Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 1 (2024)

Water Quality Index Forecasting Using Machine Learning

Arun Kumar T. D. [a*], B. E. Yogendra[b], Prashanth J. [c], Prema N. S. [d]

[a] *Assistant Professor, Department of Civil Engineering, Kalpataru Institute of Technology, Tiptur, INDIA.*

[b] *Professor, Department of Civil Engineering, Malnad College of Engineering, Hassan, INDIA.*

[c] *Assistant Professor, Department of Civil Engineering, National Institute of Technology: Silchar, Assam, INDIA.*

[d] *Associate Professor, Department of Information Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, INDIA*

***Corresponding author:** Arun Kumar T D*

Abstract

The forecasting of water quality in the Tumkur district, Karnataka state, India, was conducted using several machine learning methods, such as support vector machines, regression tree, linear regression, and neural network. The Water Quality Index (WQI) was calculated using factors like total hardness, pH, alkalinity, turbidity, chloride, dissolved solids, and conductivity. A ratio of 80:20 was used to divide the dataset into two groups for the purposes of validating and training the models. Over the past few years, water quality has been negatively impacted by multiple pollutants, making it crucial to predict and model water quality for effective pollution reduction. Advanced machine learning methods were devised for this studyto forecast the WQI. The models' effectiveness was assessed using a variety of statistical and visual assessment techniques. Among the models used, Support Vector Machine and Linear Regression exhibited superior performance, with R2 value 0.96 and 0.99 respectively for the train and test data sets. The implementation of AdaBoost, an ensemble model, for forecasting WQI also yielded excellent results, achieving R2 values of 1 and 0.91 for the training and testing data, respectively.

Keywords: Water quality index; Machine learning; Random Forest; linear regression; Support vector machine; Neural Network

1. Introduction

A growing and difficult issue is the pollution of rivers from both direct and indirect sources. Water quality degradation has a substantial impact on aquatic ecosystems as well as the availability of safe fresh water for agriculture and human use. Every development initiative has the potential to have detrimental environmental effects, and developing countries frequently experience periods of rapid economic growth. Due to the fast-expanding wealth and population, pressure on the natural fertility of soils increases, frequently over-extracting nutrients and necessitating the use of artificial fertilizers. Extra fertilizer is typically carried into rivers and the groundwater. Rivers frequently cause permanent harm to ecosystems and human health by transporting toxins to lakes and oceans. For effective, sustainable water management and preserving the health of people and the environment, water quality monitoring and assessment are crucial.

A non-dimensional index WQI is created by selecting certain water quality characteristics. Measures offer a categorical assessment of the water condition of bodies of water in the past and present.Ca2+, Mg2+, NO-3, and other commonly used factors to forecast the WQIs, like dissolved oxygen (DO), pH, temperature, and total suspended solids (TSS), are examples of frequent variables[1]. The WQI offers a significant value to direct the

policies and actions of decision-makers. However, because sub-indices are calculated inside WQI equations, the calculation of WQI is not simple.

WQI calculations have the drawbacks of taking a long time, being laborious, complicated, and inconsistent because WQIs frequently employ different formulae. This conversation may have made it clear, but there is no single WQI methodology. In this regard, the present research attempts to implement soft computing techniques to forecast the water quality of a supply system. The objective of the study is to present a trustworthy approach for predicting water quality using a machine learning technique as precisely as required. The diagrammatic representation of the present study is shown in Figure1.

The study aims to analyze water quality data to gain insights and understanding about the condition and characteristics of a particular body of water. Water quality refers to the physical, chemical, and biological properties of water, which determine its suitability for various uses such as drinking, recreation, and ecosystem health.

Analyzing water quality data involves collecting samples from different locations within the water body, such as rivers, lakes, or groundwater sources, and conducting various tests and measurements. These tests can include assessing parameters such as temperature, pH level, dissolved oxygen, turbidity, conductivity, nutrient concentrations (such as nitrates and phosphates), heavy metal content, and the presence of pollutants or contaminants. By analyzing water quality data, researchers can evaluate the overall health and ecological status of the water body. This information is crucial for making informed decisions regarding water resource management, environmental protection, and public health. It can help identify potential sources of pollution, understand the impacts of human activities on water systems, and monitor the effectiveness of water treatment and conservation measures.

The analysis of water quality data often involves statistical methods and modeling techniques to identify trends, patterns, and correlations between different variables. This allows scientists and policymakers to assess the current state of water quality, predict future changes, and develop appropriate strategies for remediation or protection. Furthermore, water quality analysis is important in assessing compliance with regulatory standards and guidelines set by governmental bodies. It provides a basis for determining if water bodies meet the necessary criteria for designated uses, such as drinking water sources or recreational areas. Additionally, water quality data analysis can contribute to the early detection of potential health risks and the implementation of appropriate mitigation measures.

