Title: Machine Learning to Access and Ensure Safe Drinking Water Supply: A Systematic Review

Feng Feng[1,2†], Zhenru Chen[3†], Jianyuan Ni[4], Yuanxun Zhang[1], Yuan Feng[5], Yunchao Xie[6*], Chiqian Zhang[7*]

Affiliations: [1]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri 65211, United States

[2]Department of Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, United States

[3]Department of Mechanical and Aerospace Engineering, University of Missouri, Columbia, Missouri 65211, United States

[4]Department of Computer Science, Lamar University, Beaumont, Texas 77710, United States

[5]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, United States

[6]Department of Mechanical and Manufacturing Engineering, Miami University, Oxford, Ohio 45056, United States

[7]Civil Engineering Program, College of Engineering & Computer Science, Arkansas State University, Arkansas 72467, United States

[†]These authors contributed equally

[*]E-mails: czhang@astate.edu; xiey54@miamioh.edu

Abstract Drinking water is essential to public health and socioeconomic growth. Therefore, assessing and ensuring drinking water supply is a critical task in modern society. Conventional approaches to analyzing and controlling drinking water quality are labor-intensive and costly with a low throughput. Machine learning (ML) is an alternative, promising technique for assessing and ensuring safe drinking water supply. Existing reviews have summarized the applications of ML in safe drinking water supply from different aspects. However, a state-of-the-art, comprehensive review is missing that focuses on applying ML to monitor, simulate, predict, and control drinking water quality, especially in engineered water systems. This review, therefore, critically compiles the applications of ML in assessing and ensuring water quality in engineered water systems. To be comprehensive, we also cover the applications of ML in other drinking-water-related settings such as water sources and water purification processes. We explain the basic mechanics and workflows of ML and focus on the applications of ML to access and control key factors or etiologies in drinking water from the physical, chemical, and microbiological aspects. Those factors or etiologies affect water quality and public health, such as water pipeline failures, disinfectant by-products, heavy metals, opportunistic pathogens, biofilms, and antimicrobial resistance genes. We then present a macroscopic illustration to display the distribution of ML models across research topics in safe drinking water supply. Neural-network-based and regression-based models are the top two models frequently used in the field of drinking water supply. We finally discuss the challenges and outlooks for the applications of machine learning in safe drinking water supply. Filling the gap between the water research and the AI research communities and using AI to solve the global drinking water crisis should be the main focus of future research. This is the first review summarizing the feasibility and applications of ML in assessing and ensuring water quality in municipal engineered water systems as well as related water environments.

Keywords: Drinking water quality; Engineered water systems; Artificial intelligence; Opportunistic pathogens; Disinfection byproducts; Heavy metals

1. Introduction

Clean and safe drinking water is vital to public health(Cabral, 2010; Cutler and Miller, 2005; Li and Wu, 2019). Most people obtain drinking water from public water systems. For instance, over 90% of people in the United States (U.S.) access drinking water from approximately 150,000 public water systems(U.S. EPA, 2022). Public water systems provide water for human consumption through engineered water systems (including drinking water distribution systems or DWDSs and building premise plumbing systems) (Zhang and Lu, 2021b). Drinking-water-related disease outbreaks and chronic diseases cause many hospitalizations and deaths, leading to significant socioeconomic losses (Benedict et al., 2017; Craun et al., 2010; Lee et al., 2023). Therefore, assessing and ensuring water quality, especially in engineered water systems, is critical to the health of the end consumers (WHO, 2011).

Assessing and ensuring water quality in engineered water systems are complex(Li and Wu, 2019). Multiple ever-changing variables affect water quality in engineered water systems such as the quality of water sources, treatment processes, water pipe materials, system configuration and length, natural disasters, and geographical factors (Delpla et al., 2009; Li and Wu, 2019; Proctor et al., 2020). For instance, the effluent at water utilities many have high quality, while the quality deteriorates in engineered water systems because of microbial (re)growth, the formation of disinfection by-products (DBPs), pipe failures (e.g., breaks and leaks), and the detachment of heavy metals from pipes(Li et al., 2019; Liu et al., 2016). Meanwhile, engineered water systems have various hazardous agents such as harmful microbes (especially opportunistic pathogens or OPs) (Zhang et al., 2021), DBPs(Benítez et al., 2021; Lee et al., 2013), heavy metals(Chowdhury et al., 2016; Gonzalez et al., 2013), pesticides and herbicides(Syafrudin et al., 2021)

(Mukhopadhyay et al., 2022), and other emerging contaminants (e.g., antimicrobials and microplastics) (Gogoi et al., 2018; Kirstein et al., 2021; Taheran et al., 2018). Those agents are interconnected, and controlling only one group of agents frequently fails to secure the drinking water quality. For instance, increasing disinfectant residual concentrations suppresses microbial regrowth in engineered water systems but promotes the formation of DBPs(Zhang and Lu, 2021c). By contrast, reducing the dose of disinfectant residuals in engineered water systems can mitigate the DBP issue, but microbes (including pathogens) can thrive.

Because of the complex nature of water quality in engineered water systems, assessing and ensuring drinking water quality using conventional means is challenging. Traditional methods are time-consuming, labor-intensive, inefficient (i.e., low throughput), and costly(Ahmed et al., 2019) (Zainurin et al., 2022). Artificial intelligence (AI), especially machine learning (ML), is promising to address the deficiencies in the traditional approaches to access and ensure safe drinking water supply (Richards et al., 2023). The adaptability and predictive power of ML offer significant advantages over other AI technologies (Willard et al., 2022; Zhu et al., 2022), particularly when handling drinking water quality with a dynamic and complex nature. Consequently, ML emerges as a specialized application of AI to enhance drinking water treatment and quality(Henrique Alves Ribeiro and Reynoso-Meza, 2023; Li et al., 2021; Narita et al., 2023; Speight et al., 2019).

Existing reviews have summarized the applications of ML in various aspects of drinking water quality (Ewuzie et al., 2022; Huang et al., 2021; Zhu et al., 2022), such as source water quality and contamination(Gong et al., 2023; Zanoni et al., 2022), drinking water treatment(Li et al., 2021; Lowe et al., 2022; Ortiz-Lopez et al., 2022), and drinking water quality anomaly detection (Dogo

et al., 2019). However, a state-of-the-art, comprehensive review is missing that focuses on applying ML to monitor, simulate, predict, and control drinking water quality, especially in engineered water systems. Since engineered water systems are the vital civil infrastructure delivering municipal water from water utilities to the residents and industrial/commercial consumers (WHO, 2011), summarizing such applications can help ensure drinking water quality, protect public health, and promote socioeconomic development.

In this review, we critically compile the applications of ML in assessing and ensuring water quality across all stages of drinking water treatment and distribution (especially in engineered water systems) from the physical, chemical, and microbiological aspects (Figure 1). We focus on the applications of ML to access and control key factors or etiologies in drinking water that affect water quality and public health, such as water pipeline failures, DBPs, heavy metals, OPs, biofilms, and antimicrobial resistance genes (ARGs). Since water sources and the treatment processes govern drinking water quality, we also discuss the applications of ML in monitoring source water quality and the efficiency of water treatment technologies. This is the first review focusing on the applications of ML in assessing and controlling water quality in engineered water systems, while the applications of ML in other drinking-water-related settings are also summarized.

2. Literature search strategy and inclusion/exclusion criteria

We follow previous protocols to search the literature on the applications of ML to assess and ensure water quality in engineered water systems and related settings(Tolaymat et al., 2010) (Miake-Lye et al., 2016; Mostafavifar et al., 2012; Zhang et al., 2016; Ziegelbauer et al., 2012). We mainly search Google Scholar (scholar.google.com) but also include other databases such as

ScienceDirect (sciencedirect.com), PubMed Central® (ncbi.nlm.nih.gov/pmc/), and ACS Publications (pubs.acs.org). We include only high-quality English articles published in or before December 2023 in top-tier, peer-reviewed academic journals, books, and conference proceedings. We manually review the retrieved publications and retain only those focusing on the applications of ML in assessing and ensuring drinking water quality. We also manually check articles that cited and were cited by those retrieved publications and keep the most relevant ones for further analysis.

The main literature search keywords are "machine learning," "deep learning," "artificial intelligence," "models," "modeling," "prediction," "networks," "anomaly detection," "municipal," "water sources" "water treatment," "water utilities," "water plants," "water treatment facilities," "distribution systems," "building," "premise plumbing," "pipes," "pipelines," "water mains," "drinking water," "potable water," "tap water," "municipal water," "bulk water," "disinfection," "disinfectants," "free chlorine," "chlorine," "monochloramine," "disinfectant residuals," "disinfection by-products," "heavy metals," "lead," "copper," "nitrification," "pipe failures," "water-related," "water-borne," "disease outbreaks," "public health," "microbes," "bacteria," "biofilms," "opportunistic pathogens," "opportunistic premise plumbing pathogens," "*Legionella*," "*Mycobacterium*," "antibiotic resistance," "antimicrobial resistance," "antimicrobial resistance genes," "*Cryptosporidium*," and "*Giardia*."

3. Machine learning primers

In the past decade, ML has sparked substantial progress across domains in modern society, including object detection (Erhan et al., 2014; Lin et al., 2017), autonomous driving (Almalioglu et al., 2022; Feng et al., 2023), drug delivery (Allesoe et al., 2023; Wołos et al., 2022), playing

games (Kaufmann et al., 2023; Vinyals et al., 2019), weather forecasting (Bi et al., 2023), and design-by-analogy (Jiang et al., 2021). Three key advancements in AI and computer science drive this process: I) the availability of extensive datasets, II) the development of robust computing hardware, and III) the refinement of advanced algorithms. In contrast to traditional physical and chemical theories relying on explicit formulas for problem-solving, ML tackles problems by extracting concealed insights from datasets through the learning process (Ley et al., 2022).

3.1 Categories of machine learning

On the basis of the quantity and nature of available datasets, ML can be divided into four main categories: I) supervised learning, II) unsupervised learning, III) semi-supervised learning, and IV) reinforcement learning (Goodfellow et al., 2016). Supervised learning uses both input data and corresponding labels for training. Unsupervised learning, on the other hand, deals solely with input data without labeled information. As a combination of supervised and unsupervised learning, semi-supervised learning combines mostly unlabeled datasets and a limited number of labeled ones. Reinforcement learning algorithms, such as Q-learning, enable learning by interacting with an environment and receiving feedback. These algorithms underpin the growth of AI, improving system performance through exposure to data and experience. Among the four categories, supervised learning is the most widely used and well-established in assessing and protecting drinking water quality because of its strength in prediction with labeled datasets (Cordero et al., 2021; Hong et al., 2020; Zhang et al., 2019; Zhou et al., 2019). Conversely, the application of semi-supervised learning in safe drinking water supply is scarce, and reinforcement learning is unexplored.

Supervised learning is acquiring a mapping function from the input data to the corresponding output data on the basis of a labeled set of input-output pairs or conditional distributions (Goodfellow et al., 2016). During training, the algorithm adjusts its parameters to minimize the discrepancy between predicted and actual outputs. Supervised learning is widely used in tasks such as classification and regression. Examples of supervised learning models include naïve Bayes (NB) (Gomez-Alvarez and Revetta, 2020), logistic regression (LR) (Bagriacik et al., 2018), support vector machines (SVM) (Oh et al., 2021), k-nearest neighbor (KNN) (Ghiassi et al., 2017), decision trees (DT) (Shi et al., 2022), random forests (RF) (Berglund et al., 2023; Cha et al., 2021), and extreme gradient boosting (XGB) (Park et al., 2020).

Unsupervised learning focuses on extracting patterns, structures, and relationships from the input data without labeled outputs (Goodfellow et al., 2016). Instead of targeting predefined targets, unsupervised learning algorithms (such as K-means and dimensionality reduction techniques) discover inherent structures within the data. For instance, K-means groups similar data points on the basis of their intrinsic features (Moodley and van der Haar, 2019). Dimensionality reduction techniques including principal component analysis (PCA) simplify complex datasets by preserving their essential characteristics (Peleato et al., 2018). In drinking water research, unsupervised learning is crucial in tasks such as clustering bacteria (Moodley and van der Haar, 2019; Pinto et al., 2014), simplifying data for subsequent analysis (Peleato et al., 2018), and analyzing raw data to identify key parameters and major relationships affecting water quality (Kazemi et al., 2023).

