory) for the purpose of predicting the water quality score [26].

2.7.1. LSTM Model

Recent studies using deep learning models, particularly LSTM networks, have shown significant promise in predicting water quality. For instance, the combination of LSTM and other techniques like the attention mechanism or convolutional neural networks (CNNs) has been applied successfully to predict key water quality parameters like dissolved oxygen, nitrogen, and phosphorus concentrations. These models can capture temporal dependencies and nonlinear patterns in water quality data, providing more accurate predictions than traditional models. In one study, LSTM models were used in the Nakdong River Basin to predict water quality with high accuracy [27,28]. The combination of CNN for water level and LSTM for water quality achieved a Nash–Sutcliffe efficiency value above 0.75, which indicates very good performance in predicting pollutant variations. Another case study on the Burnett River in Australia used LSTM with an attention mechanism to enhance prediction, showing that this approach improved prediction accuracy compared to a standard LSTM model .

2.7.2 Proposed NARNN-LSTM Techniques

The strong performance of the NARNN-LSTM model in water quality prediction stems from its ability to effectively handle the temporal and nonlinear characteristics of water quality data. Parameters such as pH, turbidity, and dissolved solids often exhibit time-dependent patterns influenced by seasonal and environmental changes. The NARNN component utilizes its autoregressive nature to analyze sequential relationships, feeding past outputs into future predictions, while the LSTM component excels at capturing long-term dependencies through its memory cells. Together, these features allow the model to identify trends, periodic variations, and interactions in the data, which are essential for accurate water quality forecasting.

Another significant advantage of NARNN-LSTM lies in its robustness to noise and missing data, common issues in water quality datasets. Noise, arising from sampling errors or inconsistencies, is smoothed during the learning process, enabling the model to focus on meaningful patterns. The architecture's regularization techniques, such as dropout, further improve its ability to generalize, reducing overfitting to noisy data. Additionally, LSTM's capability to handle variable time intervals without requiring extensive interpolation makes it particularly suitable for datasets with irregular or incomplete temporal records. These qualities ensure consistent and reliable predictions across diverse data conditions.

The NARNN-LSTM model is widely recognized as a prominent example of a multilayer feed-forward network. The process commences by initializing the weight value with an estimated value, which is subsequently refined by the incorporation of observed data. Consequently, the prediction process of the neural network model incorporates a certain degree of stochasticity. The network undergoes multiple training iterations with varying random initialization values, and the resulting outcomes are subsequently averaged. The identification of the number of hidden layers and nodes is a prerequisite in the NARNN-LSTM paradigm. The NARNN-LSTM time series model is described by Equation (7).

$$Y(time) = h\big(Y(time-1), Y(time-2), \cdots, Y(time-P)\big) + \varepsilon(time) \text{ ---------- (7)}$$

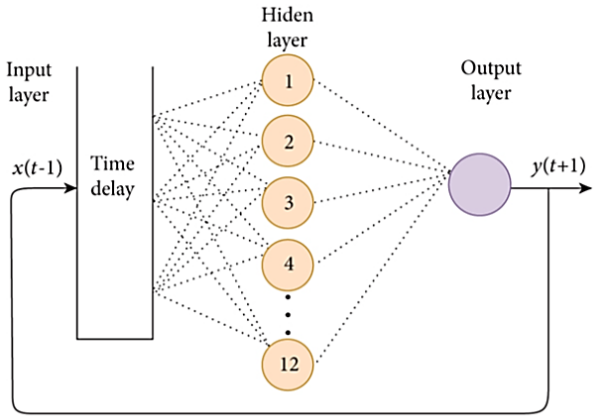

Figure 4: Calculation of the NARNN-LSTM model.

When using the p observation values of the series, y(t) is the value of the time-series data at time t. The network weights and neuron bias are optimized using the function (h). The error derived from the model at time t is known as e(time).

