Water Research 205 (2021) 117666

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Review

Machine learning in natural and engineered water systems

Ruixing Huang [a,b], Chengxue Ma [a,b], Jun Ma [b], Xiaoliu Huangfu [a,*], Qiang He [a]

[a] Key Laboratory of Eco-environments in the Three Gorges Reservoir Region, Ministry of Education, College of Environmental and Ecology, Chongqing University, Chongqing 400044, China

[b] State Key Laboratory of Urban Water Resource and Environment, School of Municipal and Environmental Engineering, Harbin Institute of Technology, Harbin 150090, China

ARTICLE INFO

Keywords:
Machine learning
Natural water systems
Engineered water systems

ABSTRACT

Water resources of desired quality and quantity are the foundation for human survival and sustainable development. To better protect the water environment and conserve water resources, efficient water management, purification, and transportation are of critical importance. In recent years, machine learning (ML) has exhibited its practicability, reliability, and high efficiency in numerous applications; furthermore, it has solved conventional and emerging problems in both natural and engineered water systems. For example, ML can predict various water quality indicators in situ and real-time by considering the complex interactions among water-related variables. ML approaches can also solve emerging pollution problems with proven rules or universal mechanisms summarized from the related research. Moreover, by applying image recognition technology to analyze the relationships between image information and physicochemical properties of the research object, ML can effectively identify and characterize specific contaminants. In view of the bright prospects of ML, this review comprehensively summarizes the development of ML applications in natural and engineered water systems. First, the concept and modeling steps of ML are briefly introduced, including data preparation, algorithm selection and model evaluation. In addition, comprehensive applications of ML in recent studies, including predicting water quality, mapping groundwater contaminants, classifying water resources, tracing contaminant sources, and evaluating pollutant toxicity in natural water systems, as well as modeling treatment techniques, assisting characterization analysis, purifying and distributing drinking water, and collecting and treating sewage water in engineered water systems, are summarized. Finally, the advantages and disadvantages of commonly used algorithms are analyzed according to their structures and mechanisms, and recommendations on the selection of ML algorithms for different studies, as well as prospects on the application and development of ML in water science are proposed. This review provides references for solving a wider range of water-related problems and brings further insights into the intelligent development of water science.

1. Introduction

Water is one of the most indispensable material resources on which the survival and development of humanity depends. However, the increasing amounts of pollutant discharge caused by human activities has been leading to increasing amount of water pollution, which poses a threat to the sustainable development of ecosystems and human society. To protect the ecological safety and human health from water pollution, a range of measures have been adopted. In natural water systems, the qualities of various waters (e.g., rivers, lakes, groundwaters, and seawaters) are under close monitoring for better management and utilization of water resources. In addition, water pollution is controlled using identification, source tracking, and toxicity evaluations of pollutants. In engineered water systems, raw water taken from natural waters is purified in drinking water treatment plants (DWTPs) using various treatment processes (e.g., coagulation, sedimentation, filtration, and disinfection) to remove the contaminants, and then this supplied to users via distribution systems. Additionally, sewage and wastewater are collected and treated in the wastewater treatment plants (WWTPs) through a series of physical, chemical, and biological processes to abate pollution toward the urban and natural environment. During the processes of natural water management, water resource utilization and polluted water treatment, a large number of corresponding studies have been conducted, gradually developing into fields related to natural and

* Corresponding author.
E-mail address: hfxl-hit@163.com (X. Huangfu).

https://doi.org/10.1016/j.watres.2021.117666
Received 31 May 2021; Received in revised form 1 September 2021; Accepted 11 September 2021
Available online 14 September 2021

engineered water systems.

Currently, "the fourth industrial revolution", combined with big data and artificial intelligence, is expected to bring immense changes to human society. Machine learning (ML) is one of the technical approaches of artificial intelligence; and it is developed using various algorithms based on mathematical and statistical knowledge. ML can predict the status of new data by summarizing the underlying relationships and rules within known data, and its prediction performance will improve with an increase in the data amount and the iteration of the algorithms (Vamathevan et al., 2019). ML can solve complex problems involving massive nonlinear processes or combinatorial spaces that cannot be solved using conventional methods, or only with great time and cost. Therefore, ML has been widely used in fields such as computer vision, speech recognition, natural language processing, robotic control, and other hot topics (Jordan and Mitchell, 2015a).

In recent years, ML has also been applied to solve a wide variety of problems in the water science field. To understand the application situation of ML in these fields, and the applicability and feasibility of various algorithms in solving water-related problems, it is necessary to review and summarize the existing studies, and some papers have reviewed relevant studies. For example, many publications have reviewed the application of artificial neural networks (ANNs) and other algorithms, such as the fuzzy inference system (FIS), evolutionary algorithms, the support vector machine (SVM), random forest (RF), decision tree (DT), and ML coupled with wavelet transformation or optimization algorithms for modeling water quality in rivers, lakes, and groundwater (Chau 2006; Che Osmi et al., 2016; Chen et al., 2020b; Ighalo et al., 2020; Maier and Dandy, 2000; Maier et al., 2010, 2014; Nicklow et al., 2010; Ostfeld and Solomatine, 2008; Raghavendra and Deka, 2014; Rajaee et al., 2020; Tiyasha, Tung and Yaseen, 2020). In addition, the applications of ML in other aspects or water-related research have also been reviewed, such as remote sensing for monitoring water quality (Sagan et al., 2020; Wagle et al., 2020), drinking water treatment (Dogo et al., 2019; Li et al., 2021), seawater desalination (Al Aani et al., 2019), and wastewater treatment and pollutant removal (Fan et al., 2018; Khataee and Kasiri, 2011; Wang et al., 2021b; Yaseen 2021). All of these reviews are beneficial for researchers to understand and expand the application of artificial intelligence in the water science field. However, many other published applications of ML, such as groundwater contaminant mapping (Podgorski and Berg, 2020), contaminant sources tracing (Balleste et al., 2020), pollutant toxicity evaluating (Wang et al., 2021d), contaminant identification (Baek et al., 2021), and others, have not been summarized and discussed. Moreover, many review papers have focused on the application of one algorithm, thus lacking a comparison of its advantages and disadvantages with other algorithms. As is well known, plenty of ML algorithms have been developed, and their performance for solving practical tasks varies. Therefore, an analysis of the applicability scope of various algorithms for dealing with different water-related problems is required.

This article provides a comprehensive review of ML applications in related fields of water science and summarizes the representative studies in recent years. First, the establishment processes of ML models, including data preparation, algorithm selection, and model evaluation, are briefly introduced. In addition, the representative applications of ML in water quality prediction and management in natural water systems, as well as technology development and operation monitoring in engineered water systems, are summarized. To be specific, in natural water systems, ML has been used to predict water quality indicators, map pollutant distribution in groundwater, classify water resources, trace contaminant sources, and evaluate pollutant toxicities. In engineered water systems, ML has been applied to optimize adsorption and oxidation processes, assist laboratory characterization analyses, and improve the operation and management of drinking water purification and distribution, as well as wastewater collection and treatment. More importantly, the advantages and disadvantages of representative algorithms are discussed, and their applicability for different data and researches is analyzed by comparing their structures and mechanisms. Finally, current research hotspots, challenges, and prospects of ML combined water utilization and pollution control are discussed.

2. An overview of machine learning

ML refers to a technology that uses a series of programmed algorithms to predict the future patterns of any raw data with the experience learned from the hidden associations within the given data through an automatic mathematical analysis (Jordan and Mitchell, 2015b). In general, to recognize the rules underlying the known data as accurately as possible, the data should first be well treated to generate the dataset. Then an appropriate ML algorithm is determined according to the characteristics of input data and the requirements of output data. The selected algorithm will then be trained with the well-prepared data and evaluated to adjust the hyper-parameters within, thus generating the desired model. Afterward, the proposed ML model is qualified to make predictions on new data. The establishment of the ML model is briefly introduced in the subsequent sections, and those who want to learn more details about ML algorithms can refer to specialized books and papers on statistics or machine learning (Hastie et al., 2008; Mohri et al., 2012).

2.1. Date preparation

An ideal ML requires appropriate and well-trained models, but the quality and quantity of data are also vitally important. Currently, we can crawl data from the internet, collect data from the literature, search data from open-source databases, or record data from the experiments, etc. However, we typically obtain raw data with missing values, errors, duplicates, or noises that should be treated with data cleansing. Then, feature engineering is conducted based on professional background knowledge to select or extract data features according to task demands. Finally, the prepared data is typically divided into a training set, a validation set (for some algorithms), and a test set. The training set is used to train the model based on the selected ML algorithm, while the validation set is applied to tune hyper-parameters to optimize the trained model. After the training process, the predictive and generalization ability of the trained model is evaluated using the test set by comparing the prediction outputs with their corresponding known results.

2.2. Algorithm selection

The rise of ML can be traced back to the 1950s when Donald Hebb proposed the Hebbian learning theory (Hebb, 1949), after which a wide range of ML algorithms were developed. In general, ML algorithms can be classified into three categories: supervised, unsupervised, and reinforcement learning (a certain classification also includes semi-supervised learning) according to the types of data and requirements of the work. Moreover, the applications of deep learning (DL) for water utilization and pollution control are also discussed due to the inseparable relationships between DL and ML. Fig. 1 outlines the reviewed ML algorithms and their applications in the fields of natural and engineered water systems.

Supervised learning algorithms are typically applied to treat labeled data to predict the values of a continuous set using regression or the category of a discrete set using classification. For the regression analysis, the least-square method (LSM) has long been used in many algorithms, as it can find the best function parameters to minimize the sum of squares of errors between the predicted and actual values. Based on the LSM, the simplest ML algorithm linear regression is typically applied when the dataset possesses a relatively small size and a linear relationship within data (Fig. 2a). When dealing with nonlinear relationships, polynomial regression, which is also based on LSM, is a better choice, as it can flexibly fit nonlinear data by adjusting the power of the variables (Fig. 2b). By using linear and polynomial regressions, rapid modeling,

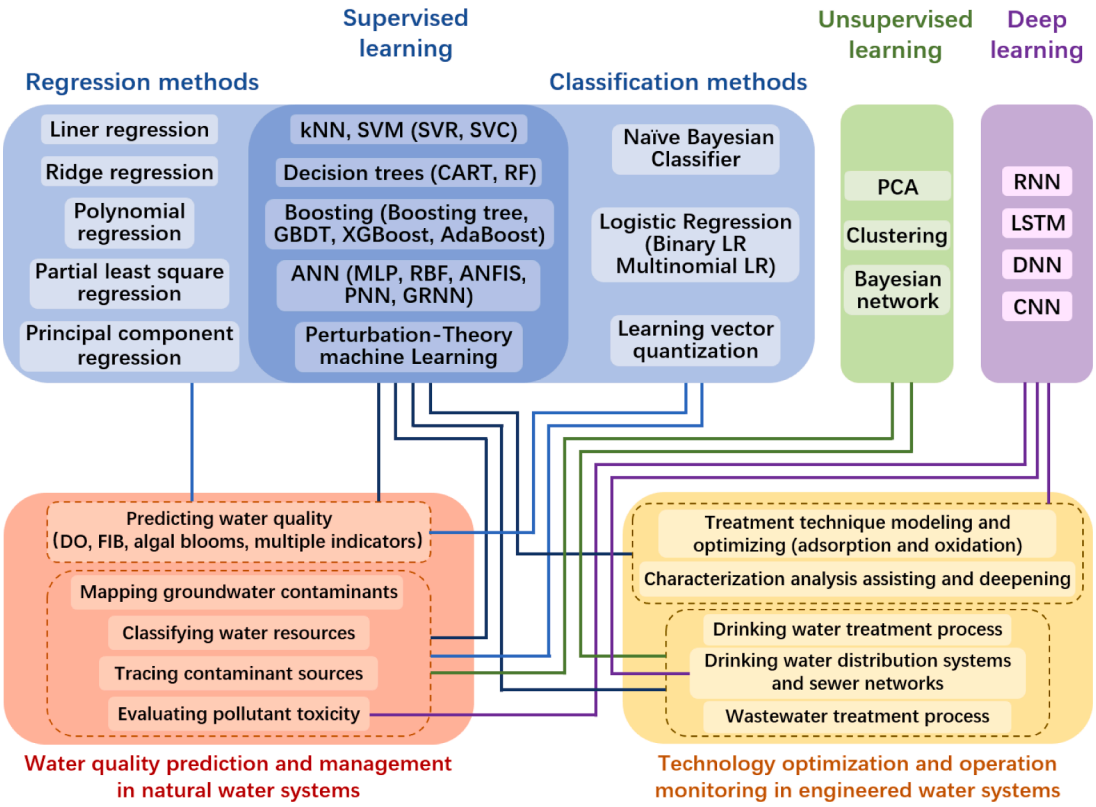

Fig. 1. Reviewed machine learning algorithms and their applications in natural and engineered water systems.

intuitive interpretation, and accurate sensitivity to outliers can be achieved. In addition to linear and polynomial regressions, the ridge regression, Lasso regression, and ElasticNet regression are also common regression algorithms used in other fields. As for the classification tasks, the naïve Bayesian (NB) classifier and logistic regression (LR) have long been used. The NB classifier can calculate the required posterior probability based on the existing prior probability of the event; then, the newly gained probability is updated to perform the subsequent tasks. The LR applies the sigmoid function to normalize the predicted values, thus calculating the probability of an event and comparing it with the selected threshold value (usually 0.5) to generate the predicted binary outcomes (yes or no).

In addition to the algorithms described above specifically developed for regression or classification, many other algorithms can be used for both regression and classification. The K-nearest neighbors (kNN) can predict the value or category of a sample according to its adjacent neighbors in the feature space. For regression, the average value of the k nearest neighbors is recognized as the prediction result, while for classification, the category of the most occurring in the k nearest neighbors is outputted (Fig. 2c) (Altman, 1992). Additionally, the support vector machine (SVM) aims at finding a hyperplane to segment the samples based on the principle of maximizing the interval between two categories of samples (Fig. 2d). To solve classification or regression problems, SVM can be divided into support vector classification (SVC) and support vector regression (SVR) (Kadyrova and Pavlova, 2014). The decision tree (DT) is also a popular algorithm that has a tree structure. It is composed of a root node, several internal nodes, and leaf nodes that respectively represent the set of all samples, attribute tests, and the decision results (Fig. 2e). The DT algorithm begins its decision-making process from the root node and then compares the tested data with the characteristic nodes. The algorithm then selects the next comparison branch according to the result of the attribute test. It finally outputs the result of the leaf node as the final decision result (Myles et al., 2004). Because the DT algorithm can solve both classification and regression problems, it is also known as the classification and regression tree (CART). Moreover, to improve the performance of the DT algorithm, many derived DT algorithms, such as boosted decision trees (BDTs), gradient boosted decision trees (GBDTs), and the random forests (RF) have been developed. In particular, RF is an ensemble algorithm comprised of many DTs in which each tree randomly samples from the input data to independently generate a prediction result. After all the trees output their decisions, they vote for the most appropriate result as the final prediction of this RF model (Fig. 2f) (Breiman, 2001). Aside from the above algorithms, the artificial neural network (ANN) was the most frequently used algorithm in the reviewed studies. Perceptron is the structural unit of ANN, and it is composed of input cells and an output cell, and the weight the connections between them. Different input information can exert different effects on the output by adjusting the values of the weight connections (Fig. 2g) (Rosenblatt, 1957). The perceptron algorithm is a linear classification model that is suitable for dealing with linearly separable data. By combining multiple perceptrons and introducing hidden layers and activation functions, the multilayer perceptron (MLP) algorithm was proposed and is capable of treating multi-dimensional data (Fig. 2h) (Clark, 1991). However, MLP is a global approximation algorithm, in which all the weights in the network should be readjusted every time the sample is learned. Therefore, MLP has the disadvantage of a slow convergence velocity, thus tending to fall into the local optimum. The radial basis function neural network (RBF NN) is another common ANN algorithm (Fig. 2i). It employs the radial basis function as the activation function and only adjusts the weights connections in the specified domain. Therefore, RBF NN has the advantage of fast convergence and is immune to the local optimum problem (Lee and Chang, 2003). Moreover, the adaptive neuro-fuzzy inference system (ANFIS) is also a commonly used ANN-derived algorithm. ANFIS can be defined as a multi-layer feed-forward network that employs fuzzy inference to map the input space to the output space. ANFIS allowed the realization of a highly non-linear mapping that is considered to be superior in yielding non-linear time series compared

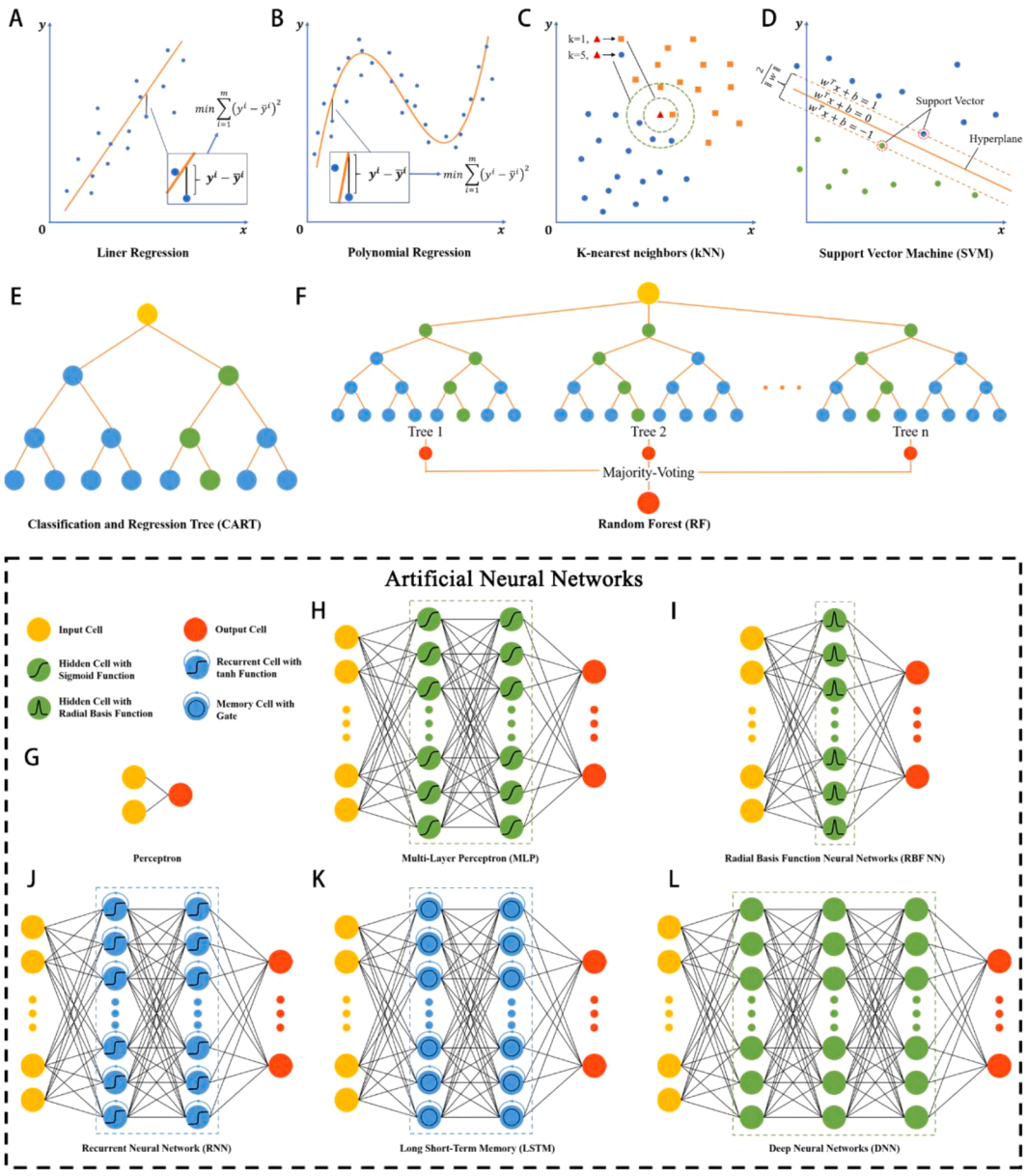

Fig. 2. A chart of the common algorithms applied in this review.

with common linear methods (Jang, 1993).

Unsupervised learning algorithms are often applied to reveal the intrinsic characteristics and rules of the unlabeled sample data. They are typically applied in dimensionality reduction, clustering, and anomaly detection (also known as outlier detection). The principal component analysis (PCA) is a representative method for unsupervised dimensionality reduction. As its name implies, PCA aims to find the most essential characteristics or generate a new characteristic to describe the original dataset, thus reducing the dimensionality of the dataset and increasing interpretability with minimized information loss (Jolliffe and Cadima, 2016). K-Means is a common method for clustering analysis. It can find a partition to organize data into differentiable groupings, by minimizing the squared error between the empirical mean and the points in cluster (Jain, 2010). Isolation forest, Gaussian distribution, and local outlier factor (LOF) are also common algorithms for anomaly detection. They are typically used to detect samples that are sparsely distributed and far from the majority of data (Ariyaluran Habeeb et al., 2019). Different from supervised learning algorithms that predict the data directly, unsupervised learning algorithms are typically applied for data pretreatment in the studies reviewed herein.

2.3. Model evaluation

Many approaches have been proposed to evaluate the performance of ML models. The evaluation parameters for regression algorithms primarily include bias, variance, the mean absolute error (MAE), the mean squared error (MSE), and R-squared (R^2) among others. Bias is the difference between the predicted result and its actual value while variance represents the degree of deviation from the mean of the total sample. MAE is the average sum of the absolute value of the bias for each sample. It can reflect the actual error of prediction by avoiding the problem where the positive and negative errors cancel each other. MSE is the average of the sum of the square of the bias for each sample. Typically, to keep the evaluation indicator and the sample values are on the same order of magnitude, MSE uses the root to obtain the root mean squared error (RMSE), which is also a commonly used measure of performance. Moreover, R^2, also known as the coefficient of determination, describes the fitting degree of the regression function to the observed

values. The value of R^2 is between 0 and 1; and the closer the R^2 value is to 1, the better the model fits the data.