3.2 Workflow chart of machine learning

3.2.1 Define the problem

Defining the problem is critical in applying ML that convents a complex challenge into a well-defined scope and purpose. To start, one should define the objectives, outline desired outcomes, and determine if the task is a regression problem (such as predicting DBP concentration) or a classification problem (such as categorizing drinking water contamination status and pipe burst localization).

3.2.2 Gather data

In any ML endeavor, the quality of data is the key to the success of the subsequent modeling. Data collection involves sourcing, gathering, and recording from various origins, such as observational studies, controlled experiments, publications, and databases. The collected data should be pertinent to the problem, accurate, and suitable for ML model development. Along with data collection, one needs to document the sources, methods, and potential biases associated with the data to ensure transparency and reproducibility.

3.2.3 Data preprocessing

Data preprocessing is critical in a ML workflow that involves cleaning (i.e., filtrating), transforming, and organizing raw data. Cleaning is identifying and rectifying errors, inaccuracies, inconsistencies, and anomalies in the collected data, which usually involves filtrating missing values, detecting outliers, removing duplicate entries, and converting data types. Normalization, as an example of data transformation, is scaling data to fit within a specific range, such as 0 to 1. Normalization equalizes the contribution of each feature to the model training and avoids any single attribute disproportionately influencing the results. Organizing raw data entails structuring and arranging the data in a manner that is optimal for the specific ML algorithms to be applied.

3.2.4 Model training

The first step in data training is to divide the dataset into training, validation, and test datasets (Hastie et al., 2009). The training dataset trains the model, the validation dataset tunes hyperparameters, and the test dataset evaluates the performance of the model. The second step is to select an appropriate model on the basis of the nature of the problem. Not all algorithms work equally well for all types of problems. During training, the model recognizes the relationship between the inputs and outputs and minimizes the difference between predicted and actual outputs by iteratively adjusting the model parameters.

3.2.5 Model evaluation

Evaluation metrics differ on the basis of the nature of the problem (Xie et al., 2023). Evaluation metrics assess the performance of classification models, revealing the effectiveness of the classification models and their ability to distinguish between classes. Accuracy refers to the proportion of correctly classified instances to the total instances. Precision measures the proportion of correctly predicted positive instances among all predicted positives. Recall gauges the proportion of correctly predicted actual positive instances. The F1-score combines precision and recall into a single metric, offering a balanced view of the accuracy of a model (Sokolova and Lapalme, 2009). The confusion matrix provides a tabular representation of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions (James et al., 2013). Receiver operating characteristic (ROC) curves illustrate the trade-off between the TP rate and the FP rate at different classification thresholds with the area under the curve (AUC) summarizing the performance of the curve (Bradley, 1997). In regression tasks, the mean squared error (MSE) and

root mean squared error (RMSE) quantify the average squared differences between predicted and actual values, and the mean absolute error (MAE) measures the average absolute differences (Willmott and Matsuura, 2005). Additionally, the coefficient of determination (R^2) indicates the proportion of variance in the target variable explained by the model (Steel and Torrie, 1960).

4. Machine learning to ensure safe drinking water supply: The physical perspective

4.1. To predict and manage drinking water production and demand

ML is promising in ensuring safe drinking water supply from the physical perspective (Table 1). Insufficient production of drinking water could constrain socioeconomic development and population growth (Grey and Sadoff, 2007). A hybrid model that combines genetic algorithms (GA) and GA artificial neural networks (GA-ANN) can predict drinking water production (Figure 2a) (Zhang et al., 2019). The model uses predictors such as temperature, chemical oxygen demand (COD), and electricity and chemical consumption. GA can optimize weights and biases, enhancing the prediction accuracy of the relationship between the inputs and outputs of the ML model. GA can also increase the tolerance to imprecisions, uncertainties, and approximations in the inputs. The hybrid GA-ANN was trained and validated with monthly data from 45 water utilities across China. The performance of GA-ANN (R^2 = 0.93) is substantially better than a three-layered ANN (R^2 = 0.71) when more training data are incorporated. The GA-ANN effectively forecasts fluctuations in water production for various scenarios, highlighting its feasibility in promptly and appropriately adjusting water treatment operations. To assess drinking water demand patterns, one can apply unsupervised learning algorithms to raw time-series data of drinking water consumption. For instance, the hierarchical K-means algorithm can classify drinking water consumption patterns (Leitão et al., 2019). In that algorithm, daily time-series demand with hourly records served as the

sample inputs within a 24-dimensional feature space to identify dense and distinctly separated temporal patterns on drinking water demand. In contrast to directly clustering the drinking water demand patterns, short-time water demand forecasting is more intriguing and has more practical merits in optimizing DWDSs. A gated recurrent unit network (GRUN) predicts short-term water demand for the upcoming 15 min and 24 h using a time step of 15 min (Guo et al., 2018). The GRUN model has a higher accuracy with a mean absolute percentage error (MAPE) of 2.06% than the conventional three-dense-layered ANN (MAPE = 2.46%) and the seasonal autoregressive integrated moving average (SARIMA) model (MAPE = 2.57%) for the 15-min prediction. For the 24-h prediction, the GRUN model also achieves more precise forecasts with MAPE ranging from 4.33% to 4.96%. Another study used three ML models to forecast both short-term and long-term water demand encompassing daily, weekly, and monthly intervals in Tehran, Iran (Ghiassi et al., 2017). The models include a dynamic artificial neural network (DAN2), a focused time-delay neural network (FTDNN), and a KNN. DAN2 based on the adaptive network principle allows the architecture to adjust dynamically responding to data-driven learning. Given its inherent design, DAN2 is promising at time series forecasting, catering to datasets characterized by evolving temporal patterns (Ghiassi et al., 2005). For instance, DAN2 achieves remarkable prediction accuracies (96% for daily, 99% for weekly, and 98% for monthly water demand forecasts) and outperforms both FTDNN and KNN (Ghiassi et al., 2017).

4.2 To monitor and predict drinking water pipeline failures

Pipeline failures in engineered water systems cause significant water loss and contaminate municipal water (Renwick et al., 2019). These failures can introduce microbes and chemicals from the surroundings into distributed water. A novel burst location identification framework by fully-

linear DenseNet (BLIFF) framework can detect pipe burst locals (Figure 2b) (Zhou et al., 2019). BLIFF relies on deep learning through the fully-linear DenseNet (FL-DenseNet) model. BLIFF supplants the convolutional layers in DenseNet with linear connections and omits pooling layers. With the real-time pressure measurements as the inputs, BLIFF generates the likelihood values of a burst for each pipe in the potential burst district. The prediction accuracies (62.35% to 98.58%) of BLIFF are almost two times the original DenseNet model. This remarkable improvement is due to the enhanced ability of the linear-connection layer to discern global features within the pressure signals. The effective deployment of deep learning methods such as BLIFF corroborates the viability of pressure values in burst localization, countering prior assertions of their insensitivity to burst events (Bakker et al., 2014; Mounce et al., 2010). Another cutting-edge paradigm, advanced meta-learning (AdvaML), can predict the failure of drinking water pipelines (Almheiri et al., 2021). AdvaML is also a deep learning model but is rooted in the ANN architecture. AdvaML comprises an input layer with 33 neurons (mirroring the 33 predictors including pipe and climate data), four hidden layers, and an output layer that yields the failure/hazard index of a pipe. AdvaML not only forecasts the risk index associated with pipe failure but also detects pivotal determinants of pipeline failure timing. Of these determinants, the number of lanes and chlorine residual concentration are paramount, collectively contributing approximately 9% to the lifespan analysis of DWDSs. Notably, AdvaML tackles the challenge of limited data, promoting the applications of deep learning to assess pipe failure. Compared with alternative models like cox-proportional hazards (Cox-PH), survival support vector machine (SSVM), and random survival forest (SRF), AdvaML has commendable performance even with scant training data because of its knowledge transfer from initial parameterization to the ultimate learning phase.

While inherent system vulnerabilities cause pipeline failures, external factors exacerbate the issue (Fan et al., 2023; Folkman, 2018). Climatic extremes and weather disasters, such as wildfires, become more frequent because of climate change and threaten drinking water infrastructures. In response, researchers leverage ML to better understand and predict the impact of these disasters on water pipes. Two ensemble ML models (RF and XGB) can assess and forecast the repercussions of calamities on water treatment infrastructures (Park et al., 2020). These models incorporate 23 variables encompassing facility specifications and operational data from 419 water utilities in South Korea. The models project the total disaster index (TDI), a metric signifying the effects or damages wrought by three predominant disasters (typhoons, heavy rainfalls, and earthquakes) on DWDSs. While both RF and XGB have commendable predictive prowess concerning the TDI, XGB slightly outperforms in most scenarios. Another study developed four models, a linear regression-based repair rate (RR) method, LR, boosted regression trees (BRT), and RF, to predict pipeline damage during an earthquake (Bagriacik et al., 2018). The models incorporate parameters such as ground shaking, permanent ground deformation, pipe material, pipe diameter, year of installation, type, and trench backfill type. Each model demonstrates unique strengths. The BRT model has the best overall predictive performance, while the LR model is instrumental in highlighting the influence of pipe materials and trench types on pipeline damage.

5. Machine learning to monitor and ensure chemical drinking water quality

5.1. To assess and control disinfection by-products in municipal water

5.1.1. To predict the concentration of disinfection by-products in drinking water

ML is useful in monitoring and ensuring chemical drinking water quality (Table 2). Drinking water disinfection is critical to ensuring microbial drinking water quality and safeguarding public health

(Zhang and Lu, 2021c). Disinfection is effective in killing pathogens, impeding microbiological recontamination, and inhibiting biofilm development in drinking water (Mazhar et al., 2020). Chlorine-based disinfectants, such as free chlorine (e.g., chlorine gas and sodium hypochlorite), bound or combined chlorine (e.g., monochloramine), and chlorine dioxide, are widely used in water treatment because of their cost-effectiveness and high efficiency (Gupta and Ali, 2013) (Jefri et al., 2022; Zhang et al., 2018). Nonetheless, when these disinfectants interact with natural organic matter (NOM) and anthropogenic compounds (such as pharmaceuticals and antimicrobials), they generate DBPs such as trihalomethanes (THMs), haloacetic acids (HAAs), haloketones (HKs), haloacetonitriles, halophenols, and halopropanoles(Li et al., 2023a) (Favere et al., 2021; Xiao et al., 2023). DBPs are harmful and even carcinogenic with significant health risks(Pandian et al., 2022; Zhou et al., 2023). Therefore, monitoring and controlling DBPs in drinking water is vital to public health(He et al., 2021; Helte et al., 2023; Redondo-Hasselerharm et al., 2022). Conventional methods for monitoring DBPs require expensive instrumentation such as gas chromatography (GC) and liquid chromatography (LC) combined with mass spectrometry (MS) and complicated pre-treatment processes. Thus, those conventional methods are labor-intensive, costly, and time-consuming, limiting the ability of water utilities to reduce DBP formation. By contrast, ML models for monitoring DBPs in drinking water are accurate, efficient, inexpensive, and easy to handle (Balogun et al., 2021; Jia et al., 2021; Podgorski and Berg, 2022).