Overall, the study of water quality data aims to provide a comprehensive understanding of the condition of water resources, facilitating informed decision-making and sustainable management practices to ensure the availability of clean and safe water for both humans and ecosystems. To do this, the researchers collect water quality data from various sources. Once the data is collected, they preprocess it to remove any noise or outliers that may affect the accuracy of the analysis. After preprocessing the data, the researchers calculate the WQI values for each sample. WQI is a numerical rating of water quality depending on several parameters, such as pH, dissolved oxygen, and turbidity. Once the WQI values have been calculated, the researchers divide the obtained dataset into two parts: a training set and a testing set. The training set is used to develop and fine-tune the regression models that will be used to analyze the water quality data. The testing set is used to validate the performance of the models. Once the dataset has been divided into training and testing sets, the researchers allocate regression models to analyze the water quality. The performance of each model is evaluated based on how well it predicts the WQI values. Finally, the researchers compare the results of the different regression models to determine which one performs the best. This data can assist them to identify the factors that most greatlyaffect water quality and develop strategies to improve it.

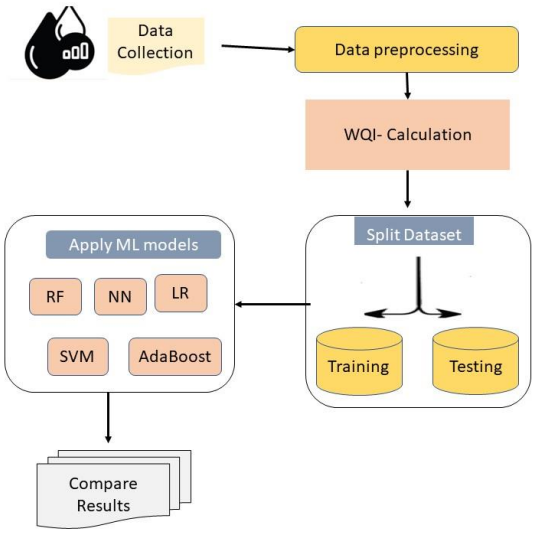

Fig. 1: Proposed WQI prediction flow diagram.

2. Literature review

Deterministic models aim to represent the various chemical and physical processes using statistical terms, incorporating variables derived from experience, examination, historical data, or empirical knowledge. To obtain suitable solutions for the model, it is common to simplify differential equations. Practical experience is often necessary before achieving optimal outcomes. Based on the model's performance, assumptions and simplifications may need to be made to solve the equations related to the model.

Statistical models, through the collection of field data, seek to derive general principles from experimental observations. The analysis and verification of hypotheses and data necessitate the careful selection of analytical methods in statistical modelling and assessment. These models typically rely on a substantial amount of field data for their research and are often highly complex. Furthermore, many statistically based models of water quality assume of a normal and linear distribution when establishing the relationship between the predictor and response variables. However, because several elements might affect water quality, conventional data processing techniques are no longer effective enough to address this problem. As a result, these parameters exhibit a complex non-linear relationship with the parameters used to predict water quality. Therefore, applying statistical approaches typically has low accuracy.

Conventionalapproaches for calculating WQI have limitations, such as the need to calculate sub-index values for various water quality parameters, which can be time-consuming and error prone. To overcome these limitations, some researchers have used machine learning (ML)-based models, which extend several benefits over traditional methods. Firstly, ML-based models generate a WQI value instantly and do not require the calculation of sub-index values. This is because they use a non-physical approach, where mathematical models are trained on water quality data to learn the relationships between various water quality parameters and the WQI. Secondly, ML algorithms offer several advantages, such as their non-linear structures, which enable them to model complex relationships between several water quality factors and the WQI. They can also handle large datasets with data at various scales and are insensitive to missing data. This makes them well-suited for analyzing water quality data, which can be complex and challenging to interpret.

However, the predictive power of machine learning algorithms depends on the accuracy and methodology of data collection and processing. The quality and quantity of data used for training and testing the models, as well as the preprocessing steps used to remove noise and outliers, can significantly affect the accuracy of the models. Therefore, careful attention must be paid to data collection and processing to ensure that the models are accurate and reliable.

