Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Water quality prediction and classification based on principal component regression and gradient boosting classifier approach

Md. Saikat Islam Khan [a,d], Nazrul Islam [b,d,*], Jia Uddin [c], Sifatul Islam [a,d], Mostofa Kamal Nasir [a,d]

[a] Department of Computer Science and Engineering, Santosh, Tangail-1902, Bangladesh
[b] Department of Information and Communication and Technology, Santosh, Tangail-1902, Bangladesh
[c] Department of Technology Studies, Endicott College, Woosong University, Daejeon, South Korea
[d] Mawlana Bhashani Science and Technology University, Santosh, Tangail 1902, Bangladesh

ARTICLE INFO

Article history:
Received 29 January 2021
Revised 11 May 2021
Accepted 3 June 2021
Available online xxxx

Keywords:
Water quality index
Principal component regression
Classification algorithm
Boxplot analysis

ABSTRACT

Estimating water quality has been one of the significant challenges faced by the world in recent decades. This paper presents a water quality prediction model utilizing the principal component regression technique. Firstly, the water quality index (WQI) is calculated using the weighted arithmetic index method. Secondly, the principal component analysis (PCA) is applied to the dataset, and the most dominant WQI parameters have been extracted. Thirdly, to predict the WQI, different regression algorithms are used to the PCA output. Finally, the Gradient Boosting Classifier is utilized to classify the water quality status. The proposed system is experimentally evaluated on a Gulshan Lake-related dataset. The results demonstrate 95% prediction accuracy for the principal component regression method and 100% classification accuracy for the Gradient Boosting Classifier method, which show credible performance compared with the state-of-art models.

1. Introduction

There are no living species on earth who can survive without water. The flow of water from rivers and lakes contributes explicitly or implicitly to both human well-being and the fisheries industry (Kar, 2019; Kar, 2013). However, water is frequently polluted because the industry has been growing every year on the back of spiralling demand, and hazardous waste is discharged into the rivers and lakes by those industries. Millions of people died every year, untold losses of income, agricultural land deteriorated due to water pollution (Dzwairo et al., 2006). In recent years, several studies have shown that the quality of groundwater has declined significantly in most countries (Adimalla, 2019; Gaikwad et al., 2020; Moon et al., 2004). In Bangladesh, many urban areas are affected by water contamination due to unplanned urbanization and industrialization. In 2015, urban pollution caused Bangladesh to lose $6.5 billion, which was 3.4% of GDP, according to the world bank report (World Bank Report, 2018), where Dhaka lost $1.44 billion, which was the 0.72% of the GDP.

Therefore, surveillance of water quality is mandatory. Though water quality can be tested using traditional techniques such as collecting the water specimens manually and then analyzed it in a laboratory (Wu and Liu, 2012). But it can be considered time-consuming and expensive. Sensors can also be regarded as another conventional approach. However, using sensors is considered costly to test all the water quality variables and often show low precision (Oelen et al., 2018). Another solution for monitoring water quality is predictive modelling using machine learning and deep learning approaches. Compared to other conventional methods, it has several advantages: lower costs, efficient in terms of time required for travel and collection, enables prediction under various phases of a system, and predicts desirable values when accessing a site is inconvenient (Sinshaw et al., 2019).

The researchers had extensively used prediction models for their studies in the water quality management system over the last few years including, the artificial neural network (Sinshaw et al., 2019; Barzegar and Moghaddam, 2016; Barzegar et al., 2020;

* Corresponding author at: Department of Information and Communication and Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail 1902, Bangladesh.
E-mail addresses: bappy.10.cse.mbstu@gmail.com (M.S. Islam Khan), nazrul.islam@ieee.org (N. Islam), jia.uddin@wsu.ac.kr (J. Uddin), sifat.acc@gmail.com (S. Islam), kamal@mbstu.ac.bd (M.K. Nasir).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2021.06.003
1319-1578/

Hameed et al., 2017; Kadam et al., 2019; Zhang et al., 2019), multiple linear regression (Choubin, 2016; Ewaid et al., 2018), least square method (Salari et al., 2018), decision tree (Saghebian et al., 2014), random forest method (Yajima and Derot, 2018), wavelet neural network approach (Xu and Liu, 2013), recurrent neural network approach (Li et al., 2019), neuro-fuzzy techniques (Aghel et al., 2019; Khadr, 2017; Kisi et al., 2019; Zhu et al., 2019), and support vector machine method (Leong et al., 2019; Mohammadpour, 2015). In recent decades, those methods have had a tremendous impact on the aquatic climate. But most of the models, including artificial neural network, wavelet neural network, recurrent neural network, and decision tree, required lots of input parameters and computational power, which are considered expensive to construct such models.

With this motivation, this paper used water quality index (WQI), a combination of different water quality metrics that demonstrates the water quality condition of a particular region, and is applied both prediction and classification models to predict WQI and classify the water quality status. Principal component regression (PCR) is used to predict WQI in this analysis, combining both supervised and unsupervised techniques. PCR's basic concept is that the principal component analysis is applied to the dataset to minimize the dimension. At the same time, a regression algorithm is used for the PCA output. Since it can solve dataset multicollinearity issues and allows fewer water quality specimens, PCR can predict the WQI more efficiently than the other techniques. The Gradient Boosting Classifier method is used in this study for the classification task. It is an ensemble technique that can operate with a small amount of data.

Besides, the paper's main contributions are as follows:

- This study's water quality index is determined using the form of a "weighted arithmetic index" method.
- For the first time, the principal component regression method is applied to predict the water quality index.
- The water quality parameters are reduced using the PCR approach, which allows using any water quality specimens.
- For classifying the water quality status, a classification model is presented.
- A boxplot analysis is also carried out on the dataset to determine the most dominant WQI parameter.

This paper has been formulated as follows: authors have discussed the literature review in Section 2. Then, Section 3 introduces the architecture of the proposed models with a proper explanation of the subsection. Additionally, the authors have discussed the experimental setup and the result analysis in Section 4. Finally, the authors have concluded the paper with limitations and future scope in Section 5.