For classification algorithms, the accuracy rate is the most used evaluation indicator, and it is the proportion of the correctly classified samples to the total number of samples. In addition, precision (P) and recall (R) are also widely used measurements for evaluating classifiers. The predicted results can be divided into true positive (TP), false positive (FP), true negative (TN), and false negative (FN) according to their real categories and the predicted categories (Table 1). P is defined as the proportion of the correctly classified positive samples in the total number of positive samples, while R is the proportion of the correctly classified positive samples in the total number of correctly classified samples. For an ideal classifier, the values of both P and R are expected to be as high as possible. However, one value will increase, while the other decreases in practical cases. Therefore, the F_β score, the weighted harmonic mean of the P and R-values, is applied to balance P and R. F_β is presented as Eq. (1), where β measures the relative importance of R to P.

$$F_\beta = \frac{(1+\beta^2) \times P \times R}{(\beta^2 \times P) + R}, \quad (1)$$

where R matters more than P when $\beta > 1$, and conversely, P matters more than R when $\beta < 1$. When $\beta = 1$, F_β is converted to the F_1 score, that represents the balance between R and P.

3. Water quality prediction and management in natural water systems

Natural waters, including rivers, lakes, groundwaters and seas, are the most important water sources for human life and productive activities. To better manage and utilize natural water resources, different countries and regions have adopted various evaluation schemes based on a series of water quality indicators (WQIs). Physical indicators include temperature, color, turbidity, electric conductivity (EC), suspended solids (SS), and total solids (TS). Chemical indicators include pH, dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), total organic carbon (TOC), alkalinity, ammonia nitrogen, total phosphorus (TP), total nitrogen (TN).Biological indicators covering chlorophyll a (Chl-a) and fecal indicator bacteria (FIB, e.g., total coliform, *E. coli,* and *Enterococcus*) (Uddin et al., 2021). In recent years, many studies have applied ML approaches in water quality prediction and water resource management in natural water environments. Below the relevant representative applications of ML are summarized, including predicting water quality WQIs (e.g., DO, FIB, and Chl-a, or other multiple indicators), mapping pollutant distribution in groundwater, classifying water resources according to different standards, tracing contamination sources, and evaluating pollutant toxicity.

3.1. Predicting water quality

The dissolved oxygen (DO) concentration is a critical parameter for evaluating the ecological health of an aquatic environment. This concentration results from an equilibrium between oxygen-producing (e.g., photosynthesis and air diffusing) and oxygen-consuming (e.g., aerobic respiration, nitrification, and chemical oxidation) processes in an aquatic environment (Olyaie et al., 2017). Under the influence of these complex processes, it is difficult for conventional process-based models or statistical approaches to simulate the DO level. On the contrary, data-driven ML does not consider the accumulation mechanism of DO, but only analyzes the statistical and mathematical relationship between different parameters. For example, WQIs such as temperature, pH, EC, and discharge were selected as inputs to train two types of ANN models for predicting DO concentrations downstream and upstream of Fountain Creek. The RBF NN performed better with a higher R^2 value for both sampling stations than the MLP model (Ay and Kisi, 2012). However, the RBF NN was not as accurate as the MLP in predicting the DO concentration in the Mediterranean Sea along with Gaza, which might have been attributed to the small database, as RBF NN was more sensitive to the data volume (Zaqoot et al., 2009). To improve the performance of the RBF NN, the general regression neural network (GRNN) modified the structure of the RBF by replacing its weight connection between the hidden layer and the output layer with a summation layer to reduce the need for data volumes. In predicting the DO concentration in the Danube River, the performances of the MLP, GRNN, and recurrent neural network (RNN) were examined using only small amount of data. Results showed that the GRNN obtained better performance than the MLP did. However, the GRNN was inferior to the RNN model that was equipped with an extra layer to store and transform the previous input information as a decision reference when needed (Fig. 2J) (J and W, 2011). The tested RNN provided considerably better predictions of the DO with all results within an error of less than ±10% (Antanasijevic et al., 2013b). However, because the previous information was sequentially stored in RNN, too much information would make it difficult to learn from the far information, thus causing long-term dependency problems. To solve this problem, the long short-term memory (LSTM) algorithm was developed by storing previous information in the memory cells that will be opened by a gate to transfer information when needed (Fig. 2K) (Greff et al., 2017). The LSTM was applied to predict the river DO concentrations at the continental scale using the CAMELS-chem database that collected DO information from 236 minimally disturbed watersheds across the U. S. Ultimately, the proposed LSTM model achieved a satisfactory prediction with a mean and median Nash-Sutcliffe Efficiency (NSE) of 0.60 and 0.78, respectively (Fig. 3A) (Zhi et al., 2021). Due to the difference in the spatial location and scale of the study, as well as the choice of variables and the amount of data, a comparison between different studies is undoubtedly difficult. However, by utilizing a direct comparison in the same study and a cross-comparison of the different studies listed in Table S1, it was found that MLP performed well with a small dataset (data bulk < 1000), while SVM performed better with a larger dataset (data bulk > 1000). Moreover, considering that the variables used in a study for predicting DO are typically based on a time series, the LSTM is also recommended as it is designed for time-series tasks and can carry useful information from the past to the future. In addition, it relies on fewer environment variables, and this saves the cost of data collection.

Drinking water or recreational water polluted with feces has the potential to cause gastrointestinal and respiratory diseases (Haile et al., 1999). Total coliform, fecal coliform (or *E. coli*), and *Enterococcus* are typically applied as fecal indicator bacteria (FIB) to characterize fecal contamination in water. However, common approaches for quantifying FIB such as multiple-tube fermentation, membrane filtration, and specific enzymatic detection, usually require 18–72 h, while emerging techniques such as flow cytometry, ATP assays, online optical sensors, and quantitative PCR are costly (García-Alba et al., 2019). However, data-driven ML can provide real-time predictions that rely on easily available information about environmental conditions (e.g., precipitation, discharge, and tide) and the representative studies are reviewed below. For instance, an MLP model was used to fill the missing microbial data in a dataset based on related physical, chemical, and bacteriological data from the Kentucky River. The MLP had a more accurate performance than conventional imputation or regression models with smaller error values. Additionally, the proposed MLP model also accurately classified observational data into two defined ranges for fecal coliform concentrations in a river system, especially when the concentration of

Table 1
Confusion matrix.

True classification	Predicted classification	
	True	False
True	True Positive (TP)	False Negative (FN)
False	False Positive (FP)	True Negative (TN)

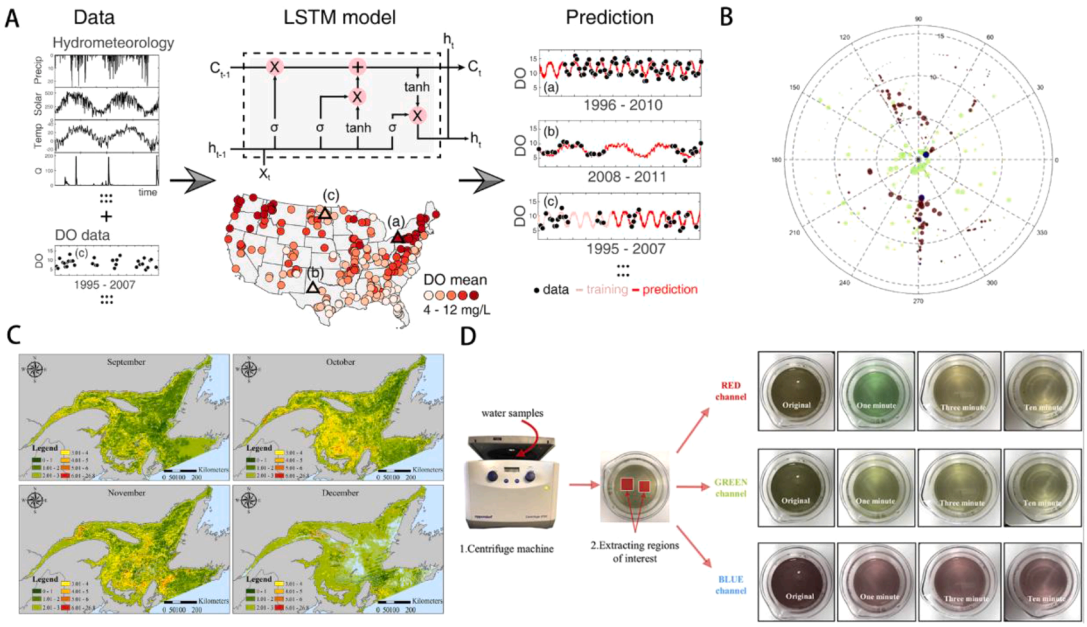

Fig. 3. The applications of ML in predicting water quality in the natural water environment. (A) The application of the LSTM model in predicting river DO concentrations across the continental United States. Reproduced from (Zhi et al. 2021) with permission. Copyright (2021) American Chemical Society. (B) Balanced data for Clarks Beach, green dot: below sample, blue dot: above sample, brown dot: ADASYN sample. Reproduced from (Xu et al. 2020) with permission. Copyright (2020) Elsevier Ltd. (C) Monthly mean Chl-a concentration derived from Aqua-MODIS by BME/SVR. Reproduced from (He et al. 2020) with permission. Copyright (2019) Elsevier Ltd. (D) RGB Channel separation diagram of a pond water sample. Reproduced from (Li et al. 2020b) with permission. Copyright (2020) Elsevier Ltd. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

FIB was below 200 CFU/100 mL (Chandramouli et al., 2007). Moreover, when the safe concentration of FIB was set as the classification threshold, ML could also be applied in water quality warnings for recreational surface waters. For example, when monitoring fecal pollution of coastal water in southern California, an MLP accurately predicted the levels of three FIBs, with FP and FN rates less than 10%, thus sufficing to make rapid and reliable decisions regarding beach closures or advisories (He and He, 2008). In addition to ANN models, some other algorithms, such as CART, RF, and the Bayesian network, have also been applied to predict FIB concentrations to provide information about water safety (Thoe et al., 2014). Though the models described above performed well with high overall accuracy, the unbalance in the dataset might have reduced the sensitivity of the predictions. Data unbalance is caused because most of the data in the dataset was typically below the guideline thresholds while the minority is exceeded. This increases the possibility of over-fitting in the majority class and information loss in the minority class when training a model (Batista et al., 2004). To solve this problem, an adaptive synthetic (ADASYN) sampling algorithm was adopted to create more samples at the boundary between the two classes using linear interpolation (Fig. 3B). Afterward, the proposed kNN, BDT, and MLP models all provided favorable predictions with relatively high sensitivity of approximately 75% and an overall accuracy of greater than 90% (Xu et al., 2020). In addition to ADASYN, recently, the synthetic minority oversampling technique (SMOTE) was also applied to solve the data imbalance problem in FIB prediction. SMOTE is a technique based on the nearest neighbor idea to synthesize new samples for the minority class. The combination with SMOTE improved the performance of the six tested algorithms. For example, RF presented a 50% higher true positive rate with respect to the baseline (Bourel et al., 2021). The proposed approach in these two studies provided new insight into data preparation for training ML models. Moreover, no matter what algorithms were applied for FIB prediction, accuracy was always used to evaluate the performance of the proposed models. However, accuracy might not be enough to assess models when the potential health effects of a high false negative (FN) rate are assumed to outweigh the risk of an unnecessary recreational water closure given by a high false positive (FP) rate. Therefore, (Stidson et al., 2012), attached more importance to the false negative (FN) rate over other metrics (Stidson et al., 2012). This was because a high FN rate meant the prediction of a dangerous situation as a safe one (Motamarri and Boccelli, 2012). Therefore, when evaluating the performance of an ML model, diversified evaluation criteria based on actual requirements should be considered. According to Table S2, MLP was the most widely used algorithm in the previous studies that modeled FIB, but algorithms based on DT, such as BDT, CART and RF, performed better in studies that solved the data imbalance problem. In addition, the model based on DT was more interpretable than the neural network model, and it was more instructive to understand and prevent FIB from exceeding the safety standard. Therefore, DT-based models combined with data imbalance processing technology are more suitable for FIB predictions.

An appropriate amount of phytoplankton in an aquatic environment can improve water quality by producing oxygen and soaking up carbon dioxide through photosynthesis (Durham et al., 2015). However, eutrophication of water bodies will promote the rapid growth of phytoplankton, thus leading to aquatic environmental problems such as algal blooms (Ma et al., 2015). Algal blooms will shade benthic primary producers from the sunlight, reduce DO levels, and accumulate toxic metabolites to threaten the ecology and human beings (Conley et al., 2009). Therefore, forecasting algal blooms or determining the key factors that control bloom production is important to prevent their deleterious effects (Recknagel et al., 1997). Various models have been proposed to achieve this purpose, including empirical models, deterministic models, time-series analysis models, and fuzzy-logic models (Yabunaka et al., 1997). ML has also been applied to predict algal blooms by identifying the environmental factors that affect the growth and accumulation of algal biomass to simulate the nonlinear processes of algal blooms (Karul et al., 2000; Lee et al., 2003). However, environmental factors that control the biomass of algae vary according to different studies. For example, flow and temperature have been reported to be predominant in determining the initiate and duration of cyanobacterial growth, while water color controls the magnitude of the population growth peak (Maier et al., 1998). Additionally, the importance of

controlling factors for different cyanobacteria genera varied with their species. Concretely, *Anabaena* and *Microcystis* had the highest sensitivity to average flow, while for *Planktolyngbya* and *Oscillatoria*, it was temperature, and for *Cylindrospermopsis*, it was water color. Therefore, it has been difficult to predict the intensity and frequency of algal blooms using a single environment variable and to control algal biomass by adjusting few influential factors due to the diversity of phytoplankton that made up algal blooms (Nelson et al., 2018). In the last few decades, the development of remote sensing technology makes has made it possible to directly observe and investigate algal blooms at large spatial and temporal scales. Recently, ML has also been applied to analyze remote sensing data (Lary et al., 2016). For example, moderate resolution imaging spectroradiometer data were applied to estimate the Chl-a concentration within the Gulf of St. Lawrence of Canada (Fig. 3C). In this study, remote sensing with 10 types of reflectance (Rrs) was combined with eight different ML algorithms to predict algal blooms. The comparison results revealed the SVR model obtained the best performance using Rrs at 412, 443, 488, 531, and 678 nm (He et al., 2020). In addition, some other remote sensing data, such as SeaWiFS, have also been applied for estimating Chl-*a* concentrations to forecast algal blooms (Campsvalls et al., 2006; Keiner 2010). In addition to predicting Chl-*a* in the form of remote sensing data, remote sensing technology can also provide hyperspectral images to simulate the concentration of algae cells. In a study that predicted cyanobacterial cell concentrations, (Pyo et al., 2021), developed a convolutional neural network (CNN) model with a convolutional block attention module (CNNan) that could emphasize valuable information and suppress insignificant ones to improve the performance of CNN for image recognition. Compared to a traditional hydrodynamic model and CNN, CNNan performed better with a higher NSE and a smaller RMSE (Table S3) (Pyo et al., 2021). As mentioned in this section, the factors that affect the growth of different algae are different, and a natural water body often contains several types of algae at the same time. Consequently, screening from all available water-related parameters to determine influential ones is often necessary prior to an algae or Chl-*a* prediction, which is time and labor-intensive. Remote sensing and image recognition technologies do not consider complex water quality parameters, but directly use color information, which is one of the major characteristics of water bodies rich in Chl-*a* or algae relative to other polluted water bodies. Therefore, algorithms based on remote sensing information or hyperspectral images are undoubtedly the optimal choice to predict the concentration of algae or Chl-*a*.

Aside from the single WQI like DO, FIB, and Chl-a introduced above, indicators like BOD, COD, TOC, TP, TN, and TS are more common in water quality evaluations, and ML has also been used to predict these water quality indicators. For instance, 13 environmental variables were adopted to develop an MLP model to simulate the daily concentrations of TOC, TN, and TP in three rivers. The results showed that concentrations of the three WQIs were all estimated with R^2 values greater than 0.75. Moreover, the future carbon and nitrogen loads under climate change scenarios were predicted, and increases in TOC and TN under notable change scenarios were reported (Holmberg et al., 2006). Recently, an RF model was used to predict the concentrations of nutrients (N and P) in both rural and urban catchments. By analyzing the impacts of each variable in the model and removing the inessential ones, the number of variables were minimized to reduce the cost of surrogate sensors. In this manner, both the low cost and high accuracy of the prediction were guaranteed (Castrillo and Garcia, 2020). In addition to rivers, the water quality of other water bodies has also been simulated, including a lake (Barzegar et al., 2020), a reservoir (Zhao et al., 2007), coastal water (Palani et al., 2008), and even stormwater runoff (Najah Ahmed et al., 2019). In particular, the water quality index, rather than WQIs, has been utilized in studies to predict the water quality in wetlands using MLP, RBF NN, SVM, and CART models. The water quality index is a single number calculated using a set of physicochemical water indicators to normalize the water quality (Mohammadpour et al., 2015, 2016). Interestingly, image analysis was also applied to determine the quantitative relationships between the image information and the water quality. In a study that analyzed animal wastewater quality, the linear regression (LR), stochastic gradient descent (SGD), and Ridge regression algorithms were applied. The RGB color index extracted from the water sample images was used to analyze the correlations between the spectral rate and the WQIs, including N, P, TS, and total coliform (Fig. 3D). The optimal R^2 values were as high as 0.98 for TS and 0.96 for total coliform, which demonstrated that this was an instructive method for rapid and cheap analysis of water quality (Li et al., 2020b). Due to random or systematic errors, data obtained from site monitoring or experimental processes were possibly affected by noise that caused uncertainty in the accurate prediction. To eliminate the effects of noise on data, the ANFIS algorithm was improved by combining it with the wavelet de-noising technique (WDT). The results indicated that WDT-ANFIS was the fastest and the most precise method for processing large volumes of non-linear and non-parametric data compared with the MLP, RBF NN, and ANFIS models (Najah Ahmed et al., 2019). Therefore, to optimize the performance of ML, more advanced analytical and data processing techniques are worthy of being considered. According to Table S4, the tested algorithms, e.g., MLP, RBF NN, SVM, and DT, performed well in predicting various WQIs in the reviewed studies, and this might have been due to the close chemical and physical relationship between the selected water quality variables and the predicted parameters. Although the performance of various algorithms seems acceptable, noise in the data is always an unavoidable problem, and this could be solved by WDT to improve the performance of ML algorithms. Moreover, image recognition technology and the DL model once again has appeared in recent studies, and this might indicate the developmental direction of predicting water quality using ML.

3.2. *Mapping groundwater contaminants*

Groundwater contamination is a public concern because of its potential health threats to billions of people globally who rely on groundwater for drinking. However, the worldwide scale of polluted regions is still unknown due to a lack of sufficient field data that reflects the spatial distribution and concentration of contaminants. To solve this problem, various research methods have been developed. In the past, geostatistical-statistical interpolation methods such as Kriging, inverse distance weighting, global polynomial interpolation, sequential Gaussian simulations, and the Thiessen polygon, have been applied to simulate the spatial distribution of contaminants in groundwater (Bindal and Singh, 2019). However, these methods depend too much on accurate field data, which is sometimes unattainable or sparse. Moreover, they do not take consider the spatial dependency of data, thus making the performance of these methods unsatisfactory (Shaji et al., 2020). In recent years, ML has also been widely applied to predict the distribution of contaminants such as arsenic (As), fluoride, manganese, and E. coli in groundwater based on analyzing the hidden relationships between contaminants and their direct or indirect influencing factors. Because of the serious health threat of As to humans, and its wide distribution throughout Earth's crust and the hydrosphere, the prediction of the arsenic spatial distribution in groundwater by ML has been highlighted (Shaji et al., 2020).