The model's ability to model complex feature interactions and seasonal dynamics further enhances its effectiveness. Water quality parameters often exhibit nonlinear relationships, such as correlations between turbidity, total dissolved solids, and pH, which are difficult for simpler models to capture. NARNN-LSTM's layered structure processes these interactions while emphasizing sequential dependencies. Moreover, the model is adept at detecting seasonal patterns, such as increased turbidity during monsoons or temperature-driven changes in dissolved oxygen levels. Empirical results, including high accuracy metrics and low residual errors, demonstrate its superior predictive capabilities. These strengths make NARNN-LSTM an optimal choice for addressing the inherent complexities of water quality data.

The present work involved the development of the NARNN model for the purpose of predicting the WQI. The NARNN model, in contrast to other Artificial Neural Network (ANN) models such as the forward neural network, is specifically designed for time series analysis and forecasting of stationary time series. The utilization of the NARNN model is proposed as a means to predict the WQI, as the parameters of the WQI exhibit characteristics of a time series. Table 3 presents the pertinent parameters for model building. Figure 4 illustrates the architecture of the NARNN model that has been built.

Table 3: Parameters of the developed NARNN Technique.

No. of hidden layers	11
No. of delays	1:9
Max. number of iterations	100
Max. number of epochs	11
No. of gradients	$1.724 * 10^3$

2.8 Monitoring a Water Quality Classification using XGBoost Techniques

XGBoost has incorporated regularization techniques, including regularized boosting, which has proven to be quite effective in mitigating the issue of overfitting. Compared to K-NN, XGBoost offers the advantage of parallel processing, resulting in significantly improved speed. In contrast to SVM and K-NN classification, Users of XGBoost can create unique optimization goals and assessment standards [29,30]. This flexibility enables the inclusion of additional dimensions in the model, thereby avoiding any constraints on data processing. During data collection, the presence of various artificial or experimental defects often leads to data loss. Hence, the use of XGBoost is attempted as a substitute for the prior classifier in order to construct a water quality classification model. Preprocessing of the acquired feature data is a necessary step prior to the formation of the model, as it serves to enhance both an accuracy and the training speed of the model. To begin with, a process known as Smoothing was conducted in order to generate a set of feature samples. XGBoost is a very effective decision tree classifier that leverages the scoring and objective function to evaluate the model's performance.

XGBoost is a classification model for water quality shows in Equation (8) & (9)

$$Object = -\frac{1}{2}\sum_{j=1}^{t}\frac{gain_j^2}{h_j+\lambda} + t^{\gamma} \quad \text{---------- (8)}$$

$$Gain(G) = \frac{1}{2}\left[\frac{g_l^2}{h_l+\lambda} + \frac{g_r^2}{h_r+\lambda} + \frac{(g_l+g_r)}{h_l+\lambda}\right] - \gamma \quad \text{---------- (9)}$$

The following stages were used in this paper to create a model for monitoring water quality:

- The feature parameter set served as XGBoost's training input.
- The most effective number of boosting rounds was obtained using the cross validation method.
- The model was analysed and assessed using the tree structure and the characteristic score.

This paper outlines the sequential process employed to create a water quality monitoring model. The input for training XGBoost consisted of the feature parameter set. The cross-validation technique was engaged to regulate the most suitable number of enhancing iterations. The utilization of the tree structure and the typical score facilitated the analysis and evaluation of the model [31].

2.9 Assessment of an Individual's Performance in a Model.

The evaluation of the effectiveness of the constructed models in predicting and classifying the water quality index is conducted in order to choose the optimal algorithm. The prediction algorithms that exhibit the highest efficiency are typically characterized by a very low Root Mean Square Error (RMSE) value. Similarly, the evaluation of the best classification model is commonly conducted by assessing its accuracy [32]. The statistical parameters employed in the analysis were as follows:

- Mean Square Error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad \text{---------- (11)}$$

- Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\sum_{i=1}^{N}\frac{(y-\hat{y})^2}{N}} \quad \text{---------- (12)}$$

Let n represent the total number of input variables. The variable x refers to the observation input data from the initial batch of training data, while y represents the observation input data from the second set of training data. R denotes the Pearson's correlation coefficient.

