Civil Engineering Beyond Limits 2 (2025) 1999

Civil Engineering Beyond Limits

journal home: www.acapublishing.com/journals/1/cebel

Research Article

Machine Learning Approaches for Water Quality Index Prediction in Paba Upazila, Rajshahi: A Comparative Study of Minitab and Advanced Algorithms

Md. Abu Sayed*,[1], Arabi Suraia Rose[2], Md.Maruf Hasan[3], Mst.Mariam Khatun[4], Harun Ar Rashid[5]

[1,3,5]Lecturer, Pundra University of Science and Technology,Bogura, Bangladesh.
[2]Research Scholar, Pundra University of Science and Technology,Bogura, Bangladesh.
[4]Research Scholar, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh.

Keywords

Water quality index, Machine learning, Minitab, Predictive modeling, Paba upazila, Rajshahi, Water quality assessment.

Abstract

Predicting the Water Quality Index (WQI), which provides communities and policymakers a measurable indicator of water quality, is crucial for efficient environmental management. The purpose of this study is to investigate numerical models for predicting the WQI values using Minitab (regression analysis) and machine learning algorithms namely Decision Tree (DTR), Random Forest (RGR), Stochastic Gradient Decent (SGD), and Support Vector Machine (SVR). This is accomplished by collecting surface and ground water from 200 locations in the Paba Upazila, Rajshahi and doing laboratory tests to determine the pH, turbidity, total dissolved solids and total solids to create an extensive dataset that reflects the water conditions in the area. The WQI is then computed using the parameters from the Brown et al. (1972) technique. According to the analysis, Minitab and SVR perform better than the others, obtaining strong classification metrics (93% accuracy, 0.94 F1-score) and remarkable prediction accuracy (r2 = 0.9503 for Minitab; r2 = 0.9443 for SVR). The intricate interactions between the several water quality indices in the study area are well captured by these models. With a data-driven strategy to monitoring and forecasting water quality in Paba Upazila, the findings offer significant insights for local water resource management. The results of the evaluation can provide a scientific basis for the conservation of the local aquatic environment, and the model created in this study can be used as a guide for similar water quality assessment work. This study also highlights the potential of integrating machine learning algorithms with statistical software such as Minitab for environmental monitoring applications, and it helps design customized solutions for water quality evaluation in comparable regions of Bangladesh.

1. Introduction

Economic growth, environmental balance, and human health all depend on having access to clean and safe water. In recent decades, the quality of surface and groundwater resources has been seriously harmed by the fast rate of urbanization, industrial growth, and agricultural intensification. Effective water resources management now depends on precise water quality evaluation and ongoing monitoring. The Water Quality Index (WQI), which combines several physico-chemical factors into a single composite score to categorize the water's suitability for different purposes, is one of the most extensively used methods for assessing water quality. The Water Quality Index (WQI) developed by Brown et al. in 1970, is one of the most well-known and frequently used techniques for measuring overall water quality. By combining multiple important water quality metrics into a single numerical value, this index provides a thorough and understandable simplified depiction of the state of the water. The equation is as follows.

$$\text{WQI}=\frac{\sum_{n}^{i=1}(W_i.Q_i)}{\sum_{n}^{i=1}(W_i)} \quad (1)$$

Here,

Q_i= quality rating of the i^{th} parameter (on a scale of 0 to 100) = $\frac{V_i - V_0}{S_i - V_0} x100$

*Corresponding Author: sayedce.1807323@std.hstu.ac.bd
Received 26 Apr 2025; Revised 06 May 2025; Accepted 06 May 2025 https://doi.org/10.36937/cebel.2025.1999

W_i= relative weight (unit weight) of the i^{th} parameter;
n = total number of water quality parameters used;
V_i= Measured value of the i^{th} parameter;
V_0= Ideal value of that parameter;

S_i= Standard permissible value for the i^{th} parameter (e.g., WHO or national standards).

According to their relative significance to water quality, the Brown et al. WQI model assigns particular weights to various physico-chemical parameters, including pH, dissolved oxygen (DO), turbidity, total dissolved solids (TDS), and biochemical oxygen demand (BOD). To create a single index score, these weighted values are then combined using a defined procedure.