Geologic and geochemical settings are primary factors that affects the occurrence of As in the lithosphere. Geologic factors such as specific rocks (i.e., mafic and granitic rocks (Ayotte et al., 2006), basalt, andesite, and rhyolite (Bretzler et al., 2017)) and soils (i.e., medium-textured soils (Winkel et al., 2011; Yang et al., 2014), organic-rich deposits (Winkel et al., 2008, 2011), and Holocene fluvial sediments (Podgorski et al., 2017; Rodríguez-Lado et al., 2013; Yang et al., 2014)), were reported to be the primary causes of arsenic accumulation. Moreover, geochemical conditions, such as the level of soil pH (Lee et al., 2009; Podgorski et al., 2017; Twarakavi and Kaluarachchi, 2006; Yang et al., 2012), DO (Yang et al., 2012), and other chemical

substances including fluoride, bromide, chloride, iron, manganese, ammonia, nitrate, nitrite, phosphate, and sulfate (Lee et al., 2009; Yang et al., 2012) were also reported to be crucial factors that affect the occurrence of As in groundwater. By introducing these relevant environmental factors as variables, ML has gained satisfactory performance in predicting the spatial distribution of arsenic in groundwater around the world. These nations include the United States (Anning et al., 2012; Ayotte et al., 2017, 2016, 2006; Frederick et al., 2016; Kim et al., 2011; Meliker et al., 2008; Twarakavi and Kaluarachchi 2006; Yang et al., 2014, 2012), China (Lee et al., 2009; Rodríguez-Lado et al., 2013; Zhang et al., 2012, 2013), Canada (Dummer et al., 2015), India (Bindal and Singh, 2019; Purkait 2008), Pakistan (Podgorski et al., 2017), Burkina Faso (Bretzler et al., 2017), Bangladesh (Tan et al., 2020), Cambodia (Lado et al., 2008) and regions such as Southeast Asia (Bangladesh, Myanmar, Thailand, Laos, Cambodia, Vietnam, Sumatra, Red River, and Mekong deltas) (Cha et al., 2016; Cho et al., 2011; Chowdhury et al., 2010; Hossain and Piantanakulchai, 2013; Lado et al., 2008; Tan et al., 2020; Winkel et al., 2008, 2011), and the global (Fig. 4A) (Podgorski and Berg, 2020). In these studies, logistic regression (LR) was the most frequently used algorithm, with an accuracy rate of approximately 70% being achieved in corresponding studies. However, LR has shortcomings in dealing with outliers, large feature space, and the multicollinearity of variables, thus causing problems of under-fitting or low accuracy (Mood and Carina, 2010). In contrast, the CART is immune to the multicollinearity, as the decision at each node of the tree is made based on a single feature (Myles et al., 2004). In addition, the CART can also solve the problem of outliers and missing data by accommodating surrogates. Therefore, the CART is capable of making more accurate predictions (Breiman et al., 1984). For example, the CART model achieved an accuracy of 78.25% in predicting the distribution of groundwater arsenic contamination in Bangladesh (Hossain and Piantanakulchai, 2013). Moreover, the CART has also been used to investigate the factors governing elevated groundwater arsenic concentrations across the western and the eastern United States, and the significance of aridity and pH was revealed (Frederick et al., 2016). To improve the prediction ability of the CART, a weak-learner ensemble model called the boosted regression tree (BRT) was recently developed by combining the CART with an adaptive boosting method (Elith et al., 2008). The BRT model captured the corresponding major geochemical processes with an accuracy of up to 91% in predicting the probabilities of groundwater arsenic in the central valley of California (Ayotte et al., 2016). In addition, the BRT also obtained an outstanding performance with an accuracy value of approximately 90.5% in comparison with the LR model and the RF model to predict the groundwater arsenic distribution in Bangladesh. Likewise, the RF model delivered an excellent result with a prediction accuracy of 90% in this comparison (Tan et al., 2020). Moreover, the RF model was also applied to predict the global groundwater arsenic concentrations. In this study, 52 types of parameters widely covering climate, soil, geology, and topography that were known or hypothesized to be related to the accumulation and release of arsenic were initially selected to be the predictor variables. By using a relative importance analysis, soil parameters (i.e., topsoil clay, subsoil sand, pH, and fluvisols), climate variables (i.e., precipitation, actual and potential evapotranspiration and combinations thereof with temperature), and the topographic wetness index were ultimately determined as influential variables for the distribution of arsenic (Podgorski and Berg, 2020). In addition to

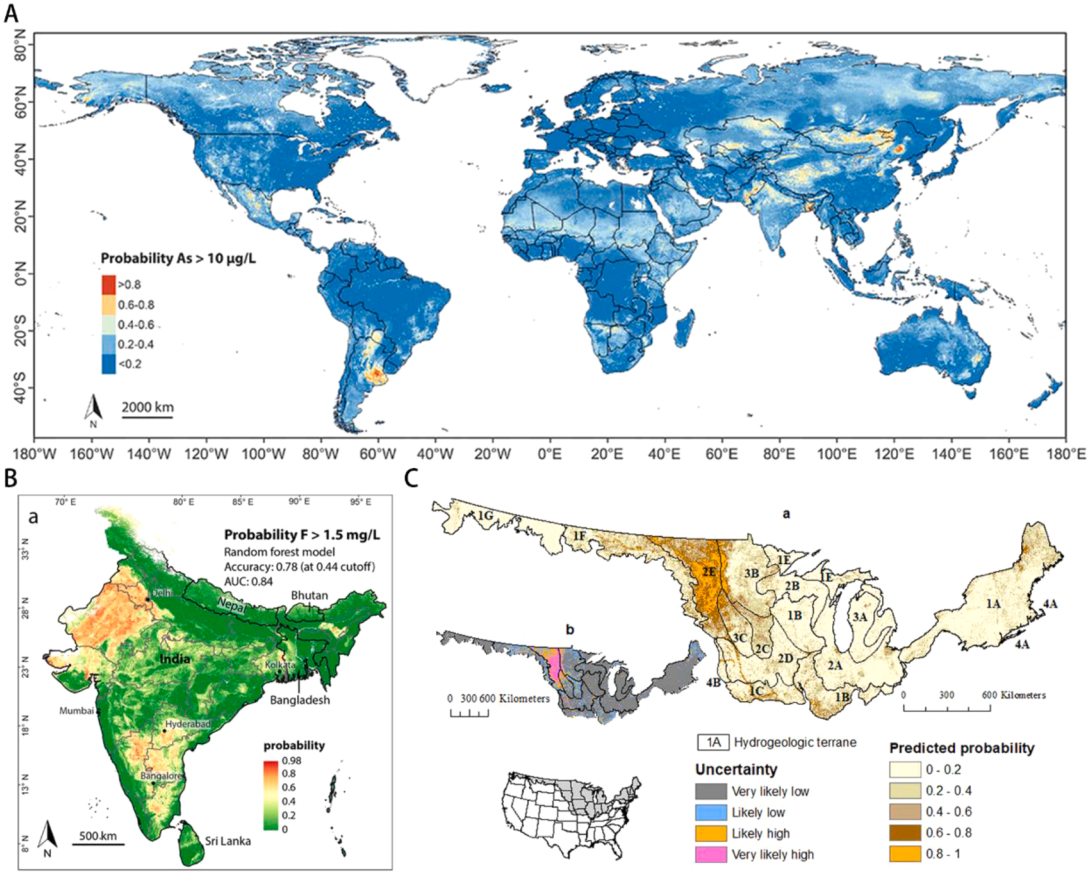

Fig. 4. The applications of ML in mapping groundwater contaminants. (A) Modeled probability of arsenic concentration in groundwater exceeding 10 µg/L for the entire globe. Reproduced from (Podgorski and Berg 2020) with permission. Copyright (2020) The American Association for the Advancement of Science. (B) Areas of aquifers with fluoride concentrations exceeding 1.5 mg/L in India, and neighboring countries of Bangladesh, Bhutan, Nepal, and Sri Lanka. Reproduced from (Podgorski et al. 2018) with permission from ACS AuthorChoice. (C) Modeled probability of high Mn indicated by Mn > 300 µg/L in the northern continental United States. Reproduced from (Erickson et al. 2021) with permission from ACS AuthorChoice.

arsenic, the RF model was also applied to predict the distribution of fluoride in groundwater throughout India (Fig. 4B) (Podgorski et al., 2018). The proposed RF model obtained a more accurate prediction with an overall prediction accuracy of 91% (Podgorski et al., 2018), compared to a previous study mapping the distribution of worldwide groundwater fluoride using the ANFIS method (Amini et al., 2008). In addition to arsenic and fluoride, the distribution of manganese in the northern continental United States was also predicted using a BRT model with a total accuracy of 83% (Fig. 4C). Environmental variables such as estimated recharge, probability of high Fe > 100 μg/L, baseflow index, and the mean annual precipitation were demonstrated to have a high relative importance on influencing the probability of high manganese (Erickson et al., 2021). Recently, a generalized additive model for location, scale, and shape (GAMLSS) regression model was applied to predict the E. coli concentration in wells across Ontario, Canada. A dataset containing E. coli concentration, well characteristics (well depth, location), and hydrogeological characteristics (bottom of well stratigraphy, and specific capacity) from 253,136 independent wells was established to train the GAMLSS model. The regression analysis revealed that bedrock wells drilled in sedimentary and igneous rock were more susceptible to contamination events, which was helpful to improve the understanding of aquifer vulnerability to contamination and for assessing water quality (White et al., 2021). In studies that mapped the groundwater contaminant distribution, it was not necessary for the ML model to accurately predict the contaminant concentration at a certain point, but to judge whether the contaminant concentration at that point exceeded the threshold (e.g., > 10 mg/L for As) according to the environmental variables. Therefore, the essence of this type of research is classification, which explains why the commonly used algorithms are LR and those based on DT. The LR algorithm is robust to small noises in the data, and will not be specially affected by slight multicollinearity, while serious multicollinearity can be solved by combining with L2 regularization. Therefore, LR performed well in dealing with environmental variables influencing the distribution of As, even though there are typically interactions among geologic and geochemical variables. However, LR covers all the features and will suffer performance decline when the feature space becomes larger. In contrast, the DT model is not limited by the size of the feature space because it can select the optimal feature according to the information gain. However, the DT model is designed to create branches for the training set, thus making it prone to over-fitting. RF overcomes this shortcoming by randomness in taking samples, determining features, and building trees. Therefore, DT-based models, especially RF, theoretically perform better than other classification algorithms. This is why DT-based models have been more popular in recent studies (Table S5).

3.3. Classifying water resources

To better manage and protect water resources, different countries and regions have adopted various standards to classify water resources according to their water quality. For example, in the National Environmental Quality Standards of Surface Water (GB3838-2002, China), water resources are divided into five categories by 24 fundamental WQIs. Each category of water resource has corresponding applications according to its water quality. However, measuring all WQIs listed in the standard is costly. Therefore, it is meaningful to classify water quality with fewer parameters that are more indicative. ML is good at evaluating the variable importance to the prediction target and has been applied to identify the key WQIs to classify water resources.

Initially, ANFIS was applied to classify the water samples of all major river basins across China using the GB3838-2002 for the classification standards. DO, COD, and NH_3-N were determined to be the influential variables, and 89.59% of the data was correctly classified (Yan et al., 2010). In addition to the GB3838-2002, other standards such as the Canadian Council of Ministers of the Environment (CCME) WQI (Canada) and the National Sanitation Foundation (NSF) WQI (U.S.) were also applied to classify water resources. For example, the SVM, the probabilistic neural network (PNN), and the kNN model were applied using the CCME WQI for the classification standards to classify the water samples of the primary aquifer in the Tehran plain. The results revealed that the kNN algorithm was the weakest classifier with the highest total number and the total value of errors. In contrast, PNN and SVM were more appropriate for a small sets of samples, and SVM presented the best performance with no errors in the calibration and validation phases (Modaresi and Araghinejad, 2014). However, in another study classifying the water of the Tiaoxi River in the Lake Taihu Watershed, the classification accuracy of SVM fell from 99.56% to 61.60% when the number of indicators was raised from one to five, indicating that it was insufficient for SVM to treat data with large features (Li et al., 2013). In addition, when classifying water samples from the Karoon River in Iran using NSF WQI as standards, the training process of SVM was reported to be more difficult and time-consuming. Therefore, the PNN model was ultimately recommended for water quality classification due to its synthetic ability to reduce the sampling costs and computation time (Dezfooli et al., 2017). To further identify a better model and the key WQIs, a more comprehensive comparison was conducted among 10 ML algorithms using big data (33,612 observations) from the major rivers and lakes in China using GB3838-2002 as the standard. Different algorithms were compared using different WQI groups as prediction indicators. Ultimately, two WQI sets (DO, COD_{Mn}, and NH_3-N; and COD_{Mn}, and NH_3-N) were identified as the key water parameters. The performances of the DT, RF, and deep cascade forest (DCF) were shown to be better than other tested algorithms, including LR, LDA, SVM, CART, NB and kNN (Fig. 5A) (Chen et al., 2020a). In addition to water quality, spatial or temporal factors can also be regarded as standards in water classification. For example, water samples from the Gomti River were classified in terms of sampling sites (spatial) and months (temporal), thereby identifying similar ones in the monitoring network to reduce sampling numbers and the annual sampling frequency. In this way, a data reduction of 92.5% without compromising the output quality was achieved (Singh et al., 2011). The nonlinear boundary problem caused by a variety of water quality parameters is one of the most distinct features of water resource classification. SVM relies on a single sample of the boundary to establish the desired separation curve, so it can deal with nonlinear decision boundaries. Therefore, compared with the application in mapping groundwater contaminants, SVM was more commonly used in classifying water resources (Tables S5 and S6). However, training support vector machine on large data is very time-consuming and sometimes it is difficult to find a suitable kernel function. Moreover, DT-based models exhibited better performance than SVM, especially in dealing with tasks with large dataset (Chen et al., 2020a). As a result, DT-based models are more suitable in classifying water resources.

3.4. Tracing contaminant sources

The accurate detection of contaminant sources in water environments is a critical step for contamination prevention and remediation. Many approaches, such as the response matrix, contaminant transport simulation, and Tikhonov regularization have been proposed to identify pollution sources (Singh et al., 2004), and ML has also joined in this attempt. For instance, an MLP was applied to simulate the time series of E. coli concentrations in a water system to locate discharge points. By inversely interpreting the transport patterns of E. coli, the source locations where E. coli was introduced into the given system were identified with an accuracy of 75% (Kim et al., 2008). Moreover, in a recent study using the RF model to identify water sample sources from three different river ecosystems, the responses of the aquatic microbial community to variations in water quality caused by pollution discharge were considered. Environmental physicochemical indices, microbiological indices, and their combination were applied as inputs to train the classifier. Microbiological indices-based models obtained the best predictions by using the abundances of the top 30 bacteria as predictor variables. With

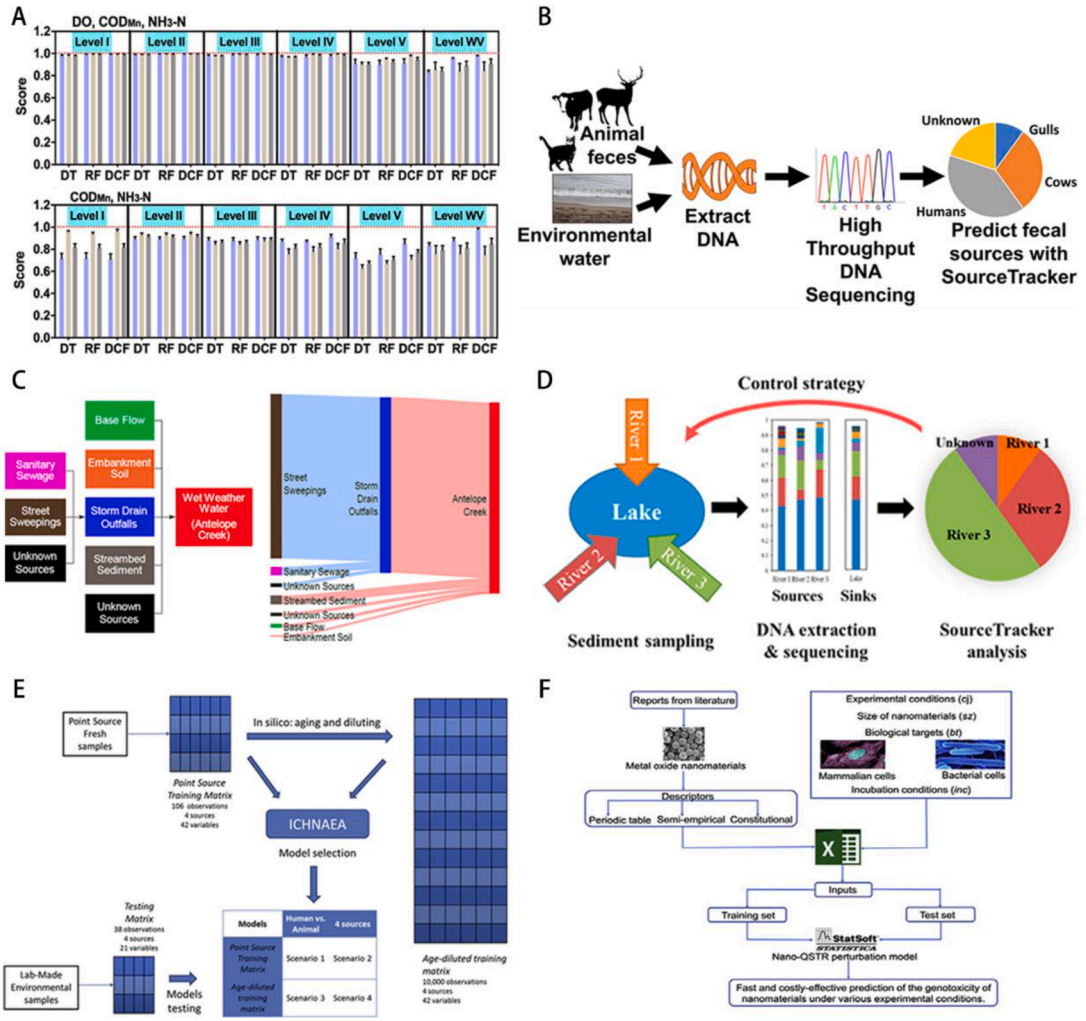

Fig. 5. The applications of ML in water resource and pollutant management in the natural water environment. (A) The surface water quality prediction performance of DT, RF, and DCF using (a) DO, COD_{Mn}, and NH_3-N; (b) COD_{Mn} and NH_3-N. Reproduced from (Chen et al. 2020a) with permission. Copyright (2020) Elsevier Ltd. (B) Schematic of a high-throughput DNA-sequencing approach for determining sources of fecal bacteria in the Lake Superior estuary. Reproduced from (Brown et al. 2017) with permission. Copyright (2017) American Chemical Society. (C) Schematic of tracking the sources of antibiotic resistance genes in an urban stream during wet weather. Reproduced from (Baral et al. 2018) with permission. Copyright (2018) American Chemical Society. (D) Schematic of tracing the sources of sediment based on the correlation between the sediment bacteria's alpha diversity, aquatic environmental variables, and aquatic sediment in Dongting Lake. Reproduced from (Zhang et al. 2019a) with permission. Copyright (2019) American Chemical Society. (E) Schematic of the computational process used to generate and validate microbial source tracking models with Ichnaea®. Reproduced from (Balleste et al. 2020) with permission. Copyright (2020) Elsevier Ltd. (F) Schematic of a perturbation theory machine learning (PTML) based QSTR approach for predicting the genotoxicity of metal oxide nanomaterials. Reproduced from (Halder et al. 2020) with permission. Copyright (2019) Elsevier Ltd.

the increasing development of gene sequencing technology, the method proposed in this study provided an economical and rapid approach to trace water sample sources based on the abundance of microbial communities (Wang et al., 2021a). Recently, ML classification programs specifically designed for contamination tracing such as SourceTracker, have gained many applications. SourceTracker can estimate the relative contribution of a specific potential source to an environmental sink based on the Bayesian approach (Knights et al., 2011). To track the sources of fecal bacteria in the Lake Superior estuary, community-based microbial source-tracking using SourceTracker was conducted. The high-throughput DNA sequencing of a fecal sample collection was used to establish a fecal library that was utilized to understand the fecal microbiome composition, as well as the marker specificity and sensitivity in several animals. It was found that fecal bacteria in the Lake Superior estuary were primarily attributed to wastewater effluent and, to a lesser extent, geese and gull wastes (Fig. 5B) (Brown et al., 2017). SourceTracker was also combined with the shotgun metagenomic technique to analyze the sources of antibiotic resistance genes (ARGs) in an urban stream during wet weather. The relative contributions of both microbes and ARGs in the sink environment were estimated using the abundances of microbial taxa and ARGs provided by shotgun metagenomics. The results revealed that storm drain outfall waters were the largest contributor of both microbes and ARGs in the urban stream, while wash-off from streets was the largest contributor of microbes and ARGs in the storm drain outfall water (Fig. 5C) (Baral et al., 2018). Moreover, SourceTracker was applied to analyze the sources of sediment pollution in Lake Dongting. A metagenomic analysis characterized the difference among community compositions of source sediment samples. Then the specific sources of sediment were identified by SourceTracker based on the inseparable relationships between sediment and adsorbed microorganisms (Fig. 5D) (Zhang et al., 2019a). Additionally, the significant relationships between sediment and microbial community were also used to investigate phosphorus sources in Lake Dongting. By analyzing phosphorus source contributions in interconnected river-lake systems using SourceTracker, a novel framework for nutrient source-tracking was established to develop effective management and

control strategies for both sediment and eutrophication in river-lake systems (Gu et al., 2020). Ichnaea® is another ML-based software utilized to improve tracking the fecal pollution sources in water. Different from SourceTracker that depends on comparing obtained gene information with a known gene library, Ichnaea® relies on library-independent markers and the abundance of fecal indicators or host-specific markers. For example, in tracking the sources of fecal pollution in a given water, different fecal indicators and source tracking markers were applied to provide host information. Ichnaea® correctly distinguished the fecal pollution of human or non-human from other several origins (Fig. 5E) (Balleste et al., 2020). Compared with the research directions introduced above, there have been fewer studies that have investigated the use of ML in tracing the source of contaminants. Despite this, most of the reviewed studies in this paper adopted ML-based contaminant tracing tools, e.g., SourceTracker and Ichnaea® (Table S7), and this not only reduced the burden of developing specialized ML models for researchers, but also provides an opportunity for researchers who are not proficient in programming and ML to use efficient contaminant tracing tools. This provides a reference for lowering the threshold of applying ML for environmental researchers.

