*Advances in Artificial Intelligence, Big Data and Algorithms*
*G. Grigoras and P. Lorenz (Eds.)*

*doi:10.3233/FAIA230931*

# Comparative Study of Algorithms Used in Water Pollution Prevention and Control

Yuwen WANG[a1], Hengming LIU[b2]
[a]*School of Environment, Hohai University, Nanjing, China*
[b]*School of Marine Technology and Environment, Dalian Ocean University, Dalian, China*

**Abstract**—Water pollution prevention and control is crucial to ensure the safety of water environment and human health, and various types of algorithms play an important role in it. We introduce the history and algorithm overview of various algorithms in water pollution prevention and control, analyze the current research status and recent research results in this field, compare and evaluate the advantages and disadvantages of various algorithms, and focus on the following algorithms: neural network, convolutional neural network, decision tree, random forest, naive Bayes, SVM, K-Means, and AdaBoost. Through the comparative analysis of these algorithms, we hope to provide a more effective method for water pollution prevention and control.

**Keywords**—algorithms, water pollution, prevention and control, comparative study

## 1. INTRODUCTION

As an essential element for sustaining life on Earth, water is critical to the health of both humans and ecosystems. The detrimental effects of water pollution on water quality, ecological balance and human well-being cannot be overemphasized. Therefore, the prevention and control of water pollution is of great importance for the protection of water resources, the maintenance of ecological balance and the protection of human health. The implementation of effective water pollution prevention and control measures can reduce the discharge of pollutants, improve water quality, protect aquatic biodiversity, and provide people with safe drinking water. In the field of water pollution prevention and control, a variety of algorithms have been widely used and play an important role. By comparing the advantages and disadvantages of different algorithms, decision makers can make an informed choice of the appropriate algorithm and optimize water pollution prevention and control strategies. In addition, the comparative study of different algorithms can promote the progress of algorithms, improve the effectiveness and efficiency of water pollution prevention and control, and promote the improvement and development of algorithms. Choosing an appropriate algorithm not only solves optimization problems mathematically and statistically, but also helps to automatically analyze and identify patterns in data. In addition, algorithms can autonomously learn,

[1] Corresponding Author: Yuwen WANG, School of Environment, Hohai University; e-mail: 973331413@qq.com
[2] Hengming LIU, School of Marine Technology and Environment, Dalian Ocean University; e-mail: 963028347@qq.com

summarize, and provide solutions to problems, using data and past experience to optimize the performance of computer programs[1]. This improves the accuracy and reliability of water pollution prevention and control, simplifies the decision-making process, and stimulates innovation in water pollution prevention and control technologies.

## 2. ALGORITHMS USED IN WATER POLLUTION PREVENTION AND CONTROL

The use of various algorithms in water pollution prevention and control dates back to the 1960s. Initially, statistical learning algorithms such as linear regression and least squares were used to monitor and predict water pollution. These algorithms were used to construct predictive and classification models to predict changes in water quality and identify pollution sources[2,3,4,5]. However, these models relied heavily on manual feature selection and design, resulting in problems of interpretability and overfitting. In the 1980s, the emergence of neural network technology led to the exploration of neural network algorithms for water quality analysis, contaminant identification, anomaly detection, and optimization[6,7,8,9,10]. These algorithms exhibited adaptive and nonlinear fitting capabilities, thereby improving accuracy and reliability. However, they faced computational and training time challenges when dealing with large datasets and complex models. Since the 1990s, new algorithms such as support vector machines, decision trees, random forests, and deep learning have been applied to water pollution control[11,12,13,14]. These algorithms have strong adaptive, nonlinear fitting, and model interpretation capabilities, making them suitable for tasks such as water quality prediction, anomaly detection, classification, and optimization. The main algorithms currently used in this field include deep learning algorithms (e.g., convolutional neural networks, recurrent neural networks, long- and short-term memory networks, self-encoders), rule-based algorithms (e.g., decision trees, random forests, gradient boosting decision trees), algorithms based on probabilistic models (e.g., simple Bayes, support vector machines, Gaussian mixture models), clustering-based algorithms (e.g., K-means, hierarchical clustering, density clustering), and algorithms based on embedded learning (e.g., AdaBoost, XGBoost, LightGBM). As shown in Table 1.

