Environ. Water Eng., 2025, 11(3), 219-227 https://doi.org/10.22034/ewe.2024.470962.1957

EWE

Environment and Water Engineering

Homepage: www.jewe.ir

ISSN: 2476-3683

Estimation of water quality index in Zohreh River using principal component analysis and artificial intelligence models

Amir Hossein Shakarami[1], Laleh Divband Hafshejani[1]✉, Parvaneh Tishehzan[1], and Hamid Abdolabadi[1]

[1]Department of Environmental Engineering, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran

ARTICLE INFO

Paper Type: Research Paper

Received: 31 July 2024
Revised: 22 August 2024
Accepted: 31 August 2024
Published: 01 September 2025

Keywords
Artificial Intelligence
Dissolved Oxygen
Prediction
Vector Machine

***Corresponding author:**
L. Divband Hafshejani
✉ ml.divband@scu.ac.ir

ABSTRACT

This research explored the root causes of hidden pollution and key factors affecting spatial changes, as well as identifying the best inputs for water quality modeling. The study used principal component analysis (PCA), artificial neural network models (MLP), gene expression programming (GEP), and support vector machine (SVM) to achieve its objectives. The dataset included 11 different parameters collected monthly over 10 water years (2012-2021) from the Zohreh River, Iran. Initially, PCA was applied to reduce parameters and calculate the Water Quality Index (WQI). Two input models (parameters before and after PCA) were then created using artificial intelligence to determine the most accurate model for predicting the WQI. The Kaiser-Meyer-Olkin measure (KMO) was 0.6524, indicating the dataset was suitable for factor analysis. Bartlett's sphericity test was also significant at the 0.05 alpha level. PCA identified five significant principal components, explaining 70.66% of the total variance. The combined SVM and PCA model showed the best prediction ability, with an R^2 of 0.889, RMSE of 0.052, and MAE of 0.038.

Highlights

- PCA reduces 11 water quality parameters to 8 key factors.
- SVM model best predicts WQI with R^2=0.911, RMSE=0.047.
- DO and turbidity significantly impact Zohreh River water quality.
- AI models combined with PCA improve WQI prediction accuracy.
- Sensitivity analysis identifies major pollutants affecting WQI.

How to cite this paper:
Shakarami, A. H., Divband Hafshejani, L., Tishehzan, P., & Abdolabadi, H. (2025). Estimation of water quality index in Zohreh River using principal component analysis and artificial intelligence models. *Environ. Water Eng.*, *11*(3), 219-227. https://doi.org/10.22034/ewe.2024.470962.1957

1. Introduction

The water of rivers, as the vital veins of the Earth, plays an essential role in providing drinking water, supporting agriculture, and fueling industry. However, human activities such as industrial agriculture, manufacturing, and wastewater discharge severely threaten the quality of this water. Contaminated water with harmful pollutants can lead to a wide range of waterborne diseases, including cholera, typhoid fever, and dysentery. Additionally, it can directly harm ecosystems, disrupt food chains, and decrease biodiversity.

Assessing river water quality is crucial for identifying these pollutants and pathogens and provides necessary guidelines for protective measures to restore and protect these delicate ecosystems. This information enables water resource managers to allocate water wisely, prevent the contamination of human drinking water, and ensure its sustainable use for other vital needs.

River water quality is evaluated using various techniques. Univariate analysis is one such technique, which entails the independent examination of specific water quality parameters, each compared to predetermined standards or guidelines to determine the current state of water quality (Ji et al., 2016).

Another approach is the water quality index, viewed as one of the best ways to gauge water quality. The water quality index assesses water quality and determines whether it is suitable for drinking, irrigation, and industrial uses by combining multiple water quality parameters into a mathematical formula (Akter et al., 2016; Shil et al., 2019). Regular monitoring of water quality produces a vast amount of complex data with numerous parameters that vary significantly over time and space. Analyzing this data to identify influencing factors and understand the current state of water quality can be challenging. Multivariate statistical methods, such as factor analysis (FA) and principal component analysis (PCA), are valuable tools for examining and integrating complex water quality data (Tripathi & Singal, 2019). In a study by Roy et al. (2024), Principal Component Analysis (PCA) was used to project the Water Quality Index (WQI) of four rivers in Dhaka. The WQI index was computed by reducing the number of water quality parameters from 12 to 7, revealing that over 70% of the samples fall into moderate or poor categories. This paper presents a statistical method for evaluating the river water quality of Dhaka, which could enhance plans for controlling river pollution. Principal component analysis was employed in a study to select the parameters used in the development of the Water Quality Index (WQI) for the Ganges River (Tripathi & Singal, 2019). This approach reduced the range of considered parameters from 28 to 9, which include dissolved oxygen, pH, electrical conductivity, biological oxygen demand, total coliform, chlorides, magnesium, sulfate, and total dissolved solids. This reduction in parameters leads to savings in time, cost, and effort for water monitoring, as well as establishing a basis for the future development of the Ganges Water Quality Index (GWQI). Using artificial intelligence as a powerful tool is revolutionizing how we manage water resources, given the increasing complexity of data on river water quality and the limitations of traditional data analysis techniques (Oruganti et al., 2023). The ability of artificial intelligence to identify trends in water quality data provides authorities with a significant opportunity to anticipate potential contamination events and take preventive actions to avert them (Maurya et al., 2024). Drought, pollution, and excessive water extraction pose significant challenges for the Zohreh River, a crucial water supply in southwestern Iran. Currently, minimal comprehensive research employs advanced AI methodologies to assess and predict the river's water quality accurately. The objective of this study is to bridge this information gap. The primary aim of this study is to utilize artificial intelligence models, including support vector machines, gene expression programming, and artificial neural networks, to forecast the water quality index of the Zohreh River, Iran. Principal component analysis (PCA) reduces the number of factors affecting water quality. This project focuses on creating a comprehensive and accurate model for assessing and predicting the river's water quality to enhance and sustainably manage local water resources.

