Eco-Environment & Health 1 (2022) 107–116

Contents lists available at ScienceDirect

Eco-Environment & Health

journal homepage: www.journals.elsevier.com/eco-environment-and-health

Review

A review of the application of machine learning in water quality evaluation

Mengyuan Zhu, Jiawei Wang, Xiao Yang, Yu Zhang, Linyu Zhang, Hongqiang Ren, Bing Wu *, Lin Ye *

State Key Laboratory of Pollution Control and Resource Reuse, School of Environment, Nanjing University, Nanjing 210023, China

ARTICLE INFO

Keywords:
Machine learning
Water quality
Evaluation
Prediction

ABSTRACT

With the rapid increase in the volume of data on the aquatic environment, machine learning has become an important tool for data analysis, classification, and prediction. Unlike traditional models used in water-related research, data-driven models based on machine learning can efficiently solve more complex nonlinear problems. In water environment research, models and conclusions derived from machine learning have been applied to the construction, monitoring, simulation, evaluation, and optimization of various water treatment and management systems. Additionally, machine learning can provide solutions for water pollution control, water quality improvement, and watershed ecosystem security management. In this review, we describe the cases in which machine learning algorithms have been applied to evaluate the water quality in different water environments, such as surface water, groundwater, drinking water, sewage, and seawater. Furthermore, we propose possible future applications of machine learning approaches to water environments.

1. Introduction

With rapid economic development, wastewater containing various pollutants is generated, posing serious threats to natural water environments. Thus, various water pollution control measures have been developed. To a large extent, water quality analysis and evaluation have substantially improved the efficiency of water pollution control [1]. To date, many methods have been developed to monitor and assess water quality worldwide, such as the multivariate statistical method [2], fuzzy inference [3], and the water quality index (WQI) [4]. For evaluating water quality, although most water quality parameters can be monitored according to the procedures defined in the relevant standards, the final water quality evaluation results may widely vary owing to the choice of parameters [5]. Considering all water quality parameters is unrealistic because it is not only expensive and technically difficult but also fails to deal with the variability in water quality [6]. However, in recent years, with the advances in machine learning methods, an increasing number of researchers believe that vast amounts of data can be successfully captured and analyzed to meet the complex and large-scale water quality evaluation requirements.

In machine learning, a branch of artificial intelligence, algorithms are used to analyze data and attempt to mine potential patterns in the data to predict new information [7,8]. As a new data analysis and processing method, machine learning has been widely used in many fields owing to its high precision, flexible customization, and convenient extensibility [9]. Complex nonlinear relational data can be easily handled with machine learning, which facilitates the discovery of the underlying mechanisms [10]. The excellent adaptability of machine learning has demonstrated its potential as a tool in the fields of environmental science and engineering in recent years. Therefore, more accurate evaluation results can be expected despite the complexity of using machine learning for water quality analysis and evaluation [11].

Water types, including drinking water, wastewater, groundwater, surface water, seawater, and freshwater, are complex [12]. These different types of water have different characteristics, leading to considerable challenges for research on their quality. Based on the findings of previous studies, machine learning can be an effective approach to addressing these challenges. As such, in this review, we summarize the advantages and disadvantages of commonly used machine learning algorithms and discuss the applications and performance of machine learning in surface water, groundwater, drinking water, wastewater, and seawater (Fig. 1).

2. Overview of machine learning

As a powerful data analysis approach, machine learning is widely used to identify patterns or make predictions based on big data generated from different scenarios. Before machine learning is applied in practice, data acquisition, appropriate algorithm selection, model training, and model validation need to be conducted. Among these processes, the choice of

* Corresponding authors.
E-mail addresses: bwu@nju.edu.cn (B. Wu), linye@nju.edu.cn (L. Ye).

https://doi.org/10.1016/j.eehl.2022.06.001
Received 1 February 2022; Received in revised form 19 May 2022; Accepted 1 June 2022
Available online 8 July 2022

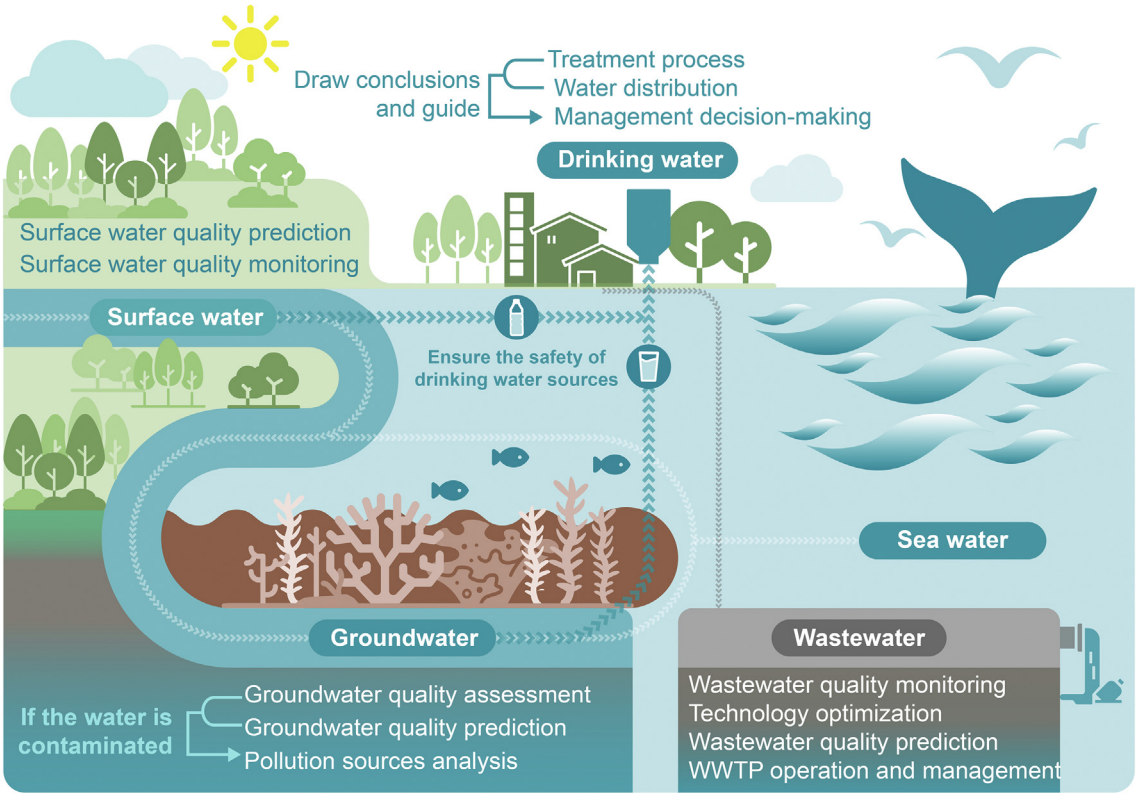

Fig. 1. Machine learning is widely used in water systems. WWTP, wasterwater treatmemt plant.

algorithm is crucial. Supervised and unsupervised learning are two main classes of machine learning technologies [13]. The main difference between these two classes is the presence of labels in the datasets. Supervised learning deduces predictive functions from the labeled training datasets. Each training instance includes input values and expected output values. Supervised learning algorithms try to identify the relationships between the input and output values and generate a predictive model to predict the result based on the corresponding input data. Supervised learning can be used for data classification and regression, and a variety of algorithms, including linear regression, artificial neural network (ANN), decision tree (DT), support vector machine (SVM), naive Bayes, k-nearest neighbor (KNN), random forest (RF), etc. have been developed.

In contrast, unsupervised learning is usually used to handle data without labels, solving various problems in pattern recognition based on unlabeled training datasets. Unsupervised learning classifies the training data into different categories according to their different characteristics, mainly based on dimensionality reduction and clustering [14]. However, the number of categories is uncertain, nor is the meaning of each category clear. Therefore, unsupervised learning is usually used for classification and association mining [13]. Principal component analysis (PCA), K-means, etc. are the commonly used unsupervised machine learning algorithms.

In addition, reinforcement learning, which refers to the generalization ability of a machine to correctly answer unlearned problems, is regarded as another class of machine learning algorithms. However, compared with the other two machine learning classes, it is seldom applied in the field of water environment.

3. Application of machine learning for different water environments

Many researchers have used machine learning to solve problems in various aspects of water treatment and management systems (Fig. 2), including real-time monitoring, prediction, pollutant source tracking, pollutant concentration estimation, water resource allocation, and water treatment technology optimization.

3.1. Applications in surface water

Municipal and industrial wastewater generated by human activities has become the main factor in deteriorating water quality in urban areas [15]. The application of machine learning in surface water quality research has become a hotspot [16,17]. A series of surface water quality prediction and analysis methods have been developed (Table 1). Many efforts have been devoted to optimizing machine learning models and improving their prediction accuracy.