Finding the optimal disinfectant dosages to simultaneously minimize the formation of DBPs and pathogen regrowth in finished water is crucial (He et al., 2021; Zhang and Lu, 2021c). Nevertheless, reaching this goal with traditional methods is time-consuming, expensive, and complex. Therefore, ML stands out as it is effective in predicting the formation of DBPs, significantly reducing capital

and human investment. Common input parameters in these models are operational parameters (temperature and time) and water quality variables, such as pH, UV_{254}, and the concentrations of dissolved organic carbon (DOC), chloride (C_{resCl^-}), bromide (C_{Br^-}), nitrite nitrogen ($C_{NO_2^- - N}$), and ammonium nitrogen ($C_{NH_4^+ - N}$) (Deng et al., 2021; Hong et al., 2020; Hu et al., 2023; Lin et al., 2020; Pan et al., 2023; Singh and Gupta, 2012). The outputs are the concentrations of DBPs, such as THMs, HAAs, and HKs

A study developed three ML models [ANN, SVM, and gene expression programming (GEP)] to forecast the formation of DBPs on the basis of a 63-point dataset (Singh and Gupta, 2012). Specifically, pH, temperature, contact time (*t*), the concentration of bromide, and DOC-normalized chlorine dose (Cl_2/DOC) are the inputs. The concentration of THMs in chlorinated river water is the output. SVM outperforms the other two models, exhibiting the highest R^2 and the lowest RMSE values. Furthermore, sensitivity analysis reveals that pH, HRT, and temperature are the top three contributors to DBP formation. In addition, radial basis function (RBF) based ANN models can predict the formation of common DBPs such as HAAs (Lin et al., 2020), THMs (Hong et al., 2020), and HKs (Deng et al., 2021) in DWDSs. For instance, a study extracted 64 representative data points from the literature to predict HAA formation using pH, temperature, DOC, UV_{254}, C_{resCl^-}, C_{Br^-}, $C_{NO_2^- - N}$, and $C_{NH_4^+ - N}$ as the inputs (Figure 3a) (Lin et al., 2020). The RBF ANN model outperforms the linear and log-linear models by 21% to 47 % in accuracy, respectively. Therefore, the RBF ANN model is promising in assessing DBP formation and optimizing disinfection. A follow-up study used 64 data points to train an RBF ANN model to predict THM formation (Hong et al., 2020). The RBF ANN model achieves accuracies between 92% and 98% and regression

coefficients ranging from 0.76 to 0.93, outperforming the linear and log-linear models and demonstrating its superiority to uncover complex non-linear patterns in THM formation. Even when trained with fewer water quality variables, a fusion of grey relation analysis with RBF ANN could provide superior prediction results. Furthermore, an RBF ANN model trained with 63 data points from tap water predicts the formation of HK (Deng et al., 2021). Both RBF ANN and back propagation (BP) ANN models outstrip the linear and log-linear models with RBF ANN displaying higher accuracies in both internal and external validations. Another study explored the application of a decision tree boost (DTB) model to predict the concentrations of THM4 and HAAs (Pan et al., 2023). The study examined correlations between water quality parameters and mixed chlorine/chloramine species. The work then selected seven variables such as NH_2Cl, $NHCl_2$ plus organic chloramines, pH, total dissolved nitrogen (TDN), nitrite, total organic carbon (TOC), and NH_4^+ as the independent variables to predict THM4 and HAAs. The DTB model demonstrates enhanced prediction accuracy with R^2 values of 0.56 for THM4 and 0.65 for HAAs, while the inclusion of organic chloramines improves the prediction precision. Additionally, a study implemented multiple ML models to predict emerging DBPs in small DWDSs across Canada by analyzing data from eleven such networks (Hu et al., 2023). The models use parameters like temperature, total chlorine residual, DOC, turbidity, pH, conductivity, and UV_{254} to predict the concentrations of THMs, HAAs, dichloroacetonitrile (DCAN), chloropicrin (CPK), and trichloropropanone (TCP). Among the evaluated models, support vector regression (SVR) and Gaussian process regression (GPR) show superior performance with SVR exhibiting the highest prediction accuracy (R^2 = 0.94) and stability for DCAN and TCP, while GPR is optimal for predicting CPK (R^2 = 0.92).

UV-visible and fluorescence spectral techniques are preferred to monitor DBPs (Krasner et al., 2006; Rodriguez et al., 2004). Fluorescence spectroscopy is sensitive to assessing the characteristics and reactivity of NOM owing to its minimal sample preparation requirement and short acquisition time (Pifer and Fairey, 2012). However, the complex high-dimensional characteristics of fluorescence spectroscopy make it difficult to predict DBP formation. The resources and time required for DBP analysis restrict the capacity of water utilities to reduce DBP formation. This situation has prompted the development of ML models that can predict DBP formation. Since DBPs form from the reaction between chlorine-based disinfectants and NOM, including NOM measurement in these models is critical. Nonetheless, NOM is a diverse group of organic molecules with complex characteristics, adding to the complexity of capturing the reactivity of NOM with chlorine-based disinfectants (Wagner and Plewa, 2017).

Autoencoder-neural networks (AE-NN) can predict the concentrations of both THMs and HAAs in river water from fluorescence spectra (Figure 3b) (Peleato et al., 2018). To manage the high dimensionality of fluorescence spectra, the researcher applied three dimension-reduction techniques [AE-NN, parallel factors analysis (PARAFAC), and PCA] before further analysis. Afterward, the researcher trained NN to identify fluorescence regions associated with DBP formation and to predict DBP concentrations. The AE-NN model has superior predictive accuracies for THMs and HAAs, achieving validation MAE values of 9.65 µg/L and 9.64 µg/L, respectively. These figures exceed the performance of PCA, which has higher errors of 13.19 µg/L for THMs and 11.92 µg/L for HAAs. Furthermore, the precision of the AE-NN model surpasses that of parallel factors analysis, which has MEA values of 20.39 µg/L for THMs and 14.00 µg/L for HAAs. In addition, convolutional neural networks (CNNs) can predict DBP concentrations

414 from fluorescence spectra without extensive data pre-processing (Peleato, 2022). Compared with
415 multilayer perceptron (MLP) and dimensionality reduction techniques, CNNs not only exhibit
416 superior prediction accuracies for THMs and HAAs but also identify the fluorescence spectra
417 regions highly associated with DBP formation.
418
419 **5.1.2. To elucidate the formation mechanisms of disinfectant by-products in drinking water**
420 ML is promising in predicting the formation of DBPs using either water quality metrics and
421 operational parameters or via online spectrum monitoring. However, even with the knowledge of
422 DBP concentrations, their removal from drinking water remains costly and inefficient (Bond et al.,
423 2011; Rodriguez et al., 2004). An effective approach for DBP control is to remove DBP precursors
424 and prevent them from reaching the clear wells in water utilities (Bond et al., 2012; Krasner et al.,
425 2013). This needs a comprehensive understanding of the mechanisms for the formation of DBPs.
426
427 A multiple linear regression (MLR) model can predict the production of chloroform (a THM
428 compound) from organic precursors (Figure 3c) (Bond and Graham, 2017). Relying on 211
429 precursors from 22 studies, the MLR model uses 19 descriptors as the inputs and chloroform yield
430 as the output. The well-trained MLR model has a promising prediction accuracy with an R^2 value
431 of 0.91 and an RMSE value of 8.93 mol/mol. Further chemical insights pinpoint that functional
432 groups, such as hydroxyl, chlorine, and carboxyl groups, significantly affect chloroform formation.
433 ML can also forecast the formation of HAAs from the interaction between organic precursors and
434 free chlorine (Cordero et al., 2021). The training dataset comprises 283 organic compounds and
435 732 chemical descriptors as the inputs with HAA yield as the output. These organic compounds
436 are converted into 2D and 3D chemical descriptors with their SMILE strings used for ML

compatibility. Three common ML models (RF, SVR, and MLP) are selected because they can handle nonlinear problems, activity cliffs, and high dimensions in addition to MLR as a benchmark. The RF model is the top performer with the lowest RMSE values of 1.05 and 1.19 for dichloroacetic acid (DCAA) and trichloroacetic acid (TCAA), respectively. The crucial predictors of TCAA formation are the number of aromatic bonds, hydrophilicity, and electrotopological descriptors related to electrostatic interactions and the atomic distribution of electronegativity.

5.1.3. To investigate alternative disinfectants to reduce disinfectant by-product formation

Since chlorine-based disinfectants produce harmful DBPs, alternative disinfectants for drinking water disinfection attract attention. An alternative disinfectant is ozone. Unlike chlorination, ozonation does not produce chlorinated THMs and HAAs. Ozone, therefore, provides a two-fold benefit: it is effective and does not generate chlorinated DBPs. However, ozonation produces various other DBPs (Mao et al., 2014; Richardson et al., 1999). The occurrence and toxicity of these ozonated DBPs are a concern (Simpson and Mitch, 2022; Srivastav et al., 2020). Therefore, while ozone does not generate chlorinated DBPs, its use necessitates careful consideration of the potential for other toxic byproducts. In this section, we review how ML models have been developed and applied in the field of ozonation. ML can help control the formation of bromate and reduce micropollutants, microbial indicators, and organic contaminants during ozonation.

ANN has two advantages when compared with MLR in controlling bromate formation during ozonation (Legube et al., 2004): First, ANN with an R^2 value of 0.98 is more accurate than MLR. Second, ANN classifies model variables (predictors) in the descending order of impact: ozone dose, $C_{NH_4^+-N}$, C_{Br^-}, pH, temperature, DOC, and alkalinity. While ANN has superior performance, the

simplicity of MLR is attractive. However, a key limitation of MLR is that its accuracy decreases with an increased number of samples because MLR cannot effectively process nonlinear components.

ML models based on routinely measured physical-chemical water quality parameters can also predict the changes in micropollutant concentration during ozonation. For instance, RF models can predict the reduction of micropollutants during ozonation(Cha et al., 2021) (Figure 3d). That study introduced four distinct RF models, each incorporating standard predictors such as pH, alkalinity, and DOC. These models have unique inclusions of fluorescence excitation-emission matrix (FEEM) data at different resolutions. These models are as FEEM-Free, FEEM-LowRes, FEEM-HighRes, and FEEM-FullRes, each varying in the resolution of FEEM data used as the unique predictors. A comparison of these four models shows that integrating FEEM data results in more accurate predictions of ozone exposures. The high-resolution FEEM data yield better predictions for microplastic abatement (R^2 = 0.904; RMSE = 6.6%). However, the improvement in prediction accuracy when using FEEM data is less substantial for predicting microplastic abatement than for predicting the oxidant exposures.

Machine-learning-quantitative-structure-property-relationship (ML-QSPR) methods can calculate the rate constant (k_{O3}) of the reactions between ozone and micropollutants (Gupta and Basant, 2016; Huang et al., 2020; Shi et al., 2022; Sudhakaran and Amy, 2013). Generally, nonlinear models outperform their linear counterparts. For instance, an MLR method (Sudhakaran and Amy, 2013) and an SVM method (Huang et al., 2020) have performance levels with R^2 values of greater than 0.75 and 0.78, respectively. Conversely, a DTB model has a notably higher R^2 value of greater

than 0.97 (Gupta and Basant, 2016). A recent study thoroughly compared several models including MLR, SVM, DT, RF, and deep neural network (DNN) for predicting logk_{O3} (Shi et al., 2022). Of these, the RF model has the highest effectiveness with a peak R^2 value of 0.91. The RF model has two primary benefits: robustness and a lower tendency toward overfitting. On the other hand, the DT model is a complex increase in its structure, subsequently raising the overfitting risk. In addition, DNN, promising at recognizing nonlinear features, underperforms in predicting logk_{O3}. A similar situation occurs when ML models predict the elimination of recalcitrant trace organic compounds by ozonation for municipal wastewater reuse (Park et al., 2015). Specifically, ANN was susceptible to overfitting issues. Incorporating PCA into ANN creates a PC-ANN workflow, which addressed these issues. PCA transforms the input variables to linearly independent variables, thereby resolving the issue of collinearity among explanatory variables. The PC-ANN method (R^2 = 0.934) surpasses the standalone ANN (R^2 = 0.914) in terms of predictive power.

5.2. To monitor and control heavy metals in drinking water

5.2.1. To monitor heavy metals in drinking water

The quality of drinking water in engineered water systems can deteriorate because of the detachment of heavy metals from the water pipes into the bulk water (i.e., leaching) (Mays, 2000; Proctor et al., 2020). The detrimental effects of heavy metal toxicity on all living beings have been well-documented (Umesh C. Gupta, 2011; Valko et al., 2005). Therefore, this section focuses on the applications of ML in assessing heavy metals in drinking water.