**Table 1.** Algorithms used in water pollution prevention and control

| Classes of algorithms | Name of the algorithm |
|---|---|
| deep Learning | convolutional neural networks<br>recurrent neural networks<br>long and short-term memory networks<br>self-encoders, etc. |
| rule-based | decision trees<br>random forests<br>gradient boosting decision trees, etc. |
| based on probabilistic models | plain Bayesian<br>support vector machine<br>Gaussian mixture models, etc. |
| based on clustering | K-Means<br>hierarchical clustering<br>density clustering, etc. |
| integrated learning based | AdaBoost, XGBoost, LightGBM, etc. |

## 3. Status of Research on Algorithms Used in Water Pollution Prevention and Control

In recent years, the field of water pollution prevention and control has witnessed a surge in the application of algorithms, thanks to advances in artificial intelligence, various algorithms, and data collection technologies. This trend has led to the increasing popularity and maturity of algorithmic approaches in addressing water pollution challenges. A wide range of different algorithms have been extensively researched and implemented in this field, with remarkable results and achievements.

### *3.1 Plain Bayesian*

In a study conducted by Abuzir[15], three different algorithms, namely J48, Plain Bayesian and MLP models were used for water quality classification. The classification accuracy of each model was analyzed and compared, taking into account the different number of features in the dataset. The experimental results showed that the simple Bayesian algorithm is well suited for small sample classification. In another study by Manaf[16], IoT technology was integrated with artificial intelligence classification to collect and classify water quality parameters such as temperature, pH, and turbidity. The simple Bayesian algorithm was used as the classifier, resulting in an accuracy rate of over 96.89%.

### *3.2 Regression Analysis*

In a study conducted by Koranga[17], eight algorithms were used for regression analysis and nine algorithms were used for classification analysis in the context of prediction and classification of water quality pollution in Nainital Lake. The results indicated that the random forest algorithm showed better performance in regression prediction. For data classification, three algorithms namely stochastic gradient descent, random forest and support vector machine were found to be more effective. In another study by Shakhari[18], a classification method was proposed for water quality data to classify and detect water pollution. The proposed method was compared with two existing classification methods, namely C-4.5 and logistic regression. The experimental results show the effectiveness of the proposed method in this field.

### *3.3 K-Nearest Neighbor (KNN)*

In a study conducted by Ramadhani[19] in Riau Province, Indonesia, the problem of water quality monitoring and classification was addressed. An improved K-nearest neighbor (MKNN) algorithm was used, which achieved a classification accuracy of 85.1%. The detection attributes used in the study included BOD, COD, $NH_3$, fecal coliform, and total coliform. Mohurle[20] focused on the issue of drinking water contamination. The study analyzed the current status of drinking water quality and the fundamentals of the K-Nearest Neighbor (KNN) classifier. The KNN classifier was used to predict and validate the accuracy of available parameters for drinking water quality indicators. Motevalli[21] used an enhanced regression tree (BRT) and the K-nearest neighbor (KNN) algorithm to generate nitrate pollution vulnerability maps for groundwater. The objective was to assess different sources of nitrate pollution and the severity of the pollution.

### 3.4 Random Forest

Grbčić[22] proposed a pollution source identification method for water supply networks based on the random forest algorithm. In the method, numerous pollution scenarios with randomly selected pollution parameters were simulated to obtain water quality time series data from sensors. The results showed that the proposed method achieved high accuracy in locating potential pollution sources. Sakaa[23] used the Minimum Optimization Support Vector Machine (SMO-SVM) and Random Forest (RF) algorithms to assess surface water quality in Algerian rivers. The study aimed at predicting the water quality values. The results showed that the RF model outperformed the SMO-SVM model in most cases and had higher prediction accuracy for water quality indicators. Victoriano[24] used dissolved oxygen (DO), pH, biochemical oxygen demand (BOD), total suspended solids (TSS), nitrate, phosphate, and coliform as attribute values to construct a predictive model using the random forest decision tree algorithm. The model was applied to predict the pollution level of rivers in the MMORS River in Bulacan Province, Philippines. The results showed a high accuracy of 99.38% in predicting water pollution levels.