Table 1. Standard Values for Drinking Water Parameters (Bangladesh & WHO)

Parameter	Symbol	Standard Permissible Limit (mg/L or unit)	Ideal Value
pH	pH	6.5 – 8.5	7
Dissolved Oxygen	DO	≥ 5.0 mg/L	14.6 mg/L
Biochemical Oxygen Demand	BOD	≤ 3.0 mg/L	0
Total Dissolved Solids	TDS	≤ 1000 mg/L	0
Turbidity	-	≤ 5 NTU	0
Total Suspended Solids	TSS	≤ 10 mg/L (surface water standard)	0
Electrical Conductivity	EC	≤ 400 µS/cm (suggested for drinking)	0
Color	-	≤ 15 TCU	0

Table 1 lists the standard acceptable values for clean drinking water parameters in Bangladesh that are in accordance with WHO criteria [1][2][3]. These values are frequently cited in national policies such Bangladesh Standards and Testing Institution (BSTI) standards. Based on widely accepted interpretations in Bangladesh and in accordance with international standards (such as WHO and DOE-BD guidelines), Table 2 shows the Water Quality Index (WQI) ranges and corresponding water quality classes [1][4][5].

Table 2. WQI Ranges and Water Quality Classification

WQI Range	Water Quality Status	Remarks / Suitability
0 – 25	Excellent	Water is of very high quality, safe for drinking
26 – 50	Good	Suitable for drinking with minor treatment
51 – 75	Moderate / Poor	Needs conventional treatment before use
76 – 100	Very Poor	Not suitable for drinking; use for irrigation
> 100	Unsuitable for use	Highly polluted; not safe for any direct use

Even with its wide range of applications, the conventional WQI calculation method can be laborious and might not adequately account for the intricate, nonlinear relationships between water quality variables. Because Conventional approaches to WQI computation frequently relies on manual weighting of parameters, which is subjective and labor-intensive (time-consuming, inherently subjective and limits adaptability to new regions etc.). In this study, author address this limitation by employing machine learning models to automate the WQI prediction process, thereby eliminating the need for manual weighting and improving both efficiency and objectivity. Data-driven methods, especially machine learning (ML), have become strong substitutes for these constraints in order to model intricate correlations between water quality measures and make more accurate and efficient predictions of WQI. ML algorithms are appropriate for dynamic environmental monitoring jobs because they can generalize predictions for unseen samples, learn from prior data, and reveal hidden patterns [6][7].

Machine learning techniques for forecasting the Water Quality Class (WQC) and WQI have been investigated in recent research. Numerous algorithms [8][9][10][11] have demonstrated encouraging outcomes in WQI prediction and WQC classification, such as Random Forest, Multi-Layer Perceptron, and Gradient Boosting. Using a mix of 19 water quality factors and nearby land use activities, Ejaz et al. (2024) [8] used sophisticated machine learning algorithms to forecast the water quality index of an industrially polluted creek in Pakistan. With just seven water quality factors, the Gradient Boost (GB) model performed the best in predicting the polluted Aik-Stream's WQI. The application of supervised machine learning algorithms to effectively forecast WQC and WQI using just four input parameters—temperature, turbidity, pH, and total dissolved solids—was investigated by Ahmed et al. in 2019 [9]. The most effective algorithms for predicting the WQI are gradient boosting, which has a learning rate of 0.1, and polynomial regression, which has a degree of 2. Their respective mean absolute errors (MAEs) are 1.9642 and 2.7273. In contrast, the multi-layer perceptron (MLP) classifies the WQC most effectively with an accuracy of 0.8507 when configured in the (3, 7) configuration. The gradient boosting classification model and the multilayer perceptron regression model performed the best among the machine learning algorithms that Jha et al. (2024) [10] offered to forecast the water quality index and water quality features. The MLP regressor model outperformed other regression models in the Nair et al. (2022) [11] study, predicting the water quality index with the lowest root mean squared error (RMSE) of 2.432. Using a variety of solo and hybrid machine learning methods, Bui et al. (2020) [12] sought to enhance the prediction of water quality indices. They discovered that the hybrid BA-RT approach performed better than the other models. Sanaa [13] evaluated machine learning algorithms on drinking water quality for better sustainability at 2022. The study evaluates the efficiency of using machine learning (ML) techniques in order to predict the quality of water. The results show that SVM and k-nearest neighbor are better according to F1-score and ROC AUC values. However, The LASSO LARS and SGD are better based on recall values.