Several studies used ML to estimate the concentration of heavy metals in drinking water. For instance, a continuous on-site, *in situ* system can estimate lead (Pb) concentrations in municipal

water (Oh et al., 2021). The system leverages the SVR algorithm, supplanting traditional mathematical models confined to analyzing stationary ions in a solid substrate. By using the radio-frequency reflection coefficient of the raw trace data, the system predicts Pb ion concentration with a resolution of 1 μg/L and an RMS prediction error of 0.71 μg/L in the presence of interfering metal ions such as Cu^{2+}, Fe^{3+}, and Zn^{2+}. Other than estimating heavy metal concentrations in individual samples, ML is promising in broader analytical applications. For instance, ML is useful in the spatial interpolation of environmental variables, significantly enhancing its performance (Li et al., 2011). This approach has been substantiated in developing spatial interpolation maps depicting the concentrations of heavy metals such as Fe, Mn, Ni, Pb, and Zn in groundwater (i.e., a water source of drinking water) (De Jesus et al., 2021). That study combined ML and geostatistical interpolation (MLGI) to leverage an ANN-based algorithm to augment the efficacy and robustness of the spatial interpolation mapping. Furthermore, the MLGI approach comprehensively assesses the carcinogenic risks of heavy metals through *in situ* measurements. The approach produces detailed spatial maps delineating metal concentrations and estimates health quotient indices (HQI) to offer a more refined risk assessment (Senoro et al., 2022). While the integration of ML algorithms elevates the efficacy and robustness of spatial interpolation, traditional interpolation techniques still have a significant role in this domain. For instance, a spherical semi-variogram model relying on the classic Kriging interpolation technique can monitor the temporospatial distribution of residual aluminum (Al) in a DWDS (Figure 4a), thereby highlighting the enduring relevance and applicability of traditional interpolation methods (Tian et al., 2020).

5.2.2. To control heavy metals in drinking water

Adsorption is proficient in mitigating heavy metal contamination in drinking water (Joseph et al., 2019; Wołowiec et al., 2019). The removal of heavy metals by adsorption has highly stochastic, non-linear, and non-stationary dynamics coupled with redundancy (Bhagat et al., 2020). Many ML techniques can enhance the precision and efficacy of predicting heavy metal adsorption dynamics. Common predictors for those ML techniques are adsorbent dosage, operating temperature, contact time, and pH, whereas the output is the removal efficiency of heavy metals. Other variables can also be incorporated into the predictive models such as the initial concentration of heavy metals, the specific surface area of metal-organic frameworks (MOFs), and the presence of anions (Abdi and Mazloom, 2022). That study predicted the adsorptive removal of arsenate [As(V)] using four ML techniques: light gradient-boosting machine (LightGBM), XGB, gradient-boosted decision trees (GBDT), and RF. In addition, each predictive model can predict the removal efficiencies of multiple heavy metals. For instance, ANN can predict the removal efficiencies of Al, Cd, Co, Cu, Fe, and Pb (Hamidian et al., 2019). The LightGBM model has the most precise predictions for As(V) adsorption (Abdi and Mazloom, 2022). However, deep-layer-structured ANN has a higher accuracy particularly when employing radial basis functions as activation functions in the hidden layer (Hamidian et al., 2019).

5.3. To monitor nitrification in engineered water systems

Nitrification is a serious issue in chloraminated engineered water systems (Sathasivan et al., 2008; Shi et al., 2020). During nitrification, ammonia-oxidizing microbes oxidize free ammonia to nitrite, and nitrite-oxidizing bacteria further oxidize nitrite to nitrate. Free ammonia thus initiates nitrification. Free ammonia appears in chloraminated engineered water systems because of excessive free ammonia dosing at water utilities, chloramine decay within water pipes, and the

552 reactions between chloramines and reducing agents. Nitrification deteriorates water quality by
553 destroying chloramine residuals, releasing free ammonia, promoting microbial (re)growth, and
554 producing nitrite and nitrate. Therefore, monitoring nitrification in engineered water systems is
555 critical to ensuring drinking water quality and protecting public health.
556
557 ML is useful in investigating drinking water nitrification. A supervised ML technique, the NB
558 classifier, relies on biomass and microbiome datasets to detect nitrification (Gomez-Alvarez and
559 Revetta, 2020). After being trained with microbiome indicators, the model has a binary
560 classification accuracy of up to 85% with an AUC of 0.825 when distinguishing between
561 nitrification and stable events. An SVR (Figure 4b) is trained with nitrate/nitrite spectra at various
562 wavelengths (predictor variables) along with corresponding nitrate/nitrite concentrations (response
563 variables) (Hossain et al., 2021). The SVR model negates the need for any chemical supplements,
564 is easy to use, and can reach a high level of precision of up to ±0.01 mg N/L.
565
566 **6. Machine learning to monitor and ensure microbiological drinking water quality**
567 **6.1. To monitor and mitigate opportunistic pathogens in municipal water with a focus on**
568 ***Legionella***
569 ML is useful in monitoring and ensuring microbial water quality in engineered water systems
570 (Table 3). The microbial community within engineered water systems is diverse, encompassing
571 general heterotrophic bacteria, protozoans (such as amoebae, ciliates, and slime molds), and OPs.
572 Dominant water-related OPs are *Legionella* (especially *L. pneumophila*), *Mycobacterium* (e.g.,
573 nontuberculosis mycobacteria or NTM and *M. avium* complex or MAC), *Pseudomonas aeruginosa*,
574 *Vermamoeba vermiformis*, *Naegleria fowleri*, and *Acanthamoeba* (Donohue et al., 2019; Isaac and

Sherchan, 2020; Lytle et al., 2021). In municipal water, OPs are the most significant parameter for microbial drinking water quality because of their frequent occurrence, high concentrations, high resistance to disinfectant residuals, and association with drinking-water-related diseases such as Legionnaires' disease (Zhang and Lu, 2021a). Therefore, closely monitoring OPs in engineered water systems is critical to assessing drinking water quality and protecting public health. *Legionella* is the most important OP in municipal water. In addition, compared with conventional microbial drinking water quality indicators such as fecal coliforms, *Legionella* is a better candidate for indicating microbial drinking water quality (Zhang and Lu, 2021a). Therefore, in this section, we focus on the applications of ML in monitoring and controlling *Legionella* in municipal water.

Studies using ML to assess the risks of OPs remain limited. An early work mitigated the proliferation of *Legionella* in premise plumbing by controlling environmental variables (Sincak et al., 2014). Using water flow and temperature as the inputs, that study presented a NN-based simulator designed using an approximate reasoning architecture (NARA) neuro-fuzzy system to predict and simulate water tank temperature profiles. The simulator emulates conditions that inhibit the spread of *Legionella* in water networks. The NARA-based simulator achieves a high fidelity in mimicking water tank temperatures with an accuracy exceeding 97%. Another study integrated both unsupervised and supervised ML techniques to correlate the spread of *Legionella* with environmental variables in retirement homes, health-related facilities, tourism-related buildings, and swimming-pool environments in Italy (Brunello et al., 2022). That study used an unsupervised learning algorithm to identify the spatiotemporal distribution of atypical *Legionella* through an ordinal regression model. The results indicate how the distribution is correlated with the healthcare facilities. Hospitals had the highest contamination cluster locations, indicating a

strong correlation between the propagation of *Legionella* and both the nature of the facilities and broader geographical characteristics. That work also used supervised ML to assess the serotypes of *Legionella* and to anticipate the corresponding contamination levels. For serogroup assessment, XGBoost, LR, and SVM Classifier were used and compared. XGBoost shows superior performance with an overall classification accuracy of 0.71. The contribution of each predictor to the final classification was evaluated by the Shapley values, which quantify the contribution of each variable to the outputs of a ML model by comparing the effect of the outputs relative to the average across all inputs. The geographical location of a sample is the most important parameter but is useful only when combined with other predictors. For contamination level prediction, all three models demonstrate low performance with the highest accuracy of 0.57 from XGBoost.

6.2. To detect *Cryptosporidium* and *Giardia* in drinking water

Cryptosporidium and *Giardia* are protozoan parasites in municipal water with substantial public health risks by causing cryptosporidiosis and giardiasis, respectively (CDC, 2021a; b). These pathogens are notably resilient to disinfectants such as chlorine, challenging water treatment (Adeyemo et al., 2019). Therefore, detecting and controlling *Cryptosporidium* and *Giardia* is critical to maintaining drinking water quality. In this section, we discuss the performance of these ML models in detecting *Cryptosporidium* and *Giardia*.

ML to detect and analyze *Cryptosporidium* and *Giardia* has high robustness and precision. For instance, deep-learning-based image classification models such as ParasNet (Xu et al., 2020) and MCellNet (Luo et al., 2021) are accurate in detecting these two parasites in drinking water. They show the power of ML in classifying parasites from the cell-level scattering images. In addition, a

linear ML model can predict the contamination of these two parasites (Figure 5a) (Ligda et al., 2020), offering a valuable tool to control waterborne diseases.

ParasNet uses an eight-layer CNN to determine whether particles in cell-level scattering images from drinking water are *Cryptosporidium* and *Giardia*. The model has superior performance compared with a traditional handcraft SVM regarding both detection accuracy and processing speed. For instance, ParasNet can reach above 95.6% detection accuracy with analysis speeds of up to 100 frame-per-second (fps). MCellNet, another image classification pipeline, uses a DNN optimized from MobileNetV2 to recognize objects (Sandler et al., 2018). MCellNet includes a convolutional layer, six inverted residual blocks (IRBs), a flattened layer, and a fully connected layer. MCellNet can process images from flow cytometry to classify *Cryptosporidium* and *Giardia*. Compared with ParasNet, MCellNet achieves a higher detection accuracy of above 99.6% with a 346-fps analysis speed. The superior accuracy and expedited analysis of MCellNet are due to the cascading six IRBs in the model.

An alternative statistical model uses linear discriminant function analysis (LDFA) to predict the contamination of *Cryptosporidium* and *Giardia* in drinking water (Ligda et al., 2020). That model uses microbiological, physicochemical, and meteorological parameters as predictors to classify the contamination of *Cryptosporidium* and *Giardia* into four categories: none, low, moderate, and high (oo)cysts concentrations. LDFA has accuracies of 75% and 69% in predicting the contamination of *Cryptosporidium* and *Giardia*, respectively.

6.3. To assess biofilm development in engineered water systems

644 Research on the applications of ML to study biofilm development in engineered water systems is
645 scarce. ML models in this research area use physical (such as hydraulic) factors to predict the
646 dynamics of biofilm development, where the heterotrophic plate count (HPC) is the common
647 output. Established ML algorithms are preferred models to study the dynamics of biofilm
648 development such as NB, RT, and RF (Ramos-Martínez et al., 2014; 2016). These algorithms have
649 high prediction accuracy and provide a deeper understanding of the impact of physical factors on
650 biofilm development in engineered water systems. For instance, a Bagging naïve Bayesian tree
651 (B-NBT) model proposes optimal flow velocities for different types of pipes to mitigate biofilm
652 development (Ramos-Martínez et al., 2014). To control biofilm accumulation in DWDSs, water
653 utilities may avoid cement pipes, implement medium- or high-flow velocities in metal pipes and
654 sustain water ages above 0.035 in plastic pipes. The 'water age' is a synthetic index derived from
655 the normalized hydraulic retention time (HRT) and the distance from the disinfection source.
656
657 Recent studies have enabled more detailed, single-cell level analyses and predictions of biofilm
658 development in engineered water systems (Berne et al., 2018). In addition, researchers have
659 expanded algorithms to incorporate deep learning methods (Jelli et al., 2023; Weigert et al., 2020).
660 These innovative approaches can enhance the understanding of biofilm dynamics in various
661 settings including engineered water systems. A study developed a refined deep learning technique
662 for the single-cell segmentation of 3D bacterial biofilms and made two breakthroughs (Figure 5b)
663 (Jelli et al., 2023). First, it expedited the annotation process of bacterial cells within 3D imagery,
664 enhancing the efficiency of data analysis. Second, it optimized the application of StarDist (Weigert
665 et al., 2020), a cutting-edge CNN-based algorithm. The refined deep learning technique achieves
666 unprecedented accuracy in biofilm segmentation, surpassing other algorithms under scrutiny.

Moreover, the technique tracks cell lineages and enables precise measurements of bacterial growth, offering unprecedented insights into biofilm structures and development. While the application of ML to assess biofilm development in engineered water systems remains to be explored, such advancements herald a promising avenue for future research of biofilm development in drinking water.