Hayes et al. combined two models, namely a daily-scale optimal dispatch model and a quasi-static two-dimensional dissolved oxygen reservoir model, to enhance the downstream water quality. They employed an environmental fluid dynamics code to create a two-dimensional numerical model that simulated the water conditions of the Mudan River [3]. In a separate study, Batur and Maktav predicted the water quality of Lake Gala (Turkey) by utilizing satellite image fusion generated on the principal component analysis method. Jaloreeet al. attempted to forecast the WQ of the Narmada River using a decision tree model and five WQ indicators[5]. Furthermore, a study recommended the utilization of the deep Bidirectional Stacked Simple Recurrent Unit model to develop a precise WQ forecasting strategy specifically designed for smart mariculture[6].

A model for forecasting WQ of China's Chao Lake was developed by Liao, Sun, and their team [7] by integrating the ANN and decision tree algorithms[7].Yan and Qian projected an affinity propagation clustering model (AP-LSSVM) based on a least-squares support vector machine[8]. This model reacts quite quickly to vacancies. This model responds to openings fast. Solanki et al. employed a deep learning network to examine and forecast the chemical eigenvalues of water, specifically dissolved oxygen and pH, which was claimed to produce more precise results than supervised learning-based techniques[9].

A novel hybrid model was created by Li et al. [10] by combining the Markov chain approach and a neural network. With the use of this model, dissolved oxygen, a crucial indication of the WQ, has been anticipated. Khan and See developed a water quality (WQ) model incorporating dissolved oxygen, chlorophyll, conductivity, and turbidity. They utilized an artificial neural network[11]. Yan et al. proposed the utilization of a genetic algorithm and particle swarm optimization algorithm to improve the predictive capabilities of the backpropagation neural network for estimating oxygen levels in lakes. The reported results indicated a significant enhancement in prediction accuracy[12].

Ahmed et al. utilized supervised machine learning techniques to assess the water quality class and the water quality index , which serves as a comprehensive indicator summarizing the overall water quality [13].To construct the WQI, they employed gradient boosting with a learning rate of 0.1 and polynomial regression with a degree of 2, yielding mean absolute errors of 1.96 and 2.72, respectively.

According to Ganga et al., in the Andhra Pradesh district of Kadapa, 70% of the data was utilised for training and 30% as test data for verification[15]. This model divides the provided locations into three classifications: Excellent, Good, and Poor for drinking. It has a 6.25% error rate and 93.75% accuracy.

3. Materials and Methods

3.1 Dataset

Tiptur taluk is about 75 km from Tumkur district and covers an area of about 758.5 sq. km. The average temperature ranges from 11℃ in winter and 38℃ during summer. Average rainfall of Tiptur town is 503 mm and its geographical area is 76,510 ha. For forecasting and evaluating the water quality in the distribution system, the municipal water supply system of Tiptur is used as a case study. Tiptur relies mostly on borewells and the Bagur-Navile tunnel's outflow to deliver water to Tumkur and Tiptur town for drinking, industrial use, and other domestic needs. The population of Tiptur town is 53,053 as per 2001 census. The town is spread over an area of 11.60 km2 with 6340 number of households in 31 wards. The decadal population growth rate is 47.2 % with a population density of 4574 persons/ km2. There are coconut oil & coir industries in the town and is predominantly agriculture based. Owing to the presence of industries, the city is expected to have a higher growth rate and floating population. The present water supply is 135 LPCD, sourced from Hemavathi canal, 6 km from Tiptur with a water treatment plant of 17.5 MLD capacities. The different water sample collected for analysis is named from N1 through N9 for the study purpose as shown in table 1 and Figure 2.The data is about 1000 samples which includes measurements ofTotal Dissolved Solids (TDS), Hardness, Alkalinity, Electrical Conductivity (EC), Dissolved Oxygen (DO), Chloride, Turbidity, and pH.

Table 1:Zone details of Tiptur city

Area number	Location details of water tank
N1	Sharada Nagar
N2	Vidya Nagar
N3	ShadakshriBadavane
N4	Govinapura
N5	Gandhi Nagar
N6	Bashaveshwar Nagar
N7	DoddaPete
N8	Goragondanahalli
N9	Halepalya

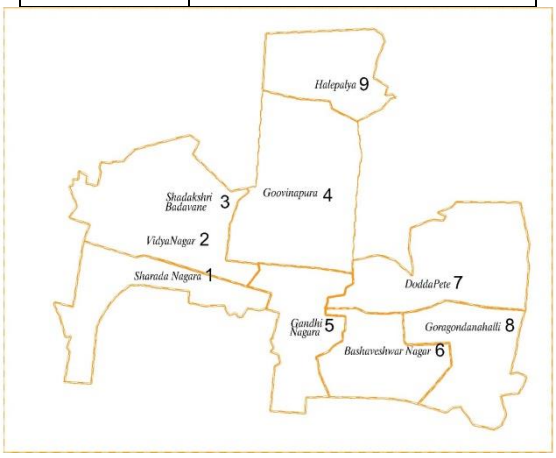

Fig. 2: Location of water sources of Tiptur

Table 2: Limits of the variables that can be used to determine WQI.