Data acquisition is a fundamental step in developing machine learning models. Both integrated and periodic water quality monitoring results can be used as benchmarks in water system management. Traditional environmental monitoring methods are widely applied by environmental agencies. However, for *in situ* monitoring, the traditional methods are limited by realistic difficulties [34]. Remote sensing technologies can meet the needs of real-time and large-scale water quality monitoring, and can also be used to reveal the migration and distribution characteristics of pollutants that are difficult to detect using conventional methods. Sagan et al. [29] found that experiment-based machine learning allowed for sophisticated optimization based on the combination of real-time monitoring sensor data and satellite data, and the accuracies of the partial least squares (PLS) regression, support vector regression (SVR), and deep neural network (DNN) models were all higher than those of traditional models. However, some water quality variables, such as the concentration of pathogens, cannot be directly measured by remote sensing, as they are not optically active or lack high-spatial-resolution hyperspectral data, but can be estimated indirectly using other measurable data [29]. Wu et al. [30] developed an attentional neural network based on a convolutional neural network (CNN) to identify clean and polluted water on the basis of water images. They conducted several comparison experiments on a water surface image dataset and verified the effectiveness of this attentional neural network. The advantage of CNN is that the reflectance image is taken as the direct input without any feature engineering and parameter tuning. Due to equipment or human reasons, some of the acquired data will inevitably be missing, wrong, or damaged, leading to a sparse matrix

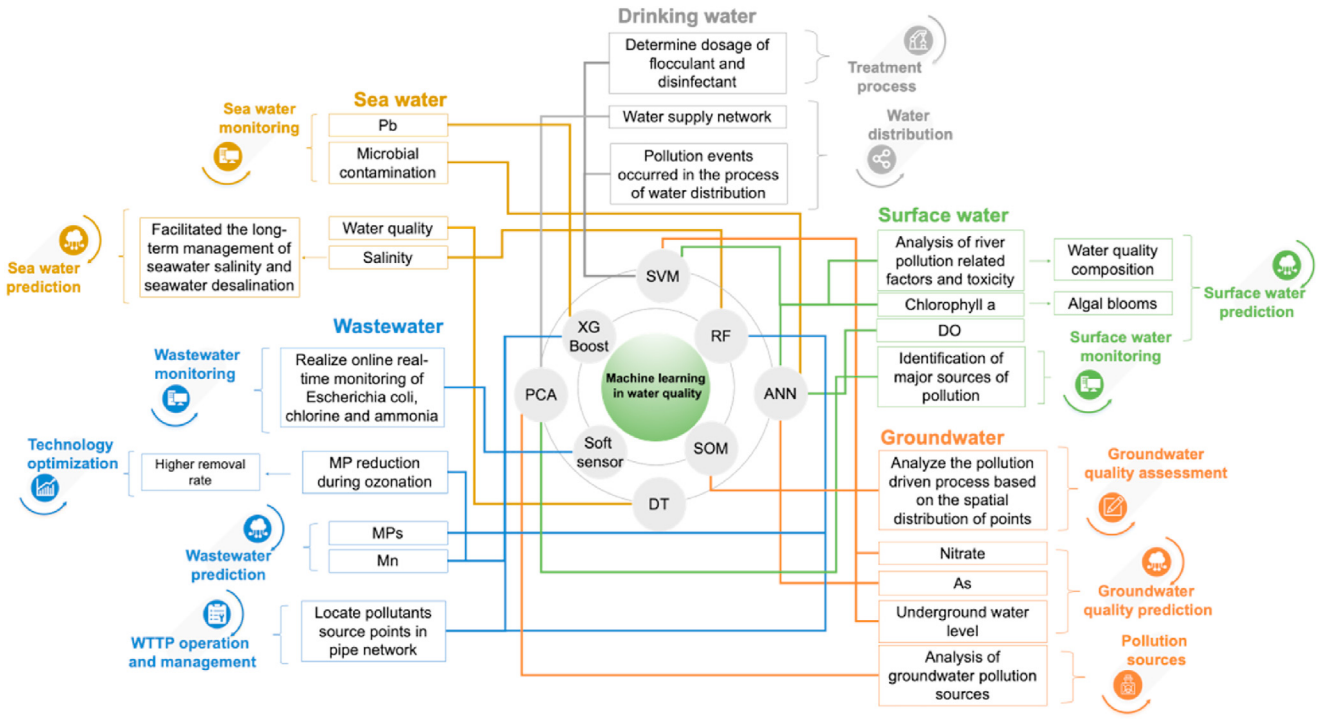

Fig. 2. Applications of different machine learning algorithms in different water treatment and management systems. SVM, support vector machine; RF, random forest; ANN, artificial neural network; SOM, self-organizing map; DT, decision tree; PCA, principal component analysis; XGBoost, extreme gradient boosting; DO, dissolved oxygen; MP, micropollutant.

and poor performance in model applications. When this happens, data cleaning, another key step in machine learning applications, becomes important. Data cleaning can be handled in different ways, including not using the set of data directly, using averages or medians, or using a combination of machine learning and matrix completion methods to supplement the raw data [35]. Ma et al. [22] proposed an approach combining DNN and deep matrix factorization (deep MF) to predict the biological oxygen demand (BOD). They verified the validity and reliability of the method using New York Harbor waters as a case study. Data cleaning improves the data quality and thus the accuracy of machine learning model applications.

For machine learning applications, the prediction accuracy is generally related to two aspects, i.e., the model selection and the quality of the training dataset. ANN and SVM have provided excellent performance in the prediction of water quality components [24,23]. In some cases, SVM may produce higher prediction accuracy and show higher generalization ability than ANN. One reason is that the optimization of model parameters in a neural network is unstable, so the accuracy of the ANN is remarkably affected by nonlinear disturbances [36]. Moreover, SVM uses an upper bound on the generalization error rather than reducing the training error [37]; therefore, it is more effective than ANN in minimizing this error. Due to the complex dynamic changes of river systems over time, the most effective way to manage rivers is to monitor water quality in real time, or to make predictions based on other data when monitoring conditions are insufficient. Researchers have verified that long short-term memory (LSTM) networks and bootstrapped wavelet neural networks (BWNN) can handle fluctuating and nonseasonal time-series water-quality data [18,19]. Some traditional statistical theories, such as the autoregressive integrated moving average (ARIMA) model, can be applied to time series prediction, but they are basically linear models [38]. This is inferior to the BWNN model, which is inspired by the self-adaption in the learning process of the ANN and the time-frequency properties of the wavelet basis functions [18], and the LSTM model, a type of recurrent neural network structure, which learns directly from time-series data [39]. LSTM and BWNN can well identify the nonlinear relationship between variables and their predicted variables, and transfer useful information from the past to the future.

The prediction accuracy of machine learning models also depends on the features used to train the models. Redundant variables will reduce the inverse power and accuracy of the model, and increase the complexity. Dissolved oxygen (DO) is one of the most widely concerned surface water quality parameters, which directly reflects the status of the aquatic ecosystem and its ability to sustain aquatic organisms. The linear polynomial neural network (PNN) model was used to predict DO concentration in the Danube River. Among 17 water quality parameters, temperature, pH, BOD, and phosphorus concentration were found to be the most important features affecting the prediction accuracy [20]. For DO concentration prediction in St. John's River, USA, among the five input features (chloride, NO_x, total dissolved solids, pH, and water temperature), pH and NO_x are strongly correlated with DO and can affect the prediction accuracy [21]. These findings are consistent with those obtained by Chen et al. [25], who reported that input parameters affected the prediction performance of the model. In addition to regular water parameters, eutrophication is another concern in surface water quality prediction. Based on the adaptive neuro-fuzzy inference system (ANFIS) model, Ly et al. [28] found that algal blooms were caused by the combined interaction of nutrients, organic matter, and environmental elements. Park et al. [27] used meteorological data and weekly water quality data to predict the concentration of chlorophyll-a in two reservoirs in the U.S., and found that SVM and ANN had approximately similar prediction accuracies. The addition of meteorological factors considerably improved the prediction accuracy. Regional hydrological and socioeconomic factors can also be added to the machine learning model so that the results may provide stronger support for the comprehensive management of the regional water environment.

The performance of a machine learning model also depends on its architecture, so analyzing the logical structure of algorithms is also a key step in the successful application of machine learning. Compared with other traditional neural network models, the advantage of PNN in determining the key model parameters mentioned above is that the

Table 1
Application of machine learning models in surface water.