6.4. To predict antimicrobial resistance risks and track the sources of antimicrobial resistance genes in drinking water

6.4.1. To predict antimicrobial resistance risks in drinking water

Antimicrobials have been extensively used since the 1920s in the medical industry, animal husbandry, and other fields (Chang et al., 2015; Hutchings et al., 2019; Prescott, 2017; WHO, 2021). More than 80% of antimicrobials used in humans and animals are not metabolized and excreted from the bodies. Antimicrobials cause the issue of antimicrobial resistant bacteria (ARB) and ARGs. ARB and ARGs can enter drinking water and cause antimicrobial resistance (AMR), significantly threatening public health (Roca et al., 2015; Walesch et al., 2023).

Detecting ARB and ARGs with conventional assays is challenging. Therefore, estimating the risks of AMR in drinking water is time-consuming and costly. A novel ML approach predicting the risks of AMR can overcome this issue (Wu et al., 2022). That approach maps the relative risk scores of AMR from continuous values into binary (0 and 1) labels using a predefined threshold(Goh et al., 2022). The study formulated the relative risks of AMR as a binary classification task and investigated multiple ML models (LR, DT, and RF) (Wu et al., 2022). Multiple parameters are included in the models such as temperature, pH, oxidation-reduction potential, electrical

conductance, resistivity, total dissolved solids, salinity, pressure, DO, turbidity, and 24-h accumulated rainfall. The results show that the RF model outperforms the other models in accuracy, precision, recall, and AUC metrics.

6.4.2. To track the sources of antimicrobial resistance genes in drinking water

The challenges in detecting ARB and ARGs in drinking water with conventional methods highlight the need for innovative solutions such as ML (Wu et al., 2022). ML effectively classifies the risks of AMR and showcases the potential of AI technology in this area. Built on this technological advancement, recent studies have employed another ML tool, SourceTracker (Knights et al., 2011), which is based on Bayesian classification algorithm, to identify the potential sources of ARGs (Chen et al., 2019; Wang et al., 2023). SourceTracker identified complex sources of ARGs and assessed their contributions to ARG pollution in a peri-urban river (Chen et al., 2019). Results show that the discharge from sewage treatment plants was the largest contributor of ARGs (81.6% to 92.1%) in the river sediments. In another work, SourceTracker identified the presence of ARGs in household drinking water and also traced their origins back to anthropogenic sources, highlighting the significant impact of human activities on drinking water quality (Figure 5c) (Wang et al., 2023). The data generated by SourceTracker have a strong Pearson correlation (r = 0.98) with the corresponding expected proportion by artificial source inputs. Source tracking analysis from that study indicates that a significant proportion of ARGs (37.1%) was from anthropogenic sources, especially wastewater treatment plants.

6.5. To study microbial communities in drinking water

712 Both supervised and unsupervised ML algorithms are useful in metagenomic studies of microbial
713 communities in drinking water (Mahajna et al., 2022). When using the NB theorem for water
714 source tracking, one could apply either maximum posterior probability (Ritter et al., 2003)
715 (evaluation metrics: $RMSE_c$) or direct averaging posterior probability (Greenberg et al., 2010)
716 (evaluation metrics: $RMSE_p$) to estimate the distribution of microbes in water sources. Direct
717 averaging of the source posterior probability yields more precise source distribution estimates with
718 $RMSE_c$ being significantly lower than $RMSE_p$. The more precise source distribution estimates are
719 because direct estimation bypasses the information loss that typically happens when frequencies
720 are first classified and then averaged (Greenberg et al., 2010). SourceTracker is a ML tool
721 estimating the proportion of contaminants (Knights et al., 2011). That tool employs the Gibbs
722 sampling technique within a Bayesian framework and is more efficient than both the
723 aforementioned NB and RF classification-based source tracking methods (Smith et al., 2010). The
724 superior performance of SourceTracker is because it can handle ambiguity in the source and sink
725 distributions and can model a sink sample as a blend of various sources. SourceTracker can track
726 the origin of bacteria in tap water (Liu et al., 2018). For instance, a study created six ML models
727 to predict microbial contamination in a watershed using data on land cover, weather, and
728 hydrologic variables (Wu et al., 2020). In that case, SourceTracker generates ground-truth data for
729 training purposes. In addition, XGBoost outperforms the other five models (KNN, NB, SVM, NN,
730 and RF) in terms of accuracy and AUC when tracking the primary sources of microbial
731 contamination.
732
733 Unsupervised ML can unveil hidden features, trends, or patterns in bacterial communities in
734 engineered water systems. For instance, alpha and beta diversity analyses can display the spatial

dynamics and temporal trends of bacterial communities in DWDSs (Pinto et al., 2014). UniFrac as another unsupervised ML tool uses a principal coordinates analysis (PCoA) coupled beta diversity measure to analyze the differences among microbial communities in environmental samples (Lozupone et al., 2011). UniFrac can effectively analyze the microbiome in drinking water (Bruno et al., 2018; Li et al., 2017; Ling et al., 2018).

7. Machine learning to detect accidental drinking water contamination

7.1. To detect anomalies and provide early warning of contamination in drinking water

The increasing use of ML in safeguarding drinking water quality has led to the development of innovative approaches to detect accidental drinking water contamination (Table 4) (Zhong et al., 2021). This section explores the advancements of these ML approaches, underscoring the versatility and potential of ML in ensuring safe drinking water supply by detecting accidental drinking water contamination.

Three studies developed various ML approaches for anomaly detection on the basis of the drinking water quality datasets from GECCO Industrial Challenges (GECCO IC) (Fehst et al., 2018; Muharemi et al., 2019; Qian et al., 2020). These studies pinpointed shifts or spotted anomalies in water quality over time. Various parameters such as pH, redox potential, electric conductivity, turbidity, and chlorine dioxide concentration are the predictors, whereas events in Boolean form are the outputs. A study developed SVM, DNN, long short-term memory (LSTM), recurrent NN (RNN), LR, simple NN, and linear discriminant analysis (LDA) to detect water quality anomaly in the dataset from 2017 GECCO IC (Muharemi et al., 2019). SVM shows the highest performance with an F1-score of 0.99 in cross-validation. Nevertheless, all the models exhibit poor performance

with the unseen test dataset with a maximum F1-score of 0.36. In the other two studies focusing on the dataset from GECCO IC 2018, LSTM demonstrates superior results, scoring a higher F1-score than traditional models such as LR and SVM with F1-scores of 0.80 and 0.78, respectively (Fehst et al., 2018; Qian et al., 2020).

The existing research, including the aforementioned key studies using the GECCO IC datasets, has made significant strides in understanding anomaly detection in drinking water through ML. A notable trend in recent research is the focus on real-time or online applications, reflecting a crucial evolution toward practical, real-world implementations. Specifically, a study implemented an LSTM-based approach to detect anomalies in water quality focusing on turbidity and conductivity (Rodriguez-Perez et al., 2020) (Figure 6a). That study highlights the efficacy of semi-supervised classification, which retains only normal values, in identifying abrupt changes and minor spikes in water quality. By contrast, supervised classification, which considers both normal and anomalous data, is more adept at identifying long-term anomalies linked to gradual changes. Notably, the LSTM-based approach surpasses regression-based ARIMA in detecting these long-term anomalies. Another study introduced an innovative stacking ensemble model designed for contamination detection (Li et al., 2022). The model uses various water quality parameters such as total chlorine, pH, electrical conductivity, temperature, TOC, and turbidity. That model integrates multiple ML base predictors with a meta-predictor, trained through cross-validation. That approach enhances the ability of the model to discern distinct features across water quality parameters. The ensemble comprises base predictors such as ANN, SVM with a linear kernel, linear regressor, extra trees, uniform weighted KNN, and a RF meta-predictor. The ensemble demonstrates superior performance in detecting contamination compared with an ANN benchmark method, achieving

781 higher accuracy, lower false positive rates, and improved F1-scores. However, these models focus
782 on single-site, one-dimensional time series data, neglecting the spatial relationships inherent in
783 multi-site sensor data. This limitation could increase false alarm rates, particularly under
784 conditions of high hydraulic variability. To address this issue, a follow-up study proposed a novel
785 unsupervised, generative-adversarial-networks-based (GANs-based) multivariate multi-site
786 contamination event detection method (Figure 6b) (Li et al., 2023b). That method effectively
787 captures spatiotemporal patterns by transforming water quality data from single and multiple sites
788 into superimposed images. The GANs-based model, comprising a generator and a discriminator,
789 evaluates the degree of abnormality at each time step by generating anomaly scores. The generator
790 is trained to map historical image data to expected current images, while the discriminator
791 differentiates between generated and actual normal images. That method is benchmarked against
792 a multivariate unsupervised method using a minimum-volume-ellipsoid (MVE)-based event
793 detection model (Oliker and Ostfeld, 2014). That method demonstrates superior performance in
794 all contamination scenarios, including enhanced detection rates and reduced false alarms,
795 particularly for sensor groups positioned at varying distances from the contamination source.
796 Another unique ML approach can rapidly signal potential contamination risks (Asheri Arnon et al.,
797 2019). That approach uses an algorithm for the early detection of drinking water contamination
798 against an unpredictable stochastic background. By extracting key features from the
799 spectrophotometric characteristics of water, the algorithm can effectively identify contamination
800 using a unique affinity measure (Asheri-Arnon et al., 2018). The measure compares the absorbance
801 spectra of different water sources, thereby amplifying the feature dissimilarity between portable
802 and contaminated water, followed by processing via SVM and post-processing. That chain of data
803 processing generates a reliable early warning for contamination events with low false positives

and high true alarm accuracy. The pre-processing stage (the affinity measure and amplification) is essential to achieving high accuracy but may be unnecessary for obtaining minimal false positives.

8. Machine learning model distribution in safe drinking water supply

We have prepared a macroscopic visual illustration to display the distribution of ML models across research topics and sub-topics in safe drinking water supply (Figure 7). To facilitate a clear but concise visual representation, we group certain ML models under broader principal categories on the basis of their foundational architecture. For instance, models such as GA-ANN, Multi-layered-ANN, and DNN share foundational characteristics inherent with ANN. Consequently, to elucidate the overarching trends in model preferences across studies, we categorize these models as "NN-based." This approach discerns the broader trends and preferences in ML model selection and also highlights the potential commonalities across research endeavors.

NN-based and regression-based models are the top two models frequently implemented. NN-based models have significant applications in accessing and controlling DBPs and managing the production and demand of drinking water. The prominent role of NN-based models in these two fields is not coincidental but rather from the synergy between the inherent characteristics of these fields and the strengths of NNs. Water management and DBP assessment often involve multifaceted, nonlinear, and high-dimensional data sets that demand a robust modeling approach (Aliashrafi et al., 2021; Ates et al., 2022; Ghobadi and Kang, 2023). Given their capability to model complex non-linear relationships and handle various intricate data, NNs are an optimal solution in these contexts. For instance, the unpredictability and variability in water demand patterns or the multifarious factors influencing DBP formation both require a model that can discern patterns from

large, intricate datasets (Ahmadpour et al., 2023; Avni et al., 2015). Furthermore, the flexibility of NNs in accommodating changing inputs makes them promising in assessing the dynamic nature of drinking water systems. The wide applications of NNs in safe drinking water supply are due to this harmonious fit between the challenges posed by these fields and the advantages of NNs.

By contrast, while regression-based models are widely applied in drinking water research, they have suboptimal performances in certain contexts (Almheiri et al., 2021; Deng et al., 2021; Hong et al., 2020; Legube et al., 2004; Lin et al., 2020; Rodriguez-Perez et al., 2020). This is not to undermine the value of regression models. However, their linear or predefined non-linear structures may limit their effectiveness, especially when compared with the adaptive and intricate abilities of NNs.

The superior performance of the NN-based models is widely acknowledged (Goodfellow et al., 2016). These general strengths become pertinent when the NN-based models are applied to the complexities of water research. First, unlike regression models which are limited by their linear or defined non-linear structures, NNs capture intricate, non-linear associations. Second, the mutable architecture of NNs allows them to modify their framework during training, optimizing alignment with the inherent data distribution. Lastly, given abundant data, NNs excel in discerning subtle data patterns because of their proficiency in processing high-dimensional input attributes, whereas regression models may have tendencies of underfitting. This proficiency of NNs is further enhanced by the use of techniques such as grid search for hyperparameter optimization, particularly crucial in fine-tuning the performance of NNs because of their complex architectures and the numerous parameters required (Daniel et al., 2023; Rodriguez-Perez et al., 2020).