2. Materials and Methods

2.1 Study area

The Zohreh River, also known as the Handijan River, is one of the longest and most plentiful rivers in Iran, flowing in the southwest of the country. This river ultimately empties into the Persian Gulf after approximately 475 km of flow across three provinces: Fars, Kohgiluyeh and Boyer-Ahmad, and Khuzestan. The Zohreh River originates in the heights of the White Mountain in Fars Province. It is the second largest independent watershed in Iran, covering an area of 17,150 km². The local population relies heavily on the Zohreh for agriculture and livestock raising. Additionally, several towns and cities depend on the Zohreh River for their drinking water. Unfortunately, the Zohreh River faces numerous challenges, including pollution, drought, and excessive water consumption. These issues can have severe consequences on the local ecology and people's lives. Therefore, it is essential to protect and restore this invaluable river through sustainable water resource management and pollution control.

2.2 Data preparation and statistical analysis

The parameters pH, electrical conductivity (µS/cm), phosphate (mg/L), nitrate (mg/L), ammonium (mg/L), total hardness (mg/L), fecal coliform (MPN), dissolved oxygen (mg/L), biochemical oxygen demand (mg/L), chemical oxygen demand (mg/L) and turbidity (NTU) were received monthly for 10 water years (2011-2021) and the Firouzabad and 720-meter Sovere stations in the Zohreh River from the Khuzestan Water and Electricity Organization. Then, the Iranian Water Quality Index was calculated for each month and each station (Khuzestan Water and Power Organization, 2021).

2.3 Principal component analysis of water quality data

Principal Component Analysis (PCA) is a mathematical method for dimensionality reduction. It uses an optimization technique to transform a set of more correlated events into a set of values of uncorrelated monotonic variables as principal components (Mukherjee et al., 2022). It is widely used as a feature extraction method. Extracting features are used in various fields, including machine learning, visualization, and signal analysis. A principal component is a linear combination of p principal variables a1, a2, ..., and an. In the PCA method, first, a matrix of measured values is formed using Eq. 1 (Chawishborwornworng et al., 2024).

$$A=\left(a_{ij}\right)_{m\times n}=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \quad (1)$$

Where, m is the number of sample stations and n is the number of water quality parameters. In the next step, the raw data are standardized using a standard score (Z-Score) to reduce the dimensionality. This process is done using Eq. 2, where aij: original value, $\bar{a}_j$: mean, s_j: standard deviation of each parameter and a'_{ij}: its normalized.

$$a'_{ij}=\frac{a_{ij}-\bar{a}_j}{s_j} \quad (2)$$

Then the correlation coefficient matrix is calculated using the normalized data and Eq. 3.

$$R=\left(r_{ij}\right)_{n\times n}=\frac{1}{m-1}\sum_{k=1}^{m}\frac{a_{ki}-\bar{a}_i}{s_i}\times\frac{a_{kj}-\bar{a}_j}{s_i}$$
$$(i,j=1,2,3\ldots..n) \quad (3)$$

The covariance or correlation matrix is calculated using Eq. 4 to determine the relationship between pairs of variables.

$$R = (r_{ij})_{n\times n} = \frac{1}{m-1}\sum_{k=1}^{m} a'_{ki} \times a'_{kj} \quad (4)$$

By decomposing the covariance/correlation matrix, the eigenvectors are obtained using Eq. 5, and the eigenvalues are obtained using Eq. 6. The eigenvectors indicate the directions in the data space that have the most variance, while the eigenvalues indicate the amount of variance in each direction.

$$\left|\begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{bmatrix} - \begin{bmatrix} \Lambda_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Lambda_{\mathrm{mn}} \end{bmatrix}\right| = 0,\ (i = 1,2,3 \ldots . n) \quad (5)$$

$$\begin{pmatrix} r_{11} - \Lambda_{\mathrm{i}} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} - \Lambda_{\mathrm{i}} \end{pmatrix} \times \begin{pmatrix} \chi_{i1} \\ \vdots \\ \chi_{in} \end{pmatrix} = 0 \quad (i = 1,2,3 \ldots . n) \quad (6)$$

where Λ and χ represent the eigenvectors and eigenvalues, respectively. The principal components are selected based on criteria such as the amount of cumulative variance using Eq. 7. These principal components preserve important information from the data and are sufficient to reduce the dimensionality of the data.

$$PC = r_{i1}a'_1 + r_{i2}a'_2 + \cdots + r_{in}a'_n \quad (i = 1,2,3 \ldots . n) \quad (7)$$

2.4 Prediction of Iran's WQI using AI models

Several artificial intelligence algorithms, including gene expression programming, artificial neural networks (deep learning), and support vector machines (machine learning), were employed to predict Iran's water quality index. Utilizing multiple models to enhance prediction accuracy and select the best one is an effective strategy in the realm of water quality prediction. When utilizing artificial intelligence models, data normalization is essential for boosting model performance and yielding more accurate results. This step is particularly crucial when handling diverse data categories that have varying scales and units of measurement. Consequently, all data (input and output) were initially normalized between zero and one using Eq. 8.

$$\mathrm{x}_i^* = \frac{(x_i - x_{min})}{(x_{max} - x_{min})} \quad (8)$$

The values x, x_{min}, and x_{max} represent the minimum and maximum values of each parameter in the data set, respectively.