### 3.5 Support Vector Machine (SVM)

Muhammad[25] focused on the increasing rate of water pollution in Malaysia. The study analyzed water quality indicators and verified and compared two different kernels of Support Vector Machine (SVM) classifier. The classification accuracy of water quality pollutants reached 91.67%. Cao [26] introduced the genetic algorithm variation factor into the particle swarm optimization (PSO) algorithm and optimized the hyperparameters using the adaptive PSO algorithm with the least squares support vector machine (LS-SVM). A water quality classification evaluation model was established and the changing trend of water quality data in the next three days was predicted. This method was found to be faster and more accurate than the traditional back-propagation neural network algorithm. Mohammadpour[27] used feed-forward backpropagation (FFBP) and radial basis function (RBF) to construct a predictive model using support vector machine (SVM) for predicting water quality indicators in a constructed wetland environment. The research results indicated that SVM prediction performed as well or better than neural networks. Koranga[28] conducted an analysis and comparison of two types of kernel functions, namely radial basis kernel function and polynomial kernel function, used in LibSVM. The study evaluated the best combination of parameters by adjusting various parameters and achieved a water quality prediction accuracy of 99.43%. Wu[29] developed a support vector regression model based on time series and applied the support vector machine method for lake eutrophication assessment and water quality prediction. The performance of this model was found to be improved compared to existing algorithms.

### 3.6 Deep Learning

Sagan[30] addressed the limitations of satellite remote sensing technology in water quality monitoring by collecting water quality data from eight lakes and rivers in the Midwestern United States. The study compared deep learning algorithms with other methods and found that data-driven deep learning outperformed other methods in terms of water quality monitoring and decision-making effectiveness. Khullar[31] proposed a deep learning based Bi-LSTM model (DLBL-WQA) for water quality prediction in the Yamuna River in India. The model used missing value interpolation, generated feature maps from input data,

improved the Bi-LSTM architecture, and optimized the loss function to reduce the training error. Comparative experiments showed that the model achieved better prediction accuracy compared to traditional methods. Singha[32] introduced a deep learning (DL) prediction model for heavy metal contamination in Indian groundwater. The model optimized the activation function of neurons and rectified linear units in the hidden layer, and applied small-batch gradient descent to ensure training data. The results showed lower prediction error compared to the ANN model, and the model helped to mitigate the overfitting problem. Solanki[33] analyzed and compared the application of deep learning algorithms with other unsupervised learning algorithms in water treatment for river pollution near Nashik, India. The research showed that deep learning showed higher accuracy and robustness in this context. Baek[34] proposed a deep learning algorithm model that combined convolutional neural network (CNN) and long short-term memory (LSTM) to simulate and predict water level and water quality data in the Nakdong River Basin in South Korea. Experimental results demonstrated the effectiveness of the proposed method in simulating water level and predicting water quality. Wang[35] addressed the nonlinear and nonstationary dynamic characteristics of water quality monitoring data by proposing a double-attention mechanism long short-term memory network (LSTM) water quality prediction method based on time series dependence and feature correlation. The method was applied to the abnormal detection of urban river water quality. Wan[36] proposed a deep learning model called SOD-VGG-LSTM to improve the accuracy of water quality prediction for non-point source (NPS) pollution. The model outperformed the RNN model in extreme value prediction accuracy.

### *3.7 Plain Fusion of Multiple Algorithms*

Elkiran[37] analyzed and compared different integration techniques for Backpropagation Neural Network (BPNN), Adaptive Neuro-Fuzzy Inference System (ANFIS), Support Vector Machine (SVM) and Linear Autoregressive Integral Moving Average (ARIMA) models, applied to water quality modeling of Yamuna River in India, it proves that the performance of ANFIS model is better. Ladjal[38] proposed a data fusion method using Principal Component Analysis (PCA) in combination with Support Vector Machines (SVM), Artificial Neural Networks (ANN) and decision templates for water quality monitoring in the Tilesdit dam area in Algeria. The achieved classification accuracy was 98%. Yusri[39] combined Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost) algorithms to develop a classification algorithm for predicting water quality classification (WQC). When the number of samples reached 2000, the average classification accuracy rate reached 90%.

