Machine Learning for Water Quality Index Forecasting

Arun Kumar Thimalapur Doddabasappaar[1], Bilegowdanamane Earappa Yogendra[1], Prashanth Janardhan[2], Prema Nisana Siddegowda[3,*]

[1] Department of Civil Engineering, Kalpataru Institute of Technology, Tiptur, India

[2] Department of Civil Engineering, National Institute of Technology of Silchar, Assam, India

[3]Department of Information Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India

Received 19 September 2023; received in revised form 11 January 2024; accepted 12 January 2024

DOI: https://doi.org/10.46604/emsi.2024.12870

Abstract

This study aims to forecast water quality in the Tumkur district, Karnataka state, India, to increase pollution levels. Various machine learning techniques, including support vector machines, regression trees, linear regression, and neural networks, are employed. The Water Quality Index (WQI) is determined using parameters such as total hardness, pH, alkalinity, turbidity, chloride, dissolved solids, and conductivity. The dataset is split into training and testing sets (80:20) to assess model performance. Support Vector Machines and Linear Regression outperform other models, achieving R2 values of 0.96 and 0.99 for training and testing, respectively. This research underscores the importance of advanced machine learning techniques for accurate water quality prediction, crucial for effective pollution reduction strategies in the region.

Keywords: water quality index, machine learning, random forest, support vector machine

1. Introduction

The availability of safe fresh water for agricultural, human use, and aquatic ecosystems, is all significantly impacted by the decline of water quality. Developing nations regularly undergo times of fast economic expansion, and every development project has the potential to have negative environmental repercussions. The pressure on the natural fertility of soils rises as a rapidly rising population and wealth, frequently leading to over-extraction of nutrients and requiring the use of artificial fertilizers. Extra fertilizer frequently finds its way into groundwater and waterways. Rivers constantly carry contaminants to lakes and oceans, harming ecosystems and endangering human health. Therefore, monitoring and evaluating water quality are essential for efficient, sustainable water management as well as for maintaining both human and environmental health.

The unitless index WQI is derived by selecting specific water quality parameters. These measures offer a categorical assessment of the historical and current water quality of bodies of water. Examples of common variables include Ca2+, Mg2+, NO-3, and other elements frequently utilized to forecast the WQIs, such as dissolved oxygen, pH, temperature, and total suspended solids [1] The WQI is highly beneficial in guiding the decisions and actions of decision-makers. However, the calculation of WQI is not straightforward because sub-indices are computed within WQI equations.

Because WQIs typically include distinct equations, computing WQI has the drawbacks of being time-consuming, tedious, challenging, and inconsistent. There is no single WQI methodology, as this discussion may have made clear. The current study endeavors to apply soft computing approaches to predict the water quality of a supply system in this context. The goal of the work is to propose a reliable method for accurately predicting water quality using machine learning technology. Fig. 1 illustrates a diagrammatic representation of the current investigation.

* Corresponding author. E-mail address: prema.gowda@gmail.com

English language proofreader: Wan-Ting Hsu

The goal of the study is to examine water quality data to gain a deeper understanding of a specific body of water's condition and characteristics. The biological, chemical, and physical properties of water constitute its quality, determining its suitability for uses such as drinking, recreation, and ecosystem health.

To analyze water quality data, samples must be collected from various points within the water body, such as rivers, lakes, or groundwater sources, and multiple tests and measurements must be conducted. Aspects including temperature, pH level, turbidity, dissolved oxygen, conductivity, nutrient concentrations (such as nitrates and phosphates), heavy metal content, and the presence of pollutants or contaminants can all be assessed by these tests. Researchers can evaluate the overall health and ecological status of the water body by examining data on water quality. Making informed decisions about managing water resources, protecting the environment, and promoting public health requires utilizing this information. It can assist in identifying potential sources of contamination, understanding the effects of human activity on water systems, and monitoring the effectiveness of water treatment and conservation efforts.