Parameter	BIS-standard limit
TDS	500
Hardness	200
Alkalinity	200
Electrical Conductivity	300
Dissolved Oxygen	5
Chloride	250
Turbidity	5
pH	8.5

The WQI was employed to comprehend and evaluate the water quality of each water sample. The term "WQI" refers to the relative importance and influence of several water quality metrics on the quality of the water. The calculation of WQI utilized the Indian standard for drinking water (BIS, 1991) as shown in table 2.

The following steps were used to calculate WQI using the weighed arithmetic index method[16]:

i. Let's say there are "n" different water quality parameters.
ii. The quality rating (Q_n) for the nth parameter is a number that indicates how much this parameter has changed from its standard permitted value in the polluted water.
iii. Values for Q_n are provided by the relationship equation (1).

$$Q_n = 100\frac{(V_n - V_i)}{(V_s - V_i)} \qquad (1)$$

Where V_n = observed value, and V_s = standard value Ideal value is V_i.

iv. Except for key parameters like pH and dissolved oxygen, V_i=0 in most circumstances.Calculation of pH & DO quality rating ($V_i \neq 0$) is done using the equation (2) and (3).

$$Q_{pH} = 100\frac{(V_{pH} - 7.0)}{(8.5 - 1.0)} \qquad (2)$$

$$Q_{DO} = 100\frac{(V_{DO} - 14.6)}{(5.0 - 14.6)} \qquad (3)$$

v. Calculating unit weight: The recommended requirements for each of the different water quality metrics are oppositely correlated to the unit weight (w_n) for those parametersas in equation (4).

$$W_n = \frac{k}{S_n} \qquad (4)$$

w_n = unit weight for nth parameter

S_n = standard acceptable value for the nth parameter
k = proportionality constant.

vi. WQI is computedusing equation (5).

$$WQI = \sum_{n=1}^{n} q_n w_n \; / \sum_{n=1}^{n} w_n \qquad (5)$$

As graded by Mishra & Patel[17] ,the following list outlines whether WQI levels are suitable for human consumption.

WQI Range	***Classification***
0–25	Excellent
26–50	Good
51–75	Bad
76-100	Very Poor
100 & higher	Unfit

WQI application is a practical technique for determining whether water is suitable for various beneficial purposes[18]. The attribute correlation matrix of the traits in this dataset with pre-processing is shown in Figure 3. This graph demonstrates that the EC and pH are positively correlated with WQI, while the other parameters are negatively correlated.