In this study, Gene Expression Programming (GEP), an evolutionary-based artificial intelligence method, was utilized to uncover nonlinear and complex relationships among various water quality parameters in the study area. The input data included qualitative parameters such as pH, electrical conductivity, phosphate, nitrate, ammonium, total hardness, fecal coliform, dissolved oxygen, biochemical oxygen demand, chemical oxygen demand, and turbidity. Given the inherently nonlinear and complex nature of natural systems, GEP was selected as an effective tool for modeling these intricacies.

The GEP algorithm was specifically designed to extract hidden patterns between water quality parameters and the Iranian Water Quality Index (IRWQI), enabling accurate predictions within the study area. To optimize the performance of the GEP model, key parameters- including population size, number of generations, mutation probability, and evolutionary operators- were carefully fine-tuned. The population size was selected to balance computational efficiency and model accuracy. The number of generations was adjusted based on the complexity of the dataset and the need for adequate model refinement. The mutation probability was optimized to avoid local minima and enhance the exploration of complex variable relationships. In addition, evolutionary operators such as crossover were employed to combine different model components and improve predictive accuracy.

Another algorithm applied in this study was the Support Vector Machine (SVM), a supervised learning technique capable of identifying complex patterns between input and output variables using kernel functions. SVM was configured to accurately model the relationships between water quality parameters and the IRWQI. To optimize its performance, critical parameters- including the type of kernel function, the regularization parameter (C), and the gamma value- were carefully adjusted.

Additionally, a Multilayer Perceptron (MLP) neural network, a powerful artificial intelligence method, was employed to predict the water quality index. As a supervised learning model, MLP uses multiple hidden layers to capture nonlinear relationships among variables. Through the backpropagation process and weight adjustments, MLP can provide highly accurate predictions. To optimize the MLP model, key network parameters- including the number of hidden layers and neurons, the ReLU activation function, learning rate, number of training iterations (epochs), and the Adam optimization algorithm-were fine-tuned. The number of layers and neurons was selected to effectively capture complex data patterns. The ReLU function was used to mitigate gradient vanishing issues, while the learning rate was set to ensure stable and efficient model convergence. The number of training epochs was determined based on data volume and complexity to prevent overfitting. The Adam optimizer was chosen for its fast convergence and overall performance enhancement.

2.5 Prediction of Iran's WQI

In this part of the study, a combination of Principal Component Analysis (PCA) and artificial intelligence (AI) models was employed to predict Iran's Water Quality Index (IRWQI). The primary goal of integrating PCA with AI models is to reduce the dimensionality of input data, thereby enhancing the predictive performance of the models. The modeling results obtained from AI methods- Gene Expression Programming (GEP), Multilayer Perceptron (MLP), and Support Vector Machine (SVM)- combined with PCA (referred to as GEP+PCA, MLP+PCA, and SVM+PCA) were compared with those from the standalone models (GEP, MLP, and SVM). This comparison was conducted to evaluate and validate the effectiveness of PCA in improving the accuracy and efficiency of water quality index prediction.

2.6 Evaluation of the accuracy of ai models in predicting Iran's WQI

In this study, the performance of various artificial intelligence models in predicting Iran's Water Quality Index (IRWQI) was evaluated using common statistical indicators, including Root Mean Square Error (RMSE) (Eq. 9), Mean Absolute Error

(MAE) (Eq. 10), and Coefficient of Determination (R^2) (Eq. 11) (Divband Hafshejani et al., 2022).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i^{exp} - Y_i^{pred})^2} \quad (9)$$

$$MAE = \frac{\sum_{i=1}^{n}\left|Y_i^{exp} - Y_i^{pred}\right|}{n} \quad (10)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i^{exp} - Y_i^{pred}\right)^2}{\sum_{i=1}^{n}\left(Y_I^{exp} - Y_{ave}^{exp}\right)^2} \quad (11)$$

In the evaluation Eqs. 9 to 11, Y_i^{pred} represents the predicted value of the water quality index at the i-th instance, Y_i^{exp} is the corresponding observed (actual) value, Y_{ave}^{exp} is the average of the actual values, and n is the total number of data points. RMSE is used to measure the deviation of the predicted values from the actual observations, where lower RMSE values indicate better model performance. The coefficient of determination (R^2) reflects the proportion of variance in the dependent variable (water quality index) that is explained by the independent variables (model inputs); higher R^2 values demonstrate stronger predictive capabilities. Meanwhile, MAE represents the average absolute difference between predicted and observed values, offering a straightforward interpretation of prediction accuracy. Together, these statistical metrics provide a comprehensive assessment of each model's ability to accurately estimate the water quality index in the study area.

2.7 Sensitivity Analysis of Iran's WQI to Water Quality Parameters

To evaluate the influence of individual water quality parameters on the prediction of Iran's WQI (IRWQI), a sensitivity analysis was conducted using Eq. 12 (Abdi & Mazloom, 2022).

$$r(I_i, \omega) = \frac{\sum_{j=1}^{n}(I_{ij} - \bar{I}_i)(\omega_j - \bar{\omega})}{\sqrt{\sum_{j=1}^{n}(I_{ij} - \bar{I}_i)^2}\sqrt{\sum_{j=1}^{n}(\omega_j - \bar{\omega})^2}} \quad (12)$$

In this equation, ωj and $\bar{\omega}$ represent the *j*-th predicted value and the average predicted value of the water quality index, respectively, Iij and Ii refer to the *i*-th observation and the mean of the *i*-th input variable, respectively. *n* denotes the total number of data points, and $r(I_i, \omega)$ is the coefficient correlation between the *i*-th input parameter and the predicted water quality index. The value of the correlation coefficient reflects the extent to which each input variable affects the predicted IRWQI. A positive correlation indicates a direct relationship between the input parameter and the output (IRWQI), whereas a negative value suggests an inverse relationship. Furthermore, the greater the absolute value of the correlation coefficient for a specific parameter, the more significant its influence on the model's output (Abdi & Mazloom, 2022; Hajirezaie et al., 2017).