- Accuracy

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \times 100\% \quad (13)$$

- Specificity

$$Specificity = \frac{TN}{TN+FB} \times 100\% \quad (14)$$

- Sensitivity

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \quad (15)$$

- Precision

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (16)$$

- F-score

$$F - score = \frac{2*preision*Sensitivity}{preision+Sensitivity} \times 100\% \quad (17)$$

The True Positive, True Negative, False Positive, and False Negative are represented by the variables TP, TN, FP, and FN, respectively. The aforementioned equations and data from river water are used to evaluate how well machine learning algorithms perform in providing water quality indicators for prediction and categorization.

2.10 Correlation Analysis

In order to investigate the relationship that exists between two or more variables, a statistical technique known as correlation analysis is utilized. The Pearson correlation coefficient approach is used, which involves assessing the strength and direction of the correlation, in order to evaluate the degree of correlation that exists between the significant parameters of the dataset that are utilized in the process of forecasting the WQI values.

$$R = \frac{n\sum(x1\times x2)-(\sum x1)(\sum x2)}{[n\sum(x1^2)-\sum(x1^2)]\times[n\sum(x1^2)-\sum(x1^2)]} \times 100\% \quad \text{---------- (18)}$$

Where:

R: The approach utilized in this study is Pearson's correlation coefficient.

x: Enter the values for the first batch of training data.

y: Please provide the input values for the second batch of training data.

n: The overall quantity of input variables.

3 Results and discussions

In this study, Physiochemical parameters such as pH, turbidity, hardness, chloride, sulphate, calcium, Total Suspended Solids (TSS), Total Dissolved Solids (TDS), iron, and copper were analyzed. The descriptive WQI and WQC are presented.

3.1 Experimental Analysis of Proposed Techniques

The primary objective of water purification is to ensure the provision of potable water that is free from contaminants. The adherence to appropriate filtering procedures is crucial due to the potential accumulation of runoff, animal excretions, and pollutants from boats and machinery in lake water. The potability of lake water is often limited but, by the implementation of appropriate filtration techniques, it is possible to obtain water that is both safe for consumption and possesses desirable taste qualities. The utilization of lake water holds significant advantages for several businesses, as it may be employed in a multitude of procedures.

Table 4: Characteristics of Raw Sample

PARAMETER	RAW SAMPLE														
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
pH (range)	7.61	7.7	6.9	7.8	7.4	7.5	6.9	6.8	7.4	7.7	7.6	7.8	6.8	7.8	7.6
Turbidity(NTU)	2.49	2.65	3.56	2.8	2.95	2.6	2.78	2.85	2.9	2.57	2.54	2.56	2.57	2.58	3.5
Hardness(mg/l)	242	252	254	251	252	240	248	246	243	250	251	256	258	253	257
Chloride (mg/l)	258	260	255	258	254	260	255	256	248	295	256	250	253	252	256
Sulphate(mg/l)	230	235	240	236	239	234	215	225	265	234	256	254	251	257	258
Calcium(mg/l)	85	79	82	84	83	81	80	81	82	86	85	82	82	86	85
TDS(mg/l)	650	660	550	574	589	620	623	587	635	659	625	654	594	653	620
TSS(mg/l)	750	756	758	779	820	824	780	789	753	745	785	754	756	785	795
Iron(mg/l)	1.378	1.2	1.4	1.6	1.25	1.35	1.45	1.56	1.58	1.38	1.96	1.5	1.45	1.58	1.25
Copper(mg/l)	1.3	1.1	1.3	1.2	1.6	1.5	1.4	1.6	1.45	1.25	1.5	1.3	1.36	1.24	1.25

The data pertaining to the collecting of water samples from the lake, as well as the subsequent analysis of these samples using ten parameters, is presented in Table 4. In this present study we have checked the feasibility of Hemidesmus indicus root as a filter media for lake water treatment. Lake water quality analysis have been conducted in the laboratory for parameters such as pH, turbidity, total dissolved solids, total suspended solids, hardness, iron, chloride, sulphate, copper, calcium. Table 5 shows that filtration of lake water samples by using Hemidesmus indicus.