Task	Algorithms	Sample size	Input parameters	Evaluation results	Reference
DO prediction	BWNN, ANN, ARIMA, BANN	370	DO	BWNN > BANN > WNN > ANN > ARIMA	[18]
DO prediction	LSTM	236	DO	The model performed well at 74% of sites (NSE ≥ 0.4)	[19]
DO prediction	PNN	1912	Cl^-, alkalinity, BOD, PO_4–P, COD, pH, temperature, NO_3–N, Ca^{2+}, P, Mg^{2+}, and EC	Good interpolation performance (R^2 = 0.82)	[20]
DO prediction	CCNN	232	DO and water quality parameters (e.g., Cl, NO_x, TDS, pH, temperature)	R^2 = 0.825 RMSE = 0.550	[21]
BOD prediction	DNN, SVR, RF	32323	Latitude, longitude, time, site actual depth, sea state, degree of turbulence at sea, wind speed, DO, temperature, salinity, total coliform, light penetration in water, chlorophyll-a, polychlorinated biphenyls plate count, NO_x–N, PO_4–P, NH_3–N, TP, pH, TSS, EC, sample depth, density, and transparency	DNN is 19.20%–25.16% lower RMSE than traditional models	[22]
EC, HCO_3^-, SO_4^{2-}, Cl, TDS, Na^+, Mg^{2+}, Ca^{2+} prediction	SVM, ANN	All data since 1960	Temperature, pH, EC, HCO_3^-, SO_4^{2-}, Cl, TDS, Na^+, Mg^{2+}, and Ca^{2+}	SVM > ANN	[23]
TN, TP prediction	SVM, ANN	660	River flow, temperature, flow travel time, rainfall, DO, TN, and TP	SVM > ANN	[24]
Water quality level prediction	DT, RF, DCF, and 10 other models	33612	pH, DO, CON_{Mn}, and NH_3–N	DT, RF, and DCF provide better predictive performance	[25]
TRP, NO_3–N, TP, NH_4–N) prediction	RF	21657	EC, turbulence, temperature, DO, pH, chlorophyll-a, and flow rate	Compared with the linear model, RMSE decreased by 60.1%	[26]
Chlorophyll-a prediction	SVM, ANN	357	Chlorophyll-a, PO_4–P, NH_3–N, NO_3–N, temperature, solar radiation, and wind speed	SVM > ANN	[27]
Algal bloom prediction	ANFIS	896	COD, BOD, TOC, TSS, TP, DTP, PO_4–P, TN, NO_3–N, NH_3–N, chlorophyll-a, temperature, precipitation, flowrate, DO, pH, EC, total coliform, and fecal coliform	ANFIS performed best in both quantitative and classification problems	[28]
Hyperparameter selection optimization	SVR	223	BGA-PC, chlorophyll-a, DO, EC, fDOM, turbidity, and pollution sediments	BGA-PC (accuracy = 0.77), chlorophyll-a (0.78), TSS (0.81), fDOM (−), turbidity (0.55) and DO (−)	[29]
Water pollution monitoring	Attention neural network	1000	Water images	The resolution accuracy of clean water was 71.2%, and that of polluted water was 73.6%	[30]
Water pollution monitoring	CNN, SVM, RF	81	Landsat8 images and water quality level	CNN (accuracy = 97.12%) > SVM (96.89%) > RF (86.21%)	[31]
Heavy metal contamination assessment	PCA	42	Cu, Mn, Cr, Zn, Pb, Cd, Ni, and Co	Areas with heavy metal pollution were identified	[32]
WQI parameters selection	PCA	240	Temperature, DO, pH, EC, BOD, NO_3–N, fecal coliform, total coliform, turbidity, alkalinity, Cl, COD, NH_3–N, total Hardness, Ca^{2+}, Mg^{2+}, Na^+, TDS, and PO_4–P	Nine key parameters were DO, pH, EC, BOD, total coliform, Cl^-, Mg, SO_4^{2-}, and TDS	[33]

DO, Dissolved oxygen; BWNN, bootstrapped wavelet neural network; ANN, artificial neural network; ARIMA, autoregressive integrated moving average; BANN, bootstrapped artificial neural network; LSTM, long short-term memory; NSE, Nash-Sutcliffe efficiency; PNN, polynomial neural network; BOD, biological oxygen demand; COD, chemical oxygen demand; EC, electrical conductivity; CCNN, cascade correlation neural network; TDS, Tsinghua/Temporary DeepSpeed; RMSE, lower root mean square error; DNN, deep neural network; SVR, support vector regression; RF, random forest; SVM, support vector machine; TP, total phosphorus; TN, total nitrogen; TRP, total reactive phosphorus; TOC, total organic carbon; TSS, total suspended solids; DTP, dissolved total phosphorus; BGA-PC, blue-green algae phycocyanin, fDOM, fluorescent dissolved organic matter, CNN, convolutional neural network; PCA, principal component analysis; WQI, water quality index.

number of hidden neurons and layers of PNN is directly determined by data, saving the time for trial [20]. The DNN model used in BOD prediction has a 19.20%–25.16% lower root mean square error (RMSE) than the traditional machine learning model [22]. This is because there are multiple layers between the input and output layers of DNN, and it uses more advanced activation functions than ANN, which is more conducive to model convergence than the sigmoid used by traditional ANN and reduces the difficulty of training [40]. When predicting water quality over time, LSTM based on time series works relatively well. It is composed of three information gates, i.e., input gate, forget gate, and

output gate, as well as two states: cell state and hidden state, which control the transmission, forgetting, and storage of sequential information [41]. BWNN can also achieve this goal but requires a sufficient number of input features to ensure accuracy [18].

3.2. Applications in groundwater

Groundwater is an important source of drinking water. As such, ensuring the safety of groundwater is essential to human health. Machine learning has extensive application prospects in groundwater analysis, including the assessment and prediction of groundwater quality and pollution sources.

In recent years, multivariate statistical analysis methods have been widely applied to analyze groundwater quality. Among them, PCA and cluster analysis are frequently used [42]. In addition, machine learning algorithms, such as SVM, DT, RF, and ANN, have also been applied for groundwater quality assessment. Related studies on groundwater quality have mainly been designed to compare the evaluation efficiency of different machine learning algorithms to identify suitable ones for specific problems. For example, Jeihouni et al. [43] compared five data mining algorithms, that is, ordinary decision tree, RF, chi-square automatic interaction detector, and iterative dichotomizer 3, to identify the key parameters affecting groundwater in semiarid areas and to determine their effect on high-quality groundwater areas in Tabriz City, Iran. Lee et al. [44] evaluated the spatial pattern of urban groundwater quality in Seoul, South Korea, by combining a self-organizing neural network and fuzzy c-means clustering. They divided the groundwater samples into three groups using a self-organizing map algorithm according to different pollution degrees and analyzed the pollution-driven process based on the spatial distribution of groundwater groups. In addition, geographic information system techniques have been frequently used to generate groundwater quality maps to more accurately identify groundwater pollution [44,45].

The complex hydrogeological conditions of groundwater, compared with those of surface water, pose additional difficulties when predicting quality change tendencies. Some researchers have used machine learning to evaluate existing regional data on a large scale or predict future water quality. Agrawal et al. [46] estimated and predicted the WQI of groundwater using a combination of particle swarm optimization (PSO) and SVM, which illustrated the feasibility of integrating these methods for groundwater prediction [47]. The action of a single pollutant in groundwater, particularly nitrate and arsenic, can be predicted. Arabgol et al. [48] adopted an SVM to predict the concentration and distribution of nitrate in groundwater. Sajedi Hosseini et al. [49] calculated the risk of nitrate pollution in groundwater using boosted regression tree, multivariate discriminant analysis, and SVM, and concluded that the risk of nitrate pollution in Iran's Lennart Plain is high. Ransom et al. [50] used machine learning to predict groundwater nitrate levels throughout the United States and found that national-scale groundwater quality can also be predicted through machine learning models. Cho et al. [51] used an ANN to predict the potential for arsenic pollution of groundwater in Cambodia, Laos, and Thailand. In addition, groundwater levels can also be predicted. Mohapatra et al. [52] predicted the groundwater level using three machine learning methods (ANFIS, DNN, and SVM) and found that DNN was most suitable for seasonal forecast and had higher accuracy and efficiency. Yadav et al. [53] adopted ensemble modeling to predict the groundwater level in Indian cities and achieved an accuracy rate of 85%.

The analysis of pollution sources is beneficial for ensuring the safety of groundwater. PCA and clustering methods are widely used in current studies. Celestino et al. [45] used PCA to reduce dimensionality, and after the dimensionality reduction data were clustered by K-means, natural variations and anthropogenic sources of hydro-geochemistry were analyzed. Chen et al. [54] applied multivariate statistical analysis and PCA to identify the key factors affecting changes in groundwater quality.

A data mining decision tree is usually used to explore groundwater quality and resources. This algorithm can learn the relationships between input variables and corresponding output variables and represent each relationship by specific rules. RF has advantages in its performance and ability to generalize rules to identify areas of high-quality groundwater for drinking. Since the continuous data set is more suitable for groundwater law induction than the discrete data set, the highest performance (accuracy of 97.10%) obtained by RF based on continuous data sets provides good decisions for groundwater resource planning and management [43]. In terms of the prediction of groundwater quality index, current research innovation tends to use the integrated model, which can combine multiple weak learners into one strong learner and improve the prediction performance. Boosting is a great integration approach. However, when combining different good models to generate models with smaller variance, attention should be paid to reducing overfitting.