3.5. Evaluating pollutant toxicity

Toxic pollutants discharged into the natural water environment will inevitably cause harm to aquatic ecosystems and ultimately poison humans through the food chain. Therefore, it is crucial to assess the toxicity of pollutants on aquatic organisms and humans. Animal testing is one of the primary approaches to evaluate the toxicity of chemical substances. However, the consumption of animal testing in time, money, and labor limits its applications. In addition, the rise of animal protectionism recently exacerbates the difficulty of implementing of animal testing (Takata et al., 2020). To reduce the costs and uncertainties of conventional toxicity evaluation methods, high-throughput computer technologies have become increasingly popular due to because of their high efficiency and convenience in data-driven toxicity assessments. The quantitative structure-activity relationship (QSAR) builds a quantitative relationship between the structural or physicochemical characteristics of chemicals and their properties or activities, including toxicity (Wu and Wang, 2018). However, it is increasingly difficult for the prior knowledge-based QSAR method to process tasks with increasing amounts of data, thus making it feasible to introduce ML to assist and improve the QSAR in predicting the toxicity of pollutants. ML-QSAR approaches collected large amounts of data from the related literature to reveal some universal rules or mechanisms of correlative chemical reactions or processes, thus guiding the modeling of similar reactions. Therefore, ML-QSAR has been widely applied for evaluating pollutant toxicity. Pesticides are organics that possess high toxicity against various organisms. In a recent study, five ML models (i.e., DT, NB, kNN, RF, and SVM) were severally combined with the QSAR to predict the concentration for 50% of the maximal effect (EC_{50}) of 639 diverse pesticides against *Daphnia magna*. A total of 365 molecular descriptors and seven molecular fingerprints (MF) were applied to describe the characteristics of the tested pesticides. MF referred to a method to encode the molecules into a mathematical format that is readable for computational programming according to their molecular structure. The best performance was obtained using the SVM model with a prediction accuracy of 0.848 in the test set verification (He et al., 2019). Endocrine-disrupting chemicals (EDCs) are another type of organic that can cause adverse effects on the normal function of homeostasis, reproduction, and metabolism by disordering the endocrine system of animals and humans (Lister and Van Der Kraak, 2001). A DL-based QSAR model was developed to classify and predict the toxicological effects of EDCs on sex-hormone binding globulin (SHBG) and estrogen receptors (ER). Datasets assessing diverse qualitative responses of the SHBG and ER to the tested EDCs were collected from the relevant literature. The proposed DL-QSAR model exhibited more satisfactory performance than MLR and SVM in simulating the toxicity of EDCs, with R^2 values of 0.86 for SHBG and 0.84 for ER (Heo et al., 2019). Moreover, chemicals can produce endocrine-disrupting effects by binding to the peroxisome proliferator-activated receptor γ (PPARγ). To identify the binding affinity of chemicals with PPARγ, recently, QSAR was also utilized to provide molecular descriptors for the construction of ML models (Wang et al., 2021d). In addition, a DNN-based DL model was developed for the toxicity classification of 1317 chemicals to screen for environmental estrogens. By introducing a novel three-dimensional molecular surface point cloud with electrostatic potential to describe chemical structures, this model recognized the active and inactive chemicals at accuracies of 82.8 and 88.9%, respectively (Wang et al., 2021c). In addition to organics, nanoparticles (NPs) are also of concern in environmental toxicology. NPs are emerging detrimental pollutants that can cause harmful effects, such as oxidative stress and lipid peroxidation on aquatic organisms by generating reactive oxygen species, due to their specific surface area or photocatalytic activity. To predict the toxicity of NPs, a perturbation theory machine learning (PTML)-based QSAR was applied to estimate the genotoxicity of metal oxide NPs. The genotoxicity information of 78 metal oxide NPs against 32 different biological targets was collected as a dataset. In addition, chemical periodic table-based descriptors, quantum chemical properties, and constitutional descriptors were adopted to describe the tested NPs. The results revealed that 97.81% of the cases in the test set were correctly classified, and the toxicity of nearly all the 78 NPs was correctly predicted in the external validation. Moreover, because the proposed PTML-QSAR model was built based on the relevant physicochemical descriptors, it could also provide reliable insights in terms of the genotoxicity mechanisms of NPs (Fig. 5F) (Halder et al., 2020). In addition to combining with QSAR, ML is sufficient to predict the toxicity of pollutants independently. For example, the 50% lethal concentrations (LC_{50}) of 400 compounds were predicted using the RF, SVM, and XGBoost algorithms based on 12 types of MF. The optimal prediction results were achieved using the SVM model with an accuracy of 92.2%, which was better than those generated by the QSAR or ANN methods (Ai et al., 2019). Additionally, the performance of kNN, SVM, ANN, RF, AdaBoost, and GBM for predicting 50% hazardous concentrations (HC_{50}) of 578 chemical substances were compared. The RF model was found to outperform the other models, including the QSAR and ECOSAR methods, as it could explain 63% of the variability of HC_{50} (Hou et al., 2020). In the reviewed studies above, various environmental variables were required to form the feature space to train the ML model, because the predicted objectives (e.g., water quality and classification, contaminant distribution and sources) were closely related to the surrounding environmental factors. However, toxicity is an attribute of the pollutant itself to the organisms. Consequently, when training a model to predict the toxicity of chemicals, the characteristics of the tested chemical itself are required to form the feature space. There were several methods adopted in the reviewed studies to describe the molecular characteristics, including molecular descriptors and molecular fingerprints, and these are also basic research tools for the QSAR methods. Therefore, ML models combined with QSAR were often used for predicting the toxicity of pollutants, and they obtained good prediction results (Table S8). However, with the deepening of studies on the prediction of pollutant toxicity, more cutting-edge descriptors, or patterns, such as higher dimensional descriptors and transformer architecture, are encouraged to be utilized for pollutant toxicity prediction.

4. Technology optimization and operation monitoring in engineered water systems

Engineered water systems refer to a series of processes that consist of taking water from natural waters, drinking water purification and distribution, sewage or wastewater collection and treatment, and water reuse or draining back to nature in modern cities and towns (Lu et al., 2016). Effective management of engineered water systems will be of

great help in providing clean water to users and preventing ecological pollution. To achieve this, ML has been applied widely in each link of engineered water systems, and related studies and applications of ML are reviewed in this section. In addition, ML has also been combined with adsorption and oxidation techniques, and the characterization analysis of pollutants, which are also necessary steps in engineered water systems. Therefore, the applications of ML in modeling adsorption and oxidation, and assisting characterization analysis, are also reviewed below.

4.1. Modeling treatment technique

Lab-scale development and optimization are fundamental for actual applications of treatment technologies for drinking water or wastewater in DWTPs and WWTPs, especially for the removal of emerging pollutants requiring specific treatment techniques. Adsorption and oxidation are common processes in both laboratory experiments and treatment plants to remove pollutants with low biodegradability or biological toxicity. Accurate simulation of treatment processes is beneficial for optimizing the reactant dosage, selecting an appropriate approach, or scaling up reactors for practical use. Some representative applications of ML for modeling the adsorption and oxidation processes to improve the pollutant removal efficiency are summarized below.

Adsorption is a popular process in wastewater treatment, especially for removing pollutants that are immune or even noxious to conventional biological treatment processes, such as heavy metals and some types of organic matter (Dabrowski, 2001). For example, in predicting the removal efficiency of Pb (II) adsorbed by red mud, the adsorbent dosage, contact time, and pH were used as variables to train an MLP model. The results were then compared with the response surface methodology (RSM). The RSM is a common method used to assess the effects of independent variables on the reaction process (Yetilmezsoy et al., 2009). The results showed that the MLP exhibited a better performance for predicting the lead removal efficiency than RSM did, with R^2 values of 0.898 for the MLP and 0.672 for the RSM (Geyikçi et al., 2012). Moreover, the performance of ML could be further improved by combining it with optimization algorithms. The genetic algorithm (GA) is a global and parallel optimization algorithm that can automatically acquire and accumulate the knowledge of a search space during the process of searching, thus controlling the searching process adaptively to obtain the best solution (Álvarez et al., 2016). In the adsorption of Cu (II) by reduced graphene oxide-supported nanoscale zero-valent iron (nZVI/rGO) magnetic nanocomposites, MLP-GA performed well with an MAE value of 1.13%, while the values were 3.64% and 7.44% for the MLP and RSM methods, respectively. Moreover, the combination of MLP and particle swarm optimization (PSO) obtained a better prediction with a minimum MAE value of 0.46%. PSO is another optimization algorithm through which global optimization is realized by group iteration based on the interaction among and update of particles (each represents a possible solution to a problem) (Kennedy and Eberhart, 1995). By using the MLP-PSO to identify and adjust the critical parameters controlling adsorption processes, the Cu (II) removal efficiency was improved by 3.15% and 8.54% compared to that of the MLP-GA model and the RSM model (Fan et al., 2017). Aside from heavy metals, ML was also applied in modeling the adsorption of some types of organic matter. For instance, triclosan (TCS) is a broad-spectrum chlorinated antibacterial ingredient that possesses high health risks and ecological toxicities. The absorption process of TCS by a novel host-guest complex (MWCNT/PEG/b-CD) was simulated using the GRNN, ANFIS, and RSM models, in which ANFIS showed a better prediction ability than the other two models. Moreover, GA was introduced to optimize the experimental design, and the maximum TCS removal efficiency was improved to 99.50% (Azqhandi et al., 2019). Synthetic dye is another type of organic compound that can cause environmental pollution (Salleh et al., 2011). Ghaedi et al. summarized in early 2017 the applications of ANN models for simulating synthetic dye adsorption. Their review concluded that four types of ML models, including MLP, ANFIS, SVM, and the hybrid with GA or PSO optimization, were primarily used in previous studies (Ghaedi and Vafaei, 2017). The better performance of the hybrid networks with optimization algorithms was confirmed in this review, as well as in a study published afterward (Jun et al., 2020).

In addition to batch experiments in a single study, the datasets generated from the published literature have also been used to develop ML algorithms. For instance, to train a deep learning neural network (DLNN), a dataset containing 200,000 sample scenarios was generated from over ten years of studies that used a carbon-based adsorbent (i.e., carbon nanotube (CNT), activated carbon, biochar, graphene, carbonaceous, or graphite) to absorb anionic, cationic, and zwitterionic ionizable organic compounds. The proposed DLNN model exhibited a strong generalization and forecasting ability for the adsorption of ionizable organic compounds by a wide range of carbonaceous materials, with R^2 values exceeding 0.98 and 0.91 for predicting the Freundlich coefficient K_F and the exponent n, respectively (Sigmund et al., 2020). Another study mined literature for the adsorption data of 165 organic chemicals on 50 biochars, 34 CNTs, 35 granular activated carbons (GAC), and 30 polymeric resins to train a neural network with the poly-parameter linear free energy relationships (pp-LFER). The results showed that the proposed NN-LFER model accurately simulated the tested adsorption processes, with R^2 values ranging from 0.86 to 0.91 (Fig. 6A) (Zhang et al., 2020). More notably, a graphical user interface was provided in these two studies for those who are not skilled in computer operation. This provides great convenience for researchers and practitioners in the fields of water purification and pollution control to select the appropriate sorbent for a given contaminant based on their requirements.

Apart from adsorption processes, advanced oxidation processes (AOPs) are also effective technologies to treat pollutants with high chemical stabilities and low biodegradabilities. The development of AOPs can be traced back to the late 19th century when H. J. H. Fenton discovered that hydrogen peroxide could generate hydroxyl radicals (OH■) catalyzed by ferrous ions to oxidize organic contaminants that then became known as the Fenton reaction (Nguyen et al., 2020). Since then, an increasing number of AOP processes have been developed based on various approaches including ozone-based, ultraviolet (UV)-based, photocatalytic, electrochemical, and physical methods (Bolton et al., 1996; Miklos et al., 2018). Currently, ML is also applied to simulate the oxidation processes for removing organic pollutants. Khataee et al. reviewed early applications of ANN in modeling homogeneous and heterogeneous nanocatalytic AOPs that included photocatalytic, photooxidative, and electrochemical treatment processes. The authors confirmed that ANN was an effective and simple approach to describe AOP processes, without considering the complex effects, such as the radiant energy balance, mass transfer, and the spatial distribution of the absorbed radiation (Khataee and Kasiri, 2010). Here the studies published after or covering content outside the scope of Khataee and Kasiri's review were summarized to present the applications of more ML models in a wider range of AOPs. For the degrading of antibiotics including amoxicillin, ampicillin, and cloxacillin with Fenton reaction, an MLP model was used to simulate the reaction processes in terms of COD removal. An accurate prediction of COD removal efficiency was achieved with an R^2 of 0.997, additionally, the sensitivity analysis showed that the H_2O_2/Fe^{2+}molar ratio was the most influential parameter on antibiotic degradation. This result pointed the direction for further improving the removal efficiency (Elmolla et al., 2010). Based on the Fenton reaction, photo-Fenton processes are developed due to the exposure of the reaction system to solar or UV light, both of which can speed up the generation of hydroxyl radicals in Fenton reactions (Kavitha and Palanivelu, 2004). During the degradation of oil-contaminated wastewater by the homogeneous photo-Fenton ($UV/H_2O_2/Fe^{2+}$) process, the MLP model accurately predicted the oil removal efficiency. In addition, pH was discovered to be the most influential parameter that affects oil degradation, compared with H_2O_2, Fe^{2+}, oil concentration, or the irradiation time and temperature

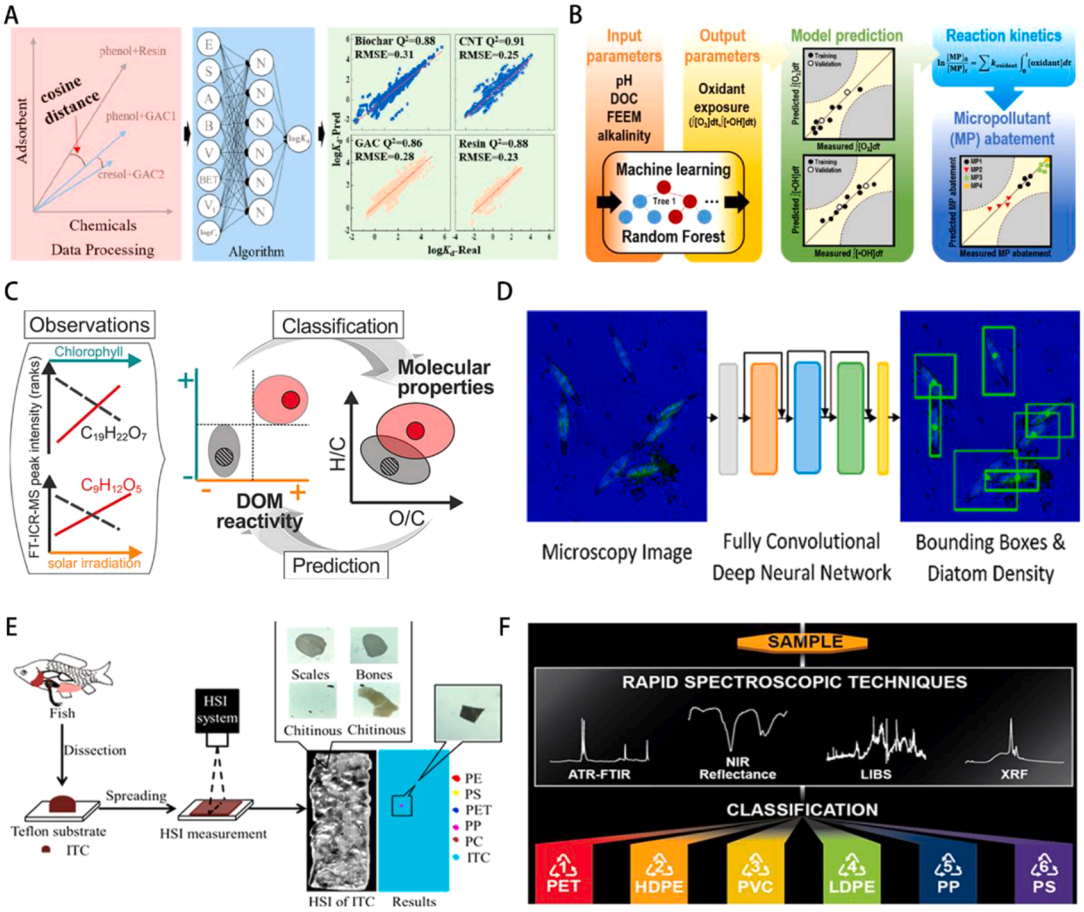

Fig. 6. The applications of ML in modeling adsorption and oxidation processes, and assisting characterization analysis. (A) Predicting aqueous adsorption of organic compounds onto biochars, carbon nanotubes, granular activated carbons, and resins with machine learning. Reproduced from (Zhang et al. 2020) with permission. Copyright (2020) American Chemical Society. (B) Schematic for prediction of oxidant exposures and micropollutant abatement during ozonation. Reproduced from (Cha et al. 2020) with permission. Copyright (2021) American Chemical Society. (C) Schematic for the understanding of DOM reactivity in freshwater. Reproduced from (Herzsprung et al. 2020) with permission. Copyright (2020) American Chemical Society. (D) Fully CNN for detection and counting of diatoms after short-term field exposure. Reproduced from (Krause et al. 2020) with permission. Copyright (2020) American Chemical Society. (E) Hyperspectral imaging-based method for rapid detection of microplastics in the intestinal tracts of fish. Reproduced from (Zhang et al. 2019c) with permission. Copyright (2019) American Chemical Society. (F) Rapid identification of marine plastic debris via spectroscopic techniques and ML classifiers. Reproduced from (Michel et al. 2020) with permission. Copyright (2020) American Chemical Society.

(Mustafa et al., 2014). In addition to facilitating the Fenton process to generate hydroxyl radicals in the photo-Fenton processes, in photocatalytic processes, light can also provide photons to excite the lone electron on a semiconductor (e.g., TiO_2 and ZnO) to create an electron-hole pair (e^--h^+). Then, a series of chain oxidative-reductive reactions will occur to degrade the contaminants on the semiconductor surface with electron-hole pairs (Chong et al., 2010). During the photocatalytic degradation of m-cresol by synthesized Mn-doped ZnO nanoparticles under visible-light irradiation, three types of MLP with different architectures were developed to simulate the reaction processes. A comparison between the predicted output and the experimental results showed that ANN trained with quick propagation and accurately simulated the degradation with an R^2 of 0.9938. Moreover, the predicted optimal parameter values were used to optimize the reaction condition, under which the actual removal efficiency of m-cresol was improved to 99% (Abdollahi et al., 2014). In addition to Fenton and photocatalytic processes, ozonation is another popular AOP technology. During the degradation of micropollutants (MPs) by the ozonation process, the RF model was applied to predict the oxidant exposures and MP abatement during ozonation. Parameters including pH, alkalinity, DOC, and fluorescence excitation-emission matrix (FEEM) data were adopted as input variables. Ultimately, the proposed RF model obtained an accurate prediction with R^2 values of 0.798 and 0.772 for O_3 and OH■ exposure, respectively, and 0.904 for MP abatement (Fig. 6B) (Cha et al., 2020). Ozonation could also be combined with other processes to enhance the ability of treating pollutants. For instance, a synthesized Dy_2O_3/TiO_2/graphite nanocomposite was used to assist the photocatalytic ozonation process for the degradation of textile dyeing in wastewater, while O_3/H_2O_2/Zr-pumice was applied in the decolorization of Rhodamine B dye. For both studies, the MLP models were applied to predict the removal efficiency, and achieved excellent accuracy with R^2 values of 0.99 and 0.9948 (Sheydaei et al., 2020; Shokoohi et al., 2020).

Data mining in AOP-related studies has also been conducted. For instance, 446 data points concerning the photo-degradation rate constants (-log(K)) of organic contaminants in a TiO_2-UV photocatalyst system were collected. Then, an MLP model was trained using a variety of reaction parameters, including ultraviolet intensity, TiO_2 dosage, initial concentration, and the MF information of 78 tested organic contaminants. The proposed MLP model showed its reliability and stability as the predicted photo-degradation rate constants were in good agreement with experimental data with an average R^2 value of 0.873 (Jiang et al., 2020). In another study, an MLP was compared with the quantitative structure-property relationship (QSPR) method to model hydroxyl radical rate constants ($k_{HO■}$) using a diverse dataset of 457 water contaminants from 27 various chemical classes. With a total of

785 molecular descriptors as variables, QSPR achieved a simulation of $k_{HO\bullet}$ with R^2 of 0.724, while the MLP performed better with R^2 of 0.847 (Borhani et al., 2016). Moreover, based on this dataset, a DNN combined with MF was developed to predict the $k_{HO\bullet}$ of other 46 organic pollutants and obtained a prediction accuracy of 0.724 (Zhong et al., 2020). The accurate prediction of reaction rate constants, such as degradation rate constants and hydroxyl radical rate constants, will undoubtedly help researchers select appropriate reaction types and reactants, as well as to design more efficient AOP-based water treatment units. The summary of Tables S9 and S10 indicates that MLP is the most frequently used algorithm for the simulation of adsorption and oxidation reactions, especially in earlier studies with small datasets. MLP is one of the most basic algorithms of an ANN, and it can approximate any nonlinear function relation with high precision. In addition, MLP has strong robustness and fault tolerance to noise datasets. Additionally, after 30 years of development, there are many MLP-based programs and software that can be used directly by researchers who are not familiar with programming. Therefore, MLP has been widely used in dealing with tasks with smaller dataset and fewer features. For example, data generated by batch experiments of a single study are reviewed herein. In these studies, optimization algorithms (e.g., GA and PSO) played important roles in improving the performance of MLP, which could be an experience worth learning. However, when dealing with tasks with large dataset and many features, for example, data generated by collecting from similar studies, an ANN with more hidden layers, that is, a DNN, has a stronger ability to model objects or abstract features, and can also simulate more complex models. Therefore, DNNs are increasingly popular in studies simulating various reaction processes, especially when the computing power of computers is becoming increasingly stronger.