2. Related works

This section demonstrated the existing literature survey. The author took the most common approaches to detect and classify the water quality, including deep neural network, recurrent neural network, neuro-fuzzy inference, and support vector regression.

For example, Barzegar et al. (2020), applied a CNN-LSTM amalgam model to predict two water quality variables, named Dissolved Oxygen (DO) and chlorophyll-a. Results indicated that the CNN-LSTM amalgam model outperformed both the individual CNN and LSTM model and the machine learning models such as SVR, Decision Tree. Oladipo et al. (2021), compared two statistical methods, including Fuzzy Logic Inference (FLI) and WQI methods, for evaluating the water quality in the Ikare community, Nigeria. They found moderate and poor water quality conditions using FLI and WQI methods, respectively. They also found that the FLI method is superior to the WQI method because of the relationship between measured values and WQI standard values. For the estimation of dissolved oxygen in aquaculture, Li et al. (2018), suggested a synthetic model by combining Sparse-autoencoder and long short-term memory networks (LSTM). Although both CNN-LSTM and Sparse-autoencoder-LSTM models showed excellent performance since they predicted only DO and chlorophyll, it may be challenging to deal with more water quality variables using such models. In another research, Asadollah et al. (2021), applied an ensemble machine learning method called Extra Tree Regression (ETR) which combines multiple week learners such as decision tree to predict WQI values in Tsuen River, Hong Kong. They applied the ETR method on ten water quality variables. Results indicated that the ETR method achieved 98% prediction accuracy, which outperformed the other state-of-the-art models such as support vector regression and decision tree. Further, Hameed et al. (2017), developed two neural artificial network techniques: a radial basis function neural network (RBFNN) and a backpropagation neural network (BNN) to predict the WQI in the tropical region of Malaysia. The WQI was measured using sub-indices equations in this study (Agamuthu and Victor, 2011). In both RBFNN and BNN strategies, the training is faster, but the prediction takes a long time, making the model slow. Bui et al. (2020), proposed a hybrid machine learning algorithm by combining the random tree and bagging (BA-RT) technique. The BA-RT method achieved 94% prediction accuracy using a 10–fold cross-validation technique, outdoing 15 standalone and hybrid algorithms. A more comprehensive study into the application of machine learning methods for modeling river water quality was performed by Rajaee et al. (2020), where they reviewed a total of 51 articles published from 2000 to 2016. According to this study, artificial neural networks and wavelet-neural networks were the most widely used methods for predicting water quality. Furthermore, Samsudin et al. (2019), developed an artificial neural network. For this study, the most significant water quality parameters were found through a spatially discriminant analysis (SDA). But these studies can barely show 71% accuracy. In another research, Yilma et al. (2018), applied an artificial neural network for predicting WQI in Ethiopia's Akaki River. In this analysis, an artificial neural network with eight hidden layers and 15 hidden neurons predict WQI with more than 90% accuracy. Also, Imani et al. (2021), applied an artificial neural network with a single hidden layer for predicting water quality resilience in São Paulo, Brazil. Applying neural networks to predict WQI required lots of water quality data, which is expensive and time-consuming. Ho et al. (2019), applied a decision tree for classifying water quality status in Klang River, Malaysia. They considered three scenarios where they used six water quality variables in the first scenario. After that, in each procedure, they removed water quality parameters such as NH3-N, pH, and SS to evaluate the decision tree algorithm's ability in different situations. They achieved 84.09%, 81.82%, and 77.27% classification accuracy in each scenario, which is higher than the 75% classification accuracy benchmark. Besides, to predict the WQI, Ahmed et al. (2019), used several supervised machine learning methods. They conducted their model on four water quality parameters. They found that by using gradient boosting and polynomial regression, the WQI is more successfully predicted where a multilayer perceptron classifies the water quality category more effectively. However, this study worked with fewer water quality parameters, but both proposed prediction and classification models did not show more than 75% accuracy. On the other hand, Wang et al. (2017), applied support vector regression to predict WQI. More than 90% of accuracy was achieved in this analysis. In this study, 22 specimens of water quality were used, which makes the model computationally costly. Li et al. (2019), proposed an amalgam model for the study of time-

series water quality data by integrating a recurrent neural network with the Dempster-Shafer Theory (DST), where the RNN is capable of analyzing time-series data effectively to predict WQI and DST, which is a probability method used to amalgamate the outcome of RNNs. It can be challenging to predict WQI using RNN and DST since specialized handling of the data is required when fitting and testing the model. Besides, Ahmed et al. (2019), proposed a neuro-fuzzy inference method based on a wavelet-de-noise technique to predict water quality parameters. Results indicated that this model outperformed the other neural network model, such as RBF and MLP. But the neuro-fuzzy inference method causes a curse of dimensionality problem, which occurs when high dimensional data is analyzed and classified.

In summary, from the above studies, most of the current approaches were based on a predictive model but did not provide any classification model. Most of the models showed less accuracy and used many water quality specimens. The proposed method is applied to address the limitations described in the current approaches above. Also, the proposed model gives a dynamic approach to use any number of water quality specimens.

3. Proposed method

The authors have presented a brief discussion on the proposed architectures for predicting and classifying WQI in this section, as showed in Fig. 1. A novel algorithm for the principal component regression method is also presented in this section. The subsections below presents descriptions of blocks related to the proposed models.

3.1. Region analysis and dataset

Gulshan Lake is situated in the northernmost part of Dhaka in Bangladesh. It is considered an essential part of Dhaka city since it is one of Dhaka city's remaining water bodies. It is surrounded by the Diplomatic Zones of Baridhara, Tejgaon Thana, and Shahjadpur. In such regions, Gulshan lake is considered a significant source of surface water recharge. Gulshan Lake is 3.8 km long, with a total surface area of about 100 hectares. It is positioned at latitudes 23° 48 North and 90° 25 East with a mean depth of 2.5 m and a density of 12–105 m^3 (Rahman and Hossain, 2019). Fig. 2 indicates the lake's location in Bangladesh. In this analysis, the water quality variables were measured every month in 2016. WQI has been determined utilizing indicators such as pH, suspended solids (SS), electrical conductivity (EC), total dissolved solids (TDS), turbidity, dissolved oxygen (DO), alkalinity, chloride, and demand for chemical oxygen (COD). In this analysis, the dataset consisted of 108 specimens. The collected samples were obtained from the Environment Department (DOE) and the Environment and Forest Ministry, Bangladesh (Dataset, 2016). Table 1 provides a detailed statistical overview of the dataset.