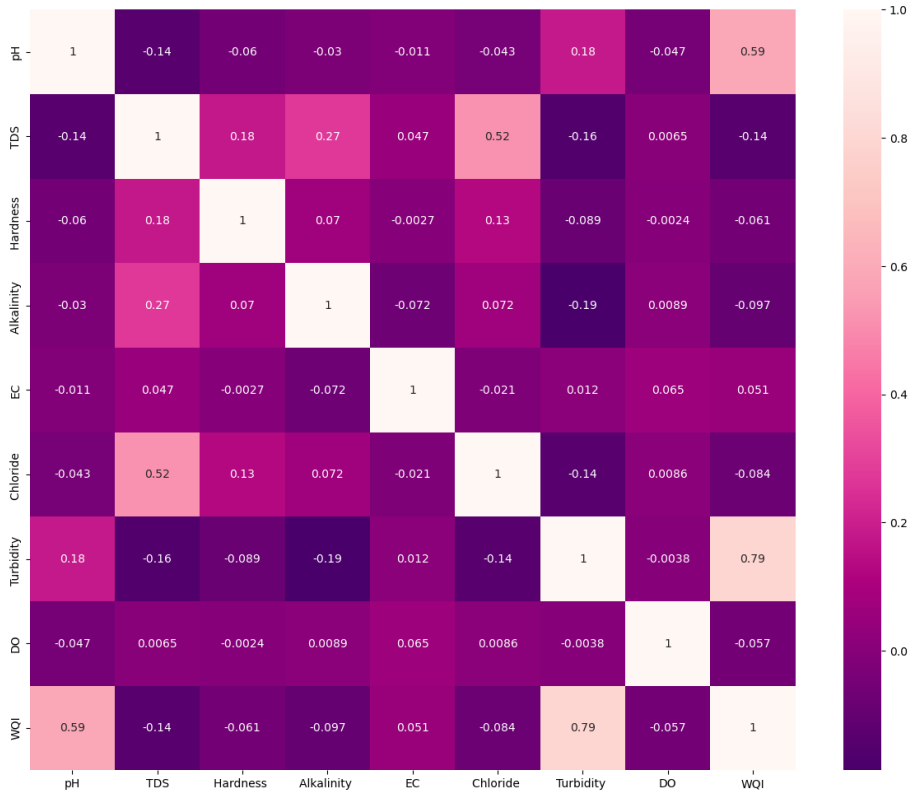

Fig. 3: Heat map of features

3.2 Methodology

The primary objective of this study is to use four regression techniques to estimate the WQI. In each trial, the usual 5-fold cross validation technique is used. With the help of this procedure, the classifiers may split the data into 4 and 1 folds, with 4 folds being used for training and 1-fold being left over for testing. Eighty percent of the data is collected for training, and the remaining twenty percent is used for testing. The models are first tested on test data after being trained on training data. As indicated above, this investigate utilizes the following regression models for forecasting WQI.

Random Forest (RF): An approach for supervised machine learning called Random Forest (RF) builds a forest and randomises it. A forest, or collection of Decision Trees, is trained using the bagging approach. To produce a reliable and accurate classification, Random Forest builds many decision trees and combines them. The ability to use the Random Forest method for both classification and regression analysis is by far its greatest benefit.

Neural Network (NN): It is a component of Artificial Intelligence (AI), a model of learning whose operation is impacted by the operation of a biological neuron. The neural network is made up of nodes, which process the data provided to them as input and transmit the results to other nodes. The activation or node value is referred to as each node's output. Weights attached to the nodes can be changed to aid network learning. These weights show how much an input may or may not influence an outcome.

Linear regression (LR): When describing the relationship between a scalar answer and one or more explanatory factors in statistics, linear regression is a linear approach. Relationships are modelled using linear predictor functions in linear regression, and the model's unidentified variables are estimated using the data. The term "linear model" is used to describe these types. The conditional median or other quantile may also be used rarely; generally speaking, the conditional average of the response is thought to be a linear relationship to the values of the variables that explain the outcome (or predictors).

Support Vector Machine (SVM): It can be characterized as a supervised machine learning technique that uses related learning techniques to analyze data for regression and classification purposes. It frequently functions as a

classification analysis tool. This approach represents each data element as a point in an m-dimensional space (m being the number of features), where each point's cost is a particular coordinate. It is discovered that a hyper-plane works best for correctly classifying the two classes. It is a formalized selective classifier, defined by a separate hyper-plane. An output that classifies the data using an ideal hyper-plane is generated using supervised training data.

AdaBoost: AdaBoost, short for Adaptive Boosting, is a machine learning algorithm commonly used for classification tasks. However, it can also be adapted for regression problems, known as AdaBoost regression. In AdaBoost regression, the algorithm combines multiple weak regression models to create a strong ensemble model.

Here's a general overview of how AdaBoost regression works:

Data preparation: Prepare your dataset for regression, ensuring it contains the input features (X) and corresponding target values (y).

Initialize weights: Assign equal weights to each data point in the training set. These weights determine the importance of each sample during the training process.

Train weak regression models: AdaBoost sequentially trains a series of weak regression models on the dataset. A weak regression model is typically a decision tree with a small depth (often called a "decision stump"), but other regression models can be used as well.

Model fitting: In each iteration, the weak regression model is fitted to the training data, with the weights emphasizing the misclassified or poorly predicted samples from previous iterations.

Weight update: After the weak regression model is trained, the weights of the data points are updated. The weights of the misclassified samples are increased, while the weights of the correctly classified samples are decreased. This adjustment focuses the subsequent weak models on the previously misclassified samples.

Model combination: The weak regression models are combined to create a strong ensemble model. The combination is achieved by assigning weights to each weak model based on its performance during training. More accurate models are given higher weights.

Predictions: To make predictions, the individual weak models are used to predict the target values for a given input. The final prediction of the ensemble model is determined by aggregating the predictions of the weak models, weighted by their importance.

Iteration and final prediction: Steps 3 to 7 are repeated for a specified number of iterations or until a desired level of performance is achieved. The final prediction of the AdaBoost regression model is obtained by combining the predictions of all the weak models.

AdaBoost regression is a powerful technique that can handle complex regression problems by combining multiple weak models. However, it is important to note that AdaBoost may be prone to over fitting if the weak models are too complex or if the dataset is noisy. In such cases, it is recommended to tune the hyper parameters or consider using other regression algorithms.