4.2. Assisting characterization analysis

In addition to modeling and optimizing of physical and chemical treatment processes, ML has also been applied in assisting characterization analyses. By processing data or images produced by various analytical instruments, ML approaches can deepen and broaden the analysis information, or reduce the workload and improve the efficiency of testers. For example, to improve the predictability of drinking water disinfection by-products (DBPs), parallel factors analysis, PCA, and autoencoders (AE) were applied to process high dimensional fluorescence data of DBPs. Then a neural network approach was used to identify fluorescence regions associated with DBP formation, thus estimating the concentrations of DBPs. The AE-NN was found to provide more accurate predictions for both trihalomethanes (THMs) and haloacetic acids (HAAs) by when compared to common organic measuring methods. The approach proposed in this work provided a promising approach to rapidly quantify DBPs or other organics by analyzing fluorescence EEM data (Peleato et al., 2018). In addition to the direct determination of DBPs, determining the relationships between the molecular properties of dissolved organic matter (DOM) and their reactivity as potential precursors of DBPs is also instructive for DBP control and management. For instance, Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) was used to provide chemical information about DOM molecules. Then the RF model was introduced to identify the molecular reactivity classes based on the chemical properties and peak intensities of the FT-ICR-MS. Finally, a classification approach based on molecular formulas rather than a prior definition of biogeochemical reactivity was applied to discriminate biogeochemical reaction types. With this approach, potential precursor DOM of DBPs could be distinguished and specifically removed in drinking water treatment processes (Fig. 6C) (Herzsprung et al., 2020). Apart from acting as DBP precursors, some organic micropollutants (MPs), including various anthropogenic pharmaceuticals and industrial chemicals are also a negligible threat to the water quality. High-resolution mass spectrometry (HRMS) has been commonly used to measure the concentration of MPs. However, the high cost of the stable isotope-labeled (SIL) standard used in HRMS restricts its economic efficiency in measuring MPs. To reduce the cost of HRMS analysis, DL and ML models were proposed to estimate the concentrations of five MPs without SIL standards, including sulpiride, metformin, benzotriazole, tebuconazole, and niflumic acid. First, 35 alternative MS subsets were selected from hundreds of MS results to examine their correlation with the SIL standard peak for determining the specific MS subset to replace SIL standards. Then four convolutional neural network (CNN) models, as well as the RF, SVM, and ANN models were applied to estimate concentrations of tested MPs by analyzing their mass spectrum using the selected MS subset. The results revealed that the CNN model, ResNet 101, achieved the best performance in estimating the concentrations of the five MPs, with an average R^2 value of 0.84 (Baek et al., 2021). The CNN algorithm has also been applied in image recognition to classify and count diatoms on fouled surfaces during short-term field exposures. The acquired images were saved as two-channel red-green-blue (RGB) images, with the phase-contrast images encoded in the blue channel and the fluorescence images encoded in the green one. The proposed DL model could analyze the tested images and distinguish diatoms from sand, dirt, and bacteria on fouled surfaces in only 30 min, while it took 180 h of manual counting (Fig. 6D) (Krause et al., 2020). Microplastics are a type of emerging harmful pollutants derived from the fragmentating of plastics. ML approaches have also been applied to identify, quantify, and classify microplastics. For example, near-infrared hyperspectral imaging (NIR-HSI) was used to characterize the intestinal tract content of crucian carps (*Carassius carassius*) that were exposed to water that contained various tested microplastics. An SVM classifier was developed to detect and classify the microplastics according to their NIR-HSI. The validation with Raman spectroscopy indicated that the SVM model performed well in classifying five different microplastics with high precision (97.19%) and recall (91.98%) values (Fig. 6E) (Zhang et al., 2019c). Moreover, the performance of the SVM, kNN, and LDA algorithms for identifying consumer plastics (CP) and marine plastic debris (MPD) was compared. Four spectroscopic techniques, including attenuated total reflectance-Fourier, transform infrared spectroscopy (ATR-FTIR), NIR reflectance spectroscopy, laser-induced breakdown spectroscopy (LIBS), and X-ray fluorescence spectroscopy (XRF), were applied to characterize the tested plastics. The comparison results revealed that ATR-FTIR combined with kNN classifiers obtained the optimal classification accuracies of 95% and 99% for CP and MPD, respectively (Fig. 6F) (Michel et al., 2020). To assist the characterization analysis, ML is primarily used to deal with data or images generated by various analysis equipment. For image processing, CNNs are undoubtedly a popular and powerful option. For data processing, it is also divided into a regression prediction (e.g., DBP and MP concentration estimation) and a classification prediction (e.g., DBP precursor and microplastic classification). Based on the above analysis of the regression and classification algorithms, it is not difficult to select an appropriate regression or classification algorithm according to the volume of both dataset and features. For assisting the characterization analysis with ML, the real difficulty is to find the combination point of ML technology and traditional analytical methods.

4.3. Purifying and distributing drinking water

In drinking water treatment plants (DWTPs), conventional treatment processes including coagulation, sedimentation, filtration, and disinfection, are designed to supply adequate and safe drinking water to users (LaPara et al., 2015). However, demand for high-quality drinking water and the gradual deterioration of raw water has brought requirements for more efficient treatment and management in DWTPs. Comprehensive and timely monitoring of water quality is essential for improving the management and automation development of DWTPs. To this end, ML has been applied to simulate water quality in a series of treatment units to provide real-time information on the operational conditions of

DWTPs (Antanasijevic et al., 2013a). Some of the relevant studies are reviewed below.

Pre-oxidation is an optional process according to the quality of the raw water. This be conducted prior to the coagulation procedure to oxidize iron, manganese, or taste and odor substances, and control microorganisms or DBP precursors in raw water (Hu et al., 2018). Following pre-oxidation, coagulants are added to coagulate the insoluble contaminants by charging neutralization, enmeshment, or sweep to form particles and flocs that can be removed during subsequent sedimentation and filtration processes (Hogg 2000). Therefore, an appropriate dosage of oxidants and coagulants is critical for the entire drinking water treatment process. However, the dosage of oxidants or coagulants is determined by complicated factors, such as NOM, turbidity, conductivity, and temperature (Heddam et al., 2012). Therefore, there is not a simple empirical approach nor an accepted mathematical model to determine the optimal dosage for both oxidants and coagulants (Hu et al., 2018). Currently, ML approaches have been applied for interpreting relationships between the raw water characteristics and the demands of oxidants or coagulants. For pre-oxidizing oxidants, an MLP model was developed to predict the $KMnO_4$ demand. In addition to an accurate simulation of the $KMnO_4$ dosing rate, the proposed model also found that turbidity had the highest impact on $KMnO_4$ demand, while the inflow rate had the least (Fig. 7A) (Godo-Pla et al., 2019). For coagulants, the MLP and ANFIS models were compared for simulating the poly aluminum chloride (PAC) dosage. The ANFIS model well predicted the PAC dosage with the PAC dosage of yesterday being the only input. However, in an emergency such as high turbidity caused by a rainstorm, the MLP output more stable predictions (Fig. 7B) (Wu and Lo, 2008). After coagulation, the sedimentation process will remove the formed particles and flocs with high density, while the filtration process removes the remaining suspended ones. An MLP model successfully predicted the post-filtration particle counts for three different size ranges in a dual-media filter of a DWTP (Griffiths and Andrews, 2011). Another MLP model was combined with GA optimization to predict the transmembrane pressure of a ceramic membrane microfiltration system. The proposed model optimized the settings for filtration time, flux, and aluminum dosage, thus leading to the minimized operational cost with a reduction of approximately 15%

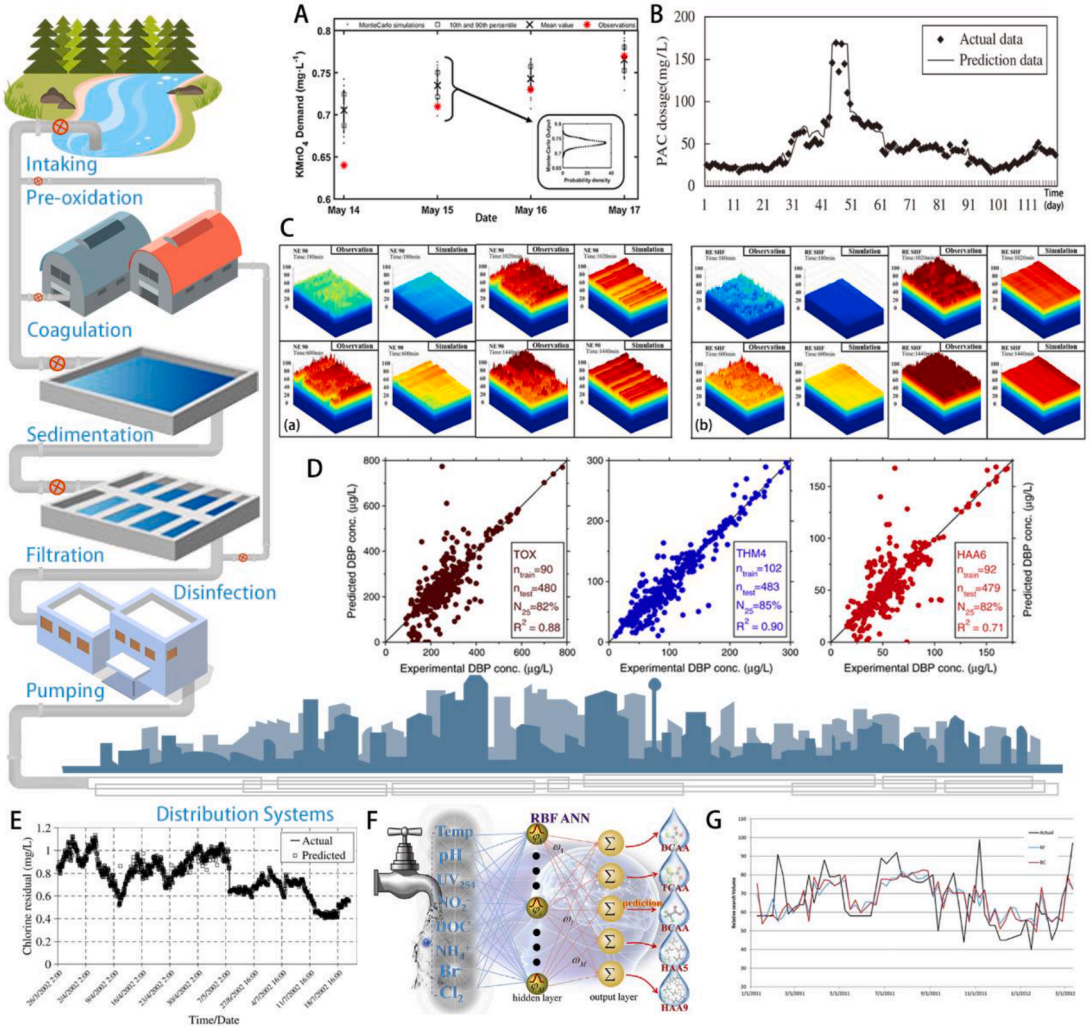

Fig. 7. The applications of ML in water purification and distribution. (A) Representation of predictions for $KMnO_4$ demand time-series. Reproduced from (Godo-Pla et al. 2019) with permission. Copyright (2019) Elsevier Ltd. (B) The optimal model of ANN for predicting real-time coagulant dosage in water treatment. Reproduced from (Wu and Lo 2008) with permission. Copyright (2008) Elsevier Ltd. (C) The comparison of observed and simulated 3D fouling images (The unit of the axis is μm). Reproduced from (Park et al. 2019) with permission. Copyright (2019) Elsevier Ltd. (D) Comparisons of ANN predictions with experimental measurements for DBP formation in conventionally treated waters. Reproduced from (Kulkarni and Chellam 2010) with permission. Copyright (2010) Elsevier Ltd. (E) Forecasts of chlorine in a water distribution system at Aldinga. Reproduced from (Bowden et al. 2006) with permission. Copyright (2006) Elsevier Ltd. (F) Schematic for predicting DBPs in a DWDS. Reproduced from (Lin et al. 2020) with permission. Copyright (2020) Elsevier Ltd. (G) Actual and predicted search volume for symptoms of gastrointestinal illness with the RF and bagged CART models. Reproduced from (Shortridge and Guikema 2014) with permission. Copyright (2014) Elsevier Ltd.

(Strugholtz et al., 2008). Membrane filtration has gained increasing popularity in water treatment processes due to its strong purification ability. However, a major problem facing the membrane filtration process is membrane fouling that would reduce the purification ability and shorten the working life of the membrane. To monitor and predict membrane fouling, an MLP model was used to simulate the membrane resistance during the nanofiltration process. The developed model correctly predicted the 93% of the experimental data with an absolute relative error < 5%, which sufficed to reduce or even replace expensive site-specific tests (Shetty and Chellam, 2003). In addition, an image analysis was also used to investigate membrane fouling. A DNN model was trained using thousands of high-resolution fouling layer images that were generated using optical coherence tomography characterizing membrane fouling during nanofiltration (NF) and reverse osmosis (RO) filtration. The proposed DNN accurately reproduced two- or three-dimensional images of the organic fouling growth with the R^2 values of 0.99 in the simulation for both fouling growth and flux decline (Fig. 7C) (Park et al., 2019). Moreover, determining the relationships between the physicochemical properties of NOM and their effects on membrane fouling is also helpful to prevent fouling events. Recently, a PCA model was used to analyze the relationships between organic matter and their fouling propensity. Then a clustering analysis was applied to classify the water resources into three groups according to their low, intermediate, or high fouling potential. The results revealed a correlation between strong fouling events and a combined increase of carbon and protein like-substances contents in water. This conclusion was helpful for membrane users to make a better selection of feed water resources to prevent membrane fouling events (Teychene et al., 2018). Following the pre-oxidation, coagulation, sedimentation, and filtration processes, disinfection is performed to further inactivate microorganisms. Chlorine is the most common disinfectant because of its high efficiency and low cost and its persistence that meets the residual disinfectant demand. However, chlorine will interact with NOM to generate DBPs, such as trihalomethanes (THMs), haloacetic acids (HAAs), and total organic halide (TOX), and these can act as human mutagens, carcinogens, and teratogens (Hamidin et al., 2008). An MLP model was developed to simulate the concentrations of these three types of DBPs in raw and treated water, and accurate predictions were obtained with R^2 values ranging from 0.71 to 0.97 (Fig. 7D) (Kulkarni and Chellam, 2010). In addition to the water quality or operational conditions in various treatment units inside a DWTP, a recent study forecasted the overall performance of DWTPs across China. The authors collected the monthly data of water quality and operational parameters from 45 DWTPs nationwide as the dataset. By using these data, they proposed an MLP model that well predicted the water production variation. This work enabled decision-makers and DWTP managers to revise production plans in response to different raw water quality, water quality standards, and market demands (Zhang et al., 2019b).

After the treatment in DWTPs, drinking water of qualified quality and quantity will be delivered to consumers via drinking water distribution systems (DWDSs). Therefore, water quality in DWDSs is also important for the entire drinking water supply system. Discoloration is one of the most disturbing problems facing drinking water producers and users. Elevated concentrations of Fe and Mn have been recognized as the major causes of drinking water discoloration (Speight et al., 2019). Therefore, predicting the accumulation potential of Fe and Mn is important for preventing discoloration. An MLP model was developed to investigate the influence of chemical properties of water on the accumulation behaviors of Fe and Mn. The results showed that free chlorine residual (FCR) played complex roles in affecting the accumulation of Fe and Mn. Concretely, a high concentration of FCR would oxidize soluble Fe^{2+} and Mn^{2+} to insoluble Fe^{3+} and Mn^{4+}, thus aggravating discoloration. On the contrary, as a disinfectant, FCR could sterilize the oxidizing bacteria to prevent biological oxidation of Fe and Mn, thus easing discoloration. Therefore, to achieve the equilibrium of chemical and biological oxidation, the FCR concentration was recommended to be 0.8 to 1.8 mg/L in the tested DWTSs (Danso-Amoakoa and Prasad, 2014). Moreover, an appropriate amount of FCR can also restrain the growth of other harmful microbial pathogens, while high dosing will lead to undesired taste, corrosion of the pipe network, and an increase in treatment costs. Therefore, it is beneficial to accurately predict FCR in DWDSs to achieve a balance among bacteriostasis, pleasant water quality, and economic cost (May et al., 2008). In study that applied the GRNN model to forecast FCR in a DWDS, the flow, temperature, and chlorine residual data were collected as variables. Then a GA method was used to optimize the division of the dataset into training, testing, and validation sets. The results showed that the GRNN model accurately predicted the FCR up to 72 h in advance with an R^2 value of 0.9617 (Fig. 7E) (Bowden et al., 2006). In addition to the FCR, the concentration of DBPs in DWDSs is another hotspot of concern related to the disinfection process. For instance, an RBF NN predicted the concentration of nine different HAAs with an accuracy ranging from 75% to 91% (Fig. 7F) (Lin et al., 2020). Moreover, a Bayesian network found that monochloramine had a negative effect, while DOC and TDS exerted positive effects on the formation of THMs (Li et al., 2020a). In addition to investigating factors that influence the water quality in DWDSs, ML has also been used to forecast the pipeline failures in the distribution networks. For instance, contamination events in DWDSs can be caused by the injection of any potential pollutants. To detect anomalous fluctuations in water quality, an MLP model was used to depict the temporal variations in water quality indicators. Then the Bayesian updating rule was recursively applied to analyze the probability of a related contamination event (Perelman et al., 2012). Moreover, the recursive Bayesian rule was also combined with GA optimization to improve the performance of the MLP-Bayesian. By applying adaptive updating dynamic thresholds to optimize the fixed thresholds in the outliers' classification module, better performance in the detection of contamination events in DWDSs was obtained (Arad et al., 2013). In addition, pipeline leakage detection and location in DWDSs are also of significance. For instance, the SVM and MLP models were respectively used to simulate the leakage size and location by analyzing the time series behavior of flow and pressure in pipe networks (Makaya and Hensel, 2016). Interestingly, Shortridge et al., compared pipe break events with weekly internet search volume for symptoms of gastrointestinal illness using data mining techniques. The results revealed a positive correlation between elevated risk on public health and pipe failures (Fig. 7G) (Shortridge and Guikema, 2014). This finding emphasized the importance of the application of ML in real-time monitoring of the functional integrity of pipelines. Like water quality predictions in natural water systems, the volume of features involved in a water quality simulation for drinking water treatment and distribution is relatively small, and water quality parameters are interconnected and affect each other through chemical and physical interactions. Therefore, NN-based models (e.g., MLP and ANFIS) are competent for most water quality prediction tasks and have been applied widely (Table S12). Additionally, LSTM based on time series and image recognition technology based on DL have also been applied for predicting membrane fouling. In fact, water quality parameters in drinking water treatment and distribution processes are all time series that provide the feasibility for LSTM to solve more problems in engineered water systems. Moreover, image recognition technology can also be used to investigate processes of visual changes such as coagulation, sedimentation, and discoloration.

4.4. Collecting and treating sewage water

Sewer networks are designed to collect domestic sewage and industrial and hospital wastewater, and transport them to WWTPs for contaminant removal. Therefore, the good condition of a sewer network plays an important role in maintaining the sanitation of an urban environment. Currently, ML approaches have been applied to improve the maintenance of sewer networks. In sewers, fluctuations in flow rate and pressure typically cause debris sedimentation that can gradually

form accumulative deposits. Bottom deposits will reduce the transport capacity of pipelines and form anaerobic conditions to cause corrosion or odor problems (Bonakdari and Larrarte, 2006). To monitor the deposition situation and predict the bed loads inside pipelines, an ANFIS model was applied to analyze the effects of sewer geometrical features and flow hydraulic characteristics (Azamathulla et al., 2012). In addition, an RF model was also used to predict the self-cleansing sewer velocity, and this allowed the flow to spontaneously scour and remove deposited sediments from the pipes. The results showed the most important variable that determines the self-cleansing velocity was the volumetric sediment concentration (Fig. 8A) (Montes et al., 2020). In addition, the multigene genetic programming technique was adopted to estimate the particle Froude number in large sewers. Both works were found to be conducive to self-cleansing sewer design (Safari and Danandeh Mehr, 2018). Recently, an ensemble procedure comprised of the Network K-function geographically weighted Poisson regression,

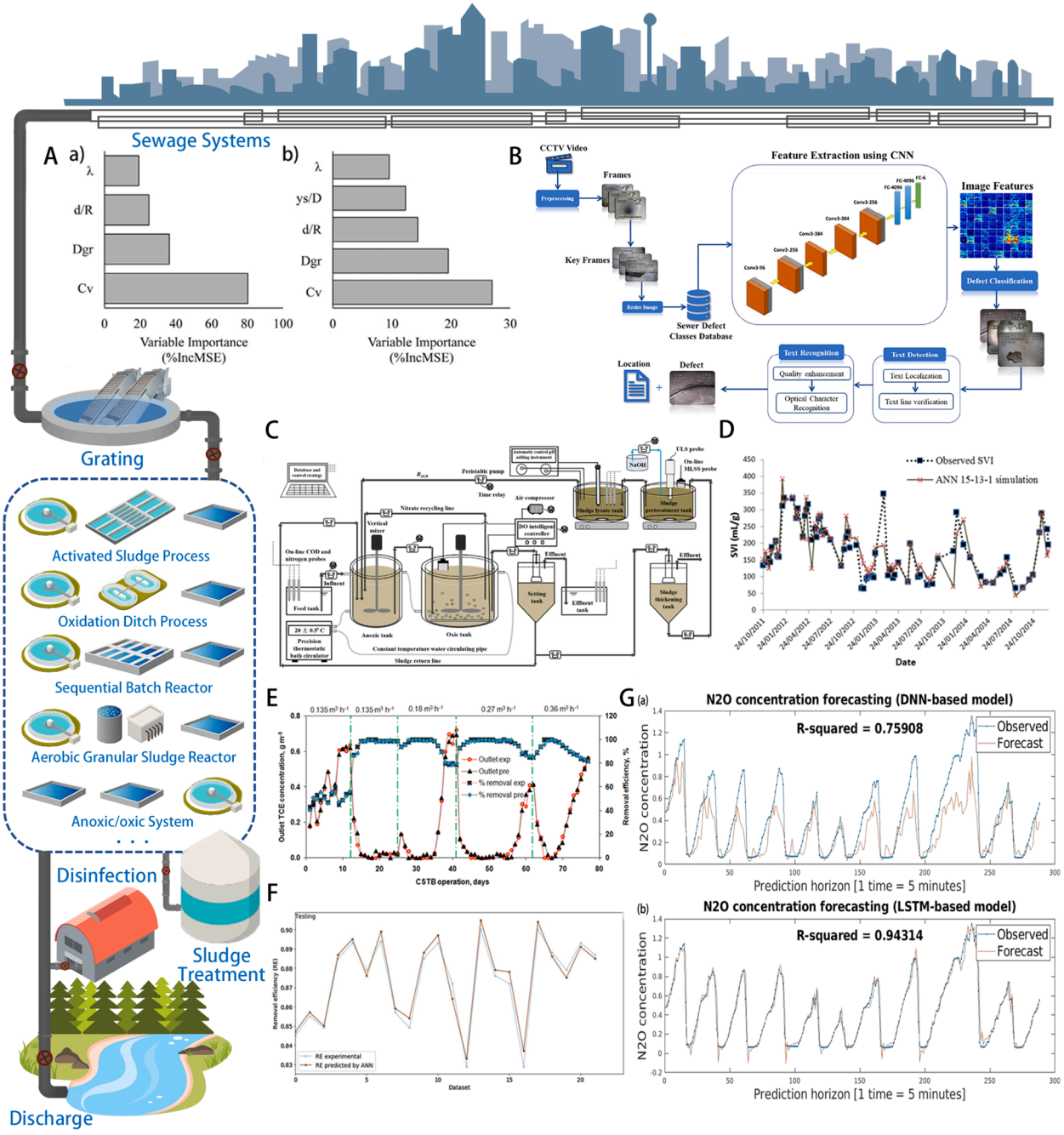

Fig. 8. The applications of ML in wastewater collection and treatment. (A) Variable importance estimated by RF model in determining self-cleansing velocity: (a) without deposited bed; (b) with a deposited bed. Reproduced from (Montes et al. 2020) with permission. Copyright (2020) Elsevier Ltd. (B) Overview of the defect classification and location recognition framework for sewer line assessment system. Reproduced from (Hassan et al. 2019) with permission. Copyright (2019) Elsevier Ltd. (C) Schematic flow diagram of the BPANN-AO + ALK/ULS system to remove ammonia nitrogen. Reproduced from (Yang et al. 2021) with permission. Copyright (2020) Elsevier Ltd. (D) Observed and predicted sludge volume index in Batna wastewater treatment plant. Reproduced from (Djeddou and Achour 2015) with permission. Copyright (2015) Larhyss Journal. (E) Comparison of experimental and ANN model predicted values for trichloroethylene concentration. Reproduced from (Baskaran et al. 2020) with permission. Copyright (2019) Elsevier Ltd. (F) Comparison of the measured and ANN predicted results for mercury removal efficiency. Reproduced from (Yaqub and Lee 2020) with permission. Copyright (2019) Elsevier Ltd. (G) Forecast N_2O concentrations from (a) the DNN-based model and (b) the LSTM-based model over the fixed prediction horizon (1 day). Reproduced from (Hwangbo et al. 2021) with permission. Copyright (2021) Elsevier Ltd.

and the RF algorithm was applied to analyze the factors that affect sewer pipe blockages. Explanatory factors, such as material type, number of service connections, self-cleaning velocity, sagging pipes, root intrusion risk, closed-circuit television (CCTV) inspection grade, and distance to restaurants showed significantly varying impacts on the blockage propensity. In addition, the RF model predicted the blockage recurrence with a 60–80% accuracy in one of the studied cities (Okwori et al., 2021). Moreover, ML has also been applied in sewer defect detection. Closed circuit television (CCTV) is also a popular tool for visual inspection of the internal conditions of pipelines. However, CCTV inspection relies heavily on human labor to perform video analysis, which is time-consuming, labor-intensive, and error-prone. In this case, the SVM model based on anomalous frame recognition and classification was proposed to improve the efficiency and accuracy of CCTV inspections (Ye et al., 2019). Moreover, better performance in defect recognition and classification was achieved by applying a CNN model, which obtained an accuracy of 96.33% (Fig. 8B) (Hassan et al., 2019).