As mentioned earlier, physico-chemical parameters, including pH, dissolved oxygen (DO), turbidity, total dissolved solids (TDS), and biochemical oxygen demand (BOD), are given particular weights according to their relative significance to water quality in the Brown et al. WQI model. A single index score is then created by integrating these weighted values using a defined formula. The conventional WQI computing method can be laborious and cannot adequately reflect the intricate, nonlinear relationships between water quality factors, despite its wide range of applications.

This study uses machine learning models and Minitab-based statistical analysis (regression analysis) to estimate the WQI as described by Brown et al. in order to address these challenges. Data on water quality, including pH, TDS, turbidity, total solids, and color, were taken from Paba Upazila, Rajshahi, Bangladesh. Comparing the predictive capabilities of different machine learning algorithms, validating the results with Minitab, and proving that it is feasible to combine contemporary data analytics with conventional frameworks for evaluating water quality are the goals. The findings of this study can help environmental authorities make informed judgments on the management of water resources in a timely manner.

2. Methodology

2.1. Study area and Data Collection

Paba Upazila, a subdistrict in the Rajshahi District of northwest Bangladesh, is where the current study is carried out. Geographically, Paba occupies an area of roughly 280.42 km² and is located between latitudes 24°19′N and 24°30′N and longitudes 88°29′E and 88°42′E. The area is a part of the Barind Tract, due to which the area is experiencing a significant decline in groundwater levels, which is distinguished by its primarily agricultural terrain, moderate to low rainfall, and reddish-brown clayey to silty soils. Owing to its close proximity to Rajshahi City Corporation and reliance on surface and groundwater resources, Paba Upazila faces mounting strain on its water resources as a result of urbanization, irrigation needs, and population increase. Figure 1. represents the map of study area. In the area, water is used for a variety of things, including drinking, small-scale industrial, irrigation, and household consumption. However, concerns over the deterioration of water quality in recent years have been raised by inappropriate waste disposal, excessive fertilizer use, and a lack of infrastructure for water treatment. In order to guarantee sustainable management of water resources and the preservation of public health, it is crucial to evaluate and forecast the WQI in this region.

Figure 1. Map of the study area (Paba Upazila, Rajshahi)

A total of 200 water samples are collected from various locations throughout Paba Upazila, including surface water (ponds, canals) and groundwater sources (tube wells, deep wells). In order to provide spatial representation in rural, residential, and peri-urban areas, the sampling locations are selected. Figure 2. represents some sampling operations on the study area.

2.2. Working Procedure

This study combines laboratory test, software analysis and machine learning models. Water samples were collected from Paba Upazila in plastic bottles of drinking products and freezed so that the properties of the materials remain same as the time of sample collection. During the sampling operation, location, source of water, time and date were labeled at the bottle. Then collected sample were tested in the laboratory to find PH, color, Turbidity, Total solids (TS), Total suspended solids (TSS) and Total dissolved solids (TDS). Table 3 represents the standard references obtained for every test and Table 4 represents sample of data of the study area.

Figure 2. Sampling operations

Table 3. Standard references for test

Test	References	Remarks
Location	Google Map	-
Sampling Process	ASTM D3370	Standard Practices for Sampling Water from Closed Conduits
	ASTM D5905	Standard Practice for Sampling Wastewater
	ASTM D1066	Practice for Sampling Industrial Wastewate
Temparature	ASTM D5469	Standard Test Method for Water Temperature Measurement
pH	ASTM D1293	Standard Test Methods for pH of Water
TDS (mg/L)	ASTM D5907	Standard Test Methods for Filterable Matter (TDS) and Nonfilterable Residue (TSS) in Water
TS (mg/L)	ASTM D5907	Standard Test Methods for Filterable Matter (TDS) and Nonfilterable Residue (TSS) in Water
Turbidity	ASTM D1889	Standard Test Method for Turbidity of Water

Subsequently, WQI is calculated by using the equation 1. The obtained dataset is then analyzed by Minitab software and ML models i.e. Decision Tree (DTR); Random Forest (RFR); Stochastic Gradient Decent (SGD); Support Vector Machine (SVM). For machine learning models, 80% data is taken for training and 20% data for testing so that maximum data can be found for training the model and validate it respectively. Data normalization is done by standardscaler and nill/null values are also being processed before implementation of model (all data were found to be applicable). Performance indicators i.e. r^2; Mean squared error (MSE); Root mean square error (RMSE) etc. are then calculated. Comparing with the ideal values of the regression performance indicators the model acceptance decision is taken and best model is so selected. Figure 3. represents the working process for this study. Based on the predicted WQI value, WQC is so measured (comparing the value with the range mentioned at Table 2.) and performance indicators for classification i.e. accuracy; precision; recall; F1_score etc. are calculated for all developed models. Then based on the two results best model is so selected and numerical model for each study is found.