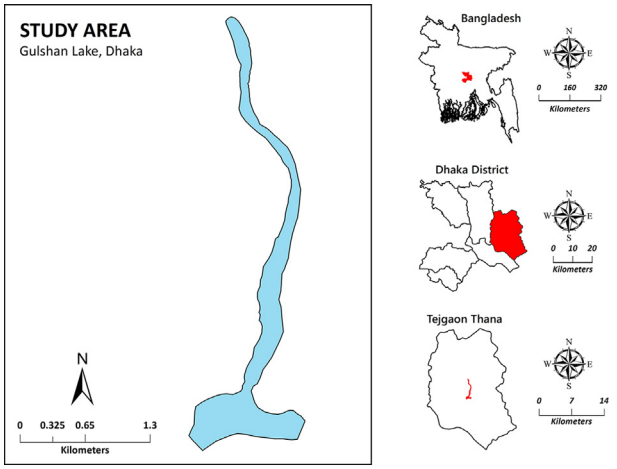

Fig. 2. Lake position in Dhaka city.

3.2. Computation of WQI

This section illustrated the way the WQI is measured. A total of nine parameters were used to estimate the WQI, including pH, DO, COD, TDS, EC, Turbidity, Chloride, SS, and Alkalinity. The water is then classified into a separate class, based on the estimated WQI [as presented in Table 2].

Here, WQI is computed using the "weighted arithmetic index method" (Tyagi et al., 2013), which was first proposed by Horton (1965). According to this technique, water quality rating (Q_j) is an integral part of the WQI and is determined using the following expression:

$$Q_j = ((M_j - l_j)/(S_j - l_j)) \times 100 \quad (1)$$

where Q_j is considered to be the quality rating of the jth water quality measurement, M_j is referred to as the measured value from the Gulshan lake, S_j is regarded as the standard value of the water coefficient recommended by WHO (World Health Organization et al., 2004), and l_j is referred to as the ideal value for water quality parameters. The ideal value considered for pH and DO, are 7 and 14.6 mg/l, respectively, and for other water quality measurements, it is equal to zero. After calculating Q_j, the unit weight W_j is estimated using the following expression:

$$W_j = 1/S_j \quad (2)$$

where, W_j is referred to as the relative unit weight, S_j is regarded as the standard value of the jth parameters, and 1 is defined as the constant of proportion. The unit weight factor (W_j) and the quality rating (Q_j) are combined to form WQI, which is expressed with the following expression:

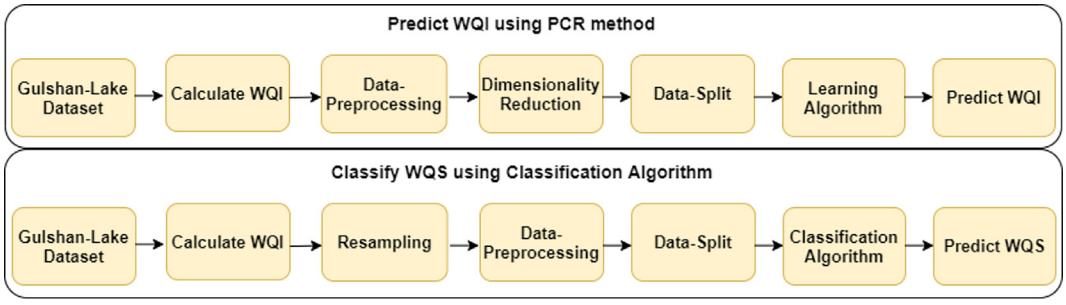

Fig. 1. Prediction and Classification of water quality index.

Table 1
Statistical description of water quality specimens.

Parameter	Count	Min	Max	Mean	SD	Weight	Relative weight
pH	108	6.85	8.27	7.35	.242	2.54	.1204
DO	108	0	15.3	5.20	3.76	4.09	.1938
COD	108	24	303	89.87	44.55	2	.0947
TDS	108	122.5	432	274.72	56.11	2.75	.1303
Turbidity	108	16	82	51.23	19.89	2.4	.1137
Chloride	108	15.8	76.9	33.43	11.03	1	.0473
SS	108	24	142	56.81	24.48	1.5	.0711
Alkalinity	108	104	210	150.68	19.82	1.6	.0758
EC	108	250	869	557.97	108.93	3.22	.1526
WQI	108	63	144	109.52	20.86		

Table 2
WQI categories and applications (Tyagi et al., 2013).

WQI Range	Water Quality Status	Applications
0–25	Excellent	Drinking, industrial and irrigation
25–50	Good	Drinking, industrial and irrigation
50–75	Medium	Industrial and irrigation
75–100	Bad	Irrigation
Above 100	Unsuitable for drinking	Treatment require before use

$$WQI = \sum W_j Q_j / \sum W_j \tag{3}$$

3.3. WQI prediction model

This section demonstrated the development of the principal component regression model to predict the water quality index. PCR comes with the idea that performed PCA on the dataset and then performed the regression model on the new PCs.

- **Data-Preprocessing:** The dataset included some null values. For handling such null values, the median method is used in this analysis. Furthermore, Min–Max scalar is used to scale the data, which makes the computation easier.
- **Dimensionality-Reduction:** For dimensionality reduction, Principal Component Analysis (PCA) is applied to the dataset, which extracts the most dominant water quality parameters. PCA is a statistical analysis that reduced a dataset's dimensionality that is influenced by the multi-correlated variables. Since PCA takes all the inter-correlated variables, it transformed them into a small number of non-correlated variables that described all the variances. The uncorrelated variables obtained from the PCA are known as principal components (PCs).
- **Data-Split:** The collected data are then divided into two sets after conducting PCA on the dataset: training and testing set with a proportion of 80 and 20 percent.
- **Learning -Algorithm:** For predicting the WQI, machine learning algorithms can be used. In this analysis, several machine learning algorithms are used, including Linear Regression, Gradient Boosting regression, Random Forest Regression, and Support Vector Regression. Such algorithms are independently paired with the PCA method and implemented on the dataset. Finally, the best prediction model is selected by comparing the performances of such models. Table 3 presents the experimental parameters used for those models.