When transported into the WWTPs, sewage and wastewater will be treated by a series of treatment processes. However, increasingly stringent sewage discharge standards have brought demands for more advanced treatment processes and more efficient management in WWTPs. Like DWTPs, timely and readily obtaining water quality information helps improve the operation and management of a WWTP. To achieve this goal, ML has been used to simulate the water quality and operation status in various treatment processes in different WWTPs.

In studies applying ML to predict the performance of WWTPs, water quality indicators such as COD, BOD, phosphorus, and ammonia have typically been adopted to evaluate the effluent quality (Khatri et al., 2019). For example, an MLP model accurately predicted the effluent TN, TP, and COD concentrations in small-scaled WWTPs in Korea's rural areas, thus realizing real-time remote supervising of WWTPs (Lee et al., 2008). Similarly, the ANN model was also applied for predicting the effluent quality of processes including a membrane bioreactor (MBR) (Giwa et al., 2016), a sequential batch reactor (SBR) (Khatri et al., 2019), and anoxic/oxic (AO) system (Yang et al., 2021), and aerobic granular sludge reactors (AGS) (Mahmod and Wahab, 2017). Additionally, some other algorithms, such as SVR (Seshan et al., 2014), ANFIS (Perendeci et al., 2009), and extreme learning machine (ELM) (Lotfi et al., 2019) have also obtained encouraging results for predicting effluent water quality. In a combined lysis-cryptic and biological nitrogen removal system, an MLP model realized an accurate simulation of the process with an R^2 value of 0.9513. By real-time adjusting the sludge lysate return ratio, the optimized system removed greater than 97–99% of the ammonia nitrogen with zero excess sludge production (Fig. 8C) (Yang et al., 2021). Another study used molecular data to train the SVR model to predict the effluent COD, ammonia, nitrate, and the removal of 3-chloroaniline. Molecular data was generated from a terminal restriction fragment length polymorphism analysis targeting the 16S rRNA and amoA genes from the sludge community. The results showed that the proposed SVR model simulated the effluent water quality with R^2 values ranging from 0.89 to 0.97 for the tested four parameters (Seshan et al., 2014). Similarly, metagenomic information was also applied to investigate the strength of associations between ARGs and different bacterial taxa using the RF model that revealed that genera including *Bacteroides, Clostridium*, and *Streptococcus* were primarily the hosts of the selected ARGs (Sun et al., 2021). Moreover, comparisons among different algorithms have been conducted. SVM was found to perform better for predicting effluent total ammonia nitrogen (TAN) than ANN (Alejo et al., 2018), while it was better for simulating the concentration of Kjeldahl nitrogen (KN) than ANFIS (Manu and Thalla, 2017). The superior ability of SVR compared to these two algorithms was also reported when predicting the performance of AGS reactors (Zaghloul et al., 2020a). In contrast to these comparisons, a multi-stage model that combined several ML algorithms was developed for predicting effluent COD, NH_4-N, and PO_4^{3-}. During the first stage of this model, the selected variables were input into the ANN, SVR, and ANFIS models. Then the outputs were combined as inputs for the subsequent ensemble algorithms, of which the best outputs were determined to be the final prediction. This approach improved the performance of ML models with a small dataset by combining the advantages of different algorithms instead of discussing and selecting the best single one (Zaghloul et al., 2020b).

In addition to the common indicators mentioned above, some other parameters related to WWTPs have also been predicted. For example, the accumulation of fat, oil, and grease in the sumps of wastewater pumping stations was monitored using CNN-based image recognition technology (Moreno-Rodenas et al., 2021). The sludge volume index (SVI) was predicted to monitor the running conditions of the activated sludge process (Fig. 8D) (Djeddou and Achour, 2015). In addition, the daily flow rates for a WWTP were simulated to better design its treatment units (Najafzadeh and Zeinolabedini, 2019). Moreover, some toxic substances present in effluent have also been predicted. For instance, in a two-phase continuous stirred tank bioreactor (CSTB) used for treating a trichloroethylene (TCE) polluted stream, an MLP model simulated the treatment process with excellent accuracy (R^2=0.9923) (Fig. 8E) (Baskaran et al., 2020). The fecal coliform and total coliform removal efficiencies were also predicted with an MLP model in a WWTP using an intermittent cycle extended aeration-sequential batch reactor (ICEAS-SBR) (Khatri et al., 2020). Additionally, ML approaches have also been used to predict the removal of heavy metals including zinc (Rahmanian et al., 2011), chromium (Yaqub et al., 2019), and mercury (Yaqub and Lee, 2020). All of these three studies provided helpful guidance for future research on the removal of other heavy metals (Fig. 8F). Moreover, a DNN and LSTM model were compared for predicting the emission of nitrous oxide (N_2O) from a WWTP in Denmark. A total of 750,000 measurements, including the influent flow rate, airflow rate, temperature, ammonium, nitrate, and DO, were collected to train the models. The higher R^2 values (0.94 versus 0.76) indicated that the LSTM performed better than the DNN (Fig. 8G). A sensitivity analysis revealed that the temperature, NO_3-N, and NH_4-N were the most important factors contributing to the N_2O concentration and N_2O emission rate. WWTPs are a great contributor to global greenhouse gas emissions because of the strong warming effects of N_2O. Therefore, this study was beneficial for understanding the production and emission mechanisms, as well as developing control strategies for N_2O to ensure the sustainable operation and development of WWTPs (Hwangbo et al., 2021). Recently, ML has also been applied to model anaerobic digestion (AD) processes. AD is a popular technique that can convert organic waste and wastewater into biomethane to harvest energy. In this study, six ML algorithms were combined with microbial gene sequencing techniques to predict the methane yield. Genomic data and their corresponding operational parameters from eight research groups were used to train the models. The RF model achieved an accuracy of 0.82 using the combination of operational parameters and genomic data. Moreover, the importance of microbial community members for methane production was first quantified, and the results showed that *Chloroflexi, Actinobacteria, Proteobacteria, Fibrobacteres*, and *Spirochaeta* were the top five most influential phyla. This study provided valuable information regarding the AD process for monitoring and controlling methane production (Long et al., 2021). According to Table S15, MLP was undoubtedly a powerful tool for predicting water quality in sewage or wastewater treatment and has been widely used the past two decades, indicating that it was not out of date due to the popularity of other emerging powerful algorithms. However, it cannot be ignored that emerging technologies bring more possibilities and efficiency to water quality prediction with ML. For example, DL makes it possible to deal with tasks with large feature spaces, while image recognition develops a non-data prediction framework. Moreover, the combination of ML and bioinformatics analysis provides richer data and an in-depth perspective for predicting biological treatment processes, which is lacking in the current studies and will require special attention in future research.

5. Discussion, conclusions, and prospects

5.1. Discussion and conclusions

The above review lists representative applications of ML in fields of water science, and exhibits the availability and efficiency of ML for solving problems concerning water utilization and pollution control. To better use ML in water-related research, previous research and algorithm applications need to be analyzed to provide other researchers with common characteristics and rules that can be used for reference.

According to the classification and analysis of the studies reviewed above, the combination of ML with water science was found to be realized primarily in the following manners (Table 2): i) Predicting the status of desired contaminants was based on analyzing their interactions with other water-related parameters. In these studies, the ML approaches could be further divided into regression and classification analyses. In studies that adopted regression analyses, the specific concentration of pollutants was predicted. For example, they were utilized for predicting the WQIs in natural and engineered systems, and simulating pollutant concentrations in treatment reaction processes. In studies that applied classification analyses, the category of the target (e.g., the pollutant concentration range or water quality classification) rather than a specific value was determined. For example, classification analyses were utilized for the mapping of contaminant distributions in groundwater and classifying water resources into different categories according to their water quality. ii) Mining big data from previous studies were used to summarize universal rules and mechanisms, thus guiding similar reactions or processes. These included evaluating the toxicity of emerging pollutants, and applying adsorption or oxidation reactions to remove pollutants or test new reactive materials. In this manner, the valuable information hidden in an increasing number of scientific studies was systematically delved to provide more theoretical guidance for similar studies; iii) Image recognition technology was applied to analyze the relationships between the image information and physicochemical properties of the research object, thus characterizing water quality, identifying specific contaminants, and detecting equipment faults in engineered water systems. This approach improved the efficiency of analysis methods with spectrum or image as output thereby reducing the difficulty in detecting sewer faults. In particular, this method provides new insights for analyzing water quality in this era when high-definition cameras and remote sensing are increasingly popular.

Though ML has been recognized as a powerful tool for solving water-related problems, different algorithms exhibited different performance in these fields. Therefore, discussing and analyzing the applicability of various commonly used algorithms in different application scenarios is informative for other researchers to choose the appropriate algorithm or further optimize an algorithm. To determine the commonly used algorithms, the usage frequency of algorithms used in 215 reviewed studies was counted, and the usage frequency of these algorithms in water-related researches was also retrieved. Results of both statistical schemes showed that MLP, ANFIS, RBF, SVM, LR, CART, and RF were commonly used ML algorithms, while LSTM, DNN, and CNN were popular DL algorithms (Text S1). Therefore, the advantages and disadvantages of these ML algorithms in different research directions were compared and analyzed, and the results are briefly presented in Table 2. In general, MLP was the most widely used algorithm in dealing with both regression and classification tasks, especially in earlier studies with small volume of data and features. ML achieved satisfactory performance for predicting pollutant concentrations in natural waters, chemical reaction processes, and water or wastewater treatment. However, MLP suffers the problem of a long training time, slow convergence, and a local optimum caused by gradient decline, although it has a strong ability in fitting nonlinear problems and robustness to data noise. In contrast, RBF NN makes up for these disadvantages of MLP and even has higher approximation accuracy and generalization ability. However, as the volume of training samples increase, the number of hidden layer neurons of RBF NN will be higher than that of MLP, which will increase

Table 2
Recommendations on the selection of ML algorithm in different research directions of water science.

Means	Applications	Algorithm recommended	Algorithm characteristics	Applicable conditions	
Regression	Predicting water quality	MLP	Simple structure; slow convergence, local optimum, black box	Fewer features	Data with big volume
	Evaluating pollutant toxicity	RBF NN	Strong generalization, fast convergence; complex structure, black box		Data with small volume
	Modeling treatment technique	CART	Interpretable, fast training; easy to over-fitting,		Data with a balanced sample size for each category
	Assisting characterization analysis	RF	Interpretable, fast training, anti- overfitting, no need for feature selection; calculation burden	More features	Data with less noisy
	Purifying and distributing drinking water	LSTM	Long-time memory; complex structure, black box, calculation burden		Data of time series
	Collecting and treating sewage water	DNN	Strong ability in fitting, feature extraction, and fault tolerance; complex structure, black box, calculation burden		Data with a huge volume
Classification	Classifying water resources	CART	As mentioned above	Fewer features	As mentioned above
	Mapping groundwater contaminants	RF	As mentioned above	More features	As mentioned above
Data mining	Evaluating pollutant toxicity	DNN	As mentioned above	Usually, the volume of data is huge and molecular descriptors are needed	
	Modeling treatment technique				
Image recognition	Predicting water quality	CNN	Automatic feature extraction	The sample is presented in the form of images	
	Assisting characterization analysis				
	Purifying and distributing drinking water				
	Collecting and treating sewage water				

Note: a) Theoretically, data mining and image recognition also belong to regression or classification, but due to their distinctive characteristics, they are discussed separately; b) That one algorithm is suitable under certain condition does not mean that other algorithms are not suitable for this condition; c) All the algorithms listed in this table can be used for both regression and classification.

the complexity and computation burden of the RBF network. Recently, DL has also been applied in dealing with problems in water science. Due to the introduction of more hidden layers, DNN has stronger abilities in nonlinear learning, feature extraction, and fault tolerance. Consequently, DNN was found to be used to handle tasks with a large volume of data and features, especially studies that applied data mining from the literature. Additionally, RNN and LSTM can deal with time series data. Moreover, through the switch of the cell gate, LSTM realizes the function of long-time memory, thus solving the problem of gradient disappearance and gradient explosion during the training process. Therefore, LSTM can solve more problems in water science, as the data in many fields of water-related research are time series data. Although the DL models seem to possess a powerful ability to solve nearly all the tasks concerning water science, they still must face the black box nature, which is an unavoidable problem for all NN-based algorithms. In contrast, the DT-based models are more interpretable, as they can show the importance of different features to generate understandable rules. In addition, DT-based models can quickly solve multi-classification tasks with large volume of features. This is the advantage over LR and SVM, the former of which cannot deal with tasks with nonlinear and large feature spaces, and it is time-consuming during training and sometimes difficult to find a suitable kernel function for the latter. Therefore, CART, especially the RF models, have been widely used for handling both regression and classification tasks, such as mapping groundwater contaminants and classifying water sources. This is because RF overcomes the problems of over-fitting and data imbalance, and it is robust to the loss of features.

Based on the above comparison and analysis of the several commonly used algorithms, recommendations on the selection of algorithms were made according to their applicability to different scenarios (Table 2). For regression tasks with a small volume of data, both MLP and RBF NN are competent. Without regard to the complexity of the model, RBF NN may be better due to its higher ability of generalization, fitting, and convergence. For classification tasks, the CART model is more suitable for data with a balanced sample size for each category because the information gain is partial to the feature with more data, thus leading to over-fitting. Otherwise, the RF model is more qualified, especially for tasks with a large volume of features. Moreover, for those tasks with a huge amount of data, DNN and LSTM are more competent.

5.2. Prospects

The development of ML in water science is still at an initial stage; however, ML has shown its feasibility and efficiency in solving complex environmental problems in many studies. To make ML technology more accessible to environmental researchers, thus solving research problems in the water science field more efficiently, joint efforts in terms of algorithm development, data curation, and interdisciplinary cooperation are required.

i) As analyzed above, there was not a perfect algorithm for all tasks necessary for the water science field. Algorithms with simple structures typically have defects in performance (e.g., MLP and CART), while those with excellent performance often possess complex structures, thus increasing the difficulty of programming and the hardware cost of operation (e.g., DNN). Therefore, according to the characteristics of the data in water-related research, such as a moderate amount of time series data, algorithms with simple structures, high performance, and strong interpretability are encouraged to be developed. Moreover, the graphical user interface (e.g., the graphical user interface designed for modeling adsorption processes) or user-friendly data analytics tools (e.g., SourceTracker) designed specifically for water-related studies can also reduce the cost and difficulties of researchers encounter when using ML techniques.

ii) Data mining is helpful to collect data from similar studies to form big data, thus revealing underlying rules or providing a data basis for other big data researchers. However, in traditional research areas, including water science, data from other studies are often difficult to obtain. Open data and the sharing of data are common ways to provide rich data sources for datasets in the application of ML. However, open source data in water science field seems to be insufficient compared to other fields where ML techniques have been applied earlier and utilized more in depth, e.g., drug research (Vamathevan et al., 2019), biological research (Camacho et al., 2018), and solid Earth geosciences (Bergen et al., 2019), for which many open source data platforms have been developed. Therefore, the concept of open source data and the sharing of data is expected to be accepted and practiced more widely in the water research community, and researchers are encouraged to share their research data without any conflict of interest or legal and regulatory restrictions.

iii) The programming and implementation of ML models depend on the researchers' computer skills and mastery of algorithms, which are difficult for most water researchers to grasp in a short amount of time. To lower the threshold for researchers to use the ML technologies, interdisciplinary communication and cooperation with data researchers are beneficial. Under this framework of cross-disciplines, data researchers can provide professional suggestions on data processing and modeling, while water researchers can interpret the output of the model with expert knowledge. Moreover, with the help of data researchers, some cutting-edge algorithms can also be used to solve problems in water science field.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgment

The authors acknowledge funding from the National Natural Science Foundation of China (52070029, 51961125104).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2021.117666.

References

Abdollahi, Y., Zakaria, A., Sairi, N.A., Matori, K.A., Masoumi, H.R., Sadrolhosseini, A.R., Jahangirian, H., 2014. Artificial neural network modelling of photodegradation in suspension of manganese doped zinc oxide nanoparticles under visible-light irradiation. ScientificWorldJournal 2014, 726101.

Ai, H., Wu, X., Zhang, L., Qi, M., Zhao, Y., Zhao, Q., Zhao, J., Liu, H., 2019. QSAR modelling study of the bioconcentration factor and toxicity of organic compounds to aquatic organisms using machine learning and ensemble methods. Ecotoxicol. Environ. Saf. 179, 71–78.

Al Aani, S., Bonny, T., Hasan, S.W., Hilal, N, 2019. Can machine language and artificial intelligence revolutionize process automation for water treatment and desalination? Desalination 458, 84–96.

Alejo, L., Atkinson, J., Guzmán-Fierro, V., Roeckel, M., 2018. Effluent composition prediction of a two-stage anaerobic digestion process: machine learning and stoichiometry techniques. Environ. Sci. Pollut. Res. 25 (21), 21149–21163.

Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. 46 (3), 175–185.

Álvarez, M.J., Ilzarbe, L., Viles, E., Tanco, M., 2016. The use of genetic algorithms in response surface methodology. Qual. Technol. Quant. Manage. 6 (3), 295–307.

Amini, M., Mueller, K., Abbaspour, K.C., Rosenberg, T., Afyuni, M., Møller, K.N., Sarr, M., Johnson, C.A., 2008. Statistical modeling of global geogenic fluoride contamination in groundwaters. Environ. Sci. Technol. 42 (10), 3662–3668.

Anning, D.W., Paul, A.P., McKinney, T.S., Huntington, J.M., Bexfield, L.M., Thiros, S.A., 2012. Predicted Nitrate and Arsenic Concentrations in Basin-Fill Aquifers of the Southwestern United States, p. 5065. Scientific Investigations Report.