To identify trends, patterns, and correlations between various variables, researchers frequently use statistical approaches and modeling techniques in the study of water quality data. This enables researchers and decision-makers to assess the current water quality, anticipate future changes, and develop effective plans for preservation or repair. Analyzing water quality is also crucial for assessing compliance with governmental regulations and rules. It provides a framework for evaluating whether water bodies meet the requirements for specific purposes, such as drinking water sources or recreational places. Analyzing water quality data can also help in the early identification of potential health issues and the implementation of effective mitigation strategies.

In general, the analysis of water quality data aims to provide a thorough comprehensive understanding of the condition of water resources, assisting with well-informed management decisions and procedures to assure the availability of clean and safe water for both people and ecosystems. Data on water quality are gathered for this purpose from numerous sources by the researchers. After gathering the data, they preprocess it to remove any noise or outliers that could compromise the accuracy of the study. The researchers calculate WQI values for each sample after preprocessing the data. WQI is a numerical measurement of water quality that considers several factors, including pH, dissolved oxygen, and turbidity. After calculating the WQI values, the researchers divide the obtained dataset into a training set and a testing set. The researchers develop and fine-tune regression models to interpret the data on water quality using the training set. They verify the performance of the models using the testing set. After separating the training and test sets from the dataset, the researchers apply regression models to examine water quality. The performance of each model is assessed based on how effectively it forecasts the WQI values. The outcomes of the various regression models are then compared to determine which one performs the best. Researchers can use this information to determine the variables that have the most effects on water quality and create strategies to enhance it.

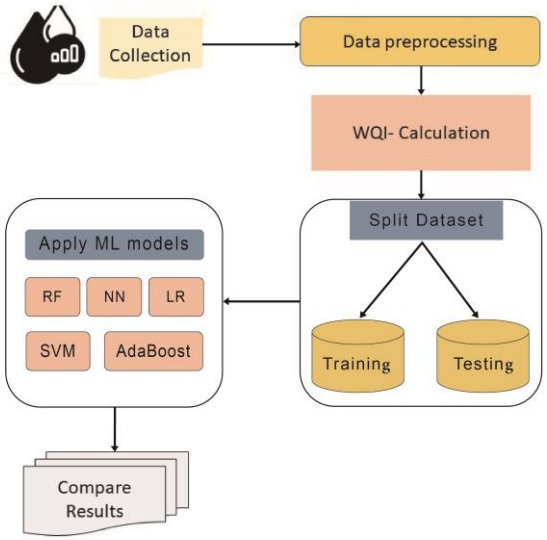


Fig. 1 Proposed WQI prediction flow diagram.

2. Literature Survey

Deterministic models utilize variables derived from experience, investigation, historical data, or empirical knowledge to statistically describe various chemical and physical processes. Differential equations are often simplified to attain appropriate solutions for the model. Practical experience is often necessary before obtaining ideal results. Assumptions and simplifications may need to be made based on the model's performance to solve the model-related equations.

Statistical models aim to extrapolate broad principles from experimental observations through the gathering of field data. The statistical modeling and evaluation process necessitates the careful selection of analytical methodologies for analyzing and verifying hypotheses and data. These models are often highly complex and rely heavily on field data for their development. Furthermore, in constructing the link between predictor and responder variables, many statistically based models of water quality assume a normal and linear distribution. However, traditional data processing methods are no longer sufficient to solve this issue due to the possibility of multiple factors affecting water quality. This results in a complex non-linear relationship between these parameters and the variables used for forecasting water quality. Consequently, the use of statistical methods often results in low accuracy.

One limitation of conventional methodologies for calculating WQI is the use of sub-index values for numerous water quality measures, which can be time-consuming and error-prone. Some researchers have utilized machine learning (ML)--based models to circumvent these limitations, which offer several advantages over conventional approaches. Firstly, ML-based models quickly generate a WQI value without requiring the calculation of sub-index values. This is attributed to the non-physical methodology they employ, in which mathematical models are trained using water quality data to elucidate the connections between different water quality metrics and the WQI. Secondly, ML algorithms offer numerous advantages, including the capability to model intricate interactions between different water quality parameters and the WQI owing to their non-linear structures. They can also handle large datasets with data at various scales and are robust to missing data. This renders them well-suited for analyzing water quality data, which can be complex and challenging to interpret.