3.3. Application in drinking water

Machine learning has been widely applied in drinking water treatment and management systems, including drinking water source management, treatment processes, water distribution, and decision making. Drinking water is typically derived from surface water or groundwater. The evaluation and prediction of source water quality based on machine learning can assist in the early warning and control of pollution. Bouamar et al. [55] evaluated the possibility of multisensor-based ANN and SVM algorithms for dynamic water quality monitoring in 2007. Both models showed acceptable recognition rates in terms of outputting the recognition rates of the two categories of water. Compared with ANN, SVM was more stable. Wu et al. [56] proposed an adaptive frequency analysis method using drinking water quality datasets from four cities in Norway. Their findings provided a basis for drinking water quality risk warning, management, and decision-making at an early stage. In addition, Liu et al. [57] used LSTM and DNN to predict time series data and established a water quality prediction model that could predict the water quality in the next six months with relatively good accuracy. Arnon et al. [58] used an SVM to implement a new scheme that could predict pollution events under unknown conditions using ultraviolet absorption. The detection rates for all four datasets obtained by the SVM were high, and the error rates were low. Most studies have been based on chemical or physical parameters, while microbial parameters have seldom been considered, especially for *Escherichia coli* (*E. coli*) [59]. The amounts of coagulants and disinfectants in drinking water plants can also be predicted using machine learning. Owing to its simple structure and strong robustness, the SVM algorithm is popular in flocculation and disinfection construction schemes. According to the residual free chlorine predicted by an SVM model, Wang et al. [60] proposed a predictive control scheme model for chemical dosage, which was more effective than the traditional proportional-integral-derivative feedback control.

The importance of drinking water supply has led scientists to focus on the normal operation of urban water supply system facilities, fault monitoring, and disaster prediction [61]. Because of the complexity of water supply systems, water from drinking water treatment plants that meet the required standards may be re-contaminated during transportation, which can be assessed using indicators of biological stability and then disinfected [62]. The cluster analysis can identify the differences in water quality between water networks [63]. Further, Tian et al. [64] used cluster analysis to identify the contribution of mixed water sources to aluminum (Al) residues in metropolitan drinking water supply systems, including Al migration and seasonal changes in drinking water. Brester et al. [65] accurately determined water quality via casting using an RF algorithm. Water supply pipeline bursting results in large-scale water losses and microbial and chemical pollution during transportation. Deep learning models can predict the potential bursting location, but with considerable uncertainty [66]. Rayaroth et al. [67] proposed a random decision tree bagging classifier based on the shuffled frog-leaping optimization technique to identify water leakage with a

minimum number of sensors at an optimal position in a water distribution network. Pipe lifespan is an important factor in water-supply management. Almheiri et al. [68] proposed an advanced meta-learning model based on a neural network and found that residual chlorine is one of the key factors affecting the service life of pipelines. In addition, pollution events that occur in water distribution systems can be predicted using an SVM algorithm [69]. Park et al. [70] quantitatively analyzed the impact of disasters on water supply systems by combining PCA, the analytic hierarchy process, RF, and extreme gradient boosting (XGBoost) models. However, the practicability of this method is limited owing to the difficulty of real-time data collection.

Water production capacity has become a factor limiting regional development and population increases. Zhang et al. [71] combined the advantages of both ANN and genetic algorithms and established a hybrid statistical model that can predict the performance of drinking water treatment plants. The hybrid statistical model can predict changes in water production under various parameter fluctuation scenarios, proving itself a useful tool for quickly and reasonably adjusting water treatment systems. Cardoso et al. [72] proposed an automatic monitoring framework for urban water management based on time-series clustering and found that the water demand was high from 3 to 6 a.m. in summer due to the irrigation of municipal and public gardens. To address the problem of short-term water demand prediction, Guo et al. [73] developed a gated recurrent unit network and set a 15-min time step to successfully predict the water demand in the next 15 min and 24 h. Ghiassi et al. [74] employed a dynamic artificial neural network (DAN2), focused time-delay neural network, KNN to predict the daily, weekly, and monthly water demands in Tehran. Among the three models, DAN2 exhibited the best performance. The prediction accuracies of the daily, weekly, and monthly models were 96%, 99%, and 98%, respectively. The application of machine learning can help solve the imbalance in water-supply systems. Accurately predicting water demand is a promising approach for effectively allocating available water resources.

In summary, the ANN and SVM are widely applied in the field of drinking water, especially in the application of large dimensions. The short computing time of the training phase (a few seconds) enables them to be applied to dynamic monitoring systems to monitor drinking water quality and safety in real time. With the increase in training techniques, ANN's recognition rate has improved significantly, despite being very sensitive to noise. In contrast, SVM is robust to noise, so the integration of ANN and SVM is gradually attracting the attention of researchers [55].

3.4. Applications in wastewater

In terms of wastewater treatment, machine learning is widely used for water quality monitoring and prediction, technology optimization, and wastewater treatment plant (WWTP) operation and management. Domestic and industrial wastewater contains various pollutants, calling for an evaluation of water quality before treatment [75]. By combining multiresolution analysis with PCA, Rosen et al. [76] provided a tool more sensitive than PCA for monitoring sewage indicators at multiple scales. The collection, processing, and analysis of big data largely rely on real-time online monitoring. A soft sensor based on the black box model was proposed for online, real-time monitoring of *E. coli* [77], which showed that the concentrations of *E. coli* substantially increased after heavy rainfall, possibly due to urban runoff resuspending sewer sediment [78,79]. Combining soft sensors with an ANN can be used to overcome the challenges of the high cost and complexity of WWTP operation and maintenance, and for online monitoring of chlorine and ammonia in real time [80,81]. Qin et al. [82] used a boosting-iterative predictor weighting-partial least squares (Boosting-IPW-PLS) method and multiple sensors to establish a water quality monitoring system equipped with a UV spectrometer and turbidimeter to monitor the chemical oxygen demand (COD) and total suspended solids. Boosting-IPW-PLS method suppressed variables unrelated to water quality by assigning small weights, and established a prediction model for wastewater quality based on weighted variables. Their test results showed that this system performed well in monitoring water quality, with a high correlation coefficient between the predicted value and the actual value.

The analysis of historical data to optimize wastewater treatment systems is a practical application of machine learning. Fang et al. [83] simulated anaerobic, anoxic, and oxic conditions with an SVM and an adaptive genetic algorithm to save land space by reducing the volume of the anoxic tank. In addition, machine learning has been used to optimize tertiary wastewater treatment, such as reverse osmosis (RO), nanofiltration (NF), ozonation, and adsorption. Cha et al. [84] applied an RF to predict micropollutant (MP) reduction during ozonation and achieved a higher removal efficiency. Machine learning based on a high-resolution fluorescence excitation-emission matrix can provide more accurate results by better calculating the complex nonlinear relationship between organic properties and oxidizer exposure. A model predicting MP removal by membrane separation is essential for the design and selection of appropriate membranes. Teychene et al. [85] used DT to reveal the specific sequence through which MP is removed by RO and NF, and found that particle size exclusion, electric repulsion, and adsorption were the main separation mechanisms used by NF and RO. In addition, XGBoost can be used to predict the removal efficiency of MP in RO and NF [86]. Sigmund et al. [87] developed two neural-network-based models that enabled practitioners to select the appropriate adsorbent for a given pollutant. Based on the results of the above cases, machine learning methods can be widely applied in the advanced treatment of wastewater containing MPs and new pollutants in the future.

Machine learning has provided predictive information for a more thorough understanding and analysis of water treatment. ANN can effectively be used to solve complex nonlinear environmental problems, especially in pollutant removal [88]. Bayat Varkeshi et al. [89] successfully constructed an ANN model that could predict COD and BOD concentrations in wastewater outflow. At present, water quality prediction models primarily aim at determining the levels of specific pollutants. For instance, after determining the photodegradation rate of tetracycline (TC) under various practical conditions, Abdi et al. [90] established CatBoost, which could accurately predict TC removal using a metal–organic framework. Baek et al. [91] constructed three different models, using RF, SVM, and ANN, to predict the removal of five different MPs. All the models were verified, and the results produced by RF proved to be the most accurate. Biological indicators can also be predicted using machine learning. Bayes approaches, including both naive Bayes and seminaïve Bayes networks, have been applied to predict pathogen removal efficiency and represent the association among pathogen reduction, operating conditions, and monitoring parameters [92]. Roguet et al. [93] used RF to predict the abundance of *Clostridiales* and *Bacteroidales* in wastewater. RF has been applied to fill the gap in the development of comprehensive evaluation and calculation methods for predicting fecal pollution sources, helping inhibit the spread of waterborne diseases [94].

The effluent quality of WWTPs can be affected by many factors, and the operation management and maintenance of WWTPs can be challenging when costs need to be controlled [95]. Therefore, machine learning can be further employed because it can provide WWTP managers with opportunities to reduce costs and improve their operations. Gomez-Munoz et al. [96] used Bayes' fundamental theorem to estimate the proportions of various costs of a WWTP, which helped with the management of the construction, regulation, and operating procedures. Toxic pollutants discharged into sewage networks can affect the regular operation of WWTPs. To prevent such cases, XGBoost and RF were used to identify pollutants and locate their source points in a wastewater network [97]. Normally, flow-measurement sensors are installed in sewage pipes. However, measurement instability caused by impurities, corrosion, and high turbidity can lead to inaccurate measurements. Deep learning can potentially improve measurement accuracy in various situations by enhancing existing sensors [98]. Ji et al. [98] used typical failures described in a historical dataset of the actual sensor settings.

Once a fault was detected, the model could adjust the process and ensure the normal operation of the WWTP.