## 4. COMPARISON OF ADVANTAGES AND DISADVANTAGES OF VARIOUS ALGORITHMS

In the realm of water pollution prevention and control, a multitude of algorithms have been developed, each possessing distinct advantages and demonstrating significant advancements. However, these algorithms also exhibit certain limitations, primarily in terms of model complexity, over-fitting or under-fitting, model interpretability, sample size requirements, and the automation of parameter optimization. Consequently, the selection of an appropriate algorithm should be contingent upon the specific characteristics of the problem at hand. For instance, convolutional neural networks prove to be a superior choice when confronted with extensive datasets, whereas decision trees

or support vector machines may be more suitable for scenarios with limited data availability. Furthermore, the combination of different algorithms can be employed to construct more intricate models, thereby enhancing the accuracy and reliability of predictions or classifications. A comprehensive comparison of the advantages and disadvantages of commonly utilized algorithms in this domain is presented in Table 2.

**Table 2.** Comparison of advantages and disadvantages of algorithms

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Large scale data handling | √ | √ | ○ | √ | √ | √ | √ | √ |
| Nonlinear problems handling | √ | √ | √ | √ | × | √ | × | √ |
| Noise sensitivity handling | ○ | ○ | × | √ | ○ | ○ | × | × |
| Low structural complexity | ○ | ○ | √ | √ | √ | √ | √ | √ |
| Good model interpretability | × | ○ | √ | √ | √ | √ | √ | √ |
| Less computational effort | × | ○ | √ | ○ | √ | √ | √ | ○ |
| Able to avoid overfitting | ○ | ○ | × | ○ | ○ | ○ | × | ○ |
| Automatic feature selection | √ | √ | √ | ○ | × | × | × | × |
| Results are globally optimal | × | × | × | × | × | √ | × | √ |
| High classification accuracy | √ | ○ | ○ | ○ | ○ | √ | ○ | ○ |

Note: 1-neural network, 2-convolutional neural network, 3- decision tree, 4- random forest, 5- naive Bayes, 6-SVM, 7- K-Means, 8- AdaBoost. "√" means good, "○" means medium, "×" means poor.

## 5. Existing Problems and Challenges

The widespread implementation of various algorithms has significantly improved the efficiency and accuracy of water pollution prevention and control. However, several challenges and problems remain that require continuous research and improvement in the following areas:

(1) Inadequate model universality: The sample data used to train these models are typically collected from specific regions or environments, limiting their applicability to other regions or environmental conditions. This lack of universality hinders the generalizability of the models.

(2) Inadequate model interpretability: Many algorithms used in water pollution prevention and control are considered black-box models, making it difficult to understand and explain their decision-making processes. This lack of interpretability raises concerns about the transparency and trustworthiness of the models.

(3) Limited real-time capability and operability: Currently, many algorithms require significant computational resources and time to train and predict, limiting their real-time capability and practicality. The extended time requirements hinder the timely response and operational efficiency of the models.

(4) Suboptimal data quality: Variability in the quality and completeness of sample data is a significant challenge. Some data may be missing, inaccurate, or incomplete, compromising the effectiveness and accuracy of the algorithms used.

## 6. Conclusion and Suggestions

The continuous development of new technologies, such as AI(artificial intelligence), has greatly enhanced the application of various algorithms in water pollution prevention and

control, showing immense potential and promising prospects. Looking ahead, several future directions for the application of algorithms in this field can be considered:

(1) Improving the interpretability of algorithms: There is a need to optimize and refine existing algorithms[40,41] to improve their interpretability. This will allow models to be more easily understood, trusted, and accepted by stakeholders involved in water pollution prevention and control efforts.

(2) Improving real-time capability and usability: Hardware acceleration techniques, such as the use of GPUs, can be used to improve computational efficiency and prediction speed. In addition, distributed computing can be used to improve the scalability and usability of algorithms, enabling real-time decision making and response.