Table 3
Experimental parameters for regression model.

Model	Parameters	Values
Muliple Linear Regression	fit_intercept	True
Support Vector Regression	kernel, C, degree	poly, 200, 3
Gradient Boosting Regression	learning_rate, n_estimators	.1, 60
Random Forest Regression	n_estimators	10

The proposed PCR method for predicting the WQI is given as Algorithm. 1.

Algorithm 1. PCR algorithm

1: **Input:** A^f is the prediction matrix of the models in f.
2: **Output:** α_j, predicting WQI
3: $B = cov(A^f)$
▷ Where, cov() is a covariance matrix
4: n = select_component(PCs)
5: $f^\beta = \beta_1 PCs_1 + \beta_2 PCs_2 + \ldots + \beta_n PCs_n$
▷ Where, $\beta = (PCs_n^T PCs_n^{-1}) PCs_n^T z$
6: $\alpha_j = \sum_{n=1}^{N} \beta_n \gamma_{n \cdot j}$
▷ Where, $\gamma_{n \cdot j}$ is the jth coefficient of the nth principal component.
7: **return** α

where the select_component() is used to select the number of components, however, it allows to use any number of components, but it must be less than or equal to the number of features used initially.

The best way to grasp the PCR model is to look at it mathematically. We can derive our regression equation using the following formula:

$$y = \beta X + \epsilon \tag{4}$$

From simple linear regression, multiple linear regression can be derived using the following expression:

$$y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \tag{5}$$

Here y is denoted as the expected value of the predictor variables. $\beta_0, \beta_1, \beta_2 \ldots \beta_k$ is the regression coefficient function associated with the independent $X_0, X_1, X_2, \ldots X_k$ variables, respectively. Then β is used to determine the interaction between the predictor variables and the independent variable. And finally, ϵ is denoted as vector of random errors. In PCR our first step is to normalize the data before executing PCA, as PCA is reactive to the data. Therefore, we used one of the pre-processing steps to normalize the data in such a way that $\sigma = 1$ and $\mu = 0$. We conduct PCA on matrix X in the next step which is our independent variables. We can obtain $X = PDV'$ by performing PCA on the independent variable X by singular value decomposition. Here, D is a diagonal matrix consisting of q non-negative singular values and q is denoted as the explanatory variables.

$$D = diagonal[\Delta_1, \Delta_2, \ldots \Delta_q] = \begin{bmatrix} \Delta_1 & 0 & \cdots & 0 \\ 0 & \Delta_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Delta_q \end{bmatrix} \tag{6}$$

That's how q × q matrix can be obtained with V as eigenvectors. The matrix $K = XV$ can be obtained in the next step using eigenvector V, where each column is represented as a major part of q. Such columns are orthogonal in the matrix and do not guarantee collinearity between the variables. We can cut the number of principal components to reduce the collinearity between the variables. In matrix V, we must cut the components to r. After cutting the components, the matrix looks like the $K_{z\times r} = X_{z\times q}V_{z\times r}$; where z is denoted as the number of observations. We then perform regression on the K matrix consisting of principal components q or r in the next step. The coefficient can be determined using the following formula after conducting regression on the principal component.

$$\beta k = (K'K)^{-1})K'y \tag{7}$$

From the Eq. (7) the formula for the calculation is:

$$y = K\beta k = X(V\beta k) = X\beta x \tag{8}$$

Finally, we can derive the following expression:

$$\beta x = V\beta k \tag{9}$$

3.4. WQS classification

This section demonstrated the development of a classification model to classify water quality status. The water quality status is divided into five groups, as suggested in Table 2.

- **Resampling:** Since most of the data belongs to group 4 and group 5, resampling is performed on the dataset to remove the data imbalance problem. After resampling the dataset, group 1, 2, 3, 4, and 5 contained a total of 10, 18, 19, 29, and 32 water quality specimens.
- **Data-Preprocessing:** In this technique, the median method is used to handle the null values and a min–max scalar to scale the data.
- **Data-Split:** The dataset is divided into two sets: training and testing with a proportion of 80 and 20 percent. Also to validate the proposed model, we have changed the random state value (like 0, 1, or 41) multiple times during splitting the dataset so that for every run, the train and test datasets would have different values. Finally we take the average accuracy from the overall run.
- **Classification-Algorithm:** For classifying the water quality status (WQS), several machine learning algorithms are used, including the Ada-Boost classifier, Random Forest classifier, Support Vector classifier, and Gradient Boosting. From such algorithms, the best algorithm is chosen by comparing the performance. Table 4 presents the experimental parameters used for those algorithms.

4. Experimental setup and result analysis

The objective of this section is to evaluate the performance of the PCR and classification model. In this section, different PCR and classification models are implemented, and the best model is selected using the statistical parameters like R^2, RMSE, MAE, Recall, Precision, and F1-score.

Table 4
Experimental parameters for classification model.

Model	Parameters	Values
Random Forest Classifier	n_estimators, Criterion	15, gini
Support Vector Classifier	C, Kernel, degree	2, linear, 1
Gradient Boosting Classifier	learning_rate, n_estimators	.1, 100
AdaBoost Classifier	learning_rate,n_estimators	1, 50

4.1. Evaluation metrics

Three statistical parameters, including R squared error (RSE), root mean squared error (RMSE) and mean absolute error (MAE), are used to test the PCR model efficiency. These criteria are expressed as follows:

$$RSE = 1 - ((explained\ variation)/(total\ variation)) \tag{10}$$

$$RMSE = \sqrt{\sum (y_{obs} - y_{pred})^2/n} \tag{11}$$

$$MAE = \sum(||y_{obs} - y_{pred}||)/n \tag{12}$$

where, y_{obs} = actual value, y_{pred} = predicted value and n = total number of samples.