Table 4. Sample dataset

Sample No	Address	Longitude	Latitude	Temperature	pH	TDS (mg/L)	TS (mg/L)	Turbidity	WQI
1	Notun Emadpur	88.68016	24.37383	29	6.4	89.458	115.25	11.13	127.7353
2	Mashkatadighi purbopara	88.67079	24.368166	29	7.5	78.587	1.0122	7.8	102.7657
3	Mashkatadighi moddhopara	88.66647	24.36829	30	8.2	74.25	102.63	2.18	53.8766
4	Dauanpara mor mashjid	88.666621	24.364963	30	7.3	989.458	1.15023	5.21	78.2072
5	Mashkatadighi purbopara	88.67322	24.3673	30	7.9	90.78	110.23	1.13	42.9962
6	Katakhali pourosova vaban	88.685897	24.36408	30	6.5	187.587	450.6	10.89	127.9344
7	Belgharia zam-a mashjid	88.40262	24.21078	29	6.1	184.36	201.36	4.13	63.4669
8	Coumahini bazar	88.695611	24.344669	30	6.2	185.94	120.24	9.45	112.6841
9	Daunpara dokkhin para	88.666317	24.361805	30	6.4	54.412	65.21	1.13	36.6607
10	Shyampur guaal para	88.665978	24.351282	30	6	144.52	163.56	8.53	103.4924

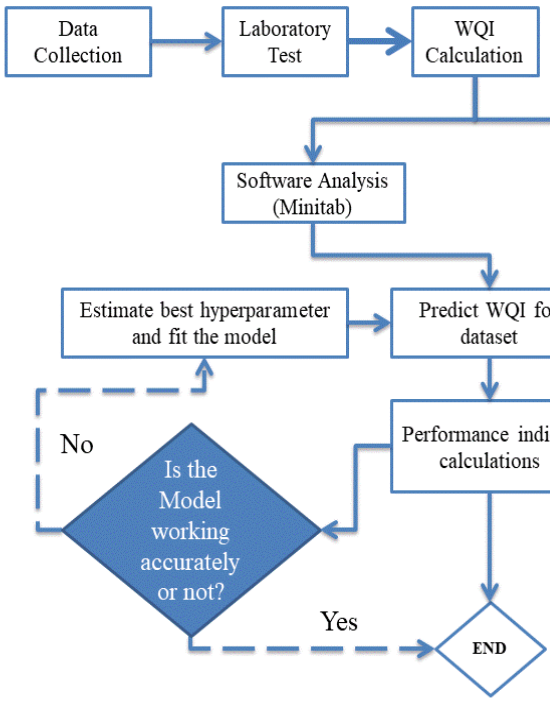

Figure 3. Workflow

2.3. Minitab Software and ML models

Minitab is a robust statistical program that is frequently used for predictive modeling, quality enhancement, and data analysis. Minitab 22 was used in this work to visualize the water quality metrics gathered from Paba Upazila, perform statistical validation, and analyze preliminary data. To assess the link between important physico-chemical characteristics, the program was utilized to build regression models and correlation matrices. In this study, author take advantages of regression analysis by MiniTab software, where WQI is so predicted using this statistical software and using the Table 2. The WQC is so classified.

Again Promising solutions are provided by machine learning (ML) models including DT, RF, SGD, and SVM, which use previous water quality data to precisely predict WQI. Table 5. represents hyperparameters found by grid search cross validation techniques. Complex patterns and connections between physicochemical parameters, including as pH, dissolved oxygen, turbidity, and heavy metal concentrations, can be found using these models. SGD provides scalability for huge datasets, SVM is excellent at handling high-dimensional data, and DT and RF deliver predictions that are robust and interpretable. In order to improve water quality monitoring and management, this study compares the performance of several ML models in order to determine which one is best for WQI prediction.