3.5. Application of machine learning in marine environments

Seawater pollution is becoming a serious problem affecting the Earth's ecosystems. Monitoring seawater pollutants with the help of machine learning provide a new solution to these issues. Bhagat et al. [99] used XGBoost to establish a lead-prediction algorithm, and trained the model using historical monitoring data from the Bramble and Deception Bay stations in Australia. They found the model performed well in selecting input parameters and predicting water quality. Goncalves et al. [100] proposed a waste mapping program based on RF and an automatic unmanned aerial vehicle system that could automatically monitor coastal plastic waste. An ensemble machine learning approach with a two-layered learning structure was proposed to predict the concentration of coastal microbial pollution in beach water [101]. To improve the accuracy of antibiotic resistance gene (ARG) prediction in beach water, Jang et al. [102] adopted an LSTM-CNN model and successfully predicted a single ARG. Mancia et al. [103] identified differentially expressed genes in dolphins exposed to marine pollutants using machine learning classification algorithms. In addition, many researchers have focused on developing surveillance technologies for algal blooms that can lead to severe contamination. Ghatkar et al. [104] trained the XGBoost model with the spectral characteristics of different water types and algal blooms to identify and distinguish the algae that cause algal blooms. Du et al. [105] evaluated the water quality along with the North Yellow and Bohai Seas using a hierarchical cluster analysis water quality evaluation method based on the Mahalanobis distance. In conclusion, machine learning methods can identify the types of seawater pollutants, determine the concentration and distribution of pollutants, and provide a relevant analysis of the status of marine organisms.

Seawater quality monitoring is essential for protecting marine ecosystems. Many researchers have applied machine learning methods to monitor seawater quality. In 2001, Alshehri et al. [106] proposed a near-shore water quality prediction model based on KNN. Sheng et al. [107] integrated BPNN, SVR, and LSTM models to establish a water quality prediction method, which significantly improved water quality prediction accuracy. Zhou et al. [108] proposed a water quality prediction method on an improved grey regression analysis algorithm and LSTM on the basis of the multivariate correlation and the time-series characteristics provided in water quality information. Du et al. [109] analyzed the data collected by a geosynchronous ocean color imager and from 1240 water quality sampling points along the coast of Zhejiang, using a water quality assessment method with a geographic neural network weighted regression model. Additionally, 75% of the world's population will face a freshwater crisis by 2050 [110]. Desalinated seawater is an important source of freshwater in areas with extreme water shortages. However, some seawater desalination difficulties remain, with the low efficiency and reliability of desalination systems being the major obstacles. Alshehri et al. [106] used a CNN model and transfer learning to classify salt particles with different concentrations in water to improve the seawater treatment performance of water treatment plants. Chawla et al. [111] predicted the salinity and development trend of the Salton Sea using regression and machine learning algorithms such as linear regression, RF, SVM, and LSTM, which facilitated the long-term management of seawater salinity and seawater desalination.

The single water quality prediction model has been thoroughly studied in the previous literature, and the integrated model has come into view in recent years. Different models have different mechanisms in the face of different input features, leading to different predictive performances. The integration model proposed by Sheng et al. [107] preferentially selects a classifier. When new data are entered, the prediction model that best fits the data is first selected before making the prediction. This is a model selection algorithm based on input features, while the XGBoost method proposed by Bhagat et al. [99] can screen input features and select 5–9 out of 21 features to be integrated with ANN and other application methods, and the information they learn will not be lost in the model training stage. XGBoost model is a promising modeling algorithm with the advantages of high accuracy and fast speed, but XGBoost as a feature selection algorithm depends on sample size.

4. Concluding remarks

Machine learning has been widely used as a powerful tool to solve problems in the water environment because it can be applied to predict water quality, optimize water resource allocation, manage water resource shortages, etc. Despite this, several challenges remain in fully applying machine learning approaches in this field to evaluate water quality: (1) Machine learning is usually dependent on large amounts of high-quality data. Obtaining sufficient data with high accuracy in water treatment and management systems is often difficult owing to the cost or technology limitations. (2) As the conditions in real water treatment and management systems can be extremely complex, the current algorithms may only be applied to specific systems, which hinders the wide application of machine learning approaches. (3) The implementation of machine learning algorithms in practical applications requires researchers to have certain professional background knowledge.

To overcome the above-mentioned challenges, the following aspects should be considered in future research and engineering practices: (1) More advanced sensors, including soft sensors, should be developed and applied in water quality monitoring to collect sufficiently accurate data to facilitate the application of machine learning approaches. (2) The feasibility and reliability of the algorithms should be improved, and more universal algorithms and models should be developed according to the water treatment and management requirements. (3) Interdisciplinary talent with knowledge in different fields should be trained to develop more advanced machine learning techniques and apply them in engineering practices.

Declaration of competing interests

The authors have declared no conflicts of interests.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (22022604 and 21976087), the Natural Science Foundation of Jiangsu Province (BK20200011), the Fundamental Research Funds for the Central Universities (021114380170), and the Excellent Research Program of Nanjing University (ZYJH005).

References

[1] J. Pan, T. Leng, Y. Liu, Shifosi reservoir water environmental assessment based on grey clustering, Prog. Environ. Sci. Eng. 857 (2013) 610–613. http://dio.org/10.4028/www.scientific.net/AMR.610-613.857.

[2] R. Alam, Z. Ahmed, S.M. Seefat, K.T.K. Nahin, Assessment of surface water quality around a landfill using multivariate statistical method, Sylhet, Bangladesh, Environ. Nanotechnol. Monit. Manag. 15 (2021), 100422, https://doi.org/10.1016/j.enmm.2020.100422.

[3] J.O. Oladipo, A.S. Akinwumiju, O.S. Aboyeji, A.A. Adelodun, Comparison between fuzzy logic and water quality index methods: a case of water quality assessment in Ikare community, Southwestern Nigeria, Environ. Chall. 3 (2021), 100038, https://doi.org/10.1016/j.envc.2021.100038.

[4] J. Wang, Z. Fu, H. Qiao, F. Liu, Assessment of eutrophication and water quality in the estuarine area of Lake Wuli, Lake Taihu, China, Sci. Total Environ. 650 (2019) 1392–1402, https://doi.org/10.1016/j.scitotenv.2018.09.137.

[5] B. Wang, Y. Wang, S. Wang, Improved water pollution index for determining spatiotemporal water quality dynamics: case study in the Erdao Songhua River Basin, China, Ecol. Indicat. 129 (2021), 107931, https://doi.org/10.1016/j.ecolind.2021.107931.

[6] F.D. Simoes, A.B. Moreira, M.C. Bisinoti, S.M.N. Gimenez, M.J.S. Yabe, Water quality index as a simple indicator of aquaculture effects on aquatic bodies, Ecol. Indicat. 8 (2008) 476–484, https://doi.org/10.1016/j.ecolind.2007.05.002.

[7] C. Ma, H.H. Zhang, X. Wang, Machine learning for Big Data analytics in plants, Trends Plant Sci. 19 (2014) 798–808, https://doi.org/10.1016/j.tplants.2014.08.004.

[8] Y. Liu, T. Zhao, W. Ju, S. Shi, Materials discovery and design using machine learning, J Materiomics. 3 (2017) 159–177, https://doi.org/10.1016/j.jmat.2017.08.002.

[9] N.K. Geetha, P. Bridjesh, Overview of machine learning and its adaptability in mechanical engineering, Mater. Today Proc. 611 (2020), https://doi.org/10.1016/j.matpr.2020.09.611.

[10] S. Messaoud, A. Bradai, S.H.R. Bukhari, P.T.A. Quang, O.B. Ahmed, M. Atri, A survey on machine learning in Internet of Things: algorithms, strategies, and applications, Internet of Things 12 (2020) 100314, https://doi.org/10.1016/j.iot.2020.100314.

[11] R.Z. Wang, J.H. Kim, M.H. Li, Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach, Sci. Total Environ. 761 (2021), 144057, https://doi.org/10.1016/j.scitotenv.2020.144057.

[12] S. Maxwell, One water: the need for more holistic thinking, analysis, and policy making in water, J. AWWA (Am. Water Works Assoc.) 107 (2015) 21–24, https://doi.org/10.5942/jawwa.2015.107.0048.

[13] M.W. Berry, A.H. Mohamed, B.W. Yap, Supervised and Unsupervised Learning for Data Science, Springer, Switzerland, 2019.

[14] L.F. Zhang, L.P. Zhang, B. Du, J.E. You, D.C. Tao, Hyperspectral image unsupervised classification by robust manifold matrix factorization, Inf. Sci. 485 (2019) 154–169, https://doi.org/10.1016/j.ins.2019.02.008.

[15] R. Mohammadpour, S. Shaharuddin, C.K. Chang, N.A. Zakaria, A. Ab Ghani, N.W. Chan, Prediction of water quality index in constructed wetlands using support vector machine, Environ. Sci. Pollut. Control Ser. 22 (2015) 6208–6219, https://doi.org/10.1007/s11356-014-3806-7.