Four classification metrics, including Accuracy, Recall, Precision, and F1-score, are used to test the classification model performance. These metrics are expressed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{16}$$

where, TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

4.2. Data and boxplot analysis

Table 1 provides a detailed statistical overview of the dataset and the factors that affect water quality. Standard deviation was maximum for TDS, EC, and COD, while the lowest standard deviation was found for DO and pH. The minimum and maximum DO values are 0 and 15.3 mg/ l, with a 5.20 mg/ l mean value. The turbidity ranged from 16 to 82 NTU with a mean value of 51.23 NTU, which means that these values are not distributed adequately (World Health Organization et al., 2004). The relative weight has

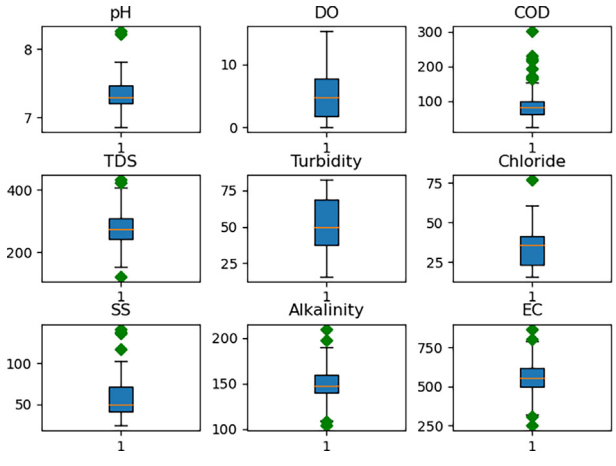

Fig. 3. Water quality variables outlier detection.

Table 5
Comparison of different PCR models for predicting WQI.

Model	R^2	RMSE	MAE
PCA+ Multiple Linear Regressor	.932	5.72	5.42
PCA+ Random Forest Regressor	.839	8.87	7.82
PCA+ Support Vector Regressor	.950	4.93	4.37
PCA+ Gradient Boosting Regressor	.722	11.6	9.15

Table 6
PCR model performance assessment using various components.

Model	Components	R^2	RMSE	MAE
PCR1	n = 1	.042	21.64	19.56
PCR2	n = 2	.411	16.97	14.57
PCR3	n = 3	.569	14.52	12.51
PCR4	n = 4	.564	14.59	12.54
PCR5	n = 5	.565	14.58	11.67
PCR6	n = 6	.709	11.92	9.44
PCR7	n = 7	.927	5.97	5.33
PCR8	n = 8	.932	5.72	5.42
PCR9	n = 9	.932	5.72	5.42

been mentioned for manually measuring the WQI. WQI was determined using Eqs. (1)–(3). From Table 1, the minimum WQI value is found 63, where the maximum value is 144, with a standard deviation of 20.86. The mean value of WQI is found 109.5, which shows that the quality of water is inappropriate for drinking and irrigation in most areas.

For establishing the intricate relationship between the water quality parameters, outlier detection carried out on the dataset. Fig. 3 demonstrates the boxplot analysis to find the outlier of the water quality parameters. From this analysis, parameters such as pH, DO, chloride, and turbidity are found normally distributed. These parameters have a positive effect on estimating WQI as they meet the standard values recommended by the WHO (World Health Organization et al., 2004). Other parameters like COD, TDS, SS, Alkalinity, and EC were extremely distorted and had high skewed values compared to the standard value. So the average value of WQI at Gulshan Lake is high due to COD, TDS, SS, Alkalinity, and EC. Therefore, it required adequate care.

4.3. PCR model result assessment

The proposed PCR method implemented using python. The results of different PCR models showed in Table 5. From this table, PCA with Support Vector Regression has achieved the highest accuracy compared to the other PCR techniques. Although other PCR models also performed well, PCA with Gradient Boosting Regression proved to be a less useful model.

Since the PCR model provides to work with fewer parameters, so we reduced the number of components instead of taking all the features. The results of taking different features showed in Table 6. For this technique, PCA with Multiple linear regression is selected since PCA is mostly related to multiple linear regression to create new principal components. Table 6 illustrated that, with nine and eight components, the PCR9 and PCR8 models showed the best performance, where PCR9 clarified all the variance. The PCR8 model gives the same result as the PCR9 model, and the number of parameters is also reduced. The R^2 value for the PCR8 model in testing steps is .932. If we reduce one more component from the PCR8 model, that model produced almost the same result as operating with all the components. The R^2 value in the PCR7 model is .927. After reducing one more component, the R^2 value reduced in the testing phases is .709. That shows less accuracy compared with the PCR7 and PCR8 models. Yet in the water samples, PCR6 still performed well. If we reduce more components from the