3. Results and Discussion

3.1 Statistical processing and analysis of data

Fig. 1 shows, for every water quality factor, the normalized data (Z-scores). Within every box, the horizontal line shows the median Z-score for that parameter. The 25th and 75th percentiles, respectively, are shown by the box's lower and upper limits. Though they are not regarded as outliers, the lines extending from the box (whiskers) mark data points deviating from the central distribution(Roy et al., 2024; Tripathi & Singal, 2019). Outliers- values quite different from the other data points- are shown as separate circles that might affect the findings of Principal Component Analysis (PCA) (Roy et al., 2024).

The box- and- whisker graphs show that the median of most parameters' normalized values is about zero. Among all the parameters, pH shows the most change; fecal coliform shows the least. Moreover, turbidity features a quite high number of outliers. Still, the general data distribution seems symmetric with rather few outliers.

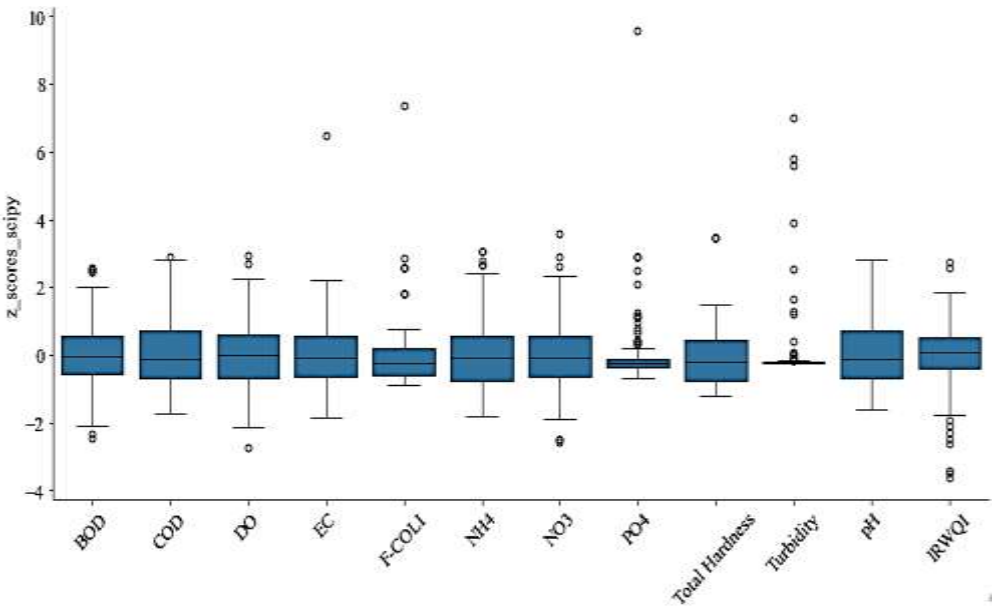

Fig. 1 Z-score variability of the dataset

Table 1 shows the correlation matrix, highlighting the interactions among every pair of water quality factors. In statistical analysis, this matrix is a useful instrument since it helps one understand the interactions among several parameters. Regarding water quality, it improves knowledge of interdependence among several indicators.

Dissolved oxygen (DO) shows a notable negative correlation with respect to parameters including nitrate, phosphate, biochemical oxygen demand (BOD), and chemical oxygen demand (COD.). This inverse link implies that DO levels drop as the organic load in the water rises- mostly from microbial activity consuming oxygen to break down organic matter. Reflecting the biological state of the water body, this correlation is a major in evaluations of water quality. Usually, the negative link between DO and pollution from nutrients and organic matter emphasizes how organic pollution influences water environments (Das et al., 2021).

Table 1 Correlation matrix of different parameters of IWQI

Parameters	BOD	COD	DO	EC	F-COLI	NH_4	NO_3	PO_4	Total Hardness	Turbidity	pH
BOD	1.00	0.09	-0.5	-0.19	-0.01	-0.30	0.06	0.01	-0.06	0.1	-0.09
COD	0.09	1.00	-0.08	0.1	-0.19	-0.26	0.21	0.16	0.01	0.38	0.02
DO	-0.5	-0.08	1.00	-0.04	-0.08	-0.18	-0.06	0.00	-0.05	-0.28	0.01
EC	-0.19	0.1	-0.04	1.00	0.04	0.14	-0.08	0.13	0.42	-0.11	-0.19
F-COLI	-0.01	-0.19	-0.08	0.04	1.00	0.22	-0.03	-0.06	-0.01	0.06	0.06
NH_4	-0.30	-0.26	-0.18	0.14	0.22	1.00	-0.44	0.02	-0.19	-0.03	-0.05
NO_3	0.06	0.21	-0.06	-0.08	-0.03	-0.44	1.00	0.27	0.30	0.18	0.1
PO_4	0.01	0.16	0.00	0.13	-0.06	0.02	0.27	1.00	0.05	0.04	0.04
Total Hardness	-0.06	0.01	-0.05	0.42	-0.01	-0.19	0.30	0.05	1.00	-0.03	-0.11
Turbidity	0.1	0.38	-0.28	-0.11	0.06	-0.03	0.18	0.04	-0.03	1.00	0.09
pH	-0.09	0.02	0.01	-0.19	0.06	-0.05	0.1	0.04	-0.11	0.09	1.00

A strong negative correlation between ammonium and nitrate shows that the nitrification process is happening, where ammonium is changed into nitrate by living organisms. This transformation is facilitated by nitrifying bacteria that convert ammonium first into nitrite and then into nitrate(Guo et al., 2013). The rate and efficiency of this process are influenced by several environmental factors, including temperature, pH, and dissolved oxygen levels.