3.3 Models' evaluation

We quantitatively assessed the models using five statistical indicators. Following is the calculation for these metrics.

Mean Square Error (MSE): The mean of the squared difference among the original and predicted values of the data set, as displayed in equation (6). It calculates the residuals' variance.

$$\mathrm{MSE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \hat{\mathrm{y}})^2 \ldots\ldots\ldots\ldots (6)$$

Root Mean Square Error (RMSE): The mean square error's square root is RMSE equation (7). The standard deviation of the errors that happen when a prediction is made based on a dataset is known as RMSE.

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{\text{N}}\sum_{\text{i=1}}^{\text{N}}(\text{y}_\text{i} - \hat{\text{y}})^2} \ldots\ldots\ldots (7)$$

Mean Absolute Error (MAE): The precise difference linking the dataset's actual and anticipated values is averaged out as MAE; as in equation (8). It calculates the dataset's residuals' average.

$$\text{MAE} = \frac{1}{\text{N}}\sum_{\text{i=1}}^{\text{N}} \left| \text{y}_\text{i} - \hat{\text{y}} \right| \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

The coefficient of determination or R-squared: The percentage of the dependent variable's variation that the linear regression model can explain is indicated by the coefficient of determination, also known as R-squared. It is a scale-free score, therefore depending on the values little or huge, the R square value will be below one. It is calculated using the equation (9).

$$\text{R}^2 = 1 - \frac{\sum(\text{y}_\text{i} - \hat{\text{y}})^2}{\sum(\text{y}_\text{i} - \bar{\text{y}})^2} \ldots\ldots\ldots\ldots (9)$$

Where $\hat{y}$ is forecast value of y and $\bar{y}$ is average value of y.

4. Results and Discussion

We applied NN, RF, LR, and SVM to our dataset. The model that performs best overall is achieved by LR (R2=0.96, RMSE=0.87) and SVM (R2=0.96, RMSE=0.85) with both training validation and testing data. Whereas the Neural network and Regression tree are the least performing with R2=0.17 with training data and with R2=0.43 for testing data respectively. The performance measures of each model for test and training data are tabulated in table 3 and 4 respectively. Figures4a, b, c and dshow the scatter plots of the NN, RF, LR, and SVM models prediction values respectively. From this figure it can be observed that the SVM and linear regression models predicted WQI values are comparable with actual values.The AdaBoost, ensemble model, produced similarly good results, with R2 values of 1 and 0.91 for the training and testing sets of data, respectively. Furthermore, the model's effectiveness was confirmed by the R2 value of 0.91 for the testing data. This value indicates that the AdaBoost model was able to explain approximately 91% of the variance in the WQI values of the testing dataset. Consequently, the model performed admirably well in generalizing its learnings from the training data to make accurate predictions on unseen data.In Figure 5, the box plot showcases the performance of the Adaboost model by comparing the actual WQI values with the predicted WQI values. The plot clearly demonstrates the superior performance of the model. Box and whisker plots are useful for comparing distributions across different groups or variables, identifying skewness or symmetry, and visualizing the spread of data. They are widely used in statistics and data analysis to provide a concise summary of numerical data.

Table 3: The models' performance using training data

Model Type	RMSE	MSE	R2	MAE
Regression Tree	3.34	11.18	0.52	1.98
Linear Regression	0.87	0.76	0.96	0.29
SVM	0.85	0.72	0.96	0.28
Neural Network	3.98	15.83	0.17	0.59
AdaBoost	0.13	0.02	1.00	0.02

Table 4: The models' performance using test data

Model Type	RMSE	MSE	R2	MAE
Regression Tree	2.32	5.37	0.43	1.39
Linear Regression	0.29	0.09	0.99	0.25
SVM	0.27	0.07	0.99	0.24
Neural Network	0.53	0.28	0.97	0.29
AdaBoost	0.93	0.87	0.91	0.45