CNN-based models represent a specialized subclass of NN-based models adept at discerning patterns in images or other forms of multi-dimensional data sets (LeCun et al., 1989; Lecun et al., 1998). Therefore, we list the CNN-based models out of the broader NN category (Figure 7). The practical implications of CNN-based models are evident in water research: They can interpret 2D fluorescence spectra and predict the formation of DBPs during drinking water disinfection (Peleato, 2022), classify microbes using cell-level scattering images from drinking water (Luo et al., 2021; Xu et al., 2020), and identify cells in 3D drinking water biofilm images (Jelli et al., 2023).

Other ensemble approaches are also widely applied in safe drinking water supply such as RF (Breiman, 2001), XGB (Chen and Guestrin, 2016), boosted decision trees (BDT) (Friedman, 2001), and stacking model (Wolpert, 1992). The core strength of these ensemble techniques is their ability to amalgamate predictions from several models, aiming to boost accuracy and diminish overfitting (Hastie et al., 2009). In water research where data can be noisy, varied, and sometimes sparse, such strategies are invaluable. Several comparative studies have delved into the performance nuances of different ensemble models. A recurring observation in these investigations is that the slight edge XGB outperforms RF (Abdi and Mazloom, 2022; Park et al., 2020; Wu et al., 2020). Furthermore, BRT outperforms RF (Bagriacik et al., 2018). Interestingly, while XGB has consistent prowess, LightGBM, another gradient boosting framework, outperforms XGB (Abdi and Mazloom, 2022). Therefore, as gradient boosting algorithms continue to evolve, newer iterations such as LightGBM might offer even more refined performance. However, while ensemble methods offer certain advantages, their efficacy is not universally dominant across scenarios. The best model is often contingent upon the nature of the problem, the characteristics of the data, and the specific

objectives of the study. Ensemble models, with their ability to amalgamate insights from multiple "weak learners," might excel in scenarios where data are diverse, noisy, or sparse (Fasel et al., 2022; Pang et al., 2018; Sluban and Lavrač, 2015). By contrast, for problems where the data structures are deeply hierarchical or when data patterns are straightforward, NN-based models or regression models are more suitable. The crucial factor is to match the ability of the models with the specific demands and characteristics of the data sets.

We include (S)ARIMA, Kriging interpolation, SaTScan, LDFA, Alpha and Beta diversity analyses, UniFrac, and MVE in statistical models (Figure 7). These models are more deterministic and often rooted in foundational principles and known relationships. For instance, SARIMA and Kriging interpolation can capture temporal and spatial patterns, respectively (Guo et al., 2018; Tian et al., 2020). Alpha and Beta diversity analyses and UniFrac quantify microbial community diversity and compositional differences (Bruno et al., 2018; Li et al., 2017; Ling et al., 2018; Lozupone et al., 2011; Pinto et al., 2014). These models typically operate under specific assumptions about the underlying data distribution or spatiotemporal relationships. By contrast, ML, especially deep learning, is more adaptive, learning patterns directly from the data without stringent assumptions (Khattak et al., 2022; Savadatti et al., 2022; Singh et al., 2023).

9. Challenges and outlooks for the applications of machine learning in safe drinking water supply

While ML has made significant advancements in drinking water research, several areas remain untapped, offering significant potential for exploration and improvement. Crucial topics, such as biofilm development, the assessment of AMR risks, and the evaluation of pathogen-related dangers

in engineered water systems, are not fully represented in current research. The untapped potential in these fields is immense, and the need to bridge the interdisciplinary divide is critical.

One significant barrier is the disconnect between water treatment experts and AI specialists. Water scientists and engineers may not be conversant with the nuances of AI, while AI technologists might lack knowledge of water treatment and supply. This knowledge gap impedes the effective deployment of ML in enhancing the safety of drinking water supply. Addressing this dichotomy is beneficial and essential, necessitating educational and collaborative efforts to build a shared understanding and to develop interdisciplinary skill sets.

Further complicating the matter is the absence of standardized tool kits tailored to the drinking water sector. Such standardization is vital for enabling consistent application across various research and implementation efforts. Uniformity in tools and approaches would not only streamline the processes but also bolster collaborative work, which is often fragmented across regions and specializations.

Advancements in ML tools must cater to the unique challenges presented by safe drinking water supply. Water quality in engineered water systems is affected by numerous factors that vary spatiotemporally, requiring ML solutions that can adapt to and learn from these dynamic conditions. Thus, future studies should customize existing ML frameworks or innovate new ones that can grapple with the complexities inherent to the water supply networks.

Looking to the horizon, the broader vision involves leveraging ML to address the global drinking water crisis. Issues such as water scarcity, the presence of emerging contaminants, and the formation of DBPs present a global challenge. ML tools have been predominantly developed with local or regional contexts, yet the drinking water crisis demands a global perspective. The ambition to harness ML for these universal challenges is not only visionary but also an imperative stride toward ensuring water security worldwide.

In pursuit of these goals, the integration of advanced ML models becomes a cornerstone in tackling the multifaceted issues tied to drinking water safety. Future endeavors should also prioritize the promotion of open-access data sharing within and beyond the drinking water research community (Zhong et al., 2021). The endeavors will enhance collaboration, drive transparency, and support the reproducibility of scientific findings, which are the bedrock of robust research. Furthermore, establishing a comprehensive comparative framework to evaluate different ML models will be instrumental in identifying the optimal solutions for the challenges in drinking water research. By embracing these strategies, we can aspire to not just bridge existing knowledge gaps but also significantly elevate the role of ML in securing safe and more sustainable water supply.

9. Conclusions

Assessing and ensuring safe drinking water supply is a global challenge with conventional approaches. ML as a novel methodology is promising in monitoring and protecting drinking water quality, especially in municipal engineered water systems. This review for the first time comprehensively summarizes the applications of ML in assessing and ensuring safe drinking water supply with a focus on water quality in engineered water systems. We compile the applications of

ML from the physical, chemical, and microbiological perspectives. From the physical perspective, ML is useful in managing drinking water production and demand and monitoring drinking water pipeline failures. From the chemical perspective, ML is promising in assessing and controlling DBPs, monitoring and mitigating heavy metals, and tracking nitrification in drinking water. From the microbiological perspective, ML can monitor and mitigate OPs, detect *Cryptosporidium* and *Giardia*, assess biofilm development, predict AMR risks, and study microbial communities in municipal water, especially in engineered water systems. In addition, ML is a useful tool in detecting accidental drinking water contamination.

Acknowledgment

This work has been supported by the Faculty Research Fund from Arkansas State University (Jonesboro, AR, U.S.A.).

Declaration of interest

No potential or actual conflict of interest exists in this work.

Uncategorized References

Abdi, J. and Mazloom, G. 2022. Machine learning approaches for predicting arsenic adsorption from water using porous metal-organic frameworks. Sci Rep 12(1), 16458.

Adeyemo, F.E., Singh, G., Reddy, P., Bux, F. and Stenstrom, T.A. 2019. Efficiency of chlorine and UV in the inactivation of Cryptosporidium and Giardia in wastewater. PLoS One 14(5), e0216040.

Ahmadpour, E., Delpla, I., Debia, M., Simard, S., Proulx, F., Sérodes, J.-B., Valois, I., Tardif, R., Haddad, S. and Rodriguez, M. 2023. Full-scale multisampling and empirical modeling of

DBPs in water and air of indoor pools. Environmental Monitoring and Assessment 195(9), 1128.
Ahmed, A.N., Othman, F.B., Afan, H.A., Ibrahim, R.K., Fai, C.M., Hossain, M.S., Ehteram, M. and Elshafie, A. 2019. Machine learning methods for better water quality prediction. Journal of Hydrology 578, 124084.
Aliashrafi, A., Zhang, Y., Groenewegen, H. and Peleato, N.M. 2021. A review of data-driven modelling in drinking water treatment. Reviews in Environmental Science and Bio/Technology 20(4), 985-1009.
Allesoe, R.L., Lundgaard, A.T., Hernandez Medina, R., Aguayo-Orozco, A., Johansen, J., Nissen, J.N., Brorsson, C., Mazzoni, G., Niu, L., Biel, J.H., Leal Rodriguez, C., Brasas, V., Webel, H., Benros, M.E., Pedersen, A.G., Chmura, P.J., Jacobsen, U.P., Mari, A., Koivula, R., Mahajan, A., Vinuela, A., Tajes, J.F., Sharma, S., Haid, M., Hong, M.G., Musholt, P.B., De Masi, F., Vogt, J., Pedersen, H.K., Gudmundsdottir, V., Jones, A., Kennedy, G., Bell, J., Thomas, E.L., Frost, G., Thomsen, H., Hansen, E., Hansen, T.H., Vestergaard, H., Muilwijk, M., Blom, M.T., t Hart, L.M., Pattou, F., Raverdy, V., Brage, S., Kokkola, T., Heggie, A., McEvoy, D., Mourby, M., Kaye, J., Hattersley, A., McDonald, T., Ridderstrale, M., Walker, M., Forgie, I., Giordano, G.N., Pavo, I., Ruetten, H., Pedersen, O., Hansen, T., Dermitzakis, E., Franks, P.W., Schwenk, J.M., Adamski, J., McCarthy, M.I., Pearson, E., Banasik, K., Rasmussen, S., Brunak, S. and Consortium, I.D. 2023. Discovery of drug-omics associations in type 2 diabetes with generative deep-learning models. Nat Biotechnol 41(3), 399-408.
Almalioglu, Y., Turan, M., Trigoni, N. and Markham, A. 2022. Deep learning-based robust positioning for all-weather autonomous driving. Nat Mach Intell 4(9), 749-760.
Almheiri, Z., Meguid, M. and Zayed, T. 2021. Failure modeling of water distribution pipelines using meta-learning algorithms. Water Res 205, 117680.
Asheri Arnon, T., Ezra, S. and Fishbain, B. 2019. Water characterization and early contamination detection in highly varying stochastic background water, based on Machine Learning methodology for processing real-time UV-Spectrophotometry. Water Research 155, 333-342.
Asheri-Arnon, T., Ezra, S. and Fishbain, B. 2018. Contamination Detection of Water with Varying Routine Backgrounds by UV-Spectrophotometry. Journal of Water Resources Planning and Management 144(9), 04018056.
Ates, N., Civelekoglu, G. and Kaplan-Bekaroglu, S.S. (2022) Water and Wastewater Management: Global Problems and Measures. Bahadir, M. and Haarstrick, A. (eds), pp. 67-82, Springer International Publishing, Cham.
Avni, N., Fishbain, B. and Shamir, U. 2015. Water consumption patterns as a basis for water demand modeling. Water Resources Research 51(10), 8165-8181.
Bagriacik, A., Davidson, R.A., Hughes, M.W., Bradley, B.A. and Cubrinovski, M. 2018. Comparison of statistical and machine learning approaches to modeling earthquake damage to water pipelines. Soil Dynamics and Earthquake Engineering 112, 76-88.
Bakker, M., Vreeburg, J.H.G., Van De Roer, M. and Rietveld, L.C. 2014. Heuristic burst detection method using flow and pressure measurements. Journal of Hydroinformatics 16(5), 1194-1209.
Balogun, A.-L., Tella, A., Baloo, L. and Adebisi, N. 2021. A review of the inter-correlation of climate change, air pollution and urban sustainability using novel machine learning algorithms and spatial information science. Urban Climate 40, 100989.
Benedict, K.M., Reses, H., Vigar, M., Roth, D.M., Roberts, V.A., Mattioli, M., Cooley, L.A., Hilborn, E.D., Wade, T.J. and Fullerton, K.E. 2017. Surveillance for waterborne disease