Table 5: Filtration of Hemidesmus indicus using raw samples

PARAMETER	RAW SAMPLE														
	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**
pH (range)	6.56	6.5	6.8	7.1	7.2	6.9	6.2	7.5	6.6	6.3	7.2	6.8	7.2	7.6	7.5
Turbidity(NTU)	1.81	1.8	2.2	2.3	2.5	1.7	1.9	2.6	3.2	3.3	3.5	1.5	1.7	1.9	3.5
Hardness(mg/l)	256	258	259	301	330	335	425	259	289	369	349	378	349	378	398
Chloride (mg/l)	212	220	225	269	359	389	385	348	379	487	483	389	347	487	496
Sulphate(mg/l)	226	229	226	228	310	315	329	305	298	297	324	326	314	300	359
Calcium(mg/l)	80	82	85	96	95	84	89	82	89	86	83	94	98	96	98
TDS(mg/l)	545	520	522	548	569	578	612	623	648	589	625	678	693	678	689
TSS(mg/l)	536	540	589	548	563	598	587	596	532	685	698	648	659	675	698
Iron(mg/l)	1.02	1.05	1.06	1.1	1.16	1.17	1.18	1.16	1.1	1.5	1.6	1.4	1.8	1.9	1.8
Copper(mg/l)	1.11	1.2	1.3	1.12	1.4	1.2	1.3	1.16	1.18	1.16	1.35	1.2	1.3	1.2	1.25

3.2 Water Quality Index Prediction Model

The proposed techniques NARNN-LSTM model, consisting of 10 hidden layers, demonstrated a favourable concert in predicting the values of the WQI. As previously mentioned, it exhibits the following attributes: The number of delays is 1:8, and the number of epochs is 12. The LSTM model that was built consists of a total of 200 hidden layers, with a maximum number of epochs set at 150.

Table 6 presents a summary of the performance parameters of the generated models for predicting WQI. It is observed that the LSTM model exhibited somewhat higher prediction accuracy for the testing data compared to the training data. Furthermore, it has been shown that the LSTM model generally exhibits a somewhat superior performance in comparison to the NARNN model, as indicated by the Mean Squared Error (MSE) values. However, in terms of value, the NARNN model has demonstrated better prediction of R%>91.83.

Table 6: Performances of the NARNN-LSTM models to predict WQI.

Models	**Training Data**		**Testing Data**	
	MSE	**R (%)**	**MSE**	**R (%)**
NARNN	0.2714	94.87	0.1242	95.18
LSTM	0.1217	91.83	0.1049	94.31

The NARNN model's histogram error is depicted in Figure 5, which may be found here. Finding faults in the goal values and the anticipated values of training and testing datasets can be accomplished with the help of the histogram metric. The overall error range has been segmented into twenty smaller bins, and the y-axis of each bin displays the number of samples that are contained inside that bin. Figure 6 presents the histogram metric as well as the mean errors that were generated by the LSTM model during the training and testing periods [33,34]. The mean error and the histogram metric are utilized in order to determine the degree to which the observed values deviate from the values that were anticipated by the training and the testing.