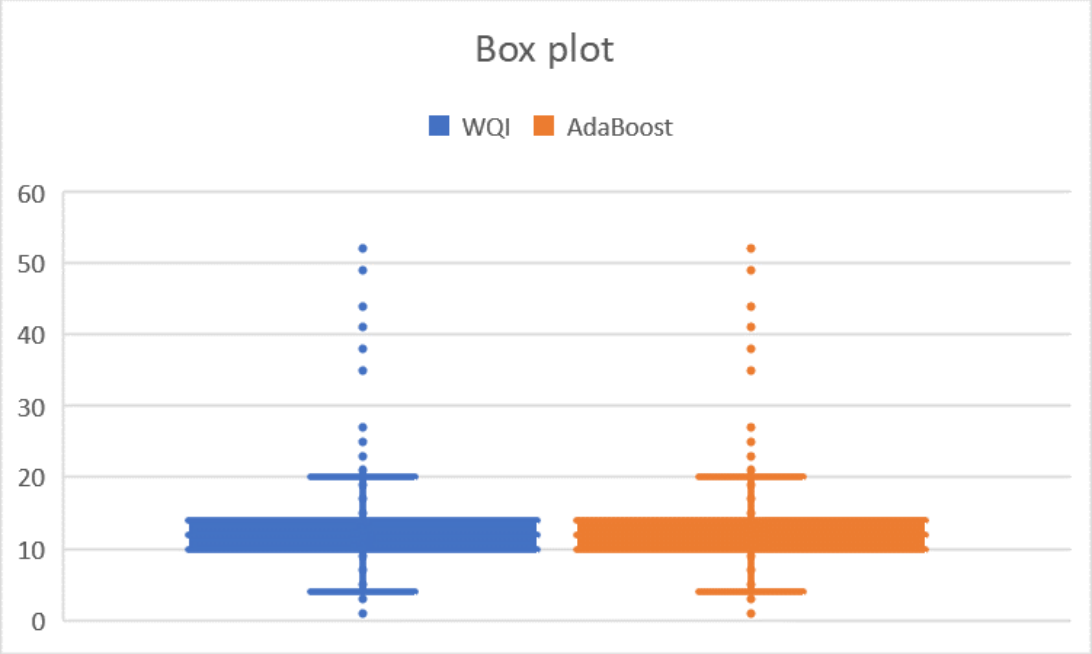

Fig.4: The box plot of actual WQI and AdaBoost WQI for training data

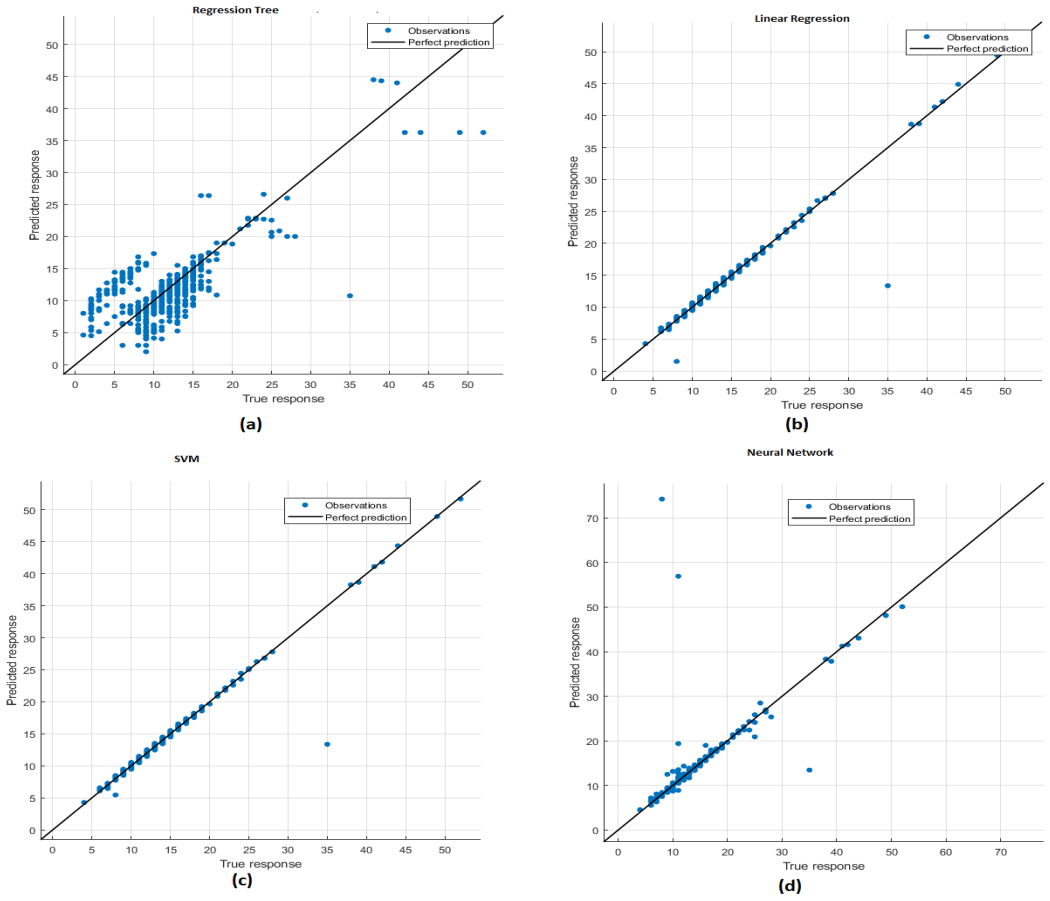

Fig. 5: Scatter plots of predicted and measured values

The outcomes demonstrated that AdaBoost, SVM and the linear regression model had achieved good results in terms of RMSE values and R2 values for both training and testing. In this study, the machine learning algorithms were tested for their ability to accurately predict monthly WQI in the Tiptur Taluk. The creation and presentation of a model based on machine learning for WQI prediction was the goal. Overall, these outcomes provide strong evidence of the AdaBoost ensemble model's capability to forecast WQI with a high degree of accuracy and reliability. The R2 values of 1 and 0.91 for the training and testing data, respectively, suggest that the model exhibits robust predictive power and can be considered a valuable tool for forecasting WQI in various scenarios.

5. Conclusions

This study evaluated how well machine learning techniques including RF, NN, LR, and SVM predicted the water quality components of an Indian water quality dataset. The most well-known dataset variables were obtained for this, including TDS, DO, EC, Nitrate, pH, and chloride. The results of the study demonstrated that the employed machine learning models were effective in forecasting water quality metrics. However, among the techniques used, SVM and AdaBoost showed the highest performance in predicting water quality components. These models exhibited better accuracy and precision compared to the others. Despite the success of the SVM and AdaBoost models, the researchers acknowledged the potential for further improvement. Therefore, they proposed conducting additional research to enhance the efficacy of the selection process. This could involve developing models that combine the suggested approach with other techniques and fuzzy neural network approaches. By integrating these additional methods, the researchers aim to create more robust and accurate models for predicting water quality components.

Overall, this study highlights the effectiveness of machine learning techniques, particularly SVM and AdaBoost, in forecasting water quality metrics. However, the researchers emphasize the need for continued research to refine and enhance the predictive models by incorporating complementary approaches.

References