Human-generated municipal and industrial wastewater is now recognized as the primary cause of declining water quality in metropolitan areas [2] Machine learning has emerged as a prominent topic in surface water quality research [2-3] There are several techniques exist for analyzing and predicting the quality of surface water. There has been considerable attention on optimizing machine learning models and improving the precision of their predictions.

However, the precision and methodology of data gathering and processing determine the effectiveness of machine learning models in making predictions. The accuracy of the models can be significantly impacted by the type and quantity of data used for training and testing, as well as by the preprocessing methods employed to eliminate noise and outliers. Consequently, significant attention must be devoted to the collection and processing of data to ensure the accuracy and reliability of the models.

Data collection is a crucial first step in creating machine learning models. The outcomes of both integrated and ad hoc water quality monitoring can serve as benchmarks for water systems management. Environmental authorities often utilize conventional environmental monitoring techniques. However, conventional approaches for in-situ monitoring are constrained by practical challenges [4] Remote sensing technologies can expose the migratory and distribution features of pollutants that are challenging to detect using conventional approaches and meet the demands of real-time and extensive water quality monitoring.

Sagan et al. [5] discovered that experiment-based machine learning enabled complex optimization based on the combination of satellite data. They found that the accuracy of the deep neural network (DNN), support vector regression (SVR), and partial least squares (PLS) models was higher than that of conventional models. However, due to their lack of optical

activity or the availability of high-resolution hyperspectral data, some water quality factors, such as the concentration of pathogens, cannot be directly detected by remote sensing [6] but can be indirectly approximated using other measurable data.

Wu et al. [7] developed an attentional neural network based on a convolutional neural network (CNN) to distinguish between clean and dirty water using water image data. They conducted several comparison studies on a collection of water surface images and confirmed the performance of this attentional neural network. CNN has the advantage of taking the reflectance image as a straight input without the need for feature engineering or parameter modification. Some of the gathered data will unavoidably be incomplete, incorrect, or corrupted for technical or human reasons, resulting in a sparse matrix and subpar performance in model applications. In such cases, data cleaning—another crucial stage in applications of machine learning—becomes crucial. Data cleaning can be accomplished in a variety of ways, such as by not directly using the data set, by employing averages or medians, or by combining machine learning and matrix completion techniques to augment the raw data [8] Ma et al. [9] proposed a method for predicting biological oxygen demand (BOD) that combines deep matrix factorization (deep MF) with DNN. They used the waters of the New York Harbor as a case study to validate and assess the method's reliability. Data cleaning increases the data quality and, consequently, enhances the precision of applications using machine learning models.

The features used to train machine learning models impact the accuracy of their predictions. Redundant variables increase the model’s complexity and diminish its predictive power and accuracy. Dissolved oxygen (DO), one of the most frequently monitored aspects of surface water quality, directly reflects the health of the aquatic ecosystem and its ability to support aquatic life. The concentration of DO in the Danube River was predicted using the linear polynomial neural network (PNN) model. Temperature, pH, BOD, and phosphorus concentration were found to be the most significant factors influencing the forecast accuracy among the 17 water quality parameters [10] Among the five input features (chloride, NOx, total dissolved solids, pH, and water temperature) for predicting DO concentration in St. John's River, USA, pH, and NOx have a substantial correlation with DO and can have an impact on prediction accuracy [11] These results align with those obtained by Chen et al. [12] who demonstrated that input parameters influenced the model's capacity for prediction. Eutrophication is an issue in surface water quality prediction in addition to typical water characteristics. Ly et al. [13] found, using the adaptive neuro-fuzzy inference system (ANFIS) model, that the interaction of nutrients, organic matter, and environmental factors contributed to algal blooms.

Auto Deep Learning has been used in the field of water quality analysis as a cutting-edge method to comprehend and evaluate water quality. To examine this crucial element, both traditional and auto-deep learning models have been heavily used. The outcomes of these comparisons indicate that traditional deep learning models outperform auto deep learning models. Conventional deep learning achieves a higher accuracy rate of 1.8% in binary class water data analysis and a 1% higher accuracy rate in multiclass water data analysis. These results emphasize the efficiency and dependability of conventional deep learning techniques for water quality analysis, emphasizing the importance of selecting the appropriate methodology based on specific needs in this field [14].