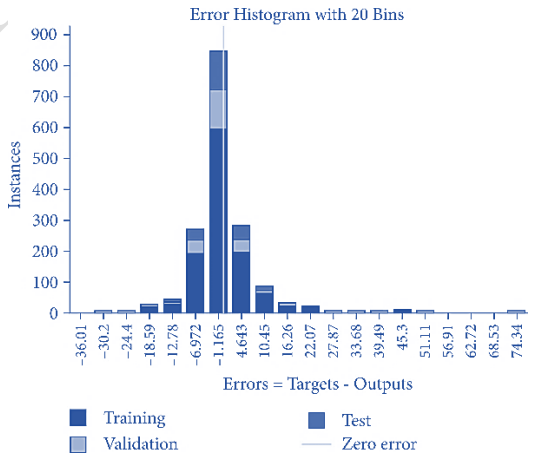

Figure 5: Histogram Model of NARNN technique

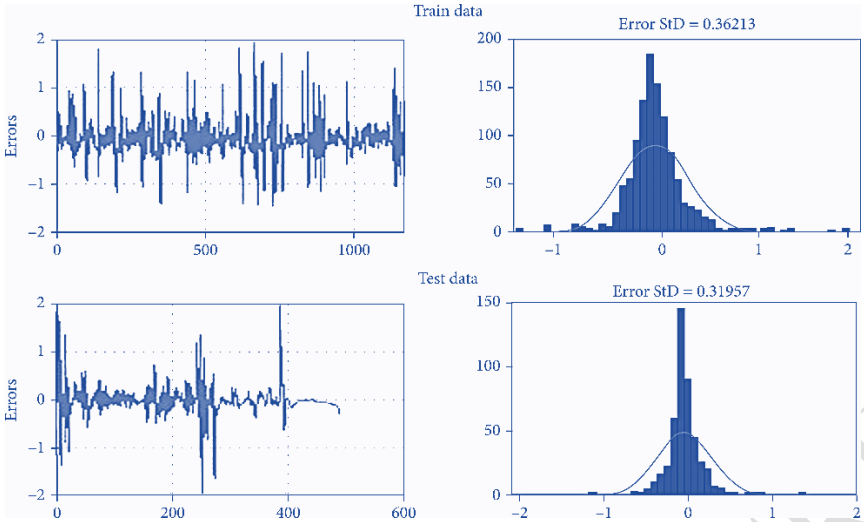

Figure 6: LSTM training and testing histogram and mean errors.

Figures 7 and 8 shows NARNN and LSTM regression charts for training, testing, and entire datasets. This graphic shows the predicted-actual relationship. The plot's "target" values are the dataset, while the "output" values are NARNN and LSTM model predictions. As seen in both panels, the projected WQI values match those estimated from measured parameters (NARNN and LSMT). Both models are highly efficient.

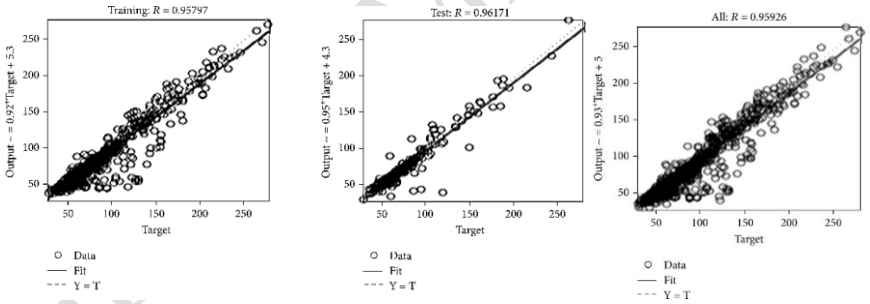

Figure 7: NARNN Regression

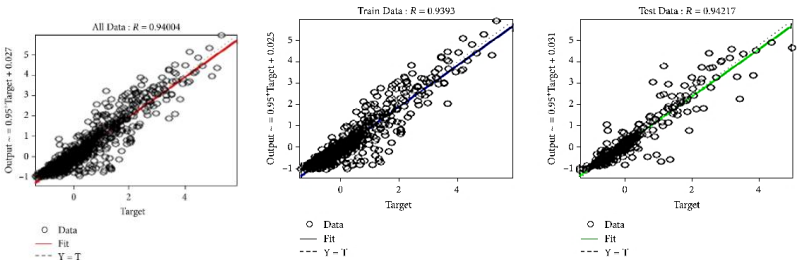

Figure 8: LSTM Regression

3.3 Prediction of Water Quality Classification

In order to obtain the result using XGBoost, the classification model that was used for the initial data 4 was employed. Table 7, which was provided, is a table that provides the attributes of the parameters that are related with the code for the model. Figure 9 presents a display of the scores obtained for each distinguishing parameter included in the model.