[16] T.M. Tung, Z.M. Yaseen Tiyasha, A survey on river water quality modelling using artificial intelligence models: 2000–2020, J. Hydrol. 585 (2020), https://doi.org/10.1016/j.jhydrol.2020.124670.

[17] N. Sharma, R. Sharma, N. Jindal, Machine learning and deep learning applications-A vision, Global Transitions Proceedings 2 (2021) 24–28, https://doi.org/10.1016/j.gltp.2021.01.004.

[18] Y. Wang, T. Zheng, Y. Zhao, J. Jiang, Y. Wang, L. Guo, P. Wang, Monthly water quality forecasting and uncertainty assessment via bootstrapped wavelet neural networks under missing data for Harbin, China, Environ. Sci. Pollut. Control Ser. 20 (2013) 8909–8923, https://doi.org/10.1007/s11356-013-1874-8.

[19] W. Zhi, D. Feng, W.P. Tsai, G. Sterle, A. Harpold, C. Shen, L. Li, From hydrometeorology to river water quality: can a deep learning model predict dissolved oxygen at the continental scale? Environ. Sci. Technol. 55 (2021) 2357–2368, https://doi.org/10.1021/acs.est.0c06783.

[20] A. Siljic Tomic, D. Antanasijevic, M. Ristic, A. Peric-Grujic, V. Pocajt, A linear and non-linear polynomial neural network modeling of dissolved oxygen content in surface water: inter- and extrapolation performance with inputs' significance analysis, Sci. Total Environ. 610–611 (2018) 1038–1046, https://doi.org/10.1016/j.scitotenv.2017.08.192.

[21] M. Zounemat-Kermani, Y. Seo, S. Kim, M.A. Ghorbani, S. Samadianfard, S. Naghshara, N.W. Kim, V.P. Singh, Can decomposition approaches always enhance soft computing models? Predicting the dissolved oxygen concentration in the St. Johns River, Florida, Appl. Sci. 9 (2019), 122534, https://doi.org/10.3390/app9122534.

[22] J. Ma, Y. Ding, J.C.P. Cheng, F. Jiang, Z. Xu, Soft detection of 5-day BOD with sparse matrix in city harbor water using deep learning techniques, Water Res. 170 (2020), 115350, https://doi.org/10.1016/j.watres.2019.115350.

[23] A. Parsaie, A.H. Nasrolahi, A.H. Haghiabi, Water quality prediction using machine learning methods, Water Qual. Res. J. 53 (2018) 3–13, https://doi.org/10.2166/wqrj.2018.025.

[24] M. Liu, J. Lu, Support vector machine-an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? Environ. Sci. Pollut. Control Ser. 21 (2014) 11036–11053, https://doi.org/10.1007/s11356-014-3046-x.

[25] K. Chen, E. Al, Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data, Water Res. 171 (2020), 115454, https://doi.org/10.1016/j.watres.2019.115454.

[26] M. Castrillo, A.L. Garcia, Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods, Water Res. 172 (2020), 115490, https://doi.org/10.1016/j.watres.2020.115490.

[27] Y. Park, K.H. Cho, J. Park, S.M. Cha, J.H. Kim, Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea, Sci. Total Environ. 502 (2015) 31–41, https://doi.org/10.1016/j.scitotenv.2014.09.005.

[28] Q.V. Ly, X.C. Nguyen, N.C. Le, T.D. Truong, T.T. Hoang, T.J. Park, T. Maqbool, J. Pyo, K.H. Cho, K.S. Lee, et al., Application of machine learning for eutrophication analysis and algal bloom prediction in an urban river: a 10-year study of the Han River, South Korea, Sci. Total Environ. 797 (2021), 149040, https://doi.org/10.1016/j.scitotenv.2021.149040.

[29] V. Sagan, K.T. Peterson, M. Maimaitijiang, P. Sidike, J. Sloan, B.A. Greeling, S. Maalouf, C. Adams, Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing, Earth Sci. Rev. 205 (2020), 103187, https://doi.org/10.1016/j.earscirev.2020.103187.

[30] Y. Wu, X. Zhang, Y. Xiao, J. Feng, Attention neural network for water image classification under IoT environment, Appl. Sci. 10 (2020), 030909, https://doi.org/10.3390/app10030909.

[31] F. Pu, C. Ding, Z. Chao, Y. Yu, X. Xu, Water-quality classification of Inland lakes using landsat8 images by convolutional neural networks, Rem. Sens. 11 (2019), 141674, https://doi.org/10.3390/rs11141674.

[32] V. Kumar, A. Sharma, R. Kumar, R. Bhardwaj, A. Kumar Thukral, J. Rodrigo-Comino, Assessment of heavy-metal pollution in three different Indian water bodies by combination of multivariate analysis and water pollution indices, Hum. Ecol. Risk Assess. 26 (2018) 1–16, https://doi.org/10.1080/10807039.2018.1497946.

[33] M. Tripathi, S.K. Singal, Use of principal component analysis for parameter selection for development of a novel water quality index: a case study of river Ganga India, Ecol. Indicat. 96 (2019) 430–436, https://doi.org/10.1016/j.ecolind.2018.09.025.

[34] W. Li, H. Fang, G. Qin, X. Tan, Z. Huang, F. Zeng, H. Du, S. Li, Concentration estimation of dissolved oxygen in Pearl River Basin using input variable selection and machine learning techniques, Sci. Total Environ. 731 (2020) 139099, https://doi.org/10.1016/j.scitotenv.2020.139099.

[35] A.R.T. Donders, G.J.M.G. Van Der Heijden, T. Stijnen, K.G.M. Moons, Review: a gentle introduction to imputation of missing values, J. Clin. Epidemiol. 59 (2006) 1087–1091, https://doi.org/10.1016/j.jclinepi.2006.01.014.

[36] R.M. Balabin, E.I. Lomakina, Support vector machine regression (SVR/LS-SVM)-an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data, Analyst 136 (2011) 1703–1712, https://doi.org/10.1039/c0an00387e.

[37] K.J. Kim, Financial time series forecasting using support vector machines, Neurocomputing 55 (2003) 307–319, https://doi.org/10.1016/S0925-2312(03)00372-2.

[38] J.S. Jia, J.Z. Zhao, H.B. Deng, J. Duan, Ecological footprint simulation and prediction by ARIMA model—a case study in Henan Province of China, Ecol. Indicat. 10 (2010) 538–544, https://doi.org/10.1016/j.ecolind.2009.06.007.

[39] C.P. Shen, A transdisciplinary review of deep learning research and its relevance for water resources scientists, Water Resour. Res. 54 (2018) 8558–8593, https://doi.org/10.1029/2018WR022643.

[40] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Network. 61 (2015) 85–117, https://doi.org/10.1016/j.neunet.2014.09.003.

[41] C.P. Shen, E. Laloy, A. Elshorbagy, A. Albert, J. Bales, F.J. Chang, S. Ganguly, K.L. Hsu, D. Kifer, Z. Fang, et al., HESS opinions: incubating deep-learning-powered hydrologic science advances as a community, Hydrol. Earth Syst. Sci. 22 (2018) 5639–5656, https://doi.org/10.5194/hess-22-5639-2018.

[42] Z.L. Hildenbrand, D.D. Carlton Jr., B.E. Fontenot, J.M. Meik, J.L. Walton, J.T. Taylor, J.B. Thacker, S. Korlie, C.P. Shelor, D. Henderson, et al., A comprehensive analysis of groundwater quality in the Barnett Shale region, Environ. Sci. Technol. 49 (2015) 8254–8262, https://doi.org/10.1021/acs.est.5b01526.

[43] M. Jeihouni, A. Toomanian, A. Mansourian, Decision tree-based data mining and rule induction for identifying high quality groundwater zones to water supply management: a novel hybrid use of data mining and GIS, Water Resour. Manag. 34 (2019) 139–154, https://doi.org/10.1007/s11269-019-02447-w.

[44] K.J. Lee, S.T. Yun, S. Yu, K.H. Kim, J.H. Lee, S.H. Lee, The combined use of self-organizing map technique and fuzzy c-means clustering to evaluate urban groundwater quality in Seoul metropolitan city, South Korea, J. Hydrol. 569 (2019) 685–697, https://doi.org/10.1016/j.jhydrol.2018.12.031.

[45] A. Marín Celestino, D. Martínez Cruz, E. Otazo Sánchez, F. Gavi Reyes, D. Vásquez Soto, Groundwater quality assessment: an improved approach to K-means clustering, principal component analysis and spatial analysis: a case study, Water 10 (2018), 040437, https://doi.org/10.3390/w10040437.

[46] P. Agrawal, A. Sinha, S. Kumar, A. Agarwal, A. Banerjee, V.G.K. Villuri, C.S.R. Annavarapu, R. Dwivedi, V.V.R. Dera, J. Sinha, et al., Exploring artificial intelligence techniques for groundwater quality assessment, Water 13 (2021) 1172, https://doi.org/10.3390/w13091172.

[47] A. El Bilali, A. Taleb, Y. Brouziyne, Groundwater quality forecasting using machine learning algorithms for irrigation purposes, Agric. Water Manag. 245 (2021), https://doi.org/10.1016/j.agwat.2020.106625.