Table 5. Hyperparametes for Machine Learning Models

Model Name	Parameters (GridSearch Cv)
DTR	min_samples_leaf=2, min_samples_split=5, random_state=42)
RFR	n_estimators=50, random_state=42
SGD	alpha=0.1, penalty='elasticnet', random_state=42
SVR	C=10, epsilon=0.5, kernel='linear'

3. Results and Discussions

The assessment of machine learning models for forecasting the Water Quality Index showed clear differences in the evaluated algorithms' performance. Figure 4. represents the regression plot for actual data and predicted data by minitab software and ML models. Minitab demonstrated the highest predictive performance (r^2 = 0.9503) with the lowest error metrics (MSE = 46.8926, RMSE = 6.8478) in capturing the intricate correlations between water quality measures and WQI, as indicated in Table 6. With r^2 values of 0.9493 and 0.9443, respectively, the SGD and SVR models likewise showed strong performance, albeit with somewhat greater error rates. While DTR followed behind with an r^2 of 0.9341 and the largest errors of any model, RFR demonstrated competitive performance (r^2= 0.9428). The numerical modeling for predicted models is tabulated at Table 7.

According to these results, ensemble-based models generally outperform single decision trees in WQI prediction. While Minitab provides reliable statistical tools for model validation and analysis, its performance is largely attributed to the quality of data preparation, model selection, and proper use of cross-validation. Advanced machine learning platforms may be more suitable for capturing complex non-linear patterns and further reducing overfitting.

Table 6. Performance metrics for regression analysis

Model Name	r^2 value	MSE	RMSE
DTR	0.9341	82.4509	9.0802
MiniTab	0.9503	46.8926	6.8478
SVR	0.9443	60.5476	7.7812
SGD	0.9493	69.1387	8.3149
RFR	0.9428	71.3483	8.4467

Table 7. Numerical model developed by each model

Model	Numerical Model		
MiniTab	WQI (MiniTab)	=	-17.9 + 6.36 pH + 0.00550 TDS (mg/L) + 0.0247 TS (mg/L) + 8.885 Turbidity
DTR	WQI (DTR)	=	30.0 + 0.68 pH + 0.00939 TDS (mg/L) + 0.0275 TS (mg/L) + 8.083 Turbidity
RFR	WQI (RFR)	=	12.8 + 2.98 pH + 0.01432 TDS (mg/L) + 0.0263 TS (mg/L) + 7.820 Turbidity
SGD	WQI (SGD)	=	-6.862 + 5.668 pH + 0.007673 TDS (mg/L) + 0.03263 TS (mg/L) + 8.071 Turbidity
SVR	WQI (SVR)	=	0.9498 + 4.337 pH + 0.001053 TDS (mg/L) + 0.003744 TS (mg/L) + 9.329 Turbidity

Again the predicted values are classified according to Table 2. Then based on WQC actual and predicted data, the performance metrics are demonstrated at Table 8. The confusion matrix is presented at Figure 5. Based on Table 5, the most reliable models are MiniTab and Support Vector Regression (SVR), both of which achieve 93% accuracy and continuously high precision (0.95), recall (0.93-0.94), and F1-score (0.93-0.94) scores. Their greater capacity to correctly classify water quality groups and forecast WQI values is demonstrated by these data. While Decision Tree Regression (DTR) and Stochastic Gradient Descent (SGD) perform comparatively poorly, with 77% and 73% accuracy, respectively, Random Forest Regression (RFR) performs pretty well with 80% accuracy.