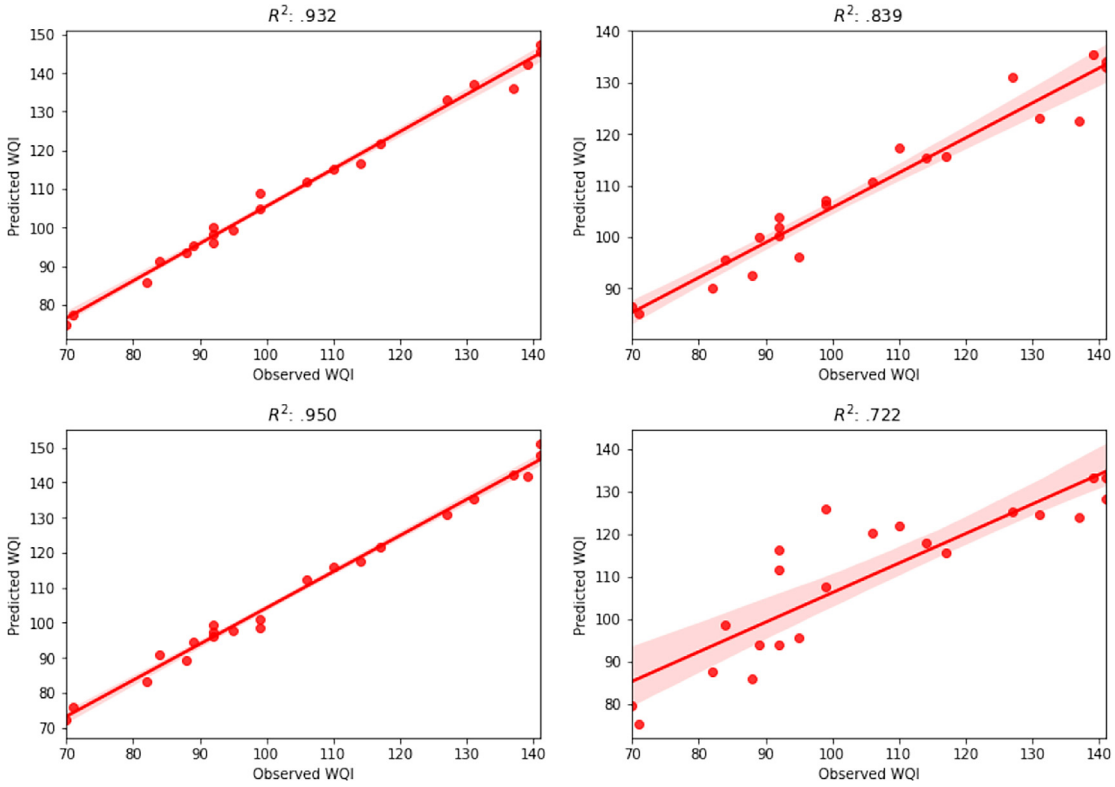

Fig. 4. Prediction and Classification of water quality indexPlot between observed and predicted WQI a. PCA+ Multiple Linear Regression model b. PCA+ Random Forest Regression model c. PCA+ Support Vector Regression model d. PCA+ Gradient Boosting Regression model.

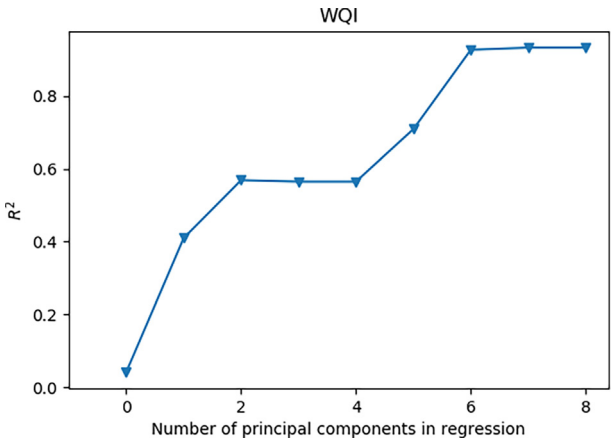

Fig. 5. Comparison of accuracy of each part in PCR model.

Table 7
Comparison of different classification model.

Model	Accuracy	Precision	Recall	F1-score
Random Forest Classifier	.91	.96	.85	.89
Support Vector Classifier	.86	.82	.91	.84
Gradient Boosting Classifier	1.0	1.0	1.0	1.0
AdaBoost Classifier	.77	.53	.60	.56

PCR model, the R^2 value is barely 50 per cent, which shows low PCR model efficiency.

The accuracy comparison of the PCR model in each principal component showed in Fig. 5. It is evident from Fig. 5 that the model performed well with six, seven and eight components. After then, it showed poor performance. Since PCR7 and PCR8 showed the same results as working with the PCR9, we could infer that the PCR method allows operating with fewer parameters instead of taking all the features.

Fig. 4 illustrated the plot between the observed and predicted WQI values for a better understanding of those models. Among them, the value appeared closer to the regression fit line in the PCA+ Support Vector Regression model because of the high training and testing accuracy.

4.4. Classification model result assessment

Different classification algorithms are implemented using python. The results of varying classification models are presented in Table 7. Among them, the Gradient Boosting Classifier has achieved the highest accuracy and proved to be an efficient model to predict water quality status. The second-best model is Random Forest Classifier, but to calculate recall, the Support Vector Classifier performs better than the Random Forest Classifier. Ada-Boost Classifier is found less effective model compared to the other techniques. The confusion matrix for those models is presented in Fig. 6. From Fig. 6, we can observe that the Gradient Boosting Classifier classify all the testing data according to the water quality level where other models misclassified some of the testing data.

4.5. Performance comparison with other studies

From Table 8, it is clear that both PCR and GBC methods have outperformed the previously developed models such as standalone machine learning (SVR, GB, DT), deep neural network (NN, MLP), and hybrid (SDA-ANN, BA-RT) in terms of predicting and classifying WQI. The best accuracy was found by Bui et al. (2020) (accuracy = 94%). However, they had not provided any classification model. Yilma et al. (2018) and Wang et al. (2017) achieved 93%