Antanasijevic, D., Pocajt, V., Povrenovic, D., Peric-Grujic, A., Ristic, M., 2013a. Modelling of dissolved oxygen content using artificial neural networks: Danube River, North Serbia, case study. Environ. Sci. Pollut. Res. 20 (12), 9006–9013.
Antanasijevic, D., Pocajt, V., Povrenovic, D., Peric-Grujic, A., Ristic, M., 2013b. Modelling of dissolved oxygen content using artificial neural networks: Danube River, North Serbia, case study. Environ. Sci. Pollut. Res. Int. 20 (12), 9006–9013.
Arad, J., Housh, M., Perelman, L., Ostfeld, A., 2013. A dynamic thresholds scheme for contaminant event detection in water distribution systems. Water Res. 47 (5), 1899–1908.
Ariyaluran Habeeb, R.A., Nasaruddin, F., Gani, A., Targio Hashem, I.A., Ahmed, E., Imran, M., 2019. Real-time big data processing for anomaly detection: a Survey. Int. J. Inf. Manage. 45, 289–307.
Ay, M., Kisi, O., 2012. Modeling of dissolved oxygen concentration using different neural network techniques in foundation Creek, El Paso County, Colorado. J. Environ. Eng. 138, 654–662.
Ayotte, J.D., Medalie, L., Qi, S.L., Backer, L.C., Nolan, B.T., 2017. Estimating the high-arsenic domestic-well population in the conterminous United States. Environ. Sci. Technol. 51 (21), 12443–12454.
Ayotte, J.D., Nolan, B.T., Gronberg, J.A., 2016. Predicting arsenic in drinking water wells of the central valley, California. Environ. Sci. Technol. 50 (14), 7555–7563.
Ayotte, J.D., Nolan, B.T., Nuckols, J.R., Cantor, K.P., Robinson, G.R., Baris, D., Hayes, L., Karagas, M., Bress, W., Silverman, D.T., Lubin, J.H., 2006. Modeling the probability of arsenic in groundwater in New England as a tool for exposure assessment. Environ. Sci. Technol. 40, 3578–3585.
Azamathulla, H.M., Ab Ghani, A., Fei, S.Y., 2012. ANFIS-based approach for predicting sediment transport in clean sewer. Appl. Soft Comput. 12 (3), 1227–1230.
Azqhandi, M.H.A., Foroughi, M., Yazdankish, E., 2019. A highly effective, recyclable, and novel host-guest nanocomposite for Triclosan removal: a comprehensive modeling and optimization-based adsorption study. J. Colloid Interface Sci. 551, 195–207.
Baek, S.S., Choi, Y., Jeon, J., Pyo, J., Park, J., Cho, K.H., 2021. Replacing the internal standard to estimate micropollutants using deep and machine learning. Water Res. 188, 116535.
Balleste, E., Belanche-Munoz, L.A., Farnleitner, A.H., Linke, R., Sommer, R., Santos, R., Monteiro, S., Maunula, L., Oristo, S., Tiehm, A.A., Stange, C., Blanch, A.R., 2020. Improving the identification of the source of faecal pollution in water using a modelling approach: from multi-source to aged and diluted samples. Water Res. 171, 115392.
Baral, D., Dvorak, B.I., Admiraal, D., Jia, S., Zhang, C., Li, X., 2018. Tracking the sources of antibiotic resistance genes in an urban stream during wet weather using shotgun metagenomic analyses. Environ. Sci. Technol. 52 (16), 9033–9044.
Barzegar, R., Aalami, M.T., Adamowski, J., 2020. Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. Stoch. Environ. Res. Risk Assess. 34 (2), 415–433.
Baskaran, D., Sinharoy, A., Paul, T., Pakshirajan, K., Rajamanickam, R., 2020. Performance evaluation and neural network modeling of trichloroethylene removal using a continuously operated two-phase partitioning bioreactor. Environ. Technol. Innov. 17, 100568.
Batista, G.E., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newslett. 6 (1), 20–29.
Bergen, K.J., Johnson, P.A., de Hoop, M.V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science 363 (6433), 6433.
Bindal, S., Singh, C.K., 2019. Predicting groundwater arsenic contamination: regions at risk in highest populated state of India. Water Res. 159, 65–76.
Bolton, J.R., Bircher, K.G., Tumas, W., Tolman, C.A., 1996. Figures-of-merit for the technical development and application of advanced oxidation processes. J. Adv. Oxid. Technol. 1, 13–17.
Bonakdari, H., Larrarte, F., 2006. Experimental and Numerical Investigation on Self Cleansing and Shear in Sewers. Vienna, Austria, pp. 19–26.
Borhani, T.N., Saniedanesh, M., Bagheri, M., Lim, J.S., 2016. QSPR prediction of the hydroxyl radical rate constant of water contaminants. Water Res. 98, 344–353.
Bourel, M., Segura, A.M., Crisci, C., Lopez, G., Sampognaro, L., Vidal, V., Kruk, C., Piccini, C., Perera, G., 2021. Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters. Water Res. 202, 117450.
Bowden, G.J., Nixon, J.B., Dandy, G.C., Maier, H.R., Holmes, M., 2006. Forecasting chlorine residuals in a water distribution system using a general regression neural network. Math. Comput. Model. 44 (5–6), 469–484.
Breiman, L., 2001. Random forest. Mach. Learn. 45, 5–32.
Breiman, L., Friedman, FJ.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA, USA.
Bretzler, A., Lalanne, F., Nikiema, J., Podgorski, J., Pfenninger, N., Berg, M., Schirmer, M., 2017. Groundwater arsenic contamination in Burkina Faso, West Africa: predicting and verifying regions at risk. Sci. Total Environ. 584-585, 958–970.
Brown, C.M., Staley, C., Wang, P., Dalzell, B., Chun, C.L., Sadowsky, M.J., 2017. A high-throughput DNA-sequencing approach for determining sources of fecal bacteria in a lake superior estuary. Environ. Sci. Technol. 51 (15), 8263–8271.
Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., Collins, J.J., 2018. Next-generation machine learning for biological networks. Cell 173 (7), 1581–1592.
Campsvalls, G., Gomezchova, L., Munozmari, J., Vilafrances, J., Amoroslopez, J., Calpemaravilla, J., 2006. Retrieval of oceanic chlorophyll concentration with relevance vector machines. Remote Sens. Environ. 105 (1), 23–33.
Castrillo, M., Garcia, A.L., 2020. Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. Water Res. 172, 115490.
Cha, D., Park, S., Kim, M.S., Kim, T., Hong, S.W., Cho, K.H., Lee, C., 2020. Prediction of oxidant exposures and micropollutant abatement during ozonation using a machine learning method. Environ. Sci. Technol. 55 (1), 709–718.
Cha, Y., Kim, Y.M., Choi, J.W., Sthiannopkao, S., Cho, K.H., 2016. Bayesian modeling approach for characterizing groundwater arsenic contamination in the Mekong River basin. Chemosphere 143, 50–56.
Chandramouli, V., Brion, G., Neelakantan, T.R., Lingireddy, S., 2007. Backfilling missing microbial concentrations in a riverine database using artificial neural networks. Water Res. 41 (1), 217–227.
Chau, K.W., 2006. A review on integration of artificial intelligence into water quality modelling. Mar. Pollut. Bull. 52 (7), 726–733.
Che Osmi, S.F., Malek, M.A., Yusoff, M., Azman, N.H., Faizal, W.M., 2016. Development of river water quality management using fuzzy techniques: a review. Int. J. River Basin Manage. 14 (2), 243–254.
Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y., Ren, H., 2020a. 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, 115454.
Chen, Y., Song, L., Liu, Y., Yang, L., Li, D., 2020b. A Review of the Artificial Neural Network Models for Water Quality Prediction. Applied Sciences 10 (17), 5776.
Cho, K.H., Sthiannopkao, S., Pachepsky, Y.A., Kim, K.-W., Kim, J.H., 2011. Prediction of contamination potential of groundwater arsenic in Cambodia, Laos, and Thailand using artificial neural network. Water Res. 45 (17), 5535–5544.
Chong, M.N., Jin, B., Chow, C.W., Saint, C., 2010. Recent developments in photocatalytic water treatment technology: a review. Water Res. 44 (10), 2997–3027.
Chowdhury, M., Alouani, A., Hossain, F., 2010. Comparison of ordinary kriging and artificial neural network for spatial mapping of arsenic contamination of groundwater. Stoch. Environ. Res. Risk Assess. 24 (1), 1–7.
Clark, J.W., 1991. Neural network modelling. Phys. Med. Biol. 36 (10), 1259–1317.
Conley, D.J., Paerl, H.W., Howarth, R.W., Howarth, R.W., Boesch, D.F., Seitzinger, S.P., Havens, K.E., Lancelot, C., Likens, G.E., 2009. Controlling eutrophication: nitrogen and phosphorus. Science 323, 1014–1015.
Dabrowski, A., 2001. Adsorption—From theory to practice. Adv. Colloid Interface Sci. 93 (1–3), 135–224.
Danso-Amoakoa, E., Prasad, T.D., 2014. ANN model to predict the influence of chemical and biological parameters on iron and manganese accumulation. Proc. Eng. 70, 409–418.
Dezfooli, D., Hosseini-Moghari, S.-.M., Ebrahimi, K., Araghinejad, S., 2017. Classification of water quality status based on minimum quality parameters: application of machine learning techniques. Model. Earth Syst. Environ. 4 (1), 311–324.
Djeddou, M., Achour, B., 2015. The use of a neural network technique for the prediction of sludge volume index in municipal wastewater treatment plant. Larhyss Journal 24, 351–370.
Dogo, E.M., Nwulu, N.I., Twala, B., Aigbavboa, C., 2019. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water J. 16 (3), 235–248.
Dummer, T.J., Yu, Z.M., Nauta, L., Murimboh, J.D., Parker, L., 2015. Geostatistical modelling of arsenic in drinking water wells and related toenail arsenic concentrations across Nova Scotia, Canada. Sci. Total Environ. 505, 1248–1258.
Durham, B.P., Sharma, S., Luo, H., Smith, C.B., Amin, S.A., Bender, S.J., Dearth, S.P., Van Mooy, B.A., Campagna, S.R., Kujawinski, E.B., 2015. Cryptic carbon and sulfur cycling between surface ocean plankton. Proc. Natl. Acad. Sci. 112, 453–457.
Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77 (4), 802–813.
Elmolla, E.S., Chaudhuri, M., Eltoukhy, M.M., 2010. The use of artificial neural network (ANN) for modeling of COD removal from antibiotic aqueous solution by the Fenton process. J. Hazard. Mater. 179 (1–3), 127–134.
Erickson, M.L., Elliott, S.M., Brown, C.J., Stackelberg, P.E., Ransom, K.M., Reddy, J.E., Cravotta, C.A., 2021. Machine-learning predictions of high arsenic and high manganese at drinking water depths of the glacial aquifer system, Northern Continental United States. Environ. Sci. Technol. 55 (9), 5791–5805.
Fan, M., Hu, J., Cao, R., Ruan, W., Wei, X., 2018. A review on experimental design for pollutants removal in water treatment with the aid of artificial intelligence. Chemosphere 200, 330–343.
Fan, M., Hu, J., Cao, R., Xiong, K., Wei, X., 2017. Modeling and prediction of copper removal from aqueous solutions by nZVI/rGO magnetic nanocomposites using ANN-GA and ANN-PSO. Sci. Rep. 7 (1), 18040.
Frederick, L., VanDerslice, J., Taddie, M., Malecki, K., Gregg, J., Faust, N., Johnson, W.P., 2016. Contrasting regional and national mechanisms for predicting elevated arsenic in private wells across the United States using classification and regression trees. Water Res. 91, 295–304.
García-Alba, J., Bárcena, J.F., Ugarteburu, C., García, A., 2019. Artificial neural networks as emulators of process-based models to analyse bathing water quality in estuaries. Water Res. 150, 283–295.
Geyikçi, F., Kılıç, E., Çoruh, S., Elevli, S., 2012. Modelling of lead adsorption from industrial sludge leachate on red mud by using RSM and ANN. Chem. Eng. J. 183, 53–59.
Ghaedi, A.M., Vafaei, A., 2017. Applications of artificial neural networks for adsorption removal of dyes from aqueous solution: a review. Adv. Colloid Interface Sci. 245, 20–39.
Giwa, A., Daer, S., Ahmed, I., Marpu, P.R., Hasan, S.W., 2016. Experimental investigation and artificial neural networks ANNs modeling of electrically-enhanced membrane bioreactor for wastewater treatment. J. Water Process Eng. 11, 88–97.
Godo-Pla, L., Emiliano, P., Valero, F., Poch, M., Sin, G., Monclús, H., 2019. Predicting the oxidant demand in full-scale drinking water treatment using an artificial neural

network: uncertainty and sensitivity analysis. Process Saf. Environ. Prot. 125, 317–327.
Greff, K., Srivastava, R.K., Koutnik, J., Steunebrink, B.R., Schmidhuber, J., 2017. LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. 28 (10), 2222–2232.
Griffiths, K.A., Andrews, R.C., 2011. Application of artificial neural networks for filtration optimization. J. Environ. Eng. 137 (11), 1040–1047.
Gu, J., Zhang, W., Li, Y., Niu, L., Wang, L., Zhang, H., 2020. Source identification of phosphorus in the river-lake interconnected system using microbial community fingerprints. Environ. Res. 186, 109498.
Haile, R.W., Witte, J.S., Gold, M., Cressey, R., McGee, C., Millikan, R.C., Glasser, A., Harawa, N., Ervin, C., Harmon, P., Harper, J., Dermand, J., Alamillo, J., Barrett, K., Nides, M., Wang, G.Y., 1999. The health effects of swimming in ocean water contaminated by storm drain runoff. Epidemiology 10 (4), 355–363.
Halder, A.K., Melo, A., Cordeiro, M., 2020. A unified in silico model based on perturbation theory for assessing the genotoxicity of metal oxide nanoparticles. Chemosphere 244, 125489.
Hamidin, N., Yu, Q.J., Connell, D.W., 2008. Human health risk assessment of chlorinated disinfection by-products in drinking water using a probabilistic approach. Water Res. 42, 3263–3274.
Hassan, S.I., Dang, L.M., Mehmood, I., Im, S., Choi, C., Kang, J., Park, Y.-S., Moon, H., 2019. Underground sewer pipe condition assessment based on convolutional neural networks. Autom. Constr. 106, 102849.
Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning. Springer.
He, J., Chen, Y., Wu, J., Stow, D.A., Christakos, G., 2020. Space-time chlorophyll-a retrieval in optically complex waters that accounts for remote sensing and modeling uncertainties and improves remote estimation accuracy. Water Res. 171, 115403.
He, L., Xiao, K., Zhou, C., Li, G., Yang, H., Li, Z., Cheng, J., 2019. Insights into pesticide toxicity against aquatic organism: QSTR models on Daphnia Magna. Ecotoxicol. Environ. Saf. 173, 285–292.
He, L.M., He, Z.L., 2008. Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern California, USA. Water Res. 42 (10–11), 2563–2573.
Hebb, D.O., 1949. The Organization of Behaviour. Wiley & Sons, New York.
Heddam, S., Bermad, A., Dechemi, N., 2012. ANFIS-based modelling for coagulant dosage in drinking water treatment plant: a case study. Environ. Monit. Assess. 184 (4), 1953–1971.
Heo, S., Safder, U., Yoo, C., 2019. Deep learning driven QSAR model for environmental toxicology: effects of endocrine disrupting chemicals on human health. Environ. Pollut. 253, 29–38.
Herzsprung, P., Wentzky, V., Kamjunke, N., von Tumpling, W., Wilske, C., Friese, K., Boehrer, B., Reemtsma, T., Rinke, K., Lechtenfeld, O.J., 2020. Improved understanding of dissolved organic matter processing in freshwater using complementary experimental and machine learning approaches. Environ. Sci. Technol. 54 (21), 13556–13565.
Hogg, R., 2000. Flocculation and dewatering. Int. J. Miner. Process. 58, 223–236.
Holmberg, M., Forsius, M., Starr, M., Huttunen, M., 2006. An application of artificial neural networks to carbon, nitrogen and phosphorus concentrations in three boreal streams and impacts of climate change. Ecol. Model. 195 (1–2), 51–60.
Hossain, M.M., Piantanakulchai, M., 2013. Groundwater arsenic contamination risk prediction using GIS and classification tree method. Eng. Geol. 156, 37–45.
Hou, P., Jolliet, O., Zhu, J., Xu, M., 2020. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. Environ. Int. 135, 105393.
Hu, J., Chu, W., Sui, M., Xu, B., Gao, N., Ding, S., 2018. Comparison of drinking water treatment processes combinations for the minimization of subsequent disinfection by-products formation during chlorination and chloramination. Chem. Eng. J. 335, 352–361.
Hwangbo, S., Al, R., Chen, X., Sin, G., 2021. Integrated model for understanding N_2O emissions from wastewater treatment plants: a deep learning approach. Environ. Sci. Technol. 55 (3), 2143–2151.
Ighalo, J.O., Adeniyi, A.G., Marques, G., 2020. Artificial intelligence for surface water quality monitoring and assessment: a systematic literature analysis. Model. Earth Syst. Environ. 7 (2), 669–681.
J, Q., W, Y., 2011. Recurrent high order neural network modeling for wastewater treatment process. J. Comput. 6, 1570–1577.
Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognit. Lett 31 (8), 651–666.
Jang, J.-S.R., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man Cybern. 23 (3), 665–685.
Jiang, Z., Hu, J., Zhang, X., Zhao, Y., Fan, X., Zhong, S., Zhang, H., Yu, X., 2020. A generalized predictive model for TiO_2-Catalyzed photo-degradation rate constants of water contaminants through artificial neural network. Environ. Res. 187, 109697.
Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philos. Trans. A Math. Phys. Eng. Sci. 374 (2065), 20150202.
Jordan, M.I., Mitchell, T.M., 2015a. Machine learning: trends, perspectives, and prospects. Science 349 (6245), 255–260.
Jordan, M.I., Mitchell, T.M., 2015b. Machine learning: trends, perspectives, and prospects. Science 4349 (6245), 255–260.
Jun, L.Y., Karri, R.R., Yon, L.S., Mubarak, N.M., Bing, C.H., Mohammad, K., Jagadish, P., Abdullah, E.C., 2020. Modeling and optimization by particle swarm embedded neural network for adsorption of methylene blue by jicama peroxidase immobilized on buckypaper/polyvinyl alcohol membrane. Environ. Res. 183, 109158.
Kadyrova, N.O., Pavlova, L.V., 2014. Statistical analysis of big data: an approach based on support vector machines for classification and regression problems. Biophysics 59 (3), 364–373.
Karul, C., Soyupak, S., Çilesiz, A.F., Akbay, N., Germen, E., 2000. Case studies on the use of neural networks in eutrophication modeling. Ecol. Modell. 134, 145–152.
Kavitha, V., Palanivelu, K., 2004. The role of ferrous ion in Fenton and photo-Fenton processes for the degradation of phenol. Chemosphere 55 (9), 1235–1243.
Keiner, L.E., 2010. Estimating oceanic chlorophyll concentrations with neural networks. Int. J. Remote Sens. 20 (1), 189–194.
Kennedy, J., Eberhart, R., 1995. Particle Swarm Optimization (PSO). Perth, Australia, pp. 1942–1948.
Khataee, A.R., Kasiri, M.B., 2010. Artificial neural networks modeling of contaminated water treatment processes by homogeneous and heterogeneous nanocatalysis. J. Mol.Catal. A 331 (1–2), 86–100.
Khataee, A.R., Kasiri, M.B., 2011. Modeling of biological water and wastewater treatment processes using artificial neural networks. CLEAN 39 (8), 742–749.
Khatri, N., Khatri, K.K., Sharma, A., 2019. Prediction of effluent quality in ICEAS-sequential batch reactor using feedforward artificial neural network. Water Sci. Technol. 80 (2), 213–222.
Khatri, N., Khatri, K.K., Sharma, A., 2020. Artificial neural network modelling of faecal coliform removal in an intermittent cycle extended aeration system-sequential batch reactor based wastewater treatment plant. J. Water Process Eng. 37, 101477.
Kim, D., Miranda, M.L., Tootoo, J., Bradley, P., Gelfand, A.E., 2011. Spatial modeling for groundwater arsenic levels in North Carolina. Environ. Sci. Technol. 45 (11), 4824–4831.
Kim, M., Choi, C.Y., Gerba, C.P., 2008. Source tracking of microbial intrusion in water systems using artificial neural networks. Water Res. 42 (4–5), 1308–1314.
Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R., Kelley, S.T., 2011. Bayesian community-wide culture-independent microbial source tracking. Nat. Methods 8, 761–763.
Krause, L.M.K., Koc, J., Rosenhahn, B., Rosenhahn, A., 2020. Fully convolutional neural network for detection and counting of diatoms on coatings after short-term field exposure. Environ. Sci. Technol. 54 (16), 10022–10030.
Kulkarni, P., Chellam, S., 2010. Disinfection by-product formation following chlorination of drinking water: artificial neural network models and changes in speciation with treatment. Sci. Total Environ. 408 (19), 4202–4210.
Lado, L.R., Polya, D., Winkel, L., Berg, M., Hegan, A., 2008. Modelling arsenic hazard in Cambodia: a geostatistical approach using ancillary data. Appl. Geochem. 23 (11), 3010–3018.
LaPara, T.M., Hope Wilkinson, K., Strait, J.M., Hozalski, R.M., Sadowksy, M.J., Hamilton, M.J., 2015. The bacterial communities of full-scale biologically active, granular activated carbon filters are stable and diverse and potentially contain novel ammonia-oxidizing microorganisms. Appl. Environ. Microbiol. 81 (19), 6864–6872.
Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., 2016. Machine learning in geosciences and remote sensing. Geosci. Front. 7, 3–10.
Lee, G.-C., Chang, S.H., 2003. Radial basis function networks applied to DNBR calculation in digital core protection systems. Ann. Nucl. Energy 30, 1516–1572.
Lee, J.H.W., Huang, Y., Dickman, M., Jayawardena, A.W., 2003. Neural network modelling of coastal algal blooms. Ecol. Modell. 159, 179–201.
Lee, J.-J., Jang, C.-S., Liu, C.-W., Liang, C.-P., Wang, S.-W., 2009. Determining the probability of arsenic in groundwater using a parsimonious model. Environ. Sci. Technol. 43 (17), 6662–6668.
Lee, M.W., Hong, S.H., Choi, H., Kim, J.-.H., Lee, D.S., Park, J.M., 2008. Real-time remote monitoring of small-scaled biological wastewater treatment plants by a multivariate statistical process control and neural network-based software sensors. Process Biochem. 43 (10), 1107–1113.
Li, L., Rong, S., Wang, R., Yu, S., 2021. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: a review. Chem. Eng. J. 405, 126673.
Li, R.A., McDonald, J.A., Sathasivan, A., Khan, S.J., 2020a. A multivariate Bayesian network analysis of water quality factors influencing trihalomethanes formation in drinking water distribution systems. Water Res. 190, 116712.
Li, W., Yang, M., Liang, Z., Zhu, Y., Mao, W., Shi, J., Chen, Y., 2013. Assessment for surface water quality in Lake Taihu Tiaoxi River Basin China based on support vector machine. Stoch. Environ. Res. Risk Assess. 27 (8), 1861–1870.
Li, Y., Wang, X., Zhao, Z., Han, S., Liu, Z., 2020b. Lagoon water quality monitoring based on digital image analysis and machine learning estimators. Water Res. 172, 115471.
Lin, H., Dai, Q., Zheng, L., Hong, H., Deng, W., Wu, F., 2020. Radial basis function artificial neural network able to accurately predict disinfection by-product levels in tap water: taking haloacetic acids as a case study. Chemosphere 248, 125999.
Lister, A.L., Van Der Kraak, G.J., 2001. Endocrine disruption: why is it so complicated? Water Qual. Res. J. 36 (2), 175–188. Canada.
Long, F., Wang, L., Cai, W., Lesnik, K., Liu, H., 2021. Predicting the performance of anaerobic digestion using machine learning algorithms and genomic data. Water Res. 199 (1), 117182.
Lotfi, K., Bonakdari, H., Ebtehaj, I., Mjalli, F.S., Zeynoddin, M., Delatolla, R., Gharabaghi, B., 2019. Predicting wastewater treatment plant quality parameters using a novel hybrid linear-nonlinear methodology. J. Environ. Manage. 240, 463–474.
Lu, S., Zhang, X., Bao, H., Skitmore, M., 2016. Review of social water cycle research in a changing environment. Renew. Sustain. Energy Rev. 63, 132–140.
Ma, J., Qin, B., Wu, P., Zhou, J., Niu, C., Deng, J., Niu, H., 2015. Controlling cyanobacterial blooms by managing nutrient ratio and limitation in a large hyper-eutrophic lake: lake Taihu, China. J. Environ. Sci. 27, 80–86.
Mahmod, N., Wahab, N.A., 2017. Dynamic modelling of aerobic granular sludge artificial neural networks. Int. Jo. Electr. Comput. Eng. 7 (3), 1568–1573.
Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ. Modell. Softw. 15, 101–124.