A positive correlation between phosphate and nitrate suggests that phosphorus commonly co-occurs with nitrogen and organic matter. These nutrients serve as key contributors to algal growth, and their simultaneous presence can trigger algal blooms. The relationship between phosphate and nitrate has long attracted the attention of researchers, as these two elements often act synergistically in fueling the rapid proliferation of algae in aquatic ecosystems. When available in excess, phosphorus frequently binds to nitrogen and organic compounds, promoting eutrophication.

Among nitrogen compounds, nitrate- a highly soluble and mobile form- is considered a primary driver of algal bloom events. Human activities, like farming runoff and wastewater releases, greatly increase nitrate levels, which harm water quality and promote unhealthy conditions for aquatic life. These conditions cause too much algae to grow, which leads to a series of environmental problems like lower oxygen levels, the release of toxic substances from blue-green algae, and disruption of the food chains in water(Dattamudi et al., 2020). Additionally, the positive correlation between total hardness and parameters like nitrate and turbidity reflects the influence of carbonate rocks on water chemistry. These rocks, commonly found in many watersheds, release calcium and magnesium ions, which contribute to both hardness and the alkaline buffering capacity of the water(Channabasava et al., 2017). Water alkalinity, primarily derived from bicarbonates, carbonates, and hydroxides of calcium, magnesium, sodium, and potassium, plays a crucial role in stabilizing pH and maintaining ecological balance.

Finally, a strong positive correlation between turbidity and the Iranian Water Quality Index (IRWQI) reveals that turbidity is a key parameter influencing overall water quality in the studied river. High turbidity levels, which are often caused by floating solids, dirt particles, and organic matter, can block light, disrupt gas exchange, and harm aquatic life, leading to poorer ecological health of the water.

3.2 Principal Component Analysis

Two crucial tests- the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's test of sphericity- were used to determine whether the data was sufficient for factor analysis prior to Principal Component Analysis (PCA) on the dataset. To show that the data are fit for PCA, the KMO value must be more than 0.5 and Bartlett's test must produce a noteworthy p-value ($p < 0.05$). The KMO value in this study was 0.6524; Bartlett's test was significant at the 0.05 level, so verifying the fit of the dataset for PCA (Chawishborwornworng et al., 2024).

Analyzing 11 water quality parameters directly would be challenging since they are related to one another and could result in complex computations and biased conclusions(Tripathi & Singal, 2019). PCA simplifies complex datasets without sacrificing much information. The underlying structure within the dataset was extracted, and data dimensionality was lowered using PCA. PCA identifies a set

of independent components with the most data variation, so facilitating the understanding of complicated datasets without much loss of information(Panigrahi et al., 2007). Table 2 shows dimensions for the first five main components, PCs. These loadings help to interpret the meanings of the principal components by showing the correlation coefficients between the original variables. Strong markers of a variable's contribution to a given component in this study were factor loadings with absolute values above 0.5.

Table 3 shows for every component the eigenvalues, variance explained, and cumulative variance associated with them. Retained for study were the first five elements, each with eigenvalues above one. They fit and simplify the dataset since together they account for almost 70% of the total variance (Ibrahim et al., 2023). Only the first five were chosen in line with the parsimony and for efficient dimensionality reduction, even if Table 4 shows that the first eight components explain over 90% of the total variance. Eight of the original eleven water quality parameters were found, based on PCA results, to be sufficient to adequately reflect the fundamental features of water quality in the study area. Although some factors, such as dissolved oxygen and nitrate, phosphate, BOD_5, and COD, show clear correlations, the final list of variables was not much shortened. This choice was based on expert opinion and literature consensus, so underlining the individual importance of every parameter in water quality evaluation(Tripathi & Singal, 2019). Consequently, every retained parameter was regarded as necessary and investigated independently to guarantee thorough assessment.

Table 2 Rotated factor loadings for water quality data

Parameters	PC1	PC2	PC3	PC4	PC5
BOD	0.0008	**-0.4947**	0.0586	0.0213	**0.5456**
COD	0.3357	-0.1816	-0.0643	**0.4899**	-0.0706
DO	-0.2419	-0.4311	0.2655	-0.0476	0.3249
EC	0.1517	0.3105	**0.5393**	0.1304	0.0675
F-COLI	0.0819	0.2660	-0.1582	-0.3580	**0.5826**
NH_4	-0.1954	**0.5169**	-0.1360	0.0227	0.2161
NO_3	**0.5187**	-0.2579	-0.0006	-0.3567	-0.2318
PO_4	-0.0172	0.0821	0.1238	**0.6293**	0.1739
Total Hardness	0.3253	0.0790	**0.5235**	-0.1482	-0.0590
Turbidity	0.3898	-0.0065	-0.4013	0.2416	0.1921
pH	0.0179	-0.0697	-0.3696	-0.0360	-0.1477

Table 3 Eigen values of Zohreh water quality data set

Principal Component	Eigen Values	Variance (%)	Cumulative Variance (%)
1	2.4901	20.61	20.61
2	1.9301	15.98	36.53
3	1.5505	12.93	49.46
4	1.3506	11.28	60.74
5	**1.1990**	**9.92**	**70.66**
6	1.0025	8.60	79.26
7	0.1220	5.93	85.19
8	0.7049	4.93	90.12
9	0.5952	3.97	94.09
10	0.4315	2.99	97.08
11	0.3443	2.91	99.99

The first principal component (PC1) accounted for 20.61% of the total variance, with a strong loading on total nitrate, suggesting that this component primarily reflects the contribution of nitrate-related processes. The second component (PC2) explained 15.98% of the total variance and was characterized by high loadings on biochemical oxygen demand (BOD_5) and ammonium, indicating a strong association with organic pollution and nitrogenous compounds. The third component (PC3) accounted for 12.93% of the total variance, with dominant loadings on electrical conductivity and total hardness, which reflect the influence of dissolved ions and overall mineral content in the water.