Table 7: XGBoost's performance in terms of classification.

Samples	**Accuracy (%)**	**Time/s**
S1	97.25	0.052
S2	99.1	0.049
S3	98.51	0.0559
S4	97.62	0.055
S5	98.26	0.053
S6	99.02	0.0545
S7	98.36	0.052
S8	97.45	0.062
S9	99.3	0.059
S10	98.71	0.0659
S11	97.82	0.065
S12	98.46	0.063
S13	99.22	0.0645
S14	98.56	0.062
S15	97.35	0.054

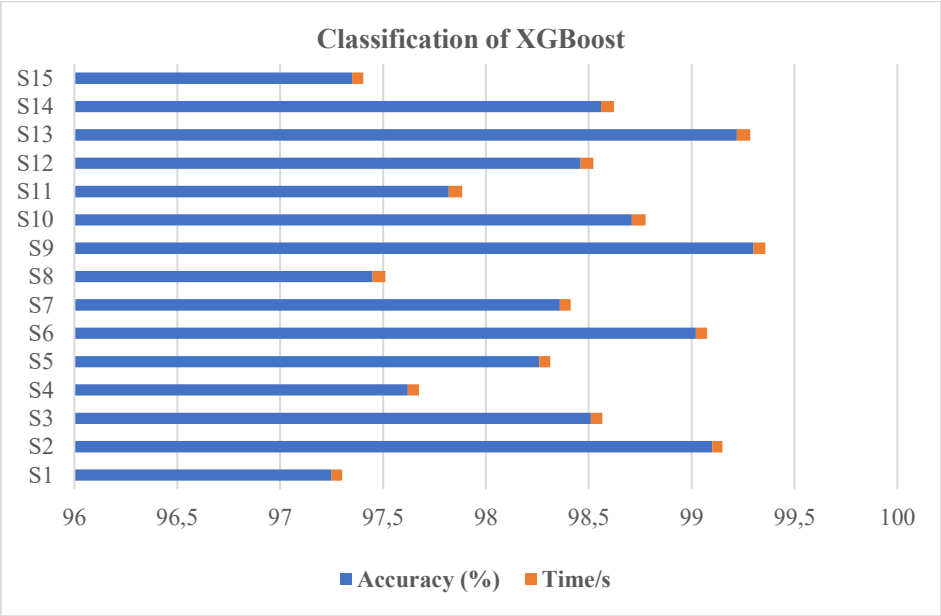

Figure 9: XGBoost's performance in terms of classification.

3.4 Comparison Analysis

We compared XGBoost to SVM, a standard classifier that has historically performed well in classification problems, to demonstrate its performance. Table 8 shows the efficacy of the employed machine learning models in forecasting WQC. XGBoost uses original data, therefore SVM must be further normalized to work best. WQC prediction Machine learning algorithm performance as shown in Figure 10.

Table 8: The efficacy of the employed machine learning models in forecasting WQC

Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-score (%)
XGBoost	97.01	99.23	97.78	94.93	98.54
SVM [11,12]	83.63	84.73	94.93	87.50	85.84
KNN[21]	75.20	77.76	91.65	78.08	81.51