[1] F. Rufino, G. Busico, E. Cuoco, T. H. Darrah, and D. Tedesco, "Evaluating the suitability of urban groundwater resources for drinking water and irrigation purposes: an integrated approach in the Agro-Aversano area of Southern Italy," *Environmental Monitoring and Assessment*, vol. 191, pp. 1-17, 2019.

[2] 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, pp. 471-483, 1998.

[3] G. Tang, J. Li, Y. 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)*, 2015, pp. 238-243.

[4] 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, pp. 2983-2989, 2018.

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

[6] J. Liu, C. Yu, Z. Hu, Y. Zhao, Y. Bai, M. Xie, *et al.*, "Accurate prediction scheme of water quality in smart mariculture with deep Bi-S-SRU learning network," *IEEE Access*, vol. 8, pp. 24784-24798, 2020.

[7] 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.

[8] L. Yan-jun and M. Qian, "AP-LSSVM modeling for water quality prediction," in *Proceedings of the 31st Chinese Control Conference*, 2012, pp. 6928-6932.

[9] A. Solanki, H. Agrawal, and K. Khare, "Predictive analysis of water quality parameters using deep learning," *International Journal of Computer Applications*, vol. 125, pp. 0975-8887, 2015.

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

[11] 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)*, 2016, pp. 1-6.

[12] 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, p. 1863, 2019.

[13] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, "Efficient water quality prediction using supervised machine learning," *Water*, vol. 11, p. 2210, 2019.

[14] L. Wang, Z. Zhu, L. Sassoubre, G. Yu, C. Liao, Q. Hu, *et al.*, "Improving the robustness of beach water quality modeling using an ensemble machine learning approach," *Science of The Total Environment*, vol. 765, p. 142760, 2021.

[15] S. Devi, "Random forest advice for water quality prediction in the regions of Kadapa district," *Int. J. Innov. Technol. Explor. Eng*, vol. 8, pp. 1-3, 2019.

[16] R. M. Brown, N. I. McClelland, R. A. Deininger, and M. F. O'Connor, "A water quality index—crashing the psychological barrier," in *Indicators of environmental quality*, ed: Springer, 1972, pp. 173-182.

[17] P. Mishra and R. Patel, "Quality of drinking water in Rourkela, Outside the steel township," *Journal of Environment and Pollution*, vol. 8, pp. 165-169, 2001.

[18] V. Kothari, S. Vij, S. Sharma, and N. Gupta, "Correlation of various water quality parameters and water quality index of districts of Uttarakhand," *Environmental and Sustainability Indicators*, vol. 9, p. 100093, 2021.