The fourth principal component (PC4) explained 11.28% of the total variance, showing strong loadings on chemical

oxygen demand (COD) and phosphate, representing persistent organic pollution and nutrient enrichment from anthropogenic sources. The fifth component (PC5) accounted for 9.29% of the total variance (note: your original said 92.9%, which seems likely to be a typo), with significant loadings on fecal coliform and BOD_5, highlighting microbial contamination associated with fecal matter and biodegradable organic substances.

The first and third components primarily relate to the mineral quality of the water, signifying the presence of soluble mineral salts. These may originate from natural geological formations such as limestone and dolomite or from anthropogenic activities like industrial discharges and agricultural runoff(Nong et al., 2019). The second component reflects the organic load in the water, indicating the presence of biodegradable organic matter, often derived from municipal and industrial wastewater or agricultural inputs(Karpagavalli et al., 2019). The fourth component, also related to organic pollution, appears to capture more persistent organic compounds and phosphate, typically associated with industrial effluents and fertilizer use. Lastly, the fifth component is associated with microbial contamination, revealing the presence of pathogenic bacteria, likely from sewage discharges, including municipal wastewater and effluents from wastewater treatment plants. These findings align with earlier studies that demonstrate the effectiveness of PCA in identifying latent factors responsible for the spatial and temporal variability in water quality (Karpagavalli et al., 2019).

Table 4 Comparing the performance of artificial intelligence models in predicting of IWQI

Model	Evaluation Metrics		
	R^2	RMSE	MAE
MLP	0.855	0.063	0.049
SVM	0.911	0.047	0.035
GEP	0.803	0.070	0.057
MLP+ PCA	0.662	0.111	0.082
SVM+ PCA	0.889	0.052	0.038
GEP+ PCA	0.756	0.077	0.058

3.3 Evaluation of the Ability of AI Models in Predicting WQI

The performance results of various AI models in predicting the Iranian Water Quality Index (IRWQIsc) are presented in Table 4. Among the models evaluated, when all 11 input variables- including pH, electrical conductivity, phosphate, nitrate, ammonium, total hardness, fecal coliform, dissolved oxygen, biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), and turbidity- were used, the Support Vector Machine (SVM) model demonstrated superior predictive accuracy compared to the Artificial Neural Network (ANN) and Gene Expression Programming (GEP) models.

The SVM model achieved a coefficient of determination (R^2) of 0.911, a Root Mean Square Error (RMSE) of 0.047, and a Mean Absolute Error (MAE) of 0.035, indicating high precision and low prediction error. SVM models are particularly effective for capturing nonlinear relationships within data. In this case, it is likely that complex, nonlinear interactions exist between the water quality index and several input variables, which the SVM model was able to accurately capture.

These findings are consistent with previous studies that have highlighted the robustness of SVM models in modeling complex and nonlinear environmental systems, especially in water quality prediction tasks(Guo et al., 2013; Tripathi & Singal, 2019). The results emphasize the potential of SVM as a reliable tool for predicting composite water quality indices, especially in data-rich and multivariate environments.

SVM, both as a standalone model and in combination with other algorithms, has proven highly effective in predicting water quality parameters, particularly due to its strong capability in modeling nonlinear relationships(Jamshidzadeh et al., 2023; Vellingiri et al., 2024). In the present study, after dimensionality reduction using PCA and the selection of 8 input parameters (with turbidity, dissolved oxygen, and DO removed), the overall predictive performance of the artificial intelligence models decreased slightly. This decline was more pronounced in the Artificial Neural Network (ANN) compared to the other models. Specifically, prior to PCA, the ANN model achieved $R^2 = 0.855$, RMSE = 0.063, and MAE = 0.049, whereas after applying PCA, these metrics declined to $R^2 = 0.662$, RMSE = 0.111, and MAE = 0.082.

In contrast, the performance decline for the Support Vector Machine (SVM) model was less significant. Even after dimensionality reduction, the SVM maintained relatively high prediction accuracy. This can be attributed to its robustness against noise and data variation, as it identifies the most optimal decision boundary within the input space. This characteristic allows the SVM to maintain predictive reliability, even when some variables are removed.

These findings align with previous research suggesting that SVM models, due to their capacity to utilize different kernel functions, are well-equipped to handle complex, nonlinear data patterns, even after dimensionality reduction(Dilmi, 2022; Khan et al., 2022). Furthermore, the results of Chawishborwornworng et al. (2024) emphasized that integrating statistical techniques such as PCA with machine learning models enhances the reliability of water quality assessments and allows for the development of more accurate composite indices.

3.4 Sensitivity analysis results

Fig. 2 displays the results of the sensitivity analysis for the Iranian water quality index in the Zohreh River. This analysis assessed how variations in each water quality parameter influenced the overall index by calculating correlation coefficients.

An increase in dissolved oxygen significantly improves water quality, highlighting its critical role as a key determinant of aquatic health. Conversely, electrical conductivity, fecal coliform, total hardness, and turbidity exhibit strong negative correlations with the water quality index, indicating that pollutants associated with organic matter, microbial contamination, and suspended solids have a substantial adverse impact on water quality.

Ammonium showed a slight positive correlation with the water quality index, possibly due to its role in certain biological processes, although this relationship requires further investigation. In contrast, COD and BOD_5 also contributed to a decline in water quality, though their effects were less pronounced compared to the aforementioned parameters.

Nitrate and phosphate demonstrated relatively weak negative correlations, suggesting a moderate impact on water quality. The influence of pH was minimal and slightly positive.