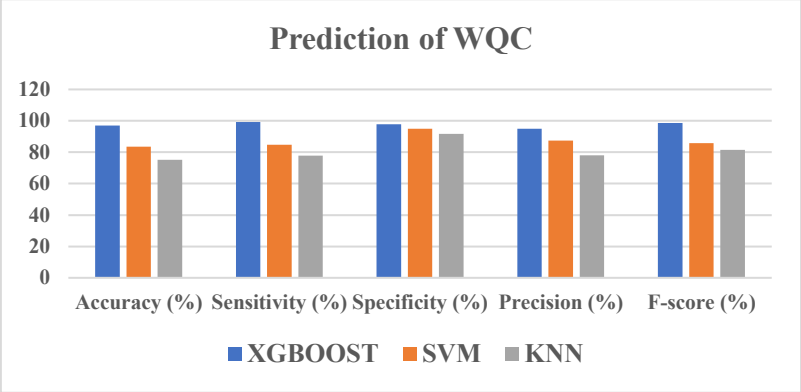

Figure 10: WQC prediction machine learning algorithm performance

4 Conclusions

The modelling and prediction of water quality play a crucial role in safeguarding the environment. The utilization of sophisticated artificial intelligence algorithms can be employed in the development of a model to assess the prospective water quality. The present study employed advanced artificial intelligence algorithms, specifically the NARNN-LSTM models, to forecast the WQI. A set of motion characteristic parameters were computed to serve as indicators for water quality assessment. During the parameter analysis, it was shown that certain features may effectively differentiate between normal and abnormal water quality conditions. The establishment of a water quality monitoring model was built on XGBoost classifiers, as proposed by this concept. Following an extensive series of studies, this model has demonstrated its ability to efficiently, precisely, and readily classify water quality. In comparison to the preceding classifier, XGBoost shown greater prominence. Nevertheless, the comprehensive procedure of monitoring water quality lacks the ability to attain complete closure. Consequently, we must depend on continuous training and the handling of human errors. As a result, the entire system falls short of achieving real-time monitoring capabilities. Furthermore, it is possible to degrade the experimental conditions and improve the resilience of the system.

Conflicts of Interest

The authors assert that there are no conflicts of interest.

Authors' Contributions

All writers made substantial contributions to the finalization of this publication.

References

1. Hmoud Al-Adhaileh, M., &Waselallah Alsaade, F. (2021) Modelling and prediction of water quality by using artificial intelligence. *Sustainability,* 13(8), 4259.
2. Torky, M., Bakhiet, A., Bakrey, M., Ismail, A. A., &Seddawy, A. I. E. (2023). Recognizing Safe Drinking Water and Predicting Water Quality Index using Machine Learning Framework. *International Journal of Advanced Computer Science and Applications,* 14(1),
3. Hong, W. J., Shamsuddin, N., Abas, E., Apong, R. A., Masri, Z., Suhaimi, H., & Noh, M. N. A (2021). Water quality monitoring with arduino based sensors. *Environments,* 8(1), 6.
4. A. Latha, R. Ganesan, B. Krishnakumari and S. Theerkadharsini(2022) Comparative Study of Organic Coagulants in Water Treatment, *ECS Transactions*. 107 ,7997-8007.
5. E. Batur and D. Maktav, (2019) 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*, 57(5) 2983–2989.
6. S. Jaloree, A. Rajput, and G. Sanjeev, (2014) Decision tree approach to build a model for water quality, *Binary Journal of Data Mining & Networking*, 4, 25–28.
7. J. Liu, C. Yu, Z. Hu (2020) Accurate prediction scheme of water quality in smart mariculture with deep Bi-S-SRU learning network," *IEEE Access*, 8, 24784–24798.
8. H. Liao and W. Sun, (2010) Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method, *Procedia Environmental Sciences*, 2, 970–979.
9. L. Yan and M. Qian, (2012) AP-LSSVM modeling for water quality prediction," in *Proceedings of the 31st Chinese Control Conference*, 6928–6932, Hefei, China, July.
10. A. Solanki, H. Agrawal, and K. Khare, (2015) Predictive analysis of water quality parameters using deep learning. *International Journal of Computers and Applications*, 125(9)29–34.
11. X. Li and J. Song, (2015) A new ANN-Markov chain methodology for water quality prediction," in *2015 International Joint Conference on Neural Networks (IJCNN)* 1–6, Killarney, Ireland, July.
12. A. A. M. Ahmed and S. M. A. Shah (2017) Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River," *Journal of King Saud University - Engineering Sciences*, 29(3) 237–243.
13. Y. Khan and C. S. See (2016) 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.
14. J. Yan, Z. Xu, Y. Yu, H. Xu, and K. Gao (2019) Application of a hybrid optimized BP network model to estimate water quality parameters of Beihai Lake in Beijing, *Applied Sciences*, 9(9) 1863.
15. H. R. Maier, A. Jain, G. C. Dandy, and K. P. Sudheer (2010). 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*, 25(8) 891–909.