outbreaks associated with drinking water-United States, 2013-2014. Morbidity and Mortality Weekly Report 66(44), 1216.
Benítez, J.S., Rodríguez, C.M. and Casas, A.F. 2021. Disinfection byproducts (DBPs) in drinking water supply systems: a systematic review. Physics and Chemistry of the Earth, Parts A/B/C 123, 102987.
Berglund, E., Vizanko, B., Kadinski, L. and Ostfeld, A. (2023) Coupling Machine Learning and Agent-Based Modeling to Characterize Contamination Sources in Water Distribution Systems for Changing Demand Regimes. World Environmental and Water Resources Congress 2023, pp. 881-890.
Berne, C., Ellison, C.K., Ducret, A. and Brun, Y.V. 2018. Bacterial adhesion at the single-cell level. Nature Reviews Microbiology 16(10), 616-627.
Bhagat, S.K., Tung, T.M. and Yaseen, Z.M. 2020. Development of artificial intelligence for modeling wastewater heavy metal removal: State of the art, application assessment and possible future research. Journal of Cleaner Production 250, 119473.
Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X. and Tian, Q. 2023. Accurate medium-range global weather forecasting with 3D neural networks. Nature 619(7970), 533-538.
Bond, T., Goslan, E.H., Parsons, S.A. and Jefferson, B. 2011. Treatment of disinfection by-product precursors. Environmental Technology 32(1), 1-25.
Bond, T. and Graham, N. 2017. Predicting chloroform production from organic precursors. Water Research 124, 167-176.
Bond, T., Templeton, M.R. and Graham, N. 2012. Precursors of nitrogenous disinfection by-products in drinking water—A critical review and analysis. Journal of Hazardous Materials 235-236, 1-16.
Bradley, A.P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7), 1145-1159.
Breiman, L. 2001. Random Forests. Machine Learning 45(1), 5-32.
Brunello, A., Civilini, M., De Martin, S., Felice, A., Franchi, M., Iacumin, L., Saccomanno, N. and Vitacolonna, N. 2022. Machine learning-assisted environmental surveillance of Legionella: A retrospective observational study in Friuli-Venezia Giulia region of Italy in the period 2002–2019. Informatics in Medicine Unlocked 28, 100803.
Bruno, A., Sandionigi, A., Bernasconi, M., Panio, A., Labra, M. and Casiraghi, M. 2018. Changes in the Drinking Water Microbiome: Effects of Water Treatments Along the Flow of Two Drinking Water Treatment Plants in a Urbanized Area, Milan (Italy). Front Microbiol 9, 2557.
Cabral, J.P. 2010. Water microbiology. Bacterial pathogens and water. International Journal of Environmental Research and Public Health 7(10), 3657-3703.
CDC 2021a. Parasites - Cryptosporidium (also known as "Crypto").
CDC 2021b. Parasites - Giardia.
Cha, D., Park, S., Kim, M.S., Kim, T., Hong, S.W., Cho, K.H. and Lee, C. 2021. Prediction of Oxidant Exposures and Micropollutant Abatement during Ozonation Using a Machine Learning Method. Environ Sci Technol 55(1), 709-718.
Chang, Q., Wang, W., Regev-Yochay, G., Lipsitch, M. and Hanage, W.P. 2015. Antibiotics in agriculture and the risk to human health: how worried should we be? Evol Appl 8(3), 240-247.
Chen, H., Bai, X., Li, Y., Jing, L., Chen, R. and Teng, Y. 2019. Source identification of antibiotic resistance genes in a peri-urban river using novel crAssphage marker genes and metagenomic signatures. Water Research 167, 115098.

Chen, T. and Guestrin, C. 2016 XGBoost: A Scalable Tree Boosting System, pp. 785–794, Association for Computing Machinery, San Francisco, California, USA.

Chowdhury, S., Mazumder, M.A.J., Al-Attas, O. and Husain, T. 2016. Heavy metals in drinking water: Occurrences, implications, and future needs in developing countries. Sci Total Environ 569-570, 476-488.

Cordero, J.A., He, K., Janya, K., Echigo, S. and Itoh, S. 2021. Predicting formation of haloacetic acids by chlorination of organic compounds using machine-learning-assisted quantitative structure-activity relationships. Journal of Hazardous Materials 408, 124466.

Craun, G.F., Brunkard, J.M., Yoder, J.S., Roberts, V.A., Carpenter, J., Wade, T., Calderon, R.L., Roberts, J.M., Beach, M.J. and Roy, S.L. 2010. Causes of outbreaks associated with drinking water in the United States from 1971 to 2006. Clinical Microbiology Reviews 23(3), 507-528.

Cutler, D. and Miller, G. 2005. The role of public health improvements in health advances: the twentieth-century United States. Demography 42, 1-22.

Daniel, I., Abhijith, G.R., Kadinski, L., Ostfeld, A. and Cominola, A. (2023) A Machine Learning-Based Surrogate Model for Coupled Hydraulic and Water Quality Simulation in Water Distribution Networks. World Environmental and Water Resources Congress 2023, pp. 817-830.

De Jesus, K.L.M., Senoro, D.B., Dela Cruz, J.C. and Chan, E.B. 2021. A Hybrid Neural Network-Particle Swarm Optimization Informed Spatial Interpolation Technique for Groundwater Quality Mapping in a Small Island Province of the Philippines. Toxics 9(11).

Delpla, I., Jung, A.-V., Baures, E., Clement, M. and Thomas, O. 2009. Impacts of climate change on surface water quality in relation to drinking water production. Environment International 35(8), 1225-1233.

Deng, Y., Zhou, X., Shen, J., Xiao, G., Hong, H., Lin, H., Wu, F. and Liao, B.-Q. 2021. New methods based on back propagation (BP) and radial basis function (RBF) artificial neural networks (ANNs) for predicting the occurrence of haloketones in tap water. Science of The Total Environment 772, 145534.

Dogo, E.M., Nwulu, N.I., Twala, B. and Aigbavboa, C. 2019. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water Journal 16(3), 235-248.

Donohue, M., King, D., Pfaller, S. and Mistry, J. 2019. The sporadic nature of *Legionella pneumophila*, *Legionella pneumophila* Sg1 and *Mycobacterium avium* occurrence within residences and office buildings across 36 states in the United States. Journal of applied microbiology 126(5), 1568-1579.

Erhan, D., Szegedy, C., Toshev, A. and Anguelov, D. 2014 Scalable Object Detection Using Deep Neural Networks, pp. 2155–2162, IEEE Computer Society.

Ewuzie, U., Bolade, O.P. and Egbedina, A.O. (2022) Current Trends and Advances in Computer-Aided Intelligent Environmental Data Engineering. Marques, G. and Ighalo, J.O. (eds), pp. 185-218, Elsevier Inc., San Diego, California, U.S.A.

Fan, X., Zhang, X., Yu, A., Speitel, M. and Yu, X. 2023. Assessment of the impacts of climat change on water supply system pipe failures. Scientific Reports 13(1), 7349.

Fasel, U., Kutz, J.N., Brunton, B.W. and Brunton, S.L. 2022. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 478(2260), 20210904.

Favere, J., Barbosa, R.G., Sleutels, T., Verstraete, W., De Gusseme, B. and Boon, N. 2021. Safeguarding the microbial water quality from source to tap. Npj Clean Water 4, 28.
Fehst, V., La, H.C., Nghiem, T.-D., Mayer, B.E., Englert, P. and Fiebig, K.-H. 2018 Automatic vs. manual feature engineering for anomaly detection of drinking-water quality, pp. 5–6, Association for Computing Machinery, Kyoto, Japan.
Feng, S., Sun, H., Yan, X., Zhu, H., Zou, Z., Shen, S. and Liu, H.X. 2023. Dense reinforcement learning for safety validation of autonomous vehicles. Nature 615(7953), 620-627.
Folkman, S. 2018. Water Main Break Rates in the USA and Canada: A Comprehensive Study, 2018. Mechanical and Aerospace Engineering Faculty Publications Paper 174.
Friedman, J.H. 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29(5), 1189-1232, 1144.
Ghiassi, M., Fa'al, F. and Abrishamchi, A. 2017. Large metropolitan water demand forecasting using DAN2, FTDNN, and KNN models: A case study of the city of Tehran, Iran. Urban Water Journal 14(6), 655-659.
Ghiassi, M., Saidane, H. and Zimbra, D.K. 2005. A dynamic artificial neural network model for forecasting time series events. International Journal of Forecasting 21(2), 341-362.
Ghobadi, F. and Kang, D. 2023. Application of Machine Learning in Water Resources Management: A Systematic Literature Review. Water 15(4), 620.
Gogoi, A., Mazumder, P., Tyagi, V.K., Chaminda, G.T., An, A.K. and Kumar, M. 2018. Occurrence and fate of emerging contaminants in water environment: a review. Groundwater for Sustainable Development 6, 169-180.
Goh, S.G., Jiang, P., Ng, C., Le, T.H., Haller, L., Chen, H., Charles, F.R., Chen, H., Liu, X., He, Y. and Gin, K.Y. 2022. A new modelling framework for assessing the relative burden of antimicrobial resistance in aquatic environments. J Hazard Mater 424(Pt C), 127621.
Gomez-Alvarez, V. and Revetta, R.P. 2020. Monitoring of Nitrification in Chloraminated Drinking Water Distribution Systems With Microbiome Bioindicators Using Supervised Machine Learning. Front Microbiol 11, 571009.
Gong, J., Guo, X., Yan, X. and Hu, C. 2023. Review of urban drinking water contamination source identification methods. Energies 16(2), 705.
Gonzalez, S., Lopez-Roldan, R. and Cortina, J.-L. 2013. Presence of metals in drinking water distribution networks due to pipe material leaching: a review. Toxicological & Environmental Chemistry 95(6), 870-889.
Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning, MIT Press.
Greenberg, J., Price, B. and Ware, A. 2010. Alternative estimate of source distribution in microbial source tracking using posterior probabilities. Water Res 44(8), 2629-2637.
Grey, D. and Sadoff, C.W. 2007. Sink or Swim? Water security for growth and development. Water Policy 9(6), 545-571.
Guo, G., Liu, S., Wu, Y., Li, J., Zhou, R. and Zhu, X. 2018. Short-Term Water Demand Forecast Based on Deep Learning Method. Journal of Water Resources Planning and Management 144(12), 04018076.
Gupta, S. and Basant, N. 2016. Modeling the reactivity of ozone and sulphate radicals towards organic chemicals in water using machine learning approaches. RSC Advances 6(110), 108448-108457.
Gupta, V.K. and Ali, I. (2013) Environmental Water. Gupta, V.K. and Ali, I. (eds), pp. 1-27, Elsevier.

Hamidian, A.H., Esfandeh, S., Zhang, Y. and Yang, M. 2019. Simulation and optimization of nanomaterials application for heavy metal removal from aqueous solutions. Inorganic and Nano-Metal Chemistry 49(7), 217-230.
Hastie, T., Tibshirani, R. and Friedman, J.H. (2009) The elements of statistical learning : data mining, inference, and prediction, Springer, New York.
He, L., Bai, L., Dionysiou, D.D., Wei, Z., Spinney, R., Chu, C., Lin, Z. and Xiao, R. 2021. Applications of computational chemistry, artificial intelligence, and machine learning in aquatic chemistry research. Chemical Engineering Journal 426, 131810.
Helte, E., Säve-Söderbergh, M., Larsson, S.C., Martling, A. and Åkesson, A. 2023. Disinfection by-products in drinking water and risk of colorectal cancer: a population-based cohort study. Journal of the National Cancer Institute, djad145.
Henrique Alves Ribeiro, V. and Reynoso-Meza, G. 2023. Multi-criteria Decision-Making Techniques for the Selection of Pareto-optimal Machine Learning Models in a Drinking-Water Quality Monitoring Problem. International Journal of Information Technology & Decision Making, 1-28.
Hong, H., Zhang, Z., Guo, A., Shen, L., Sun, H., Liang, Y., Wu, F. and Lin, H. 2020. Radial basis function artificial neural network (RBF ANN) as well as the hybrid method of RBF ANN and grey relational analysis able to well predict trihalomethanes levels in tap water. Journal of Hydrology 591, 125574.
Hossain, S., Cook, D., Chow, C.W.K. and Hewa, G.A. 2021. Development of an Optical Method to Monitor Nitrification in Drinking Water. Sensors (Basel) 21(22).
Hu, G., Mian, H.R., Mohammadiun, S., Rodriguez, M.J., Hewage, K. and Sadiq, R. 2023. Appraisal of machine learning techniques for predicting emerging disinfection byproducts in small water distribution networks. Journal of Hazardous Materials 446, 130633.
Huang, R., Ma, C., Ma, J., Huangfu, X. and He, Q. 2021. Machine learning in natural and engineered water systems. Water Res 205, 117666.
Huang, Y., Li, T., Zheng, S., Fan, L., Su, L., Zhao, Y., Xie, H.B. and Li, C. 2020. QSAR modeling for the ozonation of diverse organic compounds in water. Sci Total Environ 715, 136816.
Hutchings, M.I., Truman, A.W. and Wilkinson, B. 2019. Antibiotics: past, present and future. Current Opinion in Microbiology 51, 72-80.
Isaac, T.S. and Sherchan, S.P. 2020. Molecular detection of opportunistic premise plumbing pathogens in rural Louisiana's drinking water distribution system. Environmental Research 181, 108847.
James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013 An introduction to statistical learning with applications in R, Springer,, New York.
Jefri, U.H.N.M., Khan, A., Lim, Y.C., Lee, K.S., Liew, K.B., Kassab, Y.W., Choo, C.-Y., Al-Worafi, Y.M., Ming, L.C. and Kalusalingam, A. 2022. A systematic review on chlorine dioxide as a disinfectant. Journal of Medicine and Life 15(3), 313.
Jelli, E., Ohmura, T., Netter, N., Abt, M., Jiménez-Siebert, E., Neuhaus, K., Rode, D.K.H., Nadell, C.D. and Drescher, K. 2023. Single-cell segmentation in bacterial biofilms with an optimized deep learning method enables tracking of cell lineages and measurements of growth rates. Molecular Microbiology 119(6), 659-676.
Jia, X., O'Connor, D., Shi, Z. and Hou, D. 2021. VIRS based detection in combination with machine learning for mapping soil pollution. Environ Pollut 268(Pt A), 115845.
Jiang, S., Hu, J., Wood, K.L. and Luo, J. 2021. Data-Driven Design-By-Analogy: State-of-the-Art and Future Directions. Journal of Mechanical Design 144(2).