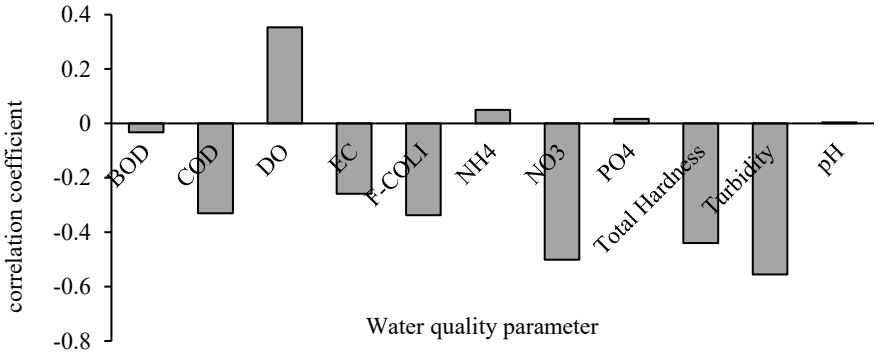

Fig. 2 The order of sensitivity of IWQI in Zohreh River to water quality parameters

Overall, the results indicate that the water quality of the Zohreh River is primarily degraded by organic, bacterial, and suspended solid pollutants. Therefore, strategies aimed at reducing these contaminants and enhancing dissolved oxygen levels could lead to significant improvements in the river's water quality.

4. Conclusion

In the current study, we employed two distinct methods for prediction: All 11 water quality parameters were fed into the AI models in the initial phase. In the second stage, with the help of PCA, potential pollution sources and important parameters were identified and entered as input to the AI models. The main results of this research are:

1. The results of KMO and Bartlett tests indicated that the available water quality data were suitable for conducting principal component analysis.

2. The use of the PCA method resulted in the creation of five principal components with an eigenvalue greater than 1.

3. Based on the factor loadings, 8 parameters were identified as the main factors affecting water quality in the study area.

4. The support vector machine model with $R^2 = 0.911$, RMSE = 0.047, and MAE = 0.035 provided the best prediction of the water quality index with the original data.

5. After applying PCA and optimizing the input parameters, the support vector machine model had a better prediction of the water quality index with $R^2 = 0.889$, RMSE = 0.052, and MAE = 0.038.

Despite providing significant results, this study also has limitations, including data limitations, which should be considered. This study only uses data from a specific region (Zohreh River) over a 10-year period. This limitation can reduce the accuracy of the generalizability of the models to other regions and time periods. It is recommended that this study be repeated in other geographical regions and for longer periods to increase the accuracy and generalizability of the proposed models.

Furthermore, in addition to PCA, other methods such as factor analysis or deep learning techniques can be investigated to reduce the dimensionality of the data.

Statements and Declarations

Acknowledgements

We are grateful to the Research Council of Shahid Chamran University of Ahvaz for financial support (GN SCU.WE1402.47794).

Data availability

The data used in this research are provided in the text of the article.

Conflicts of interest

The author of this paper declared no conflict of interest regarding the authorship or publication of this paper.

Author contribution

A.H. Shakarami: Design, Analysis, and Interpretation of data, Writing- Original draft preparation, Visualization; L. Divband Hafshejani: Conceptualization, Methodology, Design, Revision of the manuscript and Editing; P. Tishehzan: Design, Revision of the manuscript and Editing; and H. Abdolabadi: Analysis and Interpretation of data.

AI Use Declaration

This study did not incorporate artificial intelligence techniques; instead, all analyses and optimizations were conducted using conventional and widely accepted analytical methods.

References

Abdi, J., & Mazloom, G. (2022). Machine learning approaches for predicting arsenic adsorption from water using porous metal–organic frameworks. *Sci. Rep.*, *12*(1), 16458. DOI: 10.1038/s41598-022-20762-y.

Akter, T., Jhohura, F. T., Akter, F., Chowdhury, T. R., Mistry, S. K., Dey, D., Barua, M. K., Islam, M. A., & Rahman, M. (2016). Water Quality Index for measuring drinking water

quality in rural Bangladesh: a cross-sectional study. *J. Health Popul. Nut*., *35*, 1-12. DOI: 10.1186/s41043-016-0041-5.

Channabasava, S., Patil, S., Rajashekhar, M., & Vijaykumar, K. (2017). Study of ground water quality of Raichur in Industrial zone in concern to effect of industrial discharges on water quality. *Int. J. Adv. Eng. Manage. Sci., 3*(2), 239768. DOI: 10.24001/ijaems.3.2.10.

Chawishborwornworng, C., Luanwuthi, S., Umpuch, C., & Puchongkawarin, C. (2024). Bootstrap approach for quantifying the uncertainty in modeling of the water quality index using principal component analysis and artificial intelligence. *J. Saudi Soc. Agri. Sci., 23*(1), 17-33. DOI: 10.1016/j.jssas.2023.08.004.

Das, J., Karmaker, N., & Khan, R. A. (2021). Reasons and consequences of river water pollution and their remediation: In context of Bangladesh. *GSC Adv. Res. Rev., 7*(1), 023-034. DOI: 10.30574/gscarr.2021.7.1.0066.

Dattamudi, S., Kalita, P. K., Chanda, S., Alquwaizany, A., & S. Sidhu, B. (2020). Agricultural nitrogen budget for a long-term row crop production system in the Midwest USA. *Agronomy, 10*(11), 1622. DOI: 10.3390/agronomy10111622.

Dilmi, S. (2022). A combined water quality classification model based on kernel principal component analysis and machine learning techniques. *Desalin. Water Treat., 279*, 61-67. DOI: 10.5004/dwt.2022.29069.