[48] R. Arabgol, M. Sartaj, K. Asghari, Predicting nitrate concentration and its spatial distribution in groundwater resources using support vector machines (SVMs) model, Environ. Model. Assess. 21 (2015) 71–82, https://doi.org/10.1007/s10666-015-9468-0.

[49] F. Sajedi Hosseini, A. Malekian, B. Choubin, O. Rahmati, S. Cipullo, F. Coulon, B. Pradhan, A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination, Sci. Total Environ. 644 (2018) 954–962, https://doi.org/10.1016/j.scitotenv.2018.07.054.

[50] K.M. Ransom, B.T. Nolan, P.E. Stackelberg, K. Belitz, M.S. Fram, Machine learning predictions of nitrate in groundwater used for drinking supply in the conterminous United States, Sci. Total Environ. (2021), 151065, https://doi.org/10.1016/j.scitotenv.2021.151065.

[51] K.H. Cho, S. Sthiannopkao, Y.A. Pachepsky, K.W. Kim, J.H. Kim, Prediction of contamination potential of groundwater arsenic in Cambodia, Laos, and Thailand using artificial neural network, Water Res. 45 (2011) 5535–5544, https://doi.org/10.1016/j.watres.2011.08.010.

[52] J.B. Mohapatra, P. Jha, M.K. Jha, S. Biswal, Efficacy of machine learning techniques in predicting groundwater fluctuations in agro-ecological zones of India, Sci. Total Environ. 785 (2021), 147319, https://doi.org/10.1016/j.scitotenv.2021.147319.

[53] B. Yadav, P.K. Gupta, N. Patidar, S.K. Himanshu, Ensemble modelling framework for groundwater level prediction in urban areas of India, Sci. Total Environ. 712 (2020), 135539, https://doi.org/10.1016/j.scitotenv.2019.135539.
[54] T. Chen, H. Zhang, C. Sun, H. Li, Y. Gao, Multivariate statistical approaches to identify the major factors governing groundwater quality, Appl. Water Sci. 8 (2018), https://doi.org/10.1007/s13201-018-0837-0.
[55] M. Bouamar, M. Ladjal, Evaluation of the performances of ANN and SVM techniques used in water quality classification, 14th IEEE International Conference on Electronics, Circuits and Systems (2007) 1047–1050, https://doi.org/10.1109/ICECS.2007.4511173.
[56] D. Wu, H. Wang, H. Mohammed, R. Seidu, Quality risk analysis for sustainable smart water supply using data perception, IEEE Trans.Sustain. Comput. 5 (2020) 377–388, https://doi.org/10.1109/tsusc.2019.2929953.
[57] P. Liu, J. Wang, A. Sangaiah, Y. Xie, X. Yin, Analysis and prediction of water quality using LSTM deep neural networks in IoT environment, Sustainability 11 (2019), 072058, https://doi.org/10.3390/su11072058.
[58] T. Asheri Arnon, S. Ezra, B. Fishbain, Water characterization and early contamination detection in highly varying stochastic background water, based on Machine Learning methodology for processing real-time UV-Spectrophotometry, Water Res. 155 (2019) 333–342, https://doi.org/10.1016/j.watres.2019.02.027.
[59] E. Sokolova, O. Ivarsson, A. Lilliestrom, N.K. Speicher, H. Rydberg, M. Bondelind, Data-driven models for predicting microbial water quality in the drinking water source using E. coli monitoring and hydrometeorological data, Sci. Total Environ. 802 (2022) 149798, https://doi.org/10.1016/j.scitotenv.2021.149798.
[60] D. Wang, J. Shen, S. Zhu, G. Jiang, Model predictive control for chlorine dosing of drinking water treatment based on support vector machine model, Desalination Water Treat. 173 (2020) 133–141, https://doi.org/10.5004/dwt.2020.24144.
[61] M. Garrido Baserba, L. Corominas, U. Cortes, D. Rosso, M. Poch, The fourth-revolution in the water sector encounters the digital revolution, Environ. Sci. Technol. 54 (2020) 4698–4705, https://doi.org/10.1021/acs.est.9b04251.
[62] Y.X. Yuan, W.Y. Wu, J.L. Gao, K. Chang, Water quality comprehensive evaluation method for large water distribution network based on clustering analysis, J. Hydroinf. 13 (2011) 390–400, https://doi.org/10.2166/hydro.2011.021.
[63] E. Radzka, J. Jankowska, K. Rymuza, Principal component analysis and cluster analysis in multivariate assessment of water quality, Journal of Ecological Engineering 18 (2017) 92–96, https://doi.org/10.12911/22998993/68141.
[64] C. Tian, C. Feng, L. Chen, Q. Wang, Impact of water source mixture and population changes on the Al residue in megalopolitan drinking water, Water Res. 186 (2020), 116335, https://doi.org/10.1016/j.watres.2020.116335.
[65] C. Brester, I. Ryzhikov, S. Siponen, B. Jayaprakash, J. Ikonen, T. Pitkanen, I.T. Miettinen, E. Torvinen, M. Kolehmainen, Potential and limitations of a pilot-scale drinking water distribution system for bacterial community predictive modelling, Sci. Total Environ. 717 (2020), 137249, https://doi.org/10.1016/j.scitotenv.2020.137249.
[66] X. Zhou, Z. Tang, W. Xu, F. Meng, X. Chu, K. Xin, G. Fu, Deep learning identifies accurate burst locations in water distribution networks, Water Res. 166 (2019), 115058, https://doi.org/10.1016/j.watres.2019.115058.
[67] R. Rayaroth, Random bagging classifier and shuffled frog leaping based optimal sensor placement for leakage detection in WDS, Water Resour. Manag. 33 (2019) 3111–3125, https://doi.org/10.1007/s11269-019-02296-7.
[68] Z. Almheiri, M. Meguid, T. Zayed, Failure modeling of water distribution pipelines using meta-learning algorithms, Water Res. 205 (2021), 117680, https://doi.org/10.1016/j.watres.2021.117680.
[69] N. Oliker, A. Ostfeld, A coupled classification-evolutionary optimization model for contamination event detection in water distribution systems, Water Res. 51 (2014) 234–245, https://doi.org/10.1016/j.watres.2013.10.060.
[70] J. Park, J.H. Park, J.S. Choi, J.C. Joo, K. Park, H.C. Yoon, C.Y. Park, W.H. Lee, T.Y. Heo, Ensemble model development for the prediction of a disaster index in water treatment systems, Water 12 (2020), 113195, https://doi.org/10.3390/w12113195.
[71] Y. Zhang, X. Gao, K. Smith, G. Inial, S. Liu, L.B. Conil, B. Pan, Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network, Water Res. 164 (2019), 114888, https://doi.org/10.1016/j.watres.2019.114888.
[72] A. Cardoso, B. Ribeiro, P. Gil, J.A. Sá Marques, N. Simões, J. Leitão, Detecting urban water consumption patterns: a time-series clustering approach, Water Supply 19 (2019) 2323–2329, https://doi.org/10.2166/ws.2019.113.
[73] G. Guo, S. Liu, Y. Wu, J. Li, R. Zhou, X. Zhu, Short-term water demand forecast based on deep learning method, J. Water Resour. Plann. Manag. 144 (2018), https://doi.org/10.1061/(asce)wr.1943-5452.0000992.
[74] M. Ghiassi, F. Fa'al, A. Abrishamchi, Large metropolitan water demand forecasting using DAN2, FTDNN, and KNN models: a case study of the city of Tehran, Iran, Urban Water J. 14 (2016) 655–659, https://doi.org/10.1080/1573062x.2016.1223858.
[75] H. Chen, A. Chen, L. Xu, H. Xie, H. Qiao, Q. Lin, K. Cai, A deep learning CNN architecture applied in smart near-infrared analysis of water pollution for agricultural irrigation resources, Agric. Water Manag. 240 (2020), 106303, https://doi.org/10.1016/j.agwat.2020.106303.
[76] C. Rosen, J.A. Lennox, Multivariate and multiscale monitoring of wastewater treatment operation, Water Res. 35 (2001) 3402–3410, https://doi.org/10.1016/S0043-1354(01)00069-0.
[77] J. Foschi, A. Turolla, M. Antonelli, Soft sensor predictor of *E. coli* concentration based on conventional monitoring parameters for wastewater disinfection control, Water Res. 191 (2021), 116806, https://doi.org/10.1016/j.watres.2021.116806.
[78] J.M. Hathaway, W.F. Hunt, Evaluation of first flush for indicator bacteria and total suspended solids in urban stormwater runoff, Water, Air, Soil Pollut. 217 (2011) 135–147, https://doi.org/10.1007/s11270-010-0574-y.
[79] D.T. Mccarthy, A traditional first flush assessment of *E. coli* in urban stormwater runoff, Water Sci. Technol. 60 (2009) 2749–2757, https://doi.org/10.2166/wst.2009.374.
[80] F. Cecconi, D. Rosso, Soft sensing for on-line fault detection of ammonium sensors in water resource recovery facilities, Environ. Sci. Technol. 55 (2021) 10067–10076, https://doi.org/10.1021/acs.est.0c06111.
[81] M. Djerioui, M. Bouamar, M. Ladjal, A. Zerguine, Chlorine soft sensor based on extreme learning machine for water quality monitoring, Arabian J. Sci. Eng. 44 (2019) 2033–2044, https://doi.org/10.1007/s13369-018-3253-8.
[82] X. Qin, F. Gao, G. Chen, Wastewater quality monitoring system using sensor fusion and machine learning techniques, Water Res. 46 (2012) 1133–1144, https://doi.org/10.1016/j.watres.2011.12.005.
[83] F. Fang, B. Ni, W. Li, G. Sheng, H. Yu, A simulation-based integrated approach to optimize the biological nutrient removal process in a full-scale wastewater treatment plant, Chem. Eng. J. 174 (2011) 635–643, https://doi.org/10.1016/j.cej.2011.09.079.
[84] D. Cha, S. Park, M.S. Kim, T. Kim, S.W. Hong, K.H. Cho, C. Lee, Prediction of oxidant exposures and micropollutant abatement during ozonation using a machine learning method, Environ. Sci. Technol. 55 (2021) 709–718, https://doi.org/10.1021/acs.est.0c05836.
[85] B. Teychene, F. Chi, J. Chokki, G. Darracq, J. Baron, M. Joyeux, H. Gallard, Investigation of polar mobile organic compounds (PMOC) removal by reverse osmosis and nanofiltration: rejection mechanism modelling using decision tree, Water Supply 20 (2020) 975–983, https://doi.org/10.2166/ws.2020.020.
[86] N. Jeong, T.H. Chung, T. Tong, Predicting micropollutant removal by reverse osmosis and nanofiltration membranes: is machine learning viable? Environ. Sci. Technol. 55 (2021) 11348–11359, https://doi.org/10.1021/acs.est.1c04041.
[87] G. Sigmund, M. Gharasoo, T. Hueffer, T. Hofmann, Deep learning neural network approach for predicting the sorption of ionizable and polar organic pollutants to a wide range of carbonaceous materials, Environ. Sci. Technol. 54 (2020) 4583–4591, https://doi.org/10.1021/acs.est.9b06287.
[88] N. Taoufik, W. Boumya, M. Achak, H. Chennouk, R. Dewil, N. Barka, The state of art on the prediction of efficiency and modeling of the processes of pollutants removal based on machine learning, Sci. Total Environ. 807 (2022), 150554, https://doi.org/10.1016/j.scitotenv.2021.150554.
[89] M. Bayat Varkeshi, K. Mohammadi, R. Najib, BOD and COD estimation in wastewater outflow via artificial neural network, in: Recent Advances in Environmental Science from the Euro-Mediterranean and Surrounding Regions, 2018, pp. 875–876, https://doi.org/10.1007/978-3-319-70548-4_256.
[90] J. Abdi, M. Hadipoor, F. Hadavimoghaddam, A. Hemmati-Sarapardeh, Estimation of tetracycline antibiotic photodegradation from wastewater by heterogeneous metal-organic frameworks photocatalysts, Chemosphere 287 (2022), 132135, https://doi.org/10.1016/j.chemosphere.2021.132135.
[91] S.S. Baek, Y. Choi, J. Jeon, J. Pyo, J. Park, K.H. Cho, Replacing the internal standard to estimate micropollutants using deep and machine learning, Water Res. 188 (2021), 116535, https://doi.org/10.1016/j.watres.2020.116535.
[92] G. Carvajal, D.J. Roser, S.A. Sisson, A. Keegan, S.J. Khan, Modelling pathogen log10 reduction values achieved by activated sludge treatment using naive and semi naive Bayes network models, Water Res. 85 (2015) 304–315, https://doi.org/10.1016/j.watres.2015.08.035.
[93] A. Roguet, A.M. Eren, R.J. Newton, S.L. Mclellan, Fecal source identification using random forest, Microbiome 6 (2018), https://doi.org/10.1186/s40168-018-0568-3.
[94] M. Derrien, J. Vlieg, Fate, activity, and impact of ingested bacteria within the human gut microbiota, Trends Microbiol. 23 (2015) 354–366, https://doi.org/10.1016/j.tim.2015.03.002.
[95] D. Wang, S. Thunell, U. Lindberg, L. Jiang, J. Trygg, M. Tysklind, N. Souihi, A machine learning framework to improve effluent quality control in wastewater treatment plants, Sci. Total Environ. 784 (2021), 147138, https://doi.org/10.1016/j.scitotenv.2021.147138.
[96] V.M. Gomez Munoz, M.A. Porta Gandara, J.C. De Gortari, A Bayesian method to estimate proportional payments of users in a wastewater treatment plant, Water Res. 40 (2006) 175–181, https://doi.org/10.1016/j.watres.2005.11.005.
[97] M.P. Buras, F.S. Donado, Identifying and estimating the location of sources of industrial pollution in the sewage network, Sensors 21 (2021) 3426, https://doi.org/10.3390/s21103426.
[98] H.W. Ji, S.S. Yoo, B.J. Lee, D.D. Koo, J.H. Kang, Measurement of wastewater discharge in sewer pipes using image analysis, Water 12 (2020), 061771, https://doi.org/10.3390/w12061771.
[99] S.K. Bhagat, T. Tiyasha, S.M. Awadh, T. Tran Minh, A.H. Jawad, Z.M. Yaseen, Prediction of sediment heavy metal at the Australian Bays using newly developed hybrid artificial intelligence models, Environ. Pollut. 268 (2021), 115663, https://doi.org/10.1016/j.envpol.2020.115663.
[100] G. Goncalves, U. Andriolo, L. Pinto, F. Bessa, Mapping marine litter using UAS on a beach-dune system: a multidisciplinary approach, Sci. Total Environ. 706 (2020), 135742, https://doi.org/10.1016/j.scitotenv.2019.135742.
[101] L. Wang, Z. Zhu, L. Sassoubre, G. Yu, C. Liao, Q. Hu, Y. Wang, Improving the robustness of beach water quality modeling using an ensemble machine learning approach, Sci. Total Environ. 765 (2021), 142760, https://doi.org/10.1016/j.scitotenv.2020.142760.
[102] J. Jang, A. Abbas, M. Kim, J. Shin, Y.M. Kim, K.H. Cho, Prediction of antibiotic-resistance genes occurrence at a recreational beach with deep learning models,