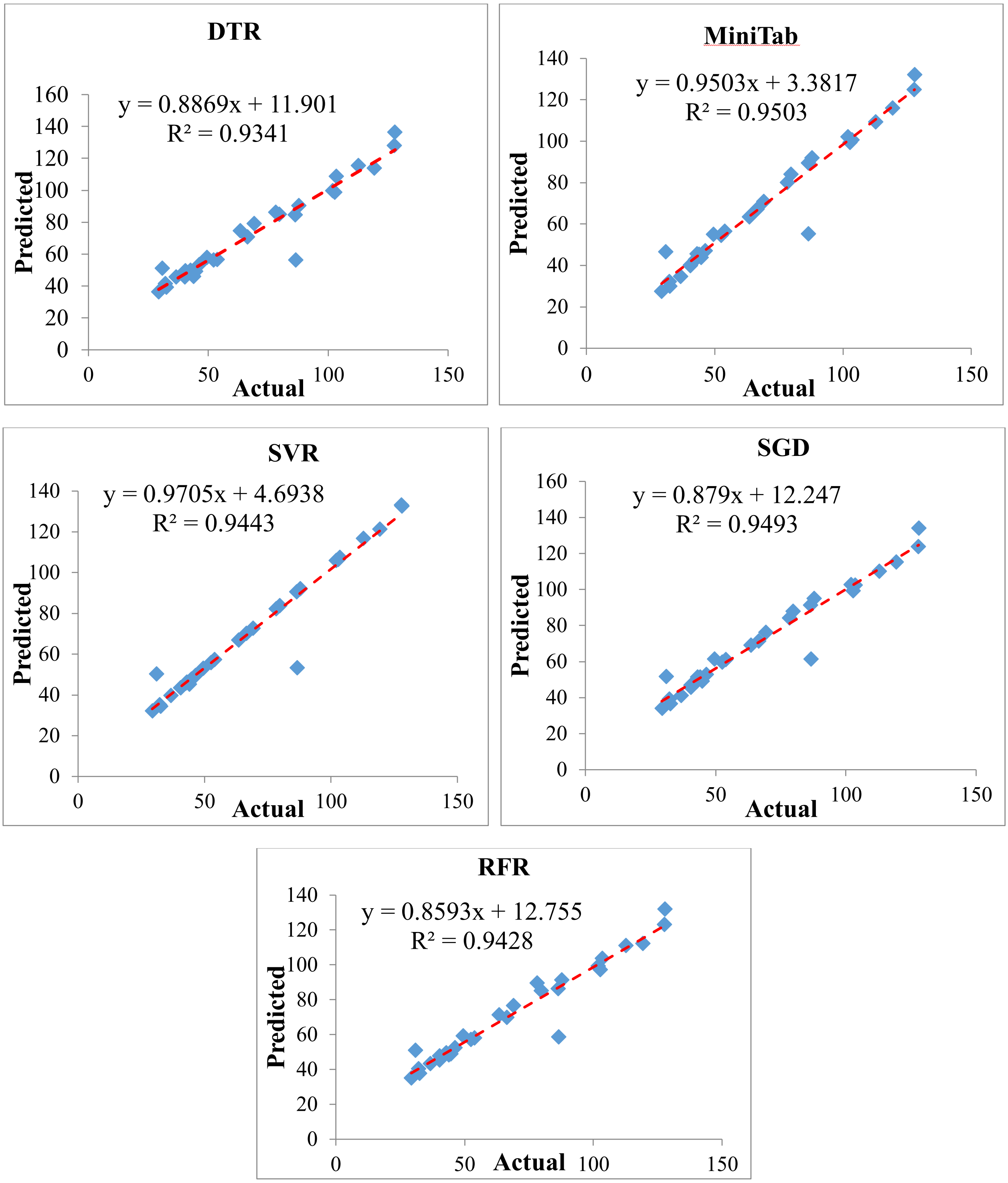

Figure 4. Regression Plot

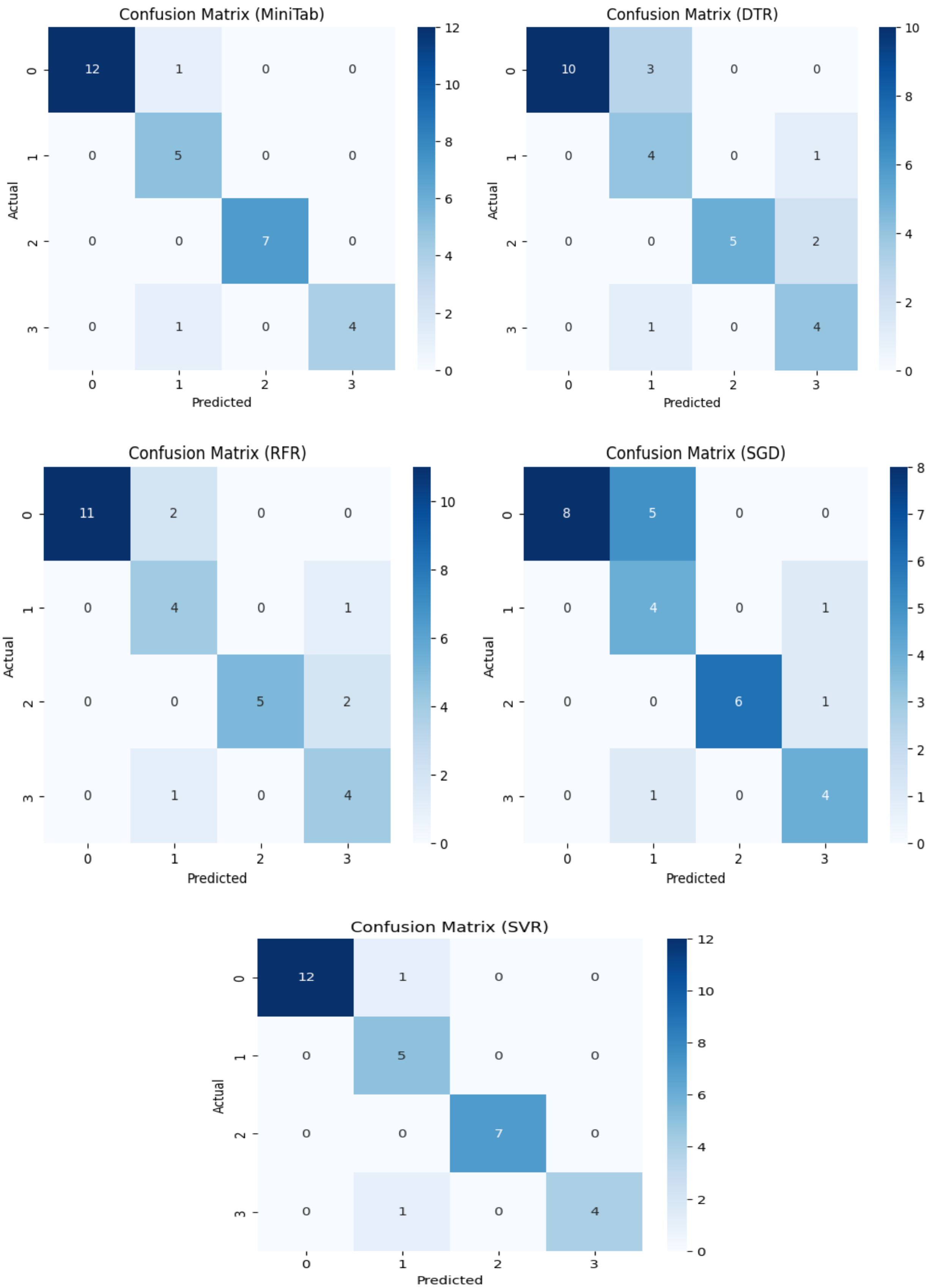

Figure 5. Confusion Matrix for each Model

These categorization metrics' performance hierarchy supports and validates the previous regression study results, which also indicated MiniTab to have the best predictive capacity. MiniTab's persistent superiority across a variety of evaluation techniques indicates that it is especially well-suited for tasks involving the assessment of water quality. This is probably because it can efficiently describe intricate, non-linear relationships in water quality data while retaining the capacity for generalization. The efficacy of kernel-based approaches in this application space is further supported by SVR's high performance. These results have significant ramifications for the management of water resources, as prompt decision-making and environmental preservation depend on precise and trustworthy WQI prediction. MiniTab and SVR are especially interesting options for integration into operational water quality monitoring systems because to their proven performance advantages.