The study [15] employed four machine learning classifier algorithms (support vector machines, Naïve Bayes, random forest, k-nearest neighbor, and gradient boosting) to identify the best classifier for predicting water quality classes using seven widely used WQI models and three new models developed by the authors. The results revealed that the k-nearest neighbor (KNN) and XGBoost algorithms excelled, achieving 100% and 99.9% correct classifications, respectively. In the work [16], groundwater quality was analyzed using a variety of statistical and graphical techniques using data on 10 water quality characteristics collected from 50 groundwater sources. The study specifically employed the Water Quality Index (WQI), geostatistical modeling, Hierarchical Cluster Analysis (HCA), Principal Component Analysis (PCA), and spatial autocorrelation analysis to assess the overall water quality and potential causes of variations across the region. According to the overall findings, the quality of groundwater in the eastern and southeastern parts of the district significantly declined. The findings of the multivariate analysis indicated significant variations between the groundwater in the district's wet and dry zones, suggesting higher groundwater mineralization in the dry zone.

3. Materials and Methods

The data set used for the study and methods used for forecasting are summarized in the following section.

3.1. Dataset

The size of Tiptur Taluk is around 75 km2, and it is located about 75 km from Tumkur District. The typical temperature ranges from 11 °C in the winter to 38 °C in the summer. Tiptur town receives 503 mm of rain on average each year, and its land area is 76,510 acres. Tiptur's municipal water delivery and distribution system serves as a case study for forecasting and assessing water quality. For supplying water to Tumkur and Tiptur town for drinking, industrial usage, and other domestic needs, Tiptur mostly relies on borewells and the outflow from the Bagur-Navile tunnel. According to the 2001 Census, Tiptur Town had 53,053 residents. The town has 6,340 households dispersed in an area of 11.60 km2 and 31 wards. With a population density of 4,574 people per km2, the decadal population growth rate is 47.2%. The town is mostly centered on agriculture and has an industry for coconut oil, coir, and related products. The city is expected to experience a faster population increase due to the presence of industries. The current water supply is 135 LPCD, derived from the Hemavathi canal, 6 km from Tiptur, and has a 17.5 MLD capacity water treatment plant. According to Table 1 and Fig. 2, the various water samples that were collected for analysis are designated N1 through N9 for the study. Total dissolved solids (TDS), hardness, alkalinity, electrical conductivity (EC), dissolved oxygen (DO), chloride, turbidity, and pH readings are all part of the data, which pertains to approximately 1000 samples.

Table 1 Zone information of Tiptur city

Area number	Locations
N1	Sharada Nagar
N2	Vidya Nagar
N3	Shadakshri Badavane
N4	Govinapura
N5	Gandhi Nagar
N6	Bashaveshwar Nagar
N7	DoddaPete
N8	Goragondanahalli
N9	Halepalya