Water Res. 196 (2021), 117001, https://doi.org/10.1016/j.watres.2021.117001.
[103] A. Mancia, J.C. Ryan, F.M. Van Dolah, J.R. Kucklick, T.K. Rowles, R.S. Wells, P.E. Rosel, A.A. Hohn, L.H. Schwacke, Machine learning approaches to investigate the impact of PCBs on the transcriptome of the common bottlenose dolphin (*Tursiops truncatus*), Mar. Environ. Res. 100 (2014) 57–67, https://doi.org/10.1016/j.marenvres.2014.03.007.
[104] J.G. Ghatkar, R.K. Singh, P. Shanmugam, Classification of algal bloom species from remote sensing data using an extreme gradient boosted decision tree model, Int. J. Rem. Sens. 40 (2019) 9412–9438, https://doi.org/10.1080/01431161.2019.1633696.
[105] X. Du, F. Shao, S. Wu, H. Zhang, S. Xu, Water quality assessment with hierarchical cluster analysis based on Mahalanobis distance, Environ. Monit. Assess. 189 (2017), https://doi.org/10.1007/s10661-017-6035-y.
[106] M. Alshehri, M. Kumar, A. Bhardwaj, S. Mishra, J. Gyani, Deep learning based approach to classify saline particles in sea water, Water 13 (2021) 1251, https://doi.org/10.3390/w13091251.
[107] L. Sheng, J. Zhou, X. Li, Y. Pan, L. Liu, Water quality prediction method based on preferred classification, IET Cyber-Physical Systems: Theory & Applicaitons 5 (2020) 176–180, https://doi.org/10.1049/iet-cps.2019.0062.
[108] J. Zhou, Y. Wang, F. Xiao, Y. Wang, L. Sun, Water quality prediction method based on IGRA and LSTM, Water 10 (2018), 091148, https://doi.org/10.3390/w10091148.
[109] Z. Du, J. Qi, S. Wu, F. Zhang, R. Liu, A spatially weighted neural network based water quality assessment method for large-scale coastal areas, Environ. Sci. Technol. 55 (2021) 2553–2563, https://doi.org/10.1021/acs.est.0c05928.
[110] S. Liyanaarachchi, L. Shu, S. Muthukumaran, V. Jegatheesan, K. Baskaran, Problems in seawater industrial desalination processes and potential sustainable solutions: a review, Rev. Environ. Sci. Biotechnol. 13 (2014) 203–214, https://doi.org/10.1007/s11157-013-9326-y.
[111] P. Chawla, X. Cao, Y. Fu, C.M. Hu, M. Wang, S. Wang, J.Z. Gao, Water quality prediction of salton sea using machine learning and big data techniques, Int. J. Environ. Anal. Chem. (2021) 1963713, https://doi.org/10.1080/03067319.2021.1963713.