Joseph, L., Jun, B.-M., Flora, J.R.V., Park, C.M. and Yoon, Y. 2019. Removal of heavy metals from water sources in the developing world using low-cost materials: A review. Chemosphere 229, 142-159.
Kaufmann, E., Bauersfeld, L., Loquercio, A., Muller, M., Koltun, V. and Scaramuzza, D. 2023. Champion-level drone racing using deep reinforcement learning. Nature 620(7976), 982-987.
Kazemi, E., Kyritsakas, G., Husband, S., Flavell, K., Speight, V. and Boxall, J. 2023. Predicting iron exceedance risk in drinking water distribution systems using machine learning. IOP Conference Series: Earth and Environmental Science 1136(1), 012047.
Khattak, A., Bukhsh, R., Aslam, S., Yafoz, A., Alghushairy, O. and Alsini, R. 2022. A Hybrid Deep Learning-Based Model for Detection of Electricity Losses Using Big Data in Power Systems. Sustainability 14(20), 13627.
Kirstein, I.V., Hensel, F., Gomiero, A., Iordachescu, L., Vianello, A., Wittgren, H.B. and Vollertsen, J. 2021. Drinking plastics?–Quantification and qualification of microplastics in drinking water distribution systems by μFTIR and Py-GCMS. Water Research 188, 116519.
Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. and Kelley, S.T. 2011. Bayesian community-wide culture-independent microbial source tracking. Nat Methods 8(9), 761-763.
Krasner, S.W., Mitch, W.A., McCurry, D.L., Hanigan, D. and Westerhoff, P. 2013. Formation, precursors, control, and occurrence of nitrosamines in drinking water: a review. Water Res 47(13), 4433-4450.
Krasner, S.W., Weinberg, H.S., Richardson, S.D., Pastor, S.J., Chinn, R., Sclimenti, M.J., Onstad, G.D. and Thruston, A.D. 2006. Occurrence of a New Generation of Disinfection Byproducts. Environmental Science & Technology 40(23), 7175-7185.
LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. and Jackel, L.D. 1989. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation 1(4), 541-551.
Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278-2324.
Lee, D., Gibson, J.M., Brown, J., Habtewold, J. and Murphy, H.M. 2023. Burden of disease from contaminated drinking water in countries with high access to safely managed water: a systematic review. Water Research 242, 120244.
Lee, J., Kim, E.-S., Roh, B.-S., Eom, S.-W. and Zoh, K.-D. 2013. Occurrence of disinfection by-products in tap water distribution systems and their associated health risk. Environmental Monitoring and Assessment 185, 7675-7691.
Legube, B., Parinet, B., Gelinet, K., Berne, F. and Croue, J.P. 2004. Modeling of bromate formation by ozonation of surface waters in drinking water treatment. Water Res 38(8), 2185-2195.
Leitão, J., Simões, N., Sá Marques, J.A., Gil, P., Ribeiro, B. and Cardoso, A. 2019. Detecting urban water consumption patterns: a time-series clustering approach. Water Supply 19(8), 2323-2329.
Ley, C., Martin, R.K., Pareek, A., Groll, A., Seil, R. and Tischer, T. 2022. Machine learning and conventional statistics: making sense of the differences. Knee Surg Sports Traumatol Arthrosc 30(3), 753-757.
Li, J., Heap, A.D., Potter, A. and Daniell, J.J. 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software 26(12), 1647-1659.

Li, J., Zhang, Z., Xiang, Y., Jiang, J. and Yin, R. 2023a. Role of UV-based advanced oxidation processes on NOM alteration and DBP formation in drinking water treatment: A state-of-the-art review. Chemosphere 311, 136870.
Li, L., Rong, S., Wang, R. and Yu, S. 2021. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. Chemical Engineering Journal 405, 126673.
Li, P. and Wu, J. 2019. Drinking water quality and public health. Exposure and Health 11(2), 73-79.
Li, Q., Yu, S., Li, L., Liu, G., Gu, Z., Liu, M., Liu, Z., Ye, Y., Xia, Q. and Ren, L. 2017. Microbial Communities Shaped by Treatment Processes in a Drinking Water Treatment Plant and Their Contribution and Threat to Drinking Water Safety. Front Microbiol 8, 2465.
Li, R.A., McDonald, J.A., Sathasivan, A. and Khan, S.J. 2019. Disinfectant residual stability leading to disinfectant decay and by-product formation in drinking water distribution systems: A systematic review. Water Res 153, 335-348.
Li, Z., Liu, H., Zhang, C. and Fu, G. 2023b. Generative adversarial networks for detecting contamination events in water distribution systems using multi-parameter, multi-site water quality monitoring. Environmental Science and Ecotechnology 14, 100231.
Li, Z., Zhang, C., Liu, H., Zhang, C., Zhao, M., Gong, Q. and Fu, G. 2022. Developing stacking ensemble models for multivariate contamination detection in water distribution systems. Science of The Total Environment 828, 154284.
Ligda, P., Claerebout, E., Kostopoulou, D., Zdragas, A., Casaert, S., Robertson, L.J. and Sotiraki, S. 2020. Cryptosporidium and Giardia in surface water and drinking water: Animal sources and towards the use of a machine-learning approach as a tool for predicting contamination. Environ Pollut 264, 114766.
Lin, H., Dai, Q., Zheng, L., Hong, H., Deng, W. and 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.
Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. 2017 Feature Pyramid Networks for Object Detection, pp. 936-944.
Ling, F., Whitaker, R., LeChevallier, M.W. and Liu, W.T. 2018. Drinking water microbiome assembly induced by water stagnation. ISME J 12(6), 1520-1531.
Liu, G., Zhang, Y., van der Mark, E., Magic-Knezev, A., Pinto, A., van den Bogert, B., Liu, W., van der Meer, W. and Medema, G. 2018. Assessing the origin of bacteria in tap water and distribution system in an unchlorinated drinking water system by SourceTracker using microbial community fingerprints. Water Res 138, 86-96.
Liu, S., Gunawan, C., Barraud, N., Rice, S.A., Harry, E.J. and Amal, R. 2016. Understanding, Monitoring, and Controlling Biofilm Growth in Drinking Water Distribution Systems. Environ Sci Technol 50(17), 8954-8976.
Lowe, M., Qin, R. and Mao, X. 2022. A review on machine learning, artificial intelligence, and smart technology in water treatment and monitoring. Water 14(9), 1384.
Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J. and Knight, R. 2011. UniFrac: an effective distance metric for microbial community comparison. ISME J 5(2), 169-172.
Luo, S., Nguyen, K.T., Nguyen, B.T.T., Feng, S., Shi, Y., Elsayed, A., Zhang, Y., Zhou, X., Wen, B., Chierchia, G., Talbot, H., Bourouina, T., Jiang, X. and Liu, A.Q. 2021. Deep learning-enabled imaging flow cytometry for high-speed Cryptosporidium and Giardia detection. Cytometry Part A 99(11), 1123-1133.

Lytle, D.A., Pfaller, S., Muhlen, C., Struewing, I., Triantafyllidou, S., White, C., Hayes, S., King, D. and Lu, J. 2021. A comprehensive evaluation of monochloramine disinfection on water quality, *Legionella* and other important microorganisms in a hospital. Water Research 189, 116656.

Mahajna, A., Dinkla, I.J.T., Euverink, G.J.W., Keesman, K.J. and Jayawardhana, B. 2022. Clean and Safe Drinking Water Systems via Metagenomics Data and Artificial Intelligence: State-of-the-Art and Future Perspective. Front Microbiol 13, 832452.

Mao, Y., Wang, X., Yang, H., Wang, H. and Xie, Y.F. 2014. Effects of ozonation on disinfection byproduct formation and speciation during subsequent chlorination. Chemosphere 117, 515-520.

Mays, L.W. (2000) Water Distribution System Handbook, McGraw-Hill Education, New York.

Mazhar, M.A., Khan, N.A., Ahmed, S., Khan, A.H., Hussain, A., Rahisuddin, Changani, F., Yousefi, M., Ahmadi, S. and Vambol, V. 2020. Chlorination disinfection by-products in municipal drinking water – A review. Journal of Cleaner Production 273, 123159.

Miake-Lye, I.M., Hempel, S., Shanman, R. and Shekelle, P.G. 2016. What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. Systematic reviews 5, article No. 28.

Moodley, T. and van der Haar, D. 2019 Comparisons in Drinking Water Systems Using K-Means and A-Priori to Find Pathogenic Bacteria Genera. Kim, K.J. and Baek, N. (eds), pp. 351-359, Springer Singapore, Singapore.

Mostafavifar, M., Wertz, J. and Borchers, J. 2012. A systematic review of the effectiveness of kinesio taping for musculoskeletal injury. The Physician and sportsmedicine 40(4), 33-40.

Mounce, S.R., Mounce, R.B. and Boxall, J.B. 2010. Novelty detection for time series data analysis in water distribution systems using support vector machines. Journal of Hydroinformatics 13(4), 672-686.

Muharemi, F., Logofătu, D. and Leon, F. 2019. Machine learning approaches for anomaly detection of water quality on a real-world data set. Journal of Information and Telecommunication 3(3), 294-307.

Mukhopadhyay, A., Duttagupta, S. and Mukherjee, A. 2022. Emerging organic contaminants in global community drinking water sources and supply: A review of occurrence, processes and remediation. Journal of Environmental Chemical Engineering 10(3), 107560.

Narita, K., Matsui, Y., Matsushita, T. and Shirasaki, N. 2023. Screening priority pesticides for drinking water quality regulation and monitoring by machine learning: Analysis of factors affecting detectability. Journal of Environmental Management 326, 116738.

Oh, S., Hossen, I., Luglio, J., Justin, G., Richie, J.E., Medeiros, H. and Lee, C.H. 2021. On-Site/In Situ Continuous Detecting ppb-Level Metal Ions in Drinking Water Using Block Loop-Gap Resonators and Machine Learning. IEEE Transactions on Instrumentation and Measurement 70, 1-9.

Oliker, N. and Ostfeld, A. 2014. Minimum volume ellipsoid classification model for contamination event detection in water distribution systems. Environmental Modelling & Software 57, 1-12.

Ortiz-Lopez, C., Bouchard, C. and Rodriguez, M. 2022. Machine learning models with potential application to predict source water quality for treatment purposes: a critical review. Environmental Technology Reviews 11(1), 118-147.

Pan, R., Zhang, T.-