Maier, H.R., Dandy, G.C., Burch, M.D., 1998. Use of artificial neural networks for modelling cyanobacteria Anabaena spp. in the River Murray, South Australia. Ecol. Modell. 105, 257–272.
Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environ. Modell. Softw. 25 (8), 891–909.
Maier, H.R., Kapelan, Z., Kasprzyk, J., Kollat, J., Matott, L.S., Cunha, M.C., Dandy, G.C., Gibbs, M.S., Keedwell, E., Marchi, A., Ostfeld, A., Savic, D., Solomatine, D.P., Vrugt, J.A., Zecchin, A.C., Minsker, B.S., Barbour, E.J., Kuczera, G., Pasha, F., Castelletti, A., Giuliani, M., Reed, P.M., 2014. Evolutionary algorithms and other metaheuristics in water resources: current status, research challenges and future directions. Environ. Modell. Softw. 62, 271–299.
Makaya, E., Hensel, O., 2016. Modelling flow dynamics in water distribution networks using artificial neural networks - A leakage detection technique. Int. J. Eng. Sci. Technol. 7 (1), 33–43.
Manu, D.S., Thalla, A.K., 2017. Artificial intelligence models for predicting the performance of biological wastewater treatment plant in the removal of Kjeldahl Nitrogen from wastewater. Appl. Water Sci. 7 (7), 3783–3791.
May, R.J., Dandy, G.C., Maier, H.R., Nixon, J.B., 2008. Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems. Environ. Modell. Softw. 23 (10–11), 1289–1299.
Meliker, J.R., AvRuskin, G.A., Slotnick, M.J., Goovaerts, P., Schottenfeld, D., Jacquez, G. M., Nriagu, J.O., 2008. Validity of spatial models of arsenic concentrations in private well water. Environ. Res. 106 (1), 42–50.
Michel, A.P.M., Morrison, A.E., Preston, V.L., Marx, C.T., Colson, B.C., White, H.K., 2020. Rapid identification of marine plastic debris via spectroscopic techniques and machine learning classifiers. Environ. Sci. Technol. 54 (17), 10630–10637.
Miklos, D.B., Remy, C., Jekel, M., Linden, K.G., Drewes, J.E., Hubner, U., 2018. Evaluation of advanced oxidation processes for water and wastewater treatment - a critical review. Water Res. 139, 118–131.
Modaresi, F., Araghinejad, S., 2014. A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. Water Resour. Manage. 28 (12), 4095–4111.
Mohammadpour, R., Shaharuddin, S., Chang, C.K., Zakaria, N.A., Ab Ghani, A., Chan, N. W., 2015. Prediction of water quality index in constructed wetlands using support vector machine. Environ. Sci. Pollut. Res. Int. 22 (8), 6208–6219.
Mohammadpour, R., Shaharuddin, S., Zakaria, N.A., Ghani, A.A., Vakili, M., Chan, N.W., 2016. Prediction of water quality index in free surface constructed wetlands. Environ. Earth Sci. 75 (2), 139.
Mohri, M., Rostamizadeh, A., Talwalkar, A., 2012. Foundations of Machine Learning. MIT Press.
Montes, C., Kapelan, Z., Saldarriaga, J., 2020. Predicting non-deposition sediment transport in sewer pipes using random forest. Water Res. 189, 116639.
Mood and Carina, 2010. Logistic regression: why we cannot do what we think we can do, and what we can do about it. Eur. Sociol. Rev. 26 (1), 67.
Moreno-Rodenas, A.M., Duinmeijer, A., Clemens, F., 2021. Deep-learning based monitoring of FOG layer dynamics in wastewater pumping stations. Water Res. 202, 117482.
Motamarri, S., Boccelli, D.L., 2012. Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms. Water Res. 46 (14), 4508–4520.
Mustafa, Y.A., Jaid, G.M., Alwared, A.I., Ebrahim, M., 2014. The use of artificial neural network (ANN) for the prediction and simulation of oil degradation in wastewater by AOP. Environ. Sci. Pollut. Res. Int. 21 (12), 7530–7537.
Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D., 2004. An introduction to decision tree modeling. J. Chemom. 18 (6), 275–285.
Najafzadeh, M., Zeinolabedini, M., 2019. Prognostication of waste water treatment plant performance using efficient soft computing models: an environmental evaluation. Measurement 138, 690–701.
Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M., Elshafie, A, 2019. Machine learning methods for better water quality prediction. J. Hydrol. 578, 124084.
Nelson, N.G., Munoz-Carpena, R., Phlips, E.J., Kaplan, D., Sucsy, P., Hendrickson, J., 2018. Revealing biotic and abiotic controls of harmful algal blooms in a shallow subtropical lake through statistical machine learning. Environ. Sci. Technol. 52 (6), 3527–3535.
Nguyen, V.-H., Smith, S.M., Wantala, K., Kajitvichyanukul, P., 2020. Photocatalytic remediation of persistent organic pollutants (POPs): a review. Arab. J. Chem. 13 (11), 8309–8337.
Nicklow, J., Reed, P., Savic, D., Dessalegne, T., Harrell, L., Chan-Hilton, A., Karamouz, M., Minsker, B., Ostfeld, A., Singh, A., Zechman, E., 2010. State of the art for genetic algorithms and beyond in water resources planning and management. J. Water Resour. Plan. Manage. 136 (4), 412–432.
Okwori, E., Viklander, M., Hedstrom, A., 2021. Spatial heterogeneity assessment of factors affecting sewer pipe blockages and predictions. Water Res. 194, 116934.
Olyaie, E., Zare Abyaneh, H., Danandeh Mehr, A., 2017. A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in Delaware River. Geosci. Front. 8 (3), 517–527.
Ostfeld, A., Solomatine, D.P., 2008. Data-driven modelling: some past experiences and new approaches. J. Hydroinform. 10 (1), 3–22.
Palani, S., Liong, S.Y., Tkalich, P., 2008. An ANN application for water quality forecasting. Mar. Pollut. Bull. 56 (9), 1586–1597.
Park, S., Baek, S.-S., Pyo, J., Pachepsky, Y., Park, J., Cho, K.H., 2019. Deep neural networks for modeling fouling growth and flux decline during NF/RO membrane filtration. J. Membr. Sci. 587, 117164.
Peleato, N.M., Legge, R.L., Andrews, R.C., 2018. Neural networks for dimensionality reduction of fluorescence spectra and prediction of drinking water disinfection by-products. Water Res. 136, 84–94.
Perelman, L., Arad, J., Housh, M., Ostfeld, A., 2012. Event detection in water distribution systems from multivariate water quality time series. Environ. Sci. Technol. 46 (15), 8212–8219.
Perendeci, A., Arslan, S., Tanyolac, A., Celebi, S.S., 2009. Effects of phase vector and history extension on prediction power of adaptive-network based fuzzy inference system (ANFIS) model for a real scale anaerobic wastewater treatment plant operating under unsteady state. Bioresour. Technol. 100 (20), 4579–4587.
Podgorski, J., Berg, M., 2020. Global threat of arsenic in groundwater. Science 368, 845–850.
Podgorski, J.E., Eqani, S.A.M.A.S., Khanam, T., Ullah, R., Shen, H., Berg, M., 2017. Extensive arsenic contamination in high-pH unconfined aquifers in the Indus Valley. Sci. Adv. 3 (8), e1700935.
Podgorski, J.E., Labhasetwar, P., Saha, D., Berg, M., 2018. Prediction modeling and mapping of groundwater fluoride contamination throughout India. Environ. Sci. Technol. 52 (17), 9889–9898.
Purkait, B., 2008. Application of artificial neural network model to study arsenic contamination in groundwater of Malda District, Eastern India. J. Environ. Inform. 12 (2), 140–149.
Pyo, J., Kim, K.H., Kim, K., Baek, S.S., Nam, G., Park, S., 2021. Cyanobacteria cell prediction using interpretable deep learning model with observed, numerical, and sensing data assemblage. Water Research 203, 117483.
Raghavendra, N.S., Deka, P.C., 2014. Support vector machine applications in the field of hydrology: a review. Appl. Soft Comput. 19, 372–386.
Rahmanian, B., Pakizeh, M., Mansoori, S.A., Abedini, R., 2011. Application of experimental design approach and artificial neural network (ANN) for the determination of potential micellar-enhanced ultrafiltration process. J. Hazard. Mater. 187 (1–3), 67–74.
Rajaee, T., Khani, S., Ravansalar, M., 2020. Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: a review. Chemomet. Intell. Lab. Syst. 200, 103978.
Recknagel, F., French, M., Harkonen, P., Yabunaka, K.-I., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecol. Modell. 96, 11–28.
Rodríguez-Lado, L., Sun, G., Berg, M., Zhang, Q., Xue, H., Zheng, Q., Johnson, C.A., 2013. Groundwater arsenic contamination throughout China.pdf. Science 341 (6148), 866–868.
Rosenblatt, F., 1957. The Perceptron: a Perceiving and Recognizing Automaton. Cornell Aeronautical Laboratory. Report 85-460-1.
Safari, M.J.S., Danandeh Mehr, A., 2018. Multigene genetic programming for sediment transport modeling in sewers for conditions of non-deposition with a bed deposit. Int. J. Sediment Res. 33 (3), 262–270.
Sagan, V., Peterson, K.T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B.A., Maalouf, S., Adams, C., 2020. 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, 103187.
Salleh, M.A.M., Mahmoud, D.K., Karim, W.A.W.A., Idris, A., 2011. Cationic and anionic dye adsorption by agricultural solid wastes: a comprehensive review. Desalination 280 (1–3), 1–13.
Seshan, H., Goyal, M.K., Falk, M.W., Wuertz, S., 2014. Support vector regression model of wastewater bioreactor performance using microbial community diversity indices: effect of stress and bioaugmentation. Water Res. 53, 282–296.
Shaji, E., Santosh, M., Sarath, K.V., Prakash, P., Deepchand, V., Divya, B.V., 2020. Arsenic contamination of groundwater: a global synopsis with focus on the Indian Peninsula. Geosci. Front. 12 (3), 101079.
Shetty, G.R., Chellam, S., 2003. Predicting membrane fouling during municipal drinking water nanofiltration using artificial neural networks. J. Membr. Sci. 217 (1–2), 69–86.
Sheydaei, M., Soleimani, D., Ayoubi-Feiz, B., 2020. Simultaneous immobilization of Dy_2O_3, graphite and TiO_2 to prepare stable nanocomposite for visible light assisted photocatalytic ozonation of a wastewater: modeling via artificial neural network. Environ. Technol. Innov. 17, 100512.
Shokoohi, R., Salari, M., Safari, R., Zolghadr Nasab, H., Shanehsaz, S., 2020. Modelling and optimisation of catalytic ozonation process assisted by ZrO_2-pumice/H_2O_2 in the degradation of Rhodamine B dye from aqueous environment. Int. J. Environ. Anal. Chem. 100, 1–25.
Shortridge, J.E., Guikema, S.D., 2014. Public health and pipe breaks in water distribution systems: analysis with internet search volume as a proxy. Water Res. 53, 26–34.
Sigmund, G., Gharasoo, M., Huffer, T., Hofmann, T., 2020. 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 (7), 4583–4591.
Singh, K.P., Basant, N., Gupta, S., 2011. Support vector machines in water quality management. Anal Chim Acta 703 (2), 152–162.
Singh, R.M., Datta, B., Jain, A., 2004. Identification of unknown groundwater pollution sources using artificial neural networks. J. Water Resour. Plan. Manage. 130 (6), 506–514.
Speight, V.L., Mounce, S.R., Boxall, J.B., 2019. Identification of the causes of drinking water discolouration from machine learning analysis of historical datasets. Environ. Sci. 5 (4), 747–755.
Stidson, R.T., Park, S.A., McPhail, C.D., 2012. Development and use of modelling techniques for real-time bathing water quality predictions. Water and Environment Journal 26 (1), 7–18.
Strugholtz, S., Panglisch, S., Gebhardt, J., Gimbel, R., 2008. Neural networks and genetic algorithms in membrane technology modelling. J. Water Supply 57 (1), 23–34.

Sun, Y., Clarke, B., Clarke, J., Li, X., 2021. Predicting antibiotic resistance gene abundance in activated sludge using shotgun metagenomics and machine learning. Water Res. 202, 117384.
Takata, M., Lin, B.L., Xue, M., Zushi, Y., Terada, A., Hosomi, M., 2020. Predicting the acute ecotoxicity of chemical substances by machine learning using graph theory. Chemosphere 238, 124604.
Tan, Z., Yang, Q., Zheng, Y., 2020. Machine learning models of groundwater arsenic spatial distribution in bangladesh: influence of holocene sediment depositional history. Environ. Sci. Technol. 54 (15), 9454–9463.
Teychene, B., Touffet, A., Baron, J., Welte, B., Joyeux, M., Gallard, H., 2018. Predicting of ultrafiltration performances by advanced data analysis. Water Res. 129, 365–374.
Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M.L., Boehm, A.B., 2014. Predicting water quality at Santa Monica Beach: evaluation of five different models for public notification of unsafe swimming conditions. Water Res. 67, 105–117.
Tiyasha, Tung, T.M., Yaseen, Z.M, 2020. A survey on river water quality modelling using artificial intelligence models: 2000–2020. J. Hydrol. 585, 124670.
Twarakavi, N.K.C., Kaluarachchi, J.J., 2006. Arsenic in the shallow ground waters of conterminous United States: assessment, health risks, and costs for MCL compliance. J. Am. Water Resour. Assoc. 42, 275–294.
Uddin, M.G., Nash, S., Olbert, A.I., 2021. A review of water quality index models and their use for assessing surface water quality. Ecol. Indic. 122, 107218.
Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., Zhao, S., 2019. Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. 18 (6), 463–477.
Wagle, N., Acharya, T.D., Lee, D.H., 2020. Comprehensive review on application of machine learning algorithms for water quality parameter estimation using remote sensing data. Sens. Mater. 32 (11), 3879–3892.
Wang, Z., Chen, J., Hong, H., 2021d. Developing QSAR Models with Defined Applicability Domains on PPARγ Binding Affinity Using Large Data Sets and Machine Learning Algorithms. Environ. Sci. Technol. 55 (10), 6857–6866.
Wang, G., Jia, Q.-S., Zhou, M., Bi, J., Qiao, J., Abusorrah, A., 2021b. Artificial neural networks for water quality soft-sensing in wastewater treatment: a review. Artif. Intell. Rev. https://doi.org/10.1007/s10462-021-10038-8. In press.
Wang, C., Mao, G., Liao, K., Ben, W., Qiao, M., Bai, Y., Qu, J., 2021a. Machine learning approach identifies water sample source based on microbial abundance. Water Res. 199 (1), 117185.
Wang, L., Zhao, L., Liu, X., Fu, J., Zhang, A., 2021c. SepPCNET: deeping learning on a 3D surface electrostatic potential point cloud for enhanced toxicity classification and its application to suspected environmental estrogens. Environ. Sci. Technol. 55 (14), 9958–9967.
White, K., Dickson-Anderson, S., Majury, A., McDermott, K., Hynds, P., Brown, R.S., Schuster-Wallace, C., 2021. Exploration of *E. coli* contamination drivers in private drinking water wells: an application of machine learning to a large, multivariable, geo-spatio-temporal dataset. Water Res. 197, 117089.
Winkel, L., Berg, M., Amini, M., Hug, S.J., Annette Johnson, C., 2008. Predicting groundwater arsenic contamination in Southeast Asia from surface parameters. Nat. Geosci. 1 (8), 536–542.
Winkel, L.H., Pham, T.K., Vi, M.L., Stengel, C., Amini, M., Nguyen, T.H., Pham, H.V., Berg, M., 2011. Arsenic pollution of groundwater in Vietnam exacerbated by deep aquifer exploitation for more than a century. Proc. Natl. Acad. Sci. U. S. A. 108 (4), 1246–1251.
Wu, G.-D., Lo, S.-L., 2008. Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. Eng. Appl. Artif. Intell. 21 (8), 1189–1195.
Wu, Y., Wang, G., 2018. Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis. Int. J. Mol. Sci. 19 (8), 2358.
Xu, T., Coco, G., Neale, M., 2020. A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning. Water Res. 177, 115788.
Yabunaka, K.-i., Hosomi, M., Murakami, A., 1997. Novel application of a backpropagation artificial neural network model formulated to predict algal bloom. Water. Sci. Tech. 36, 89–97.
Yan, H., Zou, Z., Wang, H., 2010. Adaptive neuro fuzzy inference system for classification of water quality status. J. Environ. Sci. 22 (12), 1891–1896.
Yang, N., Winkel, L.H., Johannesson, K.H., 2014. Predicting geogenic arsenic contamination in shallow groundwater of south Louisiana, United States. Environ. Sci. Technol. 48 (10), 5660–5666.
Yang, Q., Jung, H.B., Marvinney, R.G., Culbertson, C.W., Zheng, Y., 2012. Can arsenic occurrence rates in bedrock aquifers be predicted? Environ. Sci. Technol. 46 (4), 2080–2087.
Yang, S.-S., Yu, X.-L., Ding, M.-Q., He, L., Cao, G.-L., Zhao, L., Tao, Y., Pang, J.-W., Bai, S.-W., Ding, J., Ren, N.-Q., 2021. Simulating a combined lysis-cryptic and biological nitrogen removal system treating domestic wastewater at low C/N ratios using artificial neural network. Water Res. 189, 116576.
Yaqub, M., Eren, B., Eyupoglu, V., 2019. Soft computing techniques in prediction Cr(VI) removal efficiency of polymer inclusion membranes. Environ. Eng. Res. 25 (3), 418–425.
Yaqub, M., Lee, S.H., 2020. Micellar enhanced ultrafiltration (MEUF) of mercury-contaminated wastewater: experimental and artificial neural network modeling. J. Water Process Eng. 33, 101046.
Yaseen, Z.M., 2021. An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: review, challenges and solutions. Chemosphere 277, 130126.
Ye, X., Zuo, J.e., Li, R., Wang, Y., Gan, L., Yu, Z., Hu, X., 2019. Diagnosis of sewer pipe defects on image recognition of multi-features and support vector machine in a southern Chinese city. Front. Environ. Sci. Eng. 13 (2), 17.
Yetilmezsoy, K., Demirel, S., Vanderbei, R.J., 2009. Response surface modeling of Pb(II) removal from aqueous solution by *Pistacia vera* L.: box-Behnken experimental design. J. Hazard. Mater. 171 (1–3), 551–562.
Zaghloul, M.S, Hamza, R.A., Iorhemen, O.T., Tay, J.H., 2020a. Comparison of adaptive neuro-fuzzy inference systems (ANFIS) and support vector regression (SVR) for data-driven modelling of aerobic granular sludge reactors. J. Environ. Chem. Eng. 8 (3), 103742.
Zaghloul, M.S., Iorhemen, O.T., Hamza, R.A., Tay, J.H., Achari, G., 2020b. Development of an ensemble of machine learning algorithms to model aerobic granular sludge reactors. Water Res. 189, 116657.
Zaqoot, H.A., Ansari, A.K., Unar, M.A., Khan, S.H., 2009. Prediction of dissolved oxygen in the Mediterranean Sea along Gaza, Palestine - an artificial neural network approach. Water Sci. Technol. 60 (12), 3051–3059.
Zhang, K., Zhong, S., Zhang, H., 2020. Predicting aqueous adsorption of organic compounds onto biochars, carbon nanotubes, granular activated carbons, and resins with machine learning. Environ. Sci. Technol. 54 (11), 7008–7018.
Zhang, Q., Rodriguez-Lado, L., Johnson, C.A., Xue, H., Shi, J., Zheng, Q., Sun, G., 2012. Predicting the risk of arsenic contaminated groundwater in Shanxi Province, Northern China. Environ. Pollut. 165, 118–123.
Zhang, Q., Rodriguez-Lado, L., Liu, J., Johnson, C.A., Zheng, Q., Sun, G., 2013. Coupling predicted model of arsenic in groundwater with endemic arsenism occurrence in Shanxi Province, Northern China. J. Hazard. Mater. 262, 1147–1153.
Zhang, W., Gu, J., Li, Y., Lin, L., Wang, P., Wang, C., Qian, B., Wang, H., Niu, L., Wang, L., Zhang, H., Gao, Y., Zhu, M., Fang, S., 2019a. New insights into sediment transport in interconnected river–lake systems through tracing microorganisms. Environ. Sci. Technol. 53 (8), 4099–4108.
Zhang, Y., Gao, X., Smith, K., Inial, G., Liu, S., Conil, L.B., Pan, B., 2019b. Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. Water Research 164, 114888.
Zhang, Y., Wang, X., Shan, J., Zhao, J., Zhang, W., Liu, L., Wu, F., 2019c. Hyperspectral imaging based method for rapid detection of microplastics in the intestinal tracts of fish. Environ. Sci. Technol. 53 (9), 5151–5158.
Zhao, Y., Nan, J., Cui, F.-y., Guo, L., 2007. Water quality forecast through application of BP neural network at Yuqiao reservoir. J. Zhejiang Univ. 8 (9), 1482–1487.
Zhi, W., Feng, D., Tsai, W.P., Sterle, G., Harpold, A., Shen, C., Li, L., 2021. From hydrometeorology to river water quality: can a deep learning model predict dissolved oxygen at the continental scale? Environ. Sci. Technol. 55 (4), 2357–2368.
Zhong, S., Hu, J., Fan, X., Yu, X., Zhang, H., 2020. A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants. J. Hazard. Mater. 383, 121141.