16. S. Lee and D. Lee (2018) 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*, 15(7) 1322.
17. U. Shafi, R. Mumtaz, H. Anwar, A. M. Qamar, and H. Khurshid (2018) 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)*, 92–96, Islamabad, Pakistan.
18. Z. Ahmad, N. A. Rahim, A. Bahadori, and J. Zhang (2016) Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks," *International Journal of River Basin Management*, 15(1) 79–87.
19. V. Ranković, J. Radulović, I. Radojević, A. Ostojić, and L. Čomić, (2010). Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia," *Ecological Modelling*,221(8) 1239–1244.
20. N. M. Gazzaz, M. K. Yusoff, A. Z. Aris, H. Juahir, and M. F. Ramli (2012) Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors," *Marine Pollution Bulletin*, 64(11) 2409–2420.
21. H. Z. Abyaneh, (2014) Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters, *Journal of Environmental Health Science and Engineering*, 12(1) 40.
22. M. Sakizadeh, (2016) Artificial intelligence for the prediction of water quality index in groundwater systems, *Modeling Earth Systems and Environment*, 2(1) 8.
23. M. Bouamar and M. Ladjal, (2008) 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*,. 1–5, Amman, Jordan.
24. S. Maiti and R. K. Tiwari, (2014) A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction, *Environmental Earth Sciences*, 71(7) 3147–3160.
25. C. Min, (2011) An improved recurrent support vector regression algorithm for water quality prediction, *Journal of Computational Information*, 12, 4455–4462.
26. R. Das Kangabam, S. D. Bhoominathan, S. Kanagaraj, and M. Govindaraju (2017) Development of a water quality index (WQI) for the Loktak Lake in India, *Applied Water Science*, 7(6) 2907–2918.
27. G. Srivastava and P. Kumar (2013.) Water quality index with missing parameters, *International Journal of Research in Engineering and Technology*, 2(4) 609–614.
28. S. Tyagi, B. Sharma, P. Singh, and R. Dobhal (2013).Water quality assessment in terms of water quality index, *American Journal of Water Resources*, 1(3)34–38.

29. T. H. H. Aldhyani, M. Alrasheedi, A. A. Alqarni, M. Y. Alzahrani, and A. M. Bamhdi, (2020) Intelligent hybrid model to enhance time series models for predicting network traffic, *IEEE Access*, 8, 130431–130451.
30. Kistan, A.,V. Kanchana (2020) Confiscation of chemical oxygen demand from groundwater samples collected from near tanneries using activated carbon of Ricinus Communis blended with coconut shell. *Indian Journal of Environmental Protection*, 40(5) 527– 532.
31. Govindaraj, V., Manokaran, K., Sathaiya, J..Baskar, P (2023) Environmentally-Friendly Bio-Coagulants: A Cost-Effective Solution for Groundwater Pollution Treatment, Asian Journal of Water, Environment and Pollution 20(3)19-28.
32. Govindaraj, V., Asaithambi, L., Ramachandran, G..Barmavatu, P.(2024) Removal of turbidity from lake water using novel Chrysopogon zizanioides and Hemidesmus indicus, Desalination and Water Treatment 317,100245.
33. Baek k (2020). Prediction of Water Level and Water Quality Using a CNN-LSTM Combined Deep Learning Approach, *Water*, 12(12)
34. Jiang G. (2022) Water Quality Prediction Based on LSTM and Attention Mechanism, *Sustainability* 14(20) 2022.