(3) Multimodal data fusion: The integration of different types of data, such as remote sensing data, geographic information data, and water quality monitoring data, can be explored through multimodal data fusion techniques. This approach aims to improve the accuracy and reliability of algorithms by exploiting the complementary information provided by different data sources.

(4) Improve data quality: As sensor and monitoring technologies continue to advance, the quality and integrity of collected sample data is expected to improve. This improvement will facilitate the training and application of various algorithms, leading to more accurate and robust results.

(5) Automated Decision Making and Optimization: The use of automated decision making and optimization techniques will increase. For example, reinforcement learning algorithms can be used to optimize the decision-making process in water pollution prevention and control. This will allow systems to learn and adapt autonomously to different environmental conditions, improving overall efficiency and effectiveness.

## REFERENCES

[1] Wei Quanli. On Machine Learning in the Science of Artificial Intelligence [J]. Journal of Ningxia Institute of Technology: Natural Science, 1995(3):74-76.
[2] Ridoutt B G. The use of analogue and digital computers for water pollution studies[J]. Water Research, 1967, 1(4): 271-286.
[3] McKay G. Automatic classification of river quality[J]. Water Research, 1973, 7(3):429-438.
[4] Buckley E N, Gobas F A. Multiple linear regression applied to the prediction of water pollution by organic compounds[J]. Chemosphere, 1978, 7(10): 821-828.
[5] Gupta H V, Kocis V J. The application of pattern recognition techniques to the study of water pollution[J]. Water Research, 1979, 13(9): 797-808.
[6] Tomlinson R E, Sansalone J J. Artificial neural networks for the prediction of stormwater pollutant concentrations[J]. Water Research, 1988, 22(3): 351-359.
[7] Carpenter G A, Grossberg S. Self-organization of stable category recognition codes for analog input patterns[J]. Applied Optics, 1987, 26(23): 4919-4930.
[8] Reed T M, Marks II R. Neural network analysis of water quality data[J]. Journal of Water Resources Planning and Management, 1988, 114(4): 440-457.
[9] Gardner Jr. Experiences with neural networks for water quality analysis[C] //International Conference on Artificial Neural Networks. Berlin, Heidelberg :Springer, 1988: 279-284.
[10] El-Fadel M, El-Fadl K, Hashisho J. Application of artificial neural networks to the analysis of water quality data[J]. Environmental Technology Letters, 1989, 10(12):1133-1144.
[11] Reckhow D A, Simpson J M. Artificial neural networks for the prediction of wastewater treatment plant performance[J]. Water Research, 1993, 27(5):735-743.
[12] Perona J J, Diwekar U M, Badrinarayan H. A hybrid neuro-symbolic approach to modeling in environmental engineering[J]. Environmental Monitoring and Assessment, 1993, 26(3):231-254.
[13] Babovic V, Keijzer M. Neuro-fuzzy modeling of water treatment processes[J]. Journal of Water Resources Planning and Management, 1996, 122(6):419-427.