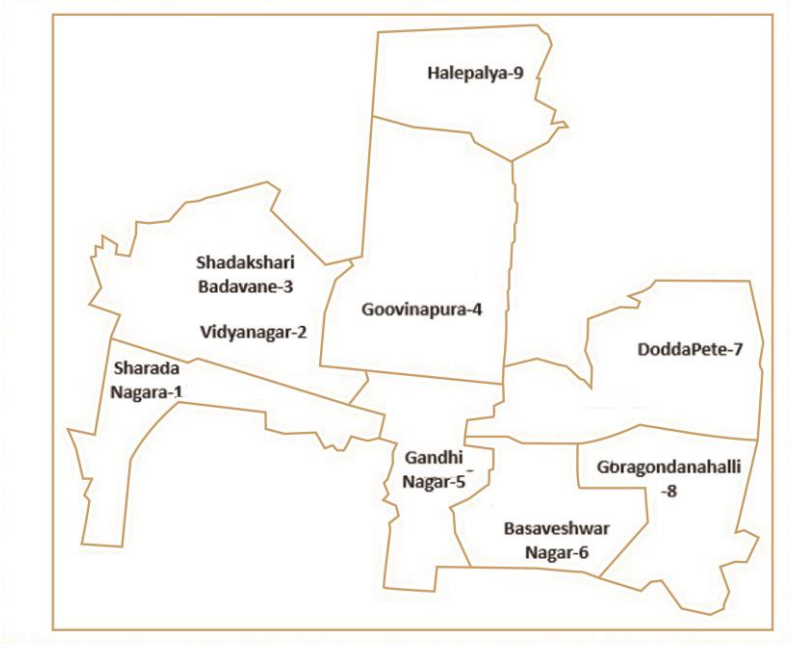


Fig. 2 Location of water sources of Tiptur

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

The WQI was used to understand and assess the water quality of each water sample. The term "WQI" refers to the relative importance and influence of various water quality metrics on water quality. The calculation of WQI utilized the Indian standard for drinking water (BIS, 1991) as shown in Table 2.

The WQI is calculated using the weighted arithmetic index method. WQI application is a practical technique for determining whether water is suitable for various beneficial purposes.

The following steps were taken to calculate WQI using the weighted arithmetic index method [17]:

Assuming there are "n" different WQ parameters. 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. Values for Q_n can be provided by:

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

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

Except for key parameters like *pH* and dissolved oxygen, V_i equals 0 in most circumstances. The *pH* & *DO* quality rating ($V_i \neq 0$) can be calculated by:

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

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

The recommended requirements for each of the different water quality metrics are oppositely correlated to the unit weight Wn for those parameters, which can be obtained by:

$$W_n = \frac{k}{S_n} \tag{4}$$

Where W_n = unit weight for nth parameter, S_n= standard acceptable value for the nth parameter, k = proportionality constant.

WQI can be derived from:

$$WQI = \frac{\sum_{i=1}^{n} q_n W_n}{\sum_{i=1}^{n} W_n} \tag{5}$$

As graded by Mishra and Patel [18], Fig. 3 lists outlines whether WQI levels are suitable for human consumption.

WQI		Classification
0-25	→	Excellent
26-50	→	Good
51-75	→	Bad
76-100	→	Very Bad
100 & above	→	Unfit

Fig. 3 Classification of water quality

The attribute correlation matrix of the traits in this dataset with pre-processing is shown in Fig. 4. This graph demonstrates that the EC and pH are positively correlated with WQI, while the other parameters are negatively correlated.

Fig. 4 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 investigation 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 randomizes 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 modeled 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 is 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.

3.3. Models' evaluation

The models are quantitatively assessed using three statistical indicators. Following is the calculation for these metrics.

(1) Mean Square Error (MSE): The mean of the squared difference among the original and predicted values of the data set, which can be acquired by calculating the residuals' variance, is as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{6}$$

(2) Root Mean Square Error (RMSE): The mean square error's square root is RMSE Eq. (7). The standard deviation of the errors that happen when a prediction is made based on a dataset is known as RMSE, which can be obtained by:

$$RMSE = \sqrt{MSE} \tag{7}$$

(3) Mean Absolute Error (MAE): The precise difference linking the dataset's actual and anticipated values is averaged out as MAE, which calculates the dataset's residuals' average, is as follows.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|\ X_i - \bar{X}\ |^2 \tag{8}$$

4. Results and Discussion

The regression models applied to the dataset are NN, RF, LR, and SVM. LR (R2=0.9602, RMSE=0.87) and SVM (R2=0.9625, RMSE=0.85) are the top-performing models overall, achieving high performance on both training and testing data. However, the Neural Network and Regression Tree exhibit lower performance with R2=0.1652 for training data and R2=0.4280 for testing data, respectively. The performance measures of each model for test and training data are tabulated in Tables 3 and 4, respectively. Figs. 5(a)-5(d) show the scatter plots of the RF, LR, SVM, and NN model's prediction values, respectively. From these figures, it can be observed that the SVM and linear regression models predicted WQI values are comparable with actual values. The time taken by the models for training and prediction is tabulated in Table 5, the SVM model takes the least time for training and prediction.