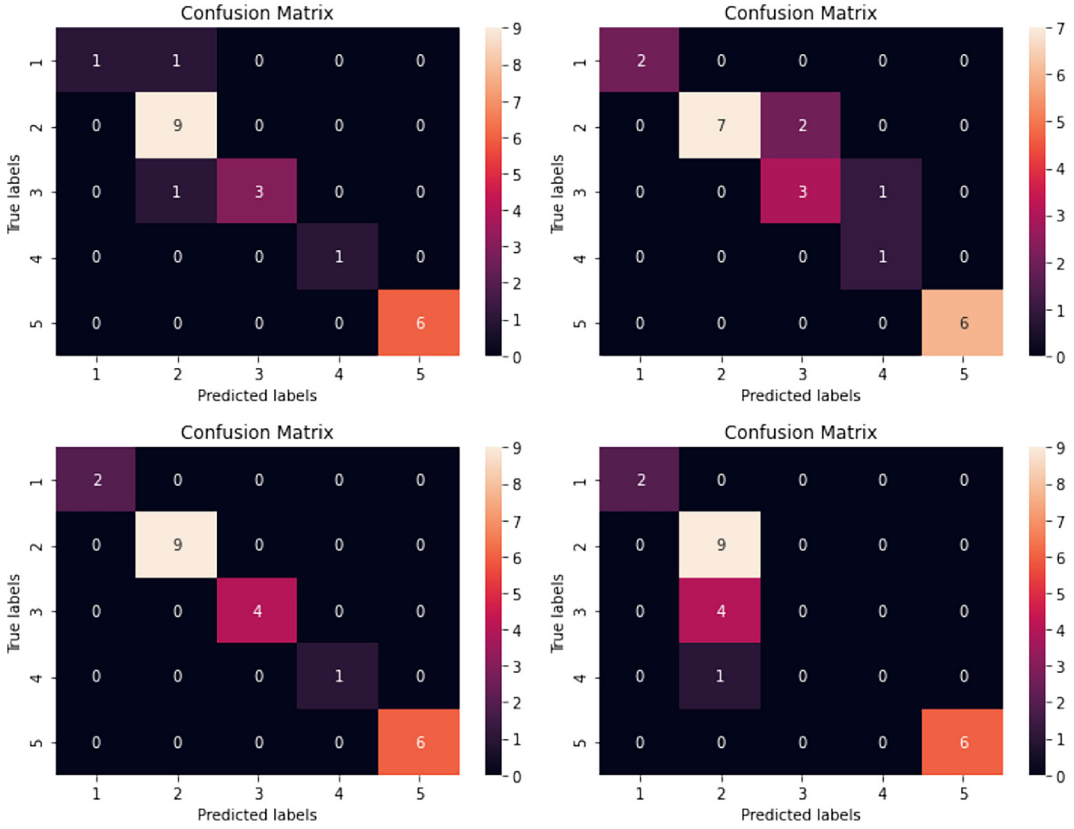

Fig. 6. Confusion matrix for classification algorithm. A. Random Forest Classifier b. Support Vector Classifier C. Gradient Boosting Classifier d. AdaBoost Classifier.

Table 8
Performance comparison between the proposed method and previous literatures work.

Name	Location	Total Specimens	Prediction Method	Classification Method	Prediction Accuracy	Classification Accuracy
Wang et al. (2017)	China	22	SVR	–	92%	–
Yilma et al. (2018)	Ethiopia	12	NN	–	93%	–
Samsudin et al. (2019)	Malaysia	13	SDA-ANN	–	71%	–
Ahmed et al. (2019)	Pakistan	4	GB	MLP	74%	85%
Ho et al. (2019)	Malaysia	6	–	DT	–	81%
Bui et al. (2020)	Iran	10	BA-RT	–	94%	–
Proposed Methods	Bangladesh	9	PCR	GBC	95%	100%

and 92% prediction accuracy using artificial neural networks and support vector regression techniques. But those techniques used 12 and 22 specimens of water that can be considered expensive since most of the specimens were examined in the laboratory. Furthermore, Samsudin et al. (2019), Ahmed et al. (2019) and Ho et al. (2019) used fewer water specimens to predict and classify WQI. Yet, their model accuracy is less than 85%. Contrary to the previous methods, the proposed models used only nine water specimens in this study and demonstrated 95% prediction accuracy and 100% classification accuracy.