[14] Hart W E, Hjaltason G R, Minsker B S. Machine learning using probabilistic networks for water resource analysis and management[J]. Journal of Water Resources Planning and Management, 1997, 123(4):202-212.
[15] Abuzir S Y, Abuzir Y S. Machine learning for water quality classification[J]. Water Quality Research Journal, 2022, 57(3): 152-164.
[16] Manaf K, Kaffah F M, Mulyana E, et al. Implementation of Naïve Bayes algorithm in IoT-based water cleanliness monitoring system[C]//IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2021, 1098(4): 042007.
[17] Koranga M, Pant P, Kumar T, et al. Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand[J]. Materials Today: Proceedings, 2022: 1706-1712.
[18] Shakhari S, Banerjee I. A multi-class classification system for continuous water quality monitoring[J]. Heliyon, 2019, 5(5): e01822.
[19] Ramadhani D, Afdal M, Rahmawita M. The Classification Status of River Water Quality in Riau Province Using Modified K-Nearest Neighbor Algorithm with STORET Modeling and Water Pollution Index[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 1783(1): 012020.
[20] Mohurle S, Devare M. A study of KNN classifier to predict water pollution index[J]. Computing in Engineering and Technology: Proceedings of ICCET 2019, 2020: 457-466.
[21] Motevalli A, Naghibi S A, Hashemi H, et al. Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater[J]. Journal of cleaner production, 2019, 228: 1248-1263.
[22] Grbčić L, Lučin I, Kranjčević L, et al. Water supply network pollution source identification by random forest algorithm[J]. Journal of Hydroinformatics, 2020, 22(6): 1521-1535.
[23] Sakaa B, Elbeltagi A, Boudibi S, et al. Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin[J]. Environmental Science and Pollution Research, 2022, 29(32): 48491-48508.
[24] Victoriano J M, Lacatan L L, Vinluan A A. Predicting river pollution using random forest decision tree with GIS model: A case study of MMORS, Philippines[J]. Int. J. Environ. Sci. Dev, 2020, 11(1): 36-42.
[25] Muhammad Z, Jailani N A J, Leh N A M, et al. Classification of Drinking Water Quality using Support Vector Machine (SVM) Algorithm[C]//2022 IEEE 12th International Conference on Control System, Computing and Engineering (ICCSCE). IEEE, 2022: 75-80.
[26] Cao S, Wang S. Design of River Water Quality Assessment and Prediction Algorithm[C]// 2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC). IEEE, 2018: 1625-1631.
[27] Mohammadpour R, Shaharuddin S, Chang C K, et al. Prediction of water quality index in constructed wetlands using support vector machine[J]. Environmental Science and Pollution Research, 2015, 22: 6208-6219.
[28] Koranga M, Pant P, Pant D, et al. SVM model to predict the water quality based on physicochemical parameters[J]. International Journal of Mathematical, Engineering and Management Sciences, 2021, 6(2): 645.
[29] Wu Guozheng. Application of support vector machine in lake eutrophication evaluation and water quality prediction[D]. Inner Mongolia Agricultural University,2008.
[30] Sagan V, Peterson K T, Maimaitijiang M, et al. Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing[J]. Earth-Science Reviews, 2020, 205: 103187.
[31] Khullar S, Singh N. Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation[J]. Environmental Science and Pollution Research, 2022, 29(9): 12875-12889.
[32] Singha S, Pasupuleti S, Singha S S, et al. Effectiveness of groundwater heavy metal pollution indices studies by deep-learning[J]. Journal of Contaminant Hydrology, 2020, 235: 103718.
[33] Solanki A, Agrawal H, Khare K. Predictive analysis of water quality parameters using deep learning[J]. International Journal of Computer Applications, 2015, 125(9): 0975-8887.
[34] Baek S S, Pyo J, Chun J A. Prediction of water level and water quality using a CNN-LSTM
combined deep learning approach[J]. Water, 2020, 12(12): 3399.
[35] Wang Lixiang. Research on urban river water quality anomaly detection method based on multi-indicator time series data[D]. Zhejiang University,2021.
[36] Wan H, Xu R, Zhang M, et al. A novel model for water quality prediction caused by non-point sources pollution based on deep learning and feature extraction methods[J]. Journal of Hydrology, 2022, 612: 128081.
[37] Elkiran G, Nourani V, Abba S I. Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach[J]. Journal of Hydrology, 2019, 577: 123962.
[38] Ladjal M, Bouamar M, Brik Y, et al. A decision fusion method based on classification models for water quality monitoring[J]. Environmental Science and Pollution Research, 2023, 30(9): 22532-22549.

[39] Yusri H I H, Hassan S L M, Halim I S A, et al. Water Quality Classification Using SVM And XGBoost Method[C]//2022 IEEE 13th Control and System Graduate Research Colloquium (ICSGRC). IEEE, 2022: 231-236.

[40] Wang Hongzhi. Research on Network Traffic Classification Based on PCA Feature Selection and Optimized ECOC[D]. Dalian Jiaotong University,2014.

[41] Wang Hongzhi, Liu Zhen, Li Donghui. Network traffic forecasting method based on multi-classification support vector machine[J]. Science and Technology Review, 2014,32(17):60-63.