Table 3 The models' performance using training data

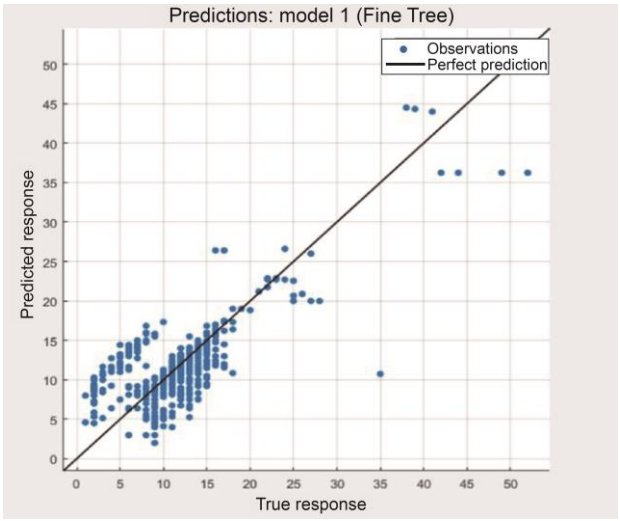
Model Type	RMSE	MSE	R^2	MAE
Regression Tree	3.34	11.18	0.5193	1.98
Linear Regression	0.87	0.76	0.9602	0.29
SVM	0.85	0.72	0.9625	0.28
Neural Network	3.98	15.83	0.1652	0.59

Table 4 The models' performance using test data

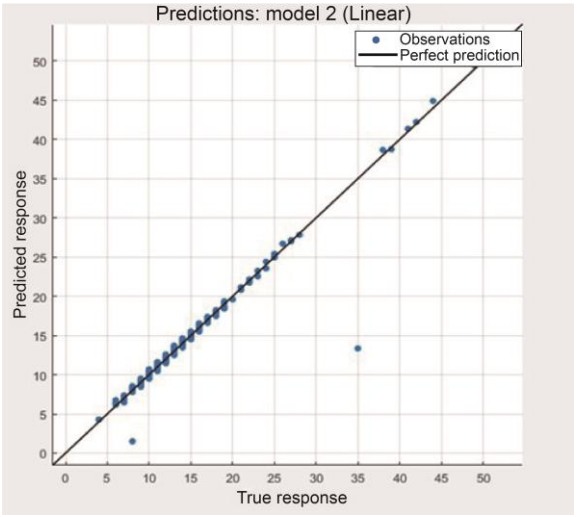
Model Type	RMSE	MSE	R^2	MAE
Regression Tree	2.32	5.37	0.4280	1.39
Linear Regression	0.29	0.09	0.9909	0.25
SVM	0.27	0.07	0.9920	0.24
Neural Network	0.53	0.28	0.9695	0.29

Table 5 The time taken by the models

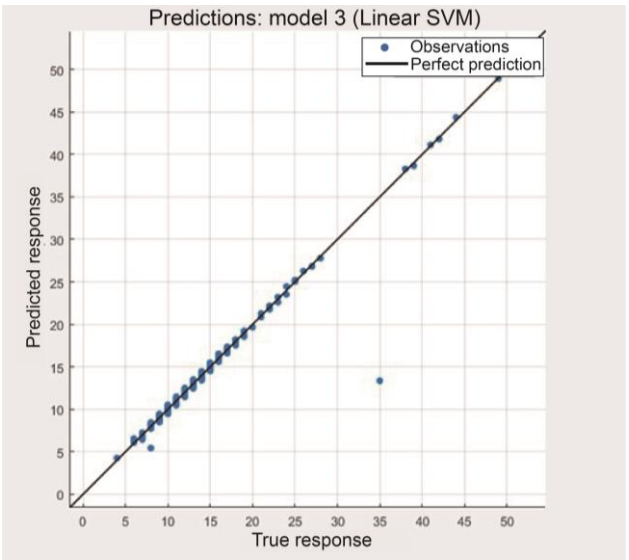
Model Type	Training time in second	Prediction Speed in observation /second
Regression Tree	8.51	3900
Linear Regression	8.75	7200
SVM	4.31	16000
Neural Network	14.46	3600