Divband Hafshejani, L., Naseri, A. A., Moradzadeh, M., Daneshvar, E., & Bhatnagar, A. (2022). Applications of soft computing techniques for prediction of pollutant removal by environmentally friendly adsorbents (case study: the nitrate adsorption on modified hydrochar). *Water Sci. Technol.*, *86*(5), 1066-1082. DOI: 10.2166/wst.2022.264.

Guo, J., Peng, Y., Wang, S., Ma, B., Ge, S., Wang, Z., Huang, H., Zhang, J., & Zhang, L. (2013). Pathways and organisms involved in ammonia oxidation and nitrous oxide emission. *Critic. Rev. Environ. Sci. Technol., 43*(21), 2213-2296. DOI: 10.1080/10643389.2012.672072.

Hajirezaie, S., Wu, X., & Peters, C. A. (2017). Scale formation in porous media and its impact on reservoir performance during water flooding. *J. Nat. Gas Sci. Eng., 39*, 188-202. DOI: 10.1016/j.jngse.2017.01.019.

Ibrahim, A., Ismail, A., Juahir, H., Iliyasu, A. B., Wailare, B. T., Mukhtar, M., & Aminu, H. (2023). Water quality modelling using principal component analysis and artificial neural network. *Mar. Pollut. Bull., 187*, 114493. DOI: 10.1016/j.marpolbul.2022.114493.

Jamshidzadeh, Z., Ehteram, M., & Shabanian, H. (2023). A new hybrid model for predicting water quality parameters. *Ain Shams Eng. J., 102510.* DOI: 10.1016/j.asej.2023.102510.

Ji, X., Dahlgren, R. A., & Zhang, M. (2016). Comparison of seven water quality assessment methods for the characterization and management of highly impaired river systems. *Environ. Monit. Assess., 188*, 1-16. DOI: 10.1007/s10661-015-5016-2.

Karpagavalli, M. S., Ramachandran, A., & Palanivelu, K. (2019). Spatial and temporal variations of water quality in Pallikaranai wetland, Chennai, India. *Int. J. Global Environ. Issues, 18*(1), 86-106. DOI: 10.1504/IJGENVI.2019.098911.

Khan, M. S. I., Islam, N., Uddin, J., Islam, S., & Nasir, M. K. (2022). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *J. King Saud Univ. Comput. Inf. Sci, 34*(8), 4773-4781. DOI: doi.org/10.1016/j.jksuci.2021.06.003.

Khuzestan Water and Power Organization. (2021). Zohreh River water quality data for Firoozabad and Sooyreh stations, 2011-2021 (Excel file). Ahvaz: Khuzestan Water and Power Organization.

Maurya, B. M., Yadav, N., Amudha, T., Satheeshkumar, J., Sangeetha, A., Parthasarathy, V., Iyer, M., Yadav, M. K., & Vellingiri, B. (2024). Artificial intelligence and machine learning algorithms in the detection of heavy metals in water and wastewater: Methodological and ethical challenges. *Chemosphere, 353*, 141474. DOI: 10.1016/j.chemosphere.2024.141474.

Mukherjee, I., Singh, U. K., & Chakma, S. (2022). Evaluation of groundwater quality for irrigation water supply using multi-criteria decision-making techniques and GIS in an agroeconomic tract of Lower Ganga basin, India. *J. Environ. Manage, 309*, 114691. DOI: 10.1016/j.jenvman.2022.114691.

Nong, X., Shao, D., Xiao, Y., & Zhong, H. (2019). Spatio-temporal characterization analysis and water quality assessment of the South-to-North Water Diversion Project of China. *Int. J. Environ. Res. Public Health, 16*(12), 2227. DOI: 10.3390/ijerph16122227.

Oruganti, R. K., Biji, A. P., Lanuyanger, T., Show, P. L., Sriariyanun, M., Upadhyayula, V. K., Gadhamshetty, V., & Bhattacharyya, D. (2023). Artificial intelligence and machine learning tools for high-performance microalgal wastewater treatment and algal biorefinery: A critical review. *Sci. Total Environ, 876*, 162797. DOI: 10.1016/j.scitotenv.2023.162797.

Panigrahi, S., Acharya, B. C., Panigrahy, R. C., Nayak, B. K., Banarjee, K., & Sarkar, S. K. (2007). Anthropogenic impact on water quality of Chilika lagoon RAMSAR site: a statistical approach. *Wetl. Ecol. Manag, 15*, 113-126. DOI: 10.1007/s11273-006-9017-3.

Roy, B. N., Roy, H., Rahman, K. S., Mahmud, F., Bhuiyan, M. M. K., Hasan, M., Bhuiyan, A.-A. K., Hasan, M., Mahbub, M. S., & Jahedi, R. M. (2024). Principal component analysis incorporated water quality index modeling for Dhaka-based rivers. *City Environ. Interact*, 23, 100150. DOI: 10.1016/j.cacint.2024.100150.

Shil, S., Singh, U. K., & Mehta, P. (2019). Water quality assessment of a tropical river using water quality index

(WQI), multivariate statistical techniques and GIS. *Appl. Water Sci., 9*, 1-21. DOI: 10.1007/s13201-019-1045-2.

Tripathi, M., & Singal, S. K. (2019). Use of principal component analysis for parameter selection for development of a novel water quality index: a case study of river Ganga India. *Ecol. Indic, 96*, 430-436. DOI: 10.1016/j.ecolind.2018.09.025.

Vellingiri, J., Kalaivanan, K., Shanmugaiah, K., & Bai, F. J. J. S. (2024). AO-SVM: a machine learning model for predicting water quality in the cauvery river. *Environ. Res. Communic., 6*(7), 075025. DOI: 10.1088/2515-7620/ad6061.