5. Conclusion

This paper demonstrated a method for predicting and classifying the water quality using machine learning algorithms. The water metrics, including PH, DO, SS, EC, Turbidity, Chloride, COD, TDS, and Alkalinity, were used in this study. For data preprocessing, the median technique used to handle the null values and min–max scalar to scale the data. For the prediction purpose, we applied the principal component regression (PCR) method. After analyzing the performance of multiple PCR models, PCA with Support Vector Regression seems to be more effective with an accuracy of 95%. However, if the number of components reduced, then PCA with the Multiple Linear Regression model proved to be more effective. For the classification purpose, the Gradient Boosting classifier used to classify the water quality status. Besides, to check the performance of the model, the proposed model is compared with several state-of-art classifiers, including AdaBoost Classifier, Support Vector Classifier, and Random Forest Classifier. Experimental results showed that the Gradient Boosting Classifier classified water quality status more efficiently. Despite the achievements outlined in this paper, some improvements are still possible, including we can collect more training samples to make the model more stable and more progress is possible on the prediction model. Those issues will be overcome in future research, perhaps by proper tuning of the PCR model and using deep neural network.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Adimalla, Narsimha, 2019. Groundwater quality for drinking and irrigation purposes and potential health risks assessment: a case study from semi-arid region of South India. Exposure and Health 11 (2), 109–123.
Agamuthu, Pariatamby, Victor, Dennis, 2011. Policy trends of extended producer responsibility in Malaysia. Waste Management & Research 29 (9), 945–953.
Aghel, B., Rezaei, A., Mohadesi, M., 2019. Modeling and prediction of water quality parameters using a hybrid particle swarm optimization–neural fuzzy approach. International Journal of Environmental Science and Technology 16 (8), 4823–4832.
Ahmed, Ali Najah et al., 2019. Machine learning methods for better water quality prediction. Journal of Hydrology 578, 124084.
Ahmed, Umair et al., 2019. Efficient water quality prediction using supervised Machine Learning. Water 11 (11), 2210.
Asadollah, Seyed Babak, Seyed, Haji, et al., 2021. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. Journal of Environmental Chemical Engineering 9, (1) 104599.
Barzegar, Rahim, Moghaddam, Asghar Asghari, 2016. Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction. Modeling Earth Systems and Environment 2 (1), 26.
Barzegar, Rahim, Mohammad Taghi, Aalami, Jan, Adamowski, 2020. Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. Stochastic Environmental Research and Risk Assessment, pp. 1–19.
Bui, Duie Tien et al., 2020. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of The Total Environment 721, 137612.
Choubin, Bahram et al., 2016. Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on largescale climate signals. Hydrological Sciences Journal 61 (6), 1001–1009.
Dataset, Gulshan Lake, 2016. Published on May 20, 2018. URL: http://www.doe.gov.bd/site/publications/5132a8d7- 68e9 - 469d - a9af - 8981306b3b9f/ Surface - and - Ground - Water-Quality-Report-2016.
Dzwairo, Bloodless et al., 2006. Assessment of the impacts of pit latrines on groundwater quality in rural areas: a case study from Marondera district, Zimbabwe. Physics and Chemistry of the Earth, Parts A/B/C 31 (15-16), 779–788.
Ewaid, Salam Hussein, Abed, Salwan Ali, Kadhum, Safaa A., 2018. Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis. Environmental Technology & Innovation 11, 390–398.
Gaikwad, Satyajit et al., 2020. Geochemical mobility of ions in groundwater from the tropical western coast of Maharashtra, India: implication to groundwater quality. Environment, Development and Sustainability 22 (3), 2591–2624.
Hameed, Mohammed et al., 2017. Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. Neural Computing and Applications 28 (1), 893–905.
Ho, Jun Yung et al., 2019. Towards a time and cost effective approach to water quality index class prediction. Journal of Hydrology 575, 148–165.
Horton, Robert K., 1965. An index number system for rating water quality. Journal of Water Pollution Control Federation 37 (3), 300–306.
Imani, Maryam, et al., 2021. A novel machine learning application: Water quality resilience prediction Model. Science of the Total Environment 768, 144459.
Kadam, A.K. et al., 2019. Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India. Modeling Earth Systems and Environment 5 (3), 951–962.
Kar, Devashish, 2013. Wetlands and Lakes of the World. Springer New Delhi, India.
Kar, 2019. Wetlands and their Fish Diversity in Assam (India). Transylvanian Review of Systematical and Ecological Research 21 (3), 47–94.
Khadr, Mosaad, 2017. Modeling of water quality parameters in Manzala lake using adaptive neuro-fuzzy inference system and stochastic models. In: Egyptian Coastal Lakes and Wetlands: Part II. Springer, pp. 47–69.
Kisi, Ozgur et al., 2019. Modeling groundwater quality parameters using hybrid neuro-fuzzy methods. Water Resources Management 33 (2), 847–861.
Leong, Wei Cong et al., 2019. Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). International Journal of River Basin Management, 1–8.
Li, Zhenbo et al., 2018. Water quality prediction model combining sparse auto-encoder and LSTM network. IFAC-PapersOnLine 51 (17), 831–836.
Li, Lei et al., 2019. Water quality prediction based on recurrent neural network and improved evidence theory: a case study of Qiantang River, China. Environmental Science and Pollution Research 26 (19), 19879–19896.
Mohammadpour, Reza et al., 2015. Prediction of water quality index in constructed wetlands using support vector machine. Environmental Science and Pollution Research 22 (8), 6208–6219.
Moon, Sang-Ki, Woo, Nam C., Lee, Kwang S., 2004. Statistical analysis of hydrographs and water-table fluctuation to estimate groundwater recharge. Journal of Hydrology 292 (1-4), 198–209.
Oelen, Allard, van Aart, Chris J., De Boer, Victor, 2018. Measuring surface water quality using a low-cost sensor kit within the context of Rural Africa. In: P-ICT4D@ WebSci.

Oladipo, Johnson O. et al., 2021. Comparison between fuzzy logic and water quality index methods: A case of water quality assessment in Ikare community, Southwestern Nigeria. Environmental Challenges 3, 100038.

Rahman, Shafkat Shamim, Hossain, Md Mahboob, 2019. Gulshan Lake, Dhaka City, Bangladesh, an onset of continuous pollution and its environmental impact: a literature review. Sustainable Water Resources Management 5 (2), 767–777.

Rajaee, Taher, Khani, Salar, Ravansalar, Masoud, 2020. Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review. Chemometrics and Intelligent Laboratory Systems 200, 103978.

Saghebian, S. Mehdi et al., 2014. Ground water quality classification by decision tree method in Ardebil region, Iran. Arabian Journal of Geosciences 7 (11), 4767–4777.

Salari, Marjan et al., 2018. Quality assessment and artificial neural networks modeling for characterization of chemical and physical parameters of potable water. Food and Chemical Toxicology 118, 212–219.

Samsudin, Mohd Saiful et al., 2019. Comparison of prediction model using spatial discriminant analysis for marine water quality index in mangrove estuarine zones. Marine Pollution Bulletin 141, 472–481.

Sinshaw, Tadesse A. et al., 2019. Artificial neural network for prediction of total nitrogen and phosphorus in US Lakes. Journal of Environmental Engineering 145 (6), 04019032.

Tyagi, Shweta et al., 2013. Water quality assessment in terms of water quality index. American Journal of Water Resources 1 (3), 34–38.

Wang, Xiaoping, Zhang, Fei, Ding, Jianli, 2017. Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. Scientific Reports 7 (1), 1–18.

World Bank Report, 2018. Published on Sept 16, 2018. URL: https: //www. dhakatribune.com/ bangladesh/ environment/ 2018/09/16/world-bank-urban-pollution-costs-bangladesh- 6–5bn-a-year.

World Health Organization, et al., 2004. Manganese in drinking-water: Background document for development of WHO Guidelines for Drinking-Water Quality. Tech. rep. World Health Organization.

Wu, Yiping, Liu, Shuguang, 2012. Modeling of land use and reservoir effects on nonpoint source pollution in a highly agricultural basin. Journal of Environmental Monitoring 14 (9), 2350–2361.

Xu, Longqin, Liu, Shuangyin, 2013. Study of short-term water quality prediction model based on wavelet neural network. Mathematical and Computer Modelling 58 (3–4), 807–813.

Yajima, Hiroshi, Derot, Jonathan, 2018. Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. Journal of Hydroinformatics 20 (1), 206–220.

Yilma, Mulugeta et al., 2018. Application of artificial neural network in water quality index prediction: a case study in Little Akaki River, Addis Ababa, Ethiopia. Modeling Earth Systems and Environment 4 (1), 175–187.

Zhang, Yanyang et al., 2019. Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. Water Research 164, 114888.

Zhu, Senlin et al., 2019. Two hybrid data-driven models for modeling water-air temperature relationship in rivers. Environmental Science and Pollution Research 26 (12), 12622–12630.