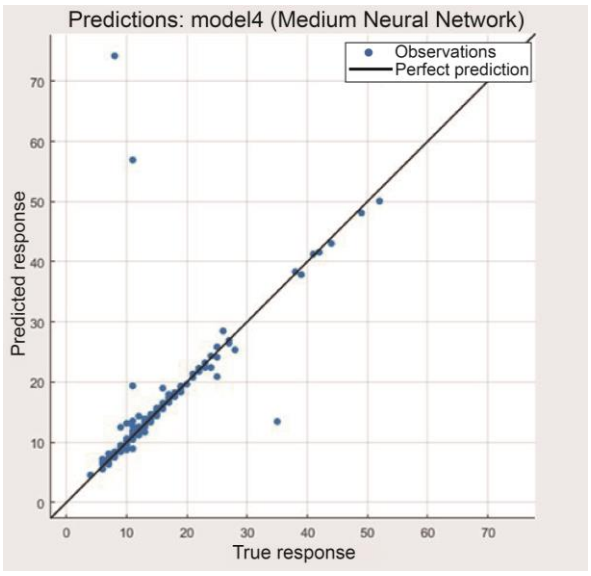
(a) Regression Tree



(b) Linear Regression



(c) SVM



(d) Neural network

Fig. 5 Scatter plots of predicted and measured values

The results showed that the linear regression model, SVM, produced good results in terms of RMSE values and R2 values for both training and testing. This study aimed to determine whether machine learning models could accurately predict monthly WQI in Tiptur Taluk. The objective was to develop and propose a machine learning-based model for WQI prediction. Overall, these results offer compelling proof that the model can predict WQI with a high level of accuracy and dependability. The model demonstrates strong predictive potential and can be considered an effective tool for WQI forecasting in various scenarios, supported by R2 values of 0.999 and 0.907 for the training and testing datasets, respectively.

5. Conclusion

This study assessed the effectiveness of machine learning methods such as RF, NN, LR, and SVM, in predicting an Indian dataset on various components of water quality. For this purpose, the most well-known dataset variables, including TDS, DO, EC, Nitrate, pH, and chloride, were collected. The findings of the study demonstrated that the machine learning models used were successful in predicting key indicators of water quality. However, among the employed techniques, SVM exhibited the best performance in predicting the elements of water quality. These models demonstrated greater accuracy and precision compared to the others. Despite the success of SVM, the researchers identified areas for improvement. Therefore, they recommended conducting further research to enhance the effectiveness of the selection procedure. As a result, they suggested undertaking more research to improve the effectiveness of the selection procedure. This may involve creating models that combine the recommended strategy with different methods and fuzzy neural network strategies. The researchers aim to develop models for forecasting water quality components that are more reliable and precise by incorporating these additional methodologies.

Conflicts of Interest

The authors say they have no conflicts of interest.

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, no. 12, pp. 1-17, November 2019.

[2] R. Mohammadpour, S. Shaharuddin, C. K. Chang, N. A. Zakaria, A. A. Ghani, and N. W. Chan, “Prediction of Water Quality Index in Constructed Wetlands Using Support Vector Machine,” Environmental Science and Pollution Research, vol. 22, no. 2, pp. 6208-6219, November 2014.

[3] T. Tiyasha, T. M. Tung, and Z. M. Yaseen, “A Survey on River Water Quality Modelling Using Artificial Intelligence Models: 2000–2020,” Journal of Hydrology, vol. 585, article no. 124670, June 2020.

[4] N. Sharma, R. Sharma, and N. Jindal, “Machine Learning and Deep Learning Applications-A Vision,” Global Transitions Proceedings, vol. 2, no. 1, pp. 24-28, January 2021.

[5] W. Li, H. Fang, G. Qin, X. Tan, Z. Huang, F. Zeng, et al., “Concentration Estimation of Dissolved Oxygen in Pearl River Basin Using Input Variable Selection and Machine Learning Techniques,” Science of The Total Environment, vol. 731, article no. 139099, August 2020.