Table 8. Performance Metrics based on WQC

Performance Parameters	Minitab	DTR	RFR	SGD	SVR
Accuracy	0.93	0.77	0.8	0.73	0.93
Precision	0.95	0.85	0.86	0.84	0.95
recall	0.93	0.77	0.8	0.73	0.94
f1-score	0.94	0.78	0.81	0.76	0.93

4. Conclusions

The performance of many machine learning models, including DTR, RFR, SGD, SVR, and MiniTab, for forecasting the WQI using regression and classification metrics was thoroughly assessed in this work. MiniTab and SVR consistently outperformed other models, as seen by their superior regression performance (r^2 = 0.9503, lowest MSE and RMSE), as well as their highest classification accuracy (93%), precision (0.95), recall (0.93–0.94), and F1-scores (0.93–0.94). Their ability to handle intricate, non-linear correlations in water quality data makes them especially dependable for predicting WQI. Because of its sensitivity to data noise and difficulties in capturing complex feature interactions, DTR and SGD performed less well than RFR, which showed intermediate performance. Strong correlations between regression and classification results support the findings' validity and imply that ensemble-based and kernel-based approaches are more appropriate for estimating water quality than simpler algorithms.

These results offer water resource managers a data-driven method for effectively evaluating water quality, which has important practical ramifications for environmental monitoring. The dataset used in this study are limited in size or geographical coverage, restricting the generalizability of the findings. Future research could investigate deep learning (Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs)) or hybrid models (e.g., CNN-LSTM or Transformer-based models) to improve forecast accuracy even more. Real-time implementation in intelligent water quality monitoring systems could also be investigated. Minitab and SVR demonstrated the strongest performance in WQI prediction, according to this study, which also provides a trustworthy instrument for managing water resources sustainably.

Declaration of Conflict of Interests

The authors declare that there is no conflict of interest. They have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1.] World Health Organization (WHO). Guidelines for Drinking-Water Quality, 4th Edition incorporating the 1st Addendum, 2017. Geneva: WHO.

[2.] Bangladesh Standards and Testing Institution (BSTI). Drinking Water Standard (BDS 1240:2013).

[3.] Department of Environment, Bangladesh (DoE). Environmental Conservation Rules, 1997.

[4.] Islam, M. S., Ahmed, M. K., & Habibullah-Al-Mamun, M. Water quality assessment along the Buriganga River using WQI and multivariate analysis. Science of the Total Environment (2015).

[5.] Brown, R. M. et al. A Water Quality Index – Do We Dare? Water & Sewage Works (1970).

[6.] Sayed, A., & Rahman, M. M. M5 soft computing techniques for assessment of soil liquefaction. Journal of Rehabilitation in Civil Engineering, 13(3) (2025) 199-214 https://doi.org/10.22075/jrce.2025.34669.2134

[7.] Rahman, M. M., Hossain, M., Sayed, A., & Thakur, S. Assessment of liquefaction potential based on the logistic regression machine learning algorithm. In Proceedings of the 7th International Conference on Civil Engineering for Sustainable Development (2024) https://doi.org/10.13140/RG.2.2.24047.65440

[8.] Ejaz, U., Khan, S. M., Jehangir, S., Ahmad, Z., Abdullah, A., Iqbal, M., Khalid, N., Nazir, A., & Svenning, J.-C. Monitoring the industrial waste polluted stream: Integrated analytics and machine learning for water quality index assessment. Journal of Cleaner Production, 450 (2024) 141877 https://doi.org/10.1016/j.jclepro.2024.141877

[9.] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. Efficient water quality prediction using supervised machine learning. Water, 11(11) (2019) 2210 https://doi.org/10.3390/w11112210

[10.] Jha, N. N., Singh, R. K., Sharma, S., & Kumar, A. Computational machine learning analytics for prediction of water quality. Computational and Numerical Analysis, 31(4s) (2024) https://doi.org/10.52783/cana.v31.942

[11.] Nair, J. P., & Vijaya, M. S. River water quality prediction and index classification using machine learning. Journal of Physics: Conference Series, 2325 (2022) 012011 https://doi.org/10.1088/1742-6596/2325/1/012011

[12.] Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of The Total Environment, 721 (2020) 137612. https://doi.org/10.1016/j.scitotenv.2020.137612

[13.] Kaddoura, S. Evaluation of machine learning algorithm on drinking water quality for better sustainability. Sustainability, 14(18) (2022) 11478. https://doi.org/10.3390/su141811478

How to Cite This Article

Sayed, M.A., Rose, A.S., Hasan, M.M., Khatun, M.M., Rashid, H.A., Machine Learning Approaches for Water Quality Index Prediction in Paba Upazila, Rajshahi: A Comparative Study of Minitab and Advanced Algorithms, Civil Engineering Beyond Limits, 2(2025), 1999. https://doi.org/10.36937/cebel.2025.1999