



PREDICTION OF WATER QUALITY INDEX (WQI) USING A NOVEL HYBRID WATER WAVE OPTIMIZED EXTREME GRADIENT BOOSTING

R.Sarala*¹, R.S. Ponmagal²

¹*Research Scholar, SRM Institute of science and technology, Chennai, India

²Association Professor, SRM Institute of science and technology, Chennai, India.

¹<http://orcid.org/0009-0007-9776-2413>, ²<http://orcid.org/0000-0002-7228-1310>

E-mail: *sr8749@srmist.edu.in, ponmagas@srmist.edu.in,

ARTICLE INFO

Article History

Received on: january 2, 2026

Reviewed on: february 08, 2026

Accepted on: July 01, 2026

Published on: July 31, 2026

Keywords:

Water Quality Index (WQI),
Outlier Detection (Box Plot
Analysis),

Min-max normalization,

Independent Linear Analysis
(ICA),

Hybrid Water Wave Optimized
extreme Gradient Boosting
(HWW-XGB).

ABSTRACT

Determining the drastic changes in water quality in this contaminated period is an urgent requirement that is crucial for the continued existence of the current and growing water demand. This is why it is essential for maintaining the water supply. Due to variability and complexity in the data, the water quality index (WQI) has been developed as a reliable way to express water quality. Analyzing the data relative to the water quality has been extremely difficult. In this investigation, the WQI prediction of water samples using the novel Hybrid Water Wave Optimized eXtreme Gradient Boosting (HWW-XGB). In this research, the data preprocessing using the Outlier Detection (Box Plot Analysis) and Min-max normalization was finished to minimize the computational time. Pre-processing is a crucial first step in the proposed system because the dataset contains inappropriate and inaccurate data. The Independent Linear Analysis (ICA) is utilized to extract the required features. The pre-processed features are removed through feature extraction, eliminating unnecessary data for the next essential phase. The collected features are fed into Hybrid Water Wave Optimized eXtreme Gradient Boosting (HWW-XGB) classifier. The effectiveness of the proposed approach was demonstrated when compared to the existing method.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The WQI is a useful measure for evaluating and tracking the available water resource quality. It gives a numerical rating that illustrates the combined impact of many water quality factors, enabling straightforward comparison and analysis of water quality data. The index considers some physical, chemical, and biological parameters to assess a water body's general health [1]. Temperature, pH level, dissolved oxygen, turbidity, total dissolved solids, biochemical oxygen demand, nitrate, and phosphate concentrations, and the presence of pathogens are some of the factors taken into account by the WQI. These variables indicate several facets of water quality, including its capacity for human use, purity, and ability to sustain aquatic life. It may assess the water's fitness for different benefits, such as drinking, leisure, or ecological health, by monitoring and examining these factors [2]. Individual ratings are produced for each parameter using the relevant scales or algorithms, and each parameter is given a weight depending on its relative relevance. An overall Water Quality Index score is then calculated using these specific ratings. Higher values on the index, which typically ranges from 0 to 100, represent usually higher water quality. Depending on the local and environmental factors, different areas may have unique WQI approaches [3] [4]. These courses provide a clearer knowledge of water quality, enabling decision-makers, resource managers, and the general public to swiftly evaluate a water body's status. Each metric has predetermined thresholds and recommendations often drawn from regulatory requirements [5]. The WQI may convey detailed water quality data clearly and intelligibly, making it a useful communication tool. It aids in locating problem regions, emphasizing possible pollution sources, and directing decision-making procedures for managing and safeguarding water resources. By comparing previous data, the index also makes it possible to analyze long-term trends, which makes it easier to spot changes in water quality over time [6].

Authorities may identify pollution sources, evaluate the efficacy of pollution management methods, and take proactive measures to protect water resources by establishing baseline water quality conditions and monitoring changes [7]. It is crucial to remember that the WQI is a broad indicator and cannot account for every facet of water quality. It gives a general image but does not take the place of careful observation and examination of particular aspects. The weights allocated to each parameter may also change based on the local circumstances and the planned use of the water resource, which might affect the index values and interpretations. Consequently, the WQI should be used in combination to other monitoring and assessment methods to offer a full picture of water quality [8] [9]. The Water Quality Index also helps to educate and raise public awareness about water quality. It allows people to understand the significance of water quality measures and their impacts on the environment and human health. When they have access to clear and understandable information, communities may actively participate in conservation initiatives, advocacy, and protection of their local water resources [10]. It could miss certain details or some pollutants. To acquire a more thorough knowledge of water quality and ecological integrity, it is often used in combination with other monitoring methods and evaluations, such as chemical analysis, biological, and environmental assessments [11]. It offers a standardized approach to assessing water quality across diverse locations by considering several characteristics and producing a composite score. The index promotes trend analysis, improves public awareness, helps with decision-making, and helps preserve and conserve water bodies for the benefit of both people and ecosystems [12-25]. In this study, we proposed the innovative Hybrid Water Wave Optimised extreme Gradient Boosting (HWW-XGB) to estimate the WQI of water samples.

1.1 CONTRIBUTIONS TO THIS STUDY ARE AS FOLLOWS

- This inquiry uses the novel HWW-XGB to predict the WQI of water samples.
- The data in this study were preprocessed using the Min-Max normalization and Outlier Detection techniques.
- The necessary features are extracted using the ICA.
- The performance metrics such as Correlation coefficient, Root Mean square error, Mean square error, Mean absolute error, & sensitivity are analyzed.

The remainder of this research is organized as follows. We provide a thorough assessment of the relevant work in Section 2. A summary of the proposed model's formulation is given in Section 3, and its results are presented in Section 4. The conclusion is mentioned in Section 5.

II. RELATED WORKS

Research [13] evaluated the uncertainties inherent in the WQI model; a statistically solid and rigorous technique. There are eight possible WQI models. Model uncertainty was estimated using a Monte Carlo simulation, and delay in WQI models was predicted using a Gaussian Process Regression approach for each sample location. It was discovered that the sub-index functions add significant uncertainty and impact the model's dependability. Research [14] developed a unique weight-integrated health hazard index and employ soft computing methods, including fuzzy logic, to identify and forecast the quality of drinkable water. This study's findings corroborate ANFIS's status as a credible method for predicting potable water's WQI, making it useful for those choosing water resource management. To effectively control and avoid river pollution, it is necessary to anticipate river WQI. Accurately predicting water quality parameters is challenging because of the significant coupling connection between distinct water quality indicators that impact each other [15]. Research [16] presented Extra Tree Regression (ETR), a ensemble model for forecasting monthly WQI values along the Lam Tsuen River in Hong Kong. Uncertainty analysis shows that the model structure is more influential on prediction accuracy than input parameters. The ETR model is an advancement over prior methods of WQI prediction since it requires fewer input parameters and yields more accurate predictions. Total Fe concentrations were average, 0.14 mg/L over the year. The WQI and the Water Pollution Index data are presented as means and standard deviations. Thirty percent of locations had WQI and Water Pollution Index values for groundwater, which made it unfit for human consumption. The total hazard index (THI) calculated the potential danger to human health from drinking groundwater [17]. The research [18] presented a Deep Learning (DL) stacking ensemble model to predict the WQI by combining existing models.

To reduce water pollution, keeping an eye on pH and turbidity was shown to be particularly important using sensitivity analysis. The greatest uncertainty is associated with the electrical conductivity and total dissolved solids in determining WQI. It allows for deploying clean manufacturing procedures to minimize pollution and enhance water quality by providing precise information to researchers, decision-makers, and water scientists. Research [19] used ML methods, such as character and ensemble learners, to predict WQI. Research [20] investigated the chemistry of view irrigate and evaluated its feasibility for industrial use in the semi-arid Guanzhong Basin in northwest China. Using multivariate statistical analysis methods, they calculated the relative importance of each characteristic to the overall quality of the water used in industrial processes. The results show that calcite, dolomite, anhydrite, and gypsum dissolve in groundwater to control the geochemistry of the water below the surface. According to the geospatial study, the northeastern parts of Xi'an and the Guanzhong Basin are home to the most polluted groundwater samples. Research [21] presented the results of a decade-long evaluation of artificial intelligence research on the WQI, intending to highlight the most promising approaches and promising models for use by future water quality. It also demonstrates that there is no universally superior Artificial intelligence (AI) model; it is necessary to the efficacy of various model combinations to determine which approach or model is most suitable for a certain country or area. AI has been more popular in recent decades, but there is still plenty of potential for competition and advancement in the water quality index. Research [22] used eight water quality criteria to evaluate a drinking water source by analyzing seven hundred and ten samples from a water-poor area of India. The whole sample was split into ten sections, each representing a distinct region of the world.

A approach to the problem of uniquely defining the quality of drinking water, which is complicated by the need to comprehend the cumulative influence of several criteria influencing water quality. A paradigm is presented that, with the correct tools, might be automated to help government organisations better comprehend and control changing water quality.

III. PROPOSED METHOD

In this research, the HWW-XGB was used to estimate the WQI of water samples. The preprocessing of the data using the Outlier Detection (Box Plot Analysis) and Min-max normalization was finished to minimize the computational time. The ICA is utilized to extract the required features. The pre-processed features are removed through feature extraction, eliminating unnecessary data for the next essential phase. Figure 1 shows the flow of the proposed model.

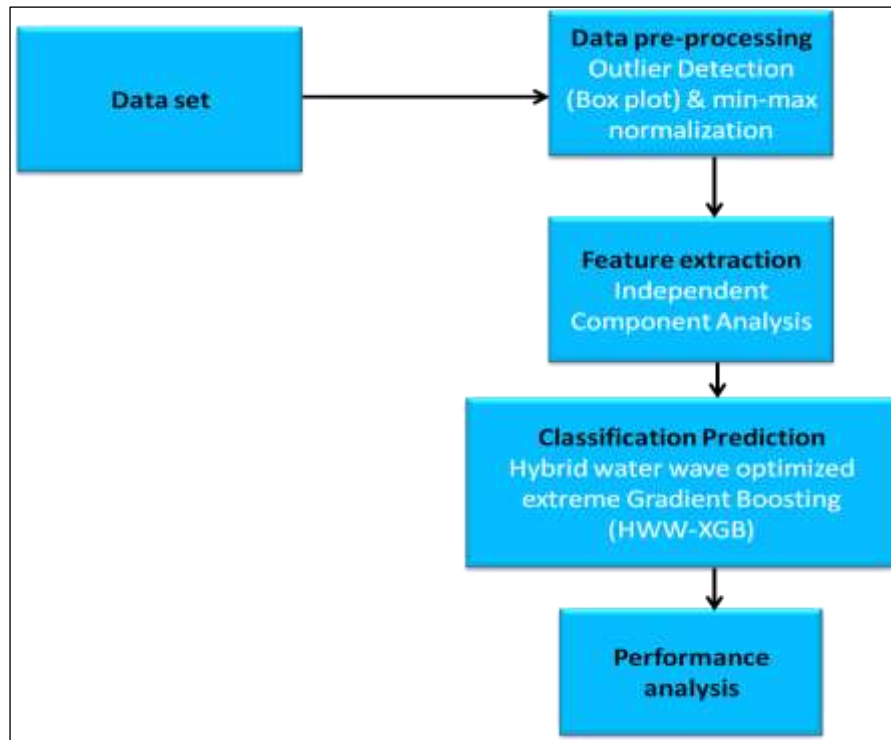


Figure 1: Flow of proposed model.
Source: Authors, (2026).

III.1 DATASET

Two strategically situated water quality monitoring stations throughout the watershed be used to collect monthly data on water quality over six years (2012-2018). There were readings for dissolved oxygen, total solids, fecal coli form, pH, phosphate, nitrate, turbidity, and electrical conductivity, among other parameters. Together, they were utilized to determine Iran's WQI. The Mazandaran Environmental Protection Organisation lab followed conventional sampling, storage, and analysis procedures. Utilising the tenfold traverse substantiation technique, the dataset was split 70:30 across the training and testing sets. Although no hard and fast rule exists, this ratio is the most often used modeling strategy for geographical and temporal forecasting [23].

III.2 DATA PREPROCESSING

The data preprocessing using the Outlier Detection (Box Plot Analysis) and Min-max normalization was finished to minimize the computational time.

III.2.1 Outer Detection

Missing data and data anomalies may significantly influence future model projections because of environmental changes and equipment failures at water quality monitoring sites. The experimental results are completed using linear interpolation, using the discovered anomalous values as missing values. It is common practice to spot outliers using modeling techniques and graphical approaches (such as box-line plots and normal distribution plots). The box-line property was utilized in this study to spot unusual results. The quantile of the data was utilized by the box-line plot method to find the extremes. This is a standard kind of statistical graph that has many applications in both research and business. Figure 2 depicts the form characterized by the box-line plot.

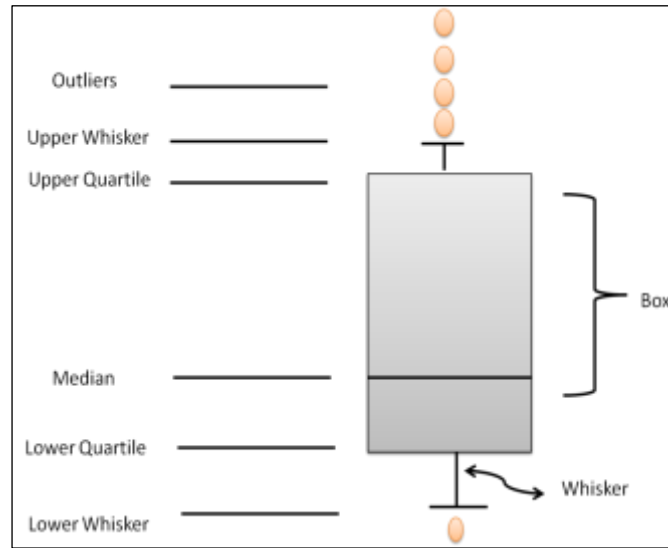


Figure 2: Structure of box-line diagram.
Source: Authors, (2026).

Figure 2 shows the median as the number that divides the upper and lower halves of the data. In contrast, the lower quartile represents the 25% quartile, and the higher quartile represents the 75% quartile. Where $Q3 - Q1$ is the quartile difference, the upper whisker formula is $Q3 + 1.5(Q3 - Q1)$, and the lower whisker formula is $Q1 - 1.5(Q3 - Q1)$. To judge whether or not a data point is an outlier, a box-line diagram may be utilized if the value of the variable in question is beyond the range covered by the upper and lower whiskers, respectively.

III.2.2 Min-Max Normalization

Preparing and enhancing datasets for subsequent analyses is the focus of data pre-processing, a step in the study of big data that aims to discover new information faster and more accurately. Putting an issue statement into practice is a preparatory step in the overall processing flow. Particularly important is the function that implementation plays in data processing techniques like soft computing, cloud computing, and the reduction or expansion of a variety of data before its usage in a later phase. Numerous normalizing methods are available, including the Min-Max method, the Z-score method, and the Decimal scaling method, to name a few. Min-max normalization is an initial processing stage, a scaling technique, or a mapping approach. It might be really helpful for assessing data or making plans. Normalization is calculated in equation (1),

$$K' = \left(\frac{k - \min \text{ value of } L}{\max \text{ value of } - \min \text{ value of } k} \right) * (P - H) + H \quad (1)$$

Where K is the Min-Max normalized set of data and $[H, P]$ is the bounding box. K is the true data range, while N is the mapped data range.

III.3 FEATURE EXTRACTION USING INDEPENDENT COMPONENT ANALYSIS

Let's say we have many independent measurements of a complex signal simultaneously. Time resolved spectroscopy allows us to get a movement of the explanatory transit's light curve at each spectrograph's resolution. Here, we refer to each time series measured using the notation $x_k(t)$, where N is the total number of time series and k is the index for that time series.

$$x_k(t) = a_{k,1}S_a(t) + S_{wn}(t) + a_{k,2}S_{sn1}(t) + a_{k,3}S_{sn2}(t) + \dots + a_{k,N_{sn}}S_{sn}(t) \quad (2)$$

As a vector sum (time dependency has been removed for simplicity).

$$x_k = a_{k,1}S_a + a_{k,2}S_{wn} + \sum_{l=3}^{N_{sn}} a_{k,l}S_{sn} \quad (3)$$

Where l is the estimated index of the source signal, we have $k = l$ unique source signals for non over full sets, which is the same as the number of time series. It's important to remember that when $k > l$, we have access to more time series than source signals inside the whole set. In such circumstances, we may use dimensionality reduction techniques for the data collection or use selection criteria to narrow down a subset of predicted source signals. We can also express Equation (2) in matrix form as

$$X = As \quad (4)$$

As an alternative notation, we may write $s = s_a + s_{wn} + s_{sn}$ to explicitly separate the astrophysical, white noise, and systematic contributions. The "mixing matrix" A consists of the weights a_k, l , and has $N \times N$ dimensions.

III.4 CLASSIFICATION USING HYBRID WATER WAVE OPTIMIZED EXTREME GRADIENT BOOSTING

The hybrid machine learning method known as Hybrid Water Wave Optimised extreme Gradient Boosting (HWW-XGB) combines the Water Wave Optimisation (WWO) algorithm with the extreme Gradient Boosting (XGB) algorithm. By merging the search and optimization skills of WWO, it seeks to improve the speed and optimization capabilities of XGB. The effective and popular gradient boosting method, XGB, excels at managing structured/tabular data and obtaining high prediction accuracy. By repeatedly optimizing an objective function that combines prediction performance and regularisation, it creates an ensemble of weak decision trees. The advantages of XGB are its quickness, scalability, and capacity for complicated nonlinear interactions. WWO is an optimization method that draws inspiration from nature and the behavior of water waves. It uses a metaheuristic algorithm to explore the search space and identify the best answers by simulating the motion of water waves. WWO has been used to tackle various optimization issues and has successfully located global optima. HWW-XGB combines the benefits of both methods to strengthen XGB's overall performance and the optimization process. It uses WWO's capabilities for exploration and exploitation to direct the search for the best hyperparameters or model configurations in XGB. HWWO-XGB aims to identify better combinations of hyperparameters that enhance model performance by searching the hyperparameter space using WWO. WWO is included in the XGB optimization process as part of the hybridization phase. A population of potential solutions is initialized in the hyperparameter space using the WWO method. Then, these possible solutions are assessed using XGB and the appropriate hyperparameter settings. Utilizing the iterative approach and the WWO tenets, the assessment findings are utilized to update and improve the solutions. In HWW-XGB, the hybridization of WWO and XGB enables more efficient exploration of the hyperparameter space, which may enhance model performance and generalization. The goal of HWW-XGB is to identify optimum or nearly optimal hyperparameter configurations for XGB models by combining the high prediction powers of XGB with the optimization capabilities of WWO.

III.4.1. Extreme gradient boosting

The Extreme Gradient Boosting (XGB) technique is a regression tree with decision-tree-like decision-making capabilities. It supports both classification and regression. This approach is a scalable and effective variation of the gradient boosting machine (GBM), which has been extensively used in computer vision, data mining, and other areas. Recently, XGB has primarily advanced in two locations as a sort of gradient boosting machine: speeding up the tree building and suggesting a new distributed technique for tree searching. The heart of XGB is optimizing the value of the goal function. Considering the dataset $D = \{(x_{-}(i), y_i)\}$ represents the x_i tumor's gene expression profile and the related y_i binary label (early stage or late stage) is used to identify the tumor. Equation (5) is the objective optimization function for the XGB model with K decision trees.

$$\hat{z}_j = \sum_{l=1}^L e_l(w_j), e_l \in E \quad (5)$$

Where f_k represents a separate tree with leaf scores, and F is the space of the regression tree. In equation (6), the loss function is shown.

$$K(e_k) = \sum k(\hat{z}_j, z_j) + \sum \Omega e_k \quad (6)$$

Différentiable loss function l in the first term assesses the discrepancy between the output $\hat{y}_i$ that anticipâtes and the actual output y_i .

$$\hat{z}_j^{(s)} = \hat{z}_j^{(s-1)} + e_s(w_j) \quad (7)$$

$$\Omega(e) = \gamma S + \frac{1}{2} ||x||^2 \quad (8)$$

Where w is the score on each leaf, and T is the number of leaf nodes. We may thus determine that from equation (9)

$$S(e_s) \approx \sum_{i=1}^S \left[\left(\sum_{j \in J_i} h_j \right) x_i + \frac{1}{2} \left(\sum_{j \in J_i} g_j + \lambda \right) x_i^2 \right] + \gamma S \quad (9)$$

Where g_i and h_i are the loss function first and second gradient statistics, respectively.

III.4.2. Water Wave Optimisation (WWO) Algorithm

The WWO method essentially models the processes of wave motion propagation, refraction, and breaking to achieve an optimal global/local search balance. Each water wave's fitness level is a function of its height and wavelength. Figure 3 shows that the water wave in shallow water is a greater fitness value, a higher wave height, and a longer wavelength than the wave in deep water, which have a lesser condition cost, a lower signal elevation, and a shorter wavelength.

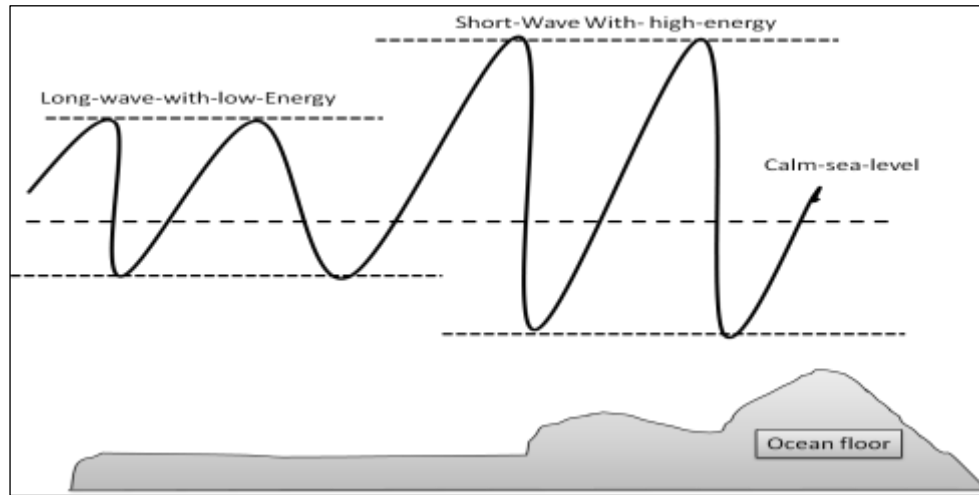


Figure 3: Variations in wave form between deep and shallow sea.
Source: Authors, (2026).

III.4.2.1 Propagation

The dissimilarity of the sea bed causes waves to propagate at varying heights and wavelengths. The initial water wave x is propagated to create a new body of water denoted by x' . To update the location in equation (10), we assume that the first water wave was at position x (d) & subsequent wet gesture was at position x (d).

$$x'(d) = x(d) + \text{rand}(-1, 1) \cdot \lambda L(d) \quad (10)$$

Where $L(d)$ denotes the span of the search space in units of d^{th} , λ is the wavelength of water wave x , and $\text{rand}(-1, 1)$ is a consistently distributed random integer between 1 and 1. If wave x' is not found within the search range, a new place is chosen randomly. The opposite is true, with wave x being kept and the wave height dropping by one to signify that the wave's energy has been expended. The wavelength λ is updated in equation (11):

$$\lambda = \lambda \cdot a^{-(f(x)-f_{\min})^{\epsilon}} / (f_{\max}^{-f} \min^{\epsilon}) \quad (11)$$

Where α is the attenuation coefficient for wavelengths, $f_{\min}$ and $f_{\max}$ are the lowest and most health standards, and ϵ is a tiny optimistic numeral chosen such that the numerator does not equal zero.

III.4.2.2 Breaking

The steepness of the wave crests increases as the wave energy rises. When the crest velocity is greater than the wave's propagation velocity, the wave will disintegrate into some independent waves. The WWO algorithm maximizes population variety by performing a breaking operation on the best water wave x^* . The equation (12) for the positional update is as follows:

$$x'(d) = x(d) + N(0,1) \cdot \beta L(d) \quad (12)$$

Where β is a broken gesture coefficient. If the fitness value of all the solitary waves acquired from the breaking operation is lower than that of water wave x^* , x^* ; otherwise, x^* be replaced with an optimum introverted gesture from the population.

III.4.2.3 Refraction

The energy of a water wave gradually decreases until it is zero as it travels through the water. It has been shown that using a refraction operation when the water wave height is zero helps the WWO algorithm converge more quickly and accurately. The refraction of water wave x is shown in equation (13).

$$x'(d) = N\left(\frac{x^*(d)+x(d)}{2}, \frac{|x^*(d)-x(d)|}{2}\right) \quad (13)$$

The function $N(\mu, \sigma)$ generates a Gaussian random number with a mean μ and standard deviation σ , where x^* is the best solution obtained at the present point. In reality, this enables wave x to gain insight from the currently best wave x^* . Since $h_{\max}$ is the current value of wave x' wave height, the equation (14) for the updated wavelength is shown below.

$$\lambda' = \lambda \frac{f(x)}{f(x')} \quad (14)$$

Algorithm 1: The WWO algorithm's process for solving problems

```

1      Initialize a water population O at random with m.
        Wavelength is set by waves.
        Breaking coefficient  $\beta$ , Reduction coefficient  $\alpha$ , and
        Coefficient of decrease in wavelength  $\lambda$ 
2      while stop the extinction criterion is not satisfied, do
3          for each water wave  $x \in O$  do
4              Use equation (10) to propagate x to a new  $w'$ ;
5              if  $f(w') < f(w)$  next
6              if  $f(w') < f(w')$  next
7                  Break  $w'$  by utilizing equation (12);
8                  Update  $w^*$  with  $w'$ ;
9          else
10             Decrease  $x.h$  by one;
11             if  $x.h = 0$  Next
12             Refract x to a new  $w'$  by utilizing equations (13) and (14);
13             Using equation (11) update the wavelengths.
14      return  $w^*$ 

```

In Algorithm 1, the WWO algorithm's solution process is described.

Source: Authors, (2026).

IV. RESULT

The suggested method's performance metrics are examined, including Mean absolute error, Root Mean square error, Mean square error, Correlation coefficient, and sensitivity. The existing techniques are adaptive neuro-fuzzy inference system (ANFIS) [24], artificial neural networks (ANN) [24], and K-nearest neighbors (KNN) [24].

IV.1 MEAN SQUARE ERROR (MSE)

The average squared difference between the predicted and actual values in a regression issue is calculated using the MSE, a commonly used statistic. It gives a quantitative assessment of a regression model's overall accuracy or goodness of fit. Mean Square Error (MSE) is represented by equation (15),

$$MSE = (1/n) * \sum (y - \hat{y})^2 \quad (15)$$

Where n denotes the number of data points or samples, $\sum$ denotes the summation symbol, y represents the actual (observed) values, and $\hat{y}$ conveys the predicted values from the regression model. Computing the squared difference between each predicted value and its corresponding actual value adds up all of these squared differences. The squared differences ensure that positive and negative mistakes equally impact the MSE. The sign is removed by taking the square of each contrast, and bigger mistakes are penalized more severely. Figure 4 and Table 1 show Mean Square Error outcomes; ANN scores 56%, KNN 39%, ANFIS 44%, and HWW-XGB 35%.

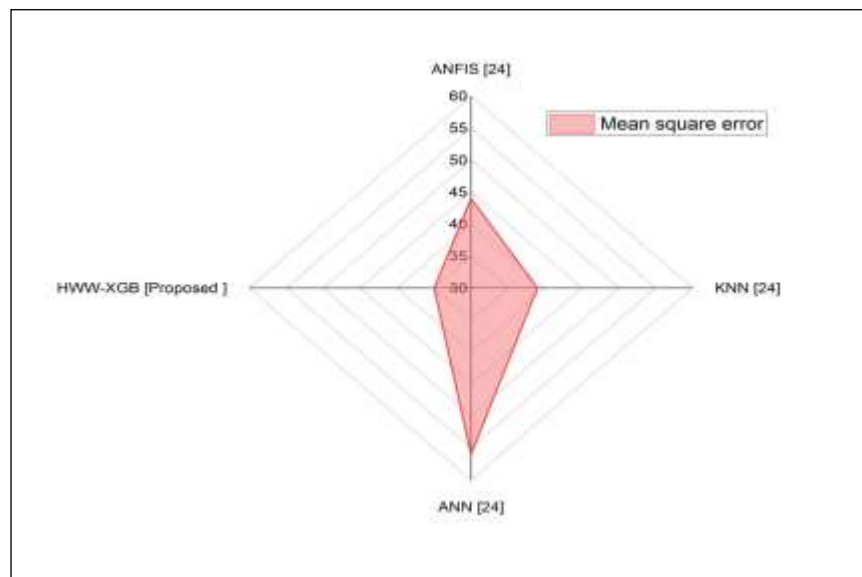


Figure 4: Results of Mean Square Error.

Source: Authors, (2026).

Table 1: Numerical Outcomes of Mean Square Error.

Methods	Mean square error
ANFIS [24]	44
KNN [24]	39
ANN [24]	56
HWW-XGB [Proposed]	35

Source: Authors, (2026).

IV.2 MEAN ABSOLUTE ERROR (MAE)

It is utilized to calculate the usual complete disparity between the predicted and actual values in a regression issue. It indicates the average size of mistakes without taking their direction into account. The MAE equation (16) is as follows:

$$MAE = (1/n) * \sum |y - \hat{y}| \quad (16)$$

Where n denotes the number of data points or samples, $\sum$ denotes the summation symbol, y represents the actual (observed) values, and $\hat{y}$ indicates the predicted values from the regression model. To determine the MAE, add up all of the absolute differences between each predicted value and its corresponding actual value. Without considering their direction, the absolute differences are utilized to estimate the size of mistakes. Because MAE does not square the differences, it handles positive and negative mistakes equally, in contrast to Mean Square Error (MSE), which does. Figure 5 and Table 2 show Root Mean Absolute Error outcomes; ANN scores 58%, KNN 71%, ANFIS 67%, and HWW-XGB 40 %.

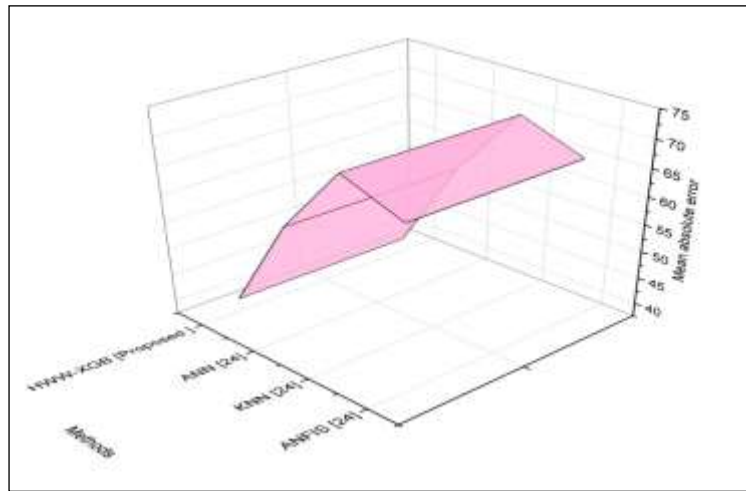


Figure 5: Results of MAE.

Source: Authors, (2026).

Table 2: Numerical Outcomes of Mean Absolute Error.

Methods	Mean absolute error
ANFIS [24]	67
KNN [24]	71
ANN [24]	58
HWW-XGB [Proposed]	40

Source: Authors, (2026).

IV.3 ROOT MEAN SQUARE ERROR

In a regression issue, the average error size between the predicted and actual values is often measured using the Root Mean Square Error (RMSE) metric. It is a variant of the MSE that scales the metric back to the dependent variable's original scale by taking the square root of the MSE. The RMSE measures the usual size of errors and is often used to evaluate a regression model's accuracy and fit quality. The RMSE equation (17) is as follows:

$$RMSE = \sqrt{(1/n) * \sum (y - \hat{y})^2} \quad (17)$$

Where n denotes the number of data points or samples, $\sum$ denotes the summation symbol, y represents the actual (observed) values, and $\hat{y}$ indicates the predicted values from a model or estimation. Figure 6 and Table 3 show Root Mean Square Error outcomes; ANN scores 51%, KNN 56%, ANFIS 65%, and HWW-XGB 42%.

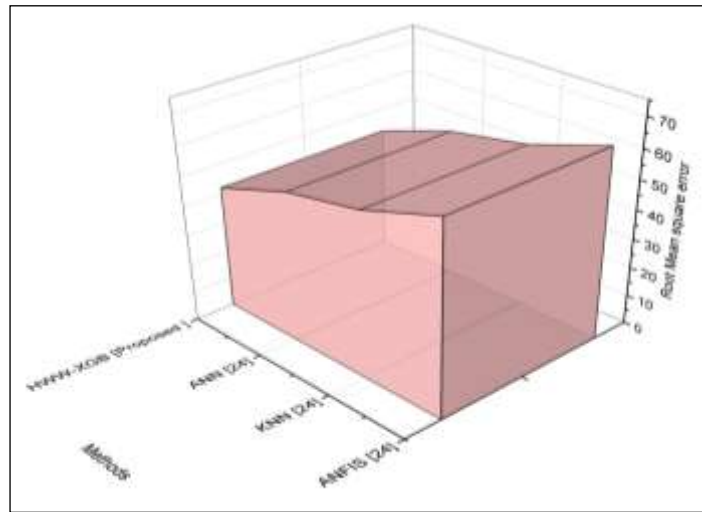


Figure 6: Results of Root Mean Square Error.
Source: Authors, (2026).

Table 3: Numerical Outcomes of RMSE.

Methods	Root Mean square error
ANFIS [24]	65
KNN [24]	56
ANN [24]	51
HWW-XGB [Proposed]	42

Source: Authors, (2026).

IV.4 CORRELATION COEFFICIENT

The degree and direction of the linear connection between two variables are shown by a statistical measure. The correlation coefficient, often known as Pearson's correlation coefficient, is represented by the equation (18):

$$r = \frac{(\sum((x - \mu_x)(y - \mu_y)))}{(\sqrt{\sum(x - \mu_x)^2} * \sqrt{\sum(y - \mu_y)^2})} \quad (18)$$

Where r denotes the Correlation coefficient, $\sum$ denotes the summation symbol, x & y denotes the individual values of the two variables being correlated, and μ_x & μ_y denotes the means of the respective variables. The correlation coefficient quantifies the degree to which the variables vary from their mean values in the same or opposing directions. In Figure 7 and Table 4, correlation coefficient outcomes, ANN scores 84%, KNN 82%, ANFIS 80%, and HWW-XGB 95%.

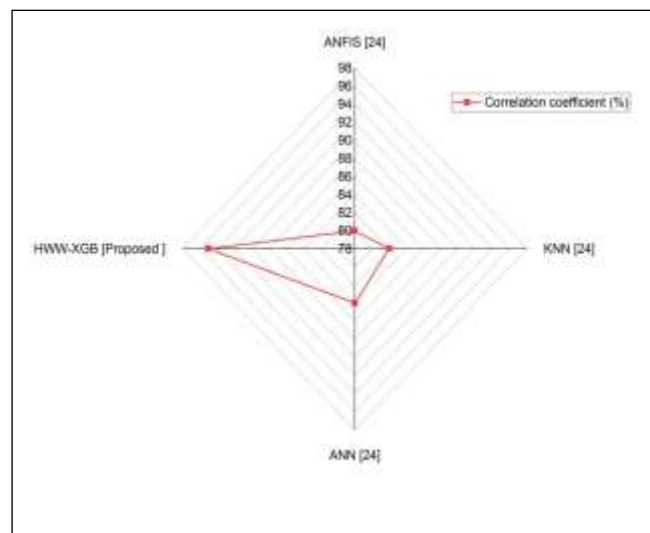


Figure 7: Results of the correlation coefficient.
Source: Authors, (2026).

Table 4: Numerical outcomes of the correlation coefficient.

Methods	Correlation coefficient (%)
ANFIS [24]	80
KNN [24]	82
ANN [24]	84
HWW-XGB [Proposed]	95

Source: Authors, (2026).

IV.5 SENSITIVITY

The sensitivity of the proposed model is its capability to identify each significant sample in a data collection. It may be calculated statistically by dividing the proportion of true positives and false negatives by the sum of true positives and false negatives. Equation (19) is used to determine the sensitivity.

$$sensitivity = \frac{TP}{TP+FN} \quad (19)$$

In Figure 8 and Table 5, Sensitivity-wise, ANN scores 85%, KNN 76%, ANFIS 74%, and HWW-XGB 93%.

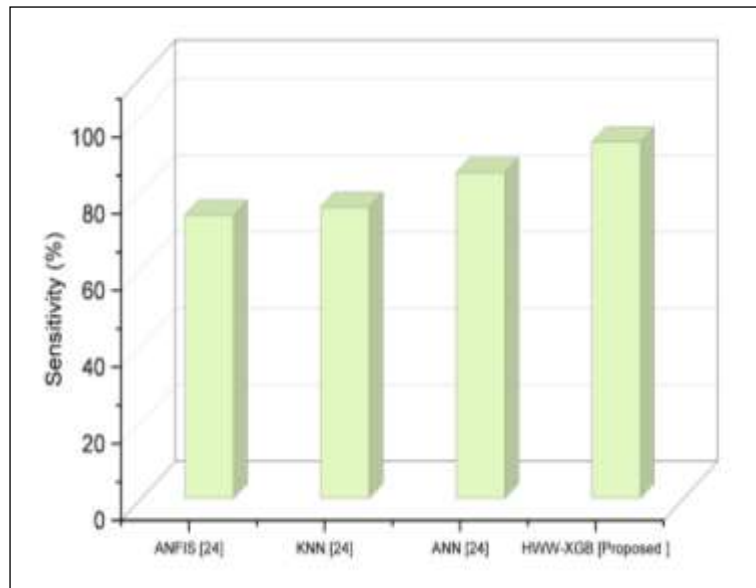


Figure 8: Results of sensitivity.

Source: Authors, (2026).

Table 5: Numerical outcomes of sensitivity.

Methods	Sensitivity (%)
ANFIS [24]	74
KNN [24]	76
ANN [24]	85
HWW-XGB [Proposed]	93

Source: Authors, (2026).

V. CONCLUSION

For assessing and monitoring the general quality of water resources, the WQI is a useful metric. It provides a numerical rating to show the combined effect of several water quality parameters, allowing for easy comparison and interpretation of water quality data. The research uses the novel HWW-XGB to estimate WQI for water samples. To save computing time, the data in this study were preprocessed using outlier detection and min-max normalization. The necessary characteristics are extracted using the ICA. The pre-processed features are eliminated by feature extraction, freeing up data for the next crucial stage. The HWW-XGB classifier receives the gathered characteristics as input. Compared to the current technique, a new approach's efficacy was shown. In the proposed method analysis, the performance metrics exhibit notable scores: Mean Square Error attains a respectable 35%, Mean Absolute Error attains a good 40%, Root Mean Square Error attains a respectable 42%, Correlation Coefficient attains a respectable 95%, and Sensitivity achieves a respectable 93%. By using the fewest characteristics feasible, rather than a surprising number of features, the future study aims to estimate the quality of water effectively.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: R.Sarala, R.S. Ponmagal.

Methodology: R.Sarala, R.S. Ponmagal.

Investigation: R.Sarala, R.S. Ponmagal.

Discussion of results: R.Sarala, R.S. Ponmagal.

Writing – Original Draft: R.Sarala, R.S. Ponmagal.

Writing – Review and Editing: R.Sarala, R.S. Ponmagal.

Resources: R.Sarala, R.S. Ponmagal.

Supervision: R.Sarala, R.S. Ponmagal.

Approval of the final text: R.Sarala, R.S. Ponmagal.

VII. REFERENCES

- [1] Derdour, A., Jodar-Abellan, A., Pardo, M.Á., Ghoneim, S.S. and Hussein, E.E., 2022. Designing Efficient and Sustainable Predictions of Water Quality Indexes at the Regional Scale Using Machine Learning Algorithms. *Water*, 14(18), p.2801.
- [2] Wong, W.Y., Al-Ani, A.K.I., Hasikin, K., Khairuddin, A.S.M., Razak, S.A., Hizaddin, H.F., Mokhtar, M.I. and Azizan, M.M., 2022. Water quality index using modified random forest technique: assessing novel input features. *CMES-Computer Modeling in Engineering & Sciences*.
- [3] Zamili, H., Bakan, G., Zubaidi, S.L. and Alawsi, M.A., 2023. Water quality index forecast using artificial neural network techniques optimized with different metaheuristic algorithms. *Modeling Earth Systems and Environment*, pp.1-11.
- [4] Agrawal, S., Jain, S.K., Khatri, A., Agarwal, M., Tripathi, A. and Hu, Y.C., 2022. Novel PSO Optimized Voting Classifier Approach for Predicting Water Quality. *Mathematical Problems in Engineering*, 2022.
- [5] Tong, S., Li, W., Chen, J., Xia, R., Lin, J., Chen, Y. and Xu, C.Y., 2023. A novel framework to improve the consistency of water quality attribution from natural and anthropogenic factors. *Journal of Environmental Management*, 342, p.118077.
- [6] Imani, M., Hasan, M.M., Bittencourt, L.F., McClymont, K. and Kapelan, Z., 2021. A novel machine learning application: Water quality resilience prediction Model. *Science of the Total Environment*, 768, p.144459.
- [7] Tiyasha, Tung, T.M. and Yaseen, Z.M., 2021. Deep learning for prediction of water quality index classification: tropical catchment environmental assessment. *Natural Resources Research*, 30(6), pp.4235-4254.
- [8] Lap, B.Q., Du Nguyen, H., Hang, P.T., Phi, N.Q., Hoang, V.T., Linh, P.G. and Hang, B.T.T., 2023. Predicting Water Quality Index (WQI) by feature selection and machine learning: A case study of An Kim Hai irrigation system. *Ecological Informatics*, p.101991.
- [9] Paul, V., Ramesh, R., Sreeja, P., Jarin, T., Kumar, P.S., Ansar, S., Ashraf, G.A., Pandey, S. and Said, Z., 2022. Hybridization of long short-term memory with Sparrow Search Optimization model for water quality index prediction. *Chemosphere*, 307, p.135762.
- [10] Patil, D., Kar, S., Shastri, V. and Gupta, R., 2023. Qualitative and health risk assessment of water using a novel weight-integrated health hazard and fuzzy-derived indices. *Sustainable Water Resources Management*, 9(2), p.55.
- [11] Lee, H., Park, S., Hang, V., Nguyen, M. and Shin, H.S., 2023. Proposal for a new customization process for a data-based water quality index using a random forest approach. *Environmental Pollution*, 323, p.121222.
- [12] Ramaraj, M. and Sivakumar, R., 2023. Remote Sensing and Nonlinear Auto-regressive Neural Network (NARNET) Based Surface Water Chemical Quality Study: A Spatio-Temporal Hybrid Novel Technique (STHNT). *Bulletin of Environmental Contamination and Toxicology*, 110(1), p.28.
- [13] Uddin, M.G., Nash, S., Rahman, A. and Olbert, A.I., 2023. A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches. *Water Research*, 229, p.119422.
- [14] Patil, D., Kar, S. and Gupta, R., 2023. Classification and Prediction of Developed Water Quality Indexes Using Soft Computing Tools. *Water Conservation Science and Engineering*, 8(1), p.16.
- [15] Chen, L., Wu, T., Wang, Z., Lin, X. and Cai, Y., 2023. A novel hybrid BPNN model based on adaptive evolutionary Artificial Bee Colony Algorithm for water quality index prediction. *Ecological Indicators*, 146, p.109882.
- [16] Asadollah, S.B.H.S., Sharafati, A., Motta, D. and Yaseen, Z.M., 2021. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *Journal of environmental chemical engineering*, 9(1), p.104599.
- [17] Biswas, P., Hossain, M. and Patra, P.K., 2023. Arsenic hydrogeochemistry, quality assessment, and associated health risks of groundwater through the novel water pollution index (WPI) and GIS approach. *Groundwater for Sustainable Development*, 21, p.100944.
- [18] Talukdar, S., Ahmed, S., Naikoo, M.W., Rahman, A., Mallik, S., Ningthoujam, S., Bera, S. and Ramana, G.V., 2023. Predicting lake water quality index with sensitivity-uncertainty analysis using deep learning algorithms. *Journal of Cleaner Production*, 406, p.136885.
- [19] Aldrees, A., Awan, H.H., Javed, M.F. and Mohamed, A.M., 2022. Prediction of water quality indexes with ensemble learners: Bagging and Boosting. *Process Safety and Environmental Protection*, 168, pp.344-361.
- [20] Nsabimana, A. and Li, P., 2023. Hydrogeochemical characterization and appraisal of groundwater quality for industrial purposes using a novel industrial water quality index (IndWQI) in the Guanzhong Basin, China. *Geochemistry*, 83(1), p.125922.

- [21] Aminu, I.I., 2022. A novel approach to predict Water Quality Index using machine learning models: A review of the methods employed and future possibilities. *Global Journal of Engineering and Technology Advances*, 13(02), pp.026-037.
- [22] Sinha, K.K., Gupta, M.K., Banerjee, M.K., Meraj, G., Singh, S.K., Kanga, S., Farooq, M., Kumar, P. and Sahu, N., 2022. Neural network-based modeling of water quality in Jodhpur, India. *Hydrology*, 9(5), p.92.
- [23] Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H. and Kazakis, N., 2020. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, 721, p.137612.
- [24] Hmoud Al-Adhaileh, M. and Waselallah Alsaade, F., 2021. Modeling and prediction of water quality by using artificial intelligence. *Sustainability*, 13(8), p.4259.
- [25] Torky, M., Bakhiet, A., Bakrey, M., Ismail, A.A. and 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).