[6] V. Sagan, K. T. Peterson, M. Maimaitijiang, P. Sidike, J. Sloan, B. A. Greeling, et al., “Monitoring Inland Water Quality Using Remote Sensing: Potential and Limitations of Spectral Indices, Bio-Optical Simulations, Machine Learning, and Cloud Computing,” Earth-Science Reviews, vol. 205, article no. 103187, June 2020.

[7] Y. Wu, X. Zhang, Y. Xiao, and J. Feng, “Attention Neural Network for Water Image Classification Under IoT Environment,” Applied Sciences, vol. 10, article no. 909, January 2020.

[8] A. R. T. Donders, G. J. M. G. Van Der Heijden, T. Stijnen, and K. G. Moons, “Review: A Gentle Introduction To Imputation of Missing Values,” Journal of Clinical Epidemiology, vol. 59, no. 10, pp. 1087-1091, October 2006.

[9] J. Ma, Y. Ding, J. C. P. Cheng, F. Jiang, and Z. Xu, “Soft Detection of 5-day BOD with Sparse Matrix in City Harbor Water Using Deep Learning Techniques,” Water Research, vol. 170, article no. 115350, March 2020.

[10] A. Š. Tomić, D. Antanasijević, M. Ristić, A. Perić-Grujić, and V. Pocajt, "A Linear and Non-Linear Polynomial Neural Network Modeling of Dissolved Oxygen Content in Surface Water: Inter-and Extrapolation Performance with Inputs' Significance Analysis," Science of The Total Environment, vol. 610, pp. 1038-1046, January 2018.

[11] M. Zounemat-Kermani, Y. Seo, S. Kim, M. A. Ghorbani, S. Samadianfard, S. Naghshara, et al., "Can Decomposition Approaches Always Enhance Soft Computing Models? Predicting The Dissolved Oxygen Concentration in the St. Johns River, Florida," Applied Sciences, vol. 9, article no. 2534, June 2019.

[12] K. Chen, H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, et al., "Comparative Analysis of Surface Water Quality Prediction Performance and Identification of Key Water Parameters Using Different Machine Learning Models Based on Big Data," Water Research, vol. 171, article no. 115454, March 2020.

[13] Q. V. Ly, X. C. Nguyen, N. C. Lê, T. D. Truong, T. H. T. Hoang, and T. J. Park, et al., "Application of Machine Learning for Eutrophication Analysis and Algal Bloom Prediction in An Urban River: A 10-year Study of The Han River, South Korea," Science of The Total Environment, vol. 797, article no. 149040, November 2021.

[14] D. V. V. Prasad, L. Y. Venkataramana, P. S. Kumar, G. Prasannamedha, S. Harshana, and S. J. Srividya, et al., "Analysis And Prediction Of Water Quality Using Deep Learning and Auto Deep Learning Techniques," Science of The Total Environment, vol. 821, article no. 153311, May 2022.

[15] M. G. Uddin, S. Nash, A. Rahman, and A. I. Olbert, "Performance Analysis of The Water Quality Index Model for Predicting Water State Using Machine Learning Techniques," Process Safety and Environmental Protection, vol. 169, pp. 808-828, January 2023.

[16] R. C. Karangoda and K. G. N. Nanayakkara, "Use of The Water Quality Index and Multivariate Analysis to Assess Groundwater Quality for Drinking Purpose in Ratnapura District, Sri Lanka," Groundwater for Sustainable Development, vol. 21, article no. 100910, May 2023.

[17] R. M. Brown, N. I. McClelland, R. A. Deininger, and M. F. O'Connor, A Water Quality Index-Crashing The Psychological Barrier, 1st ed., Boston: Springer, pp. 173-182, 1972.

[18] P. C. Mishra and R. Patel, "Quality of Drinking Water in Rourkela, Outside The Steel Township," Journal of Environment and Pollution, vol. 8, pp. 165-169, January 2001.

Copyright© by the authors. Licensee TAETI, Taiwan. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC) license (https://creativecommons.org/licenses/by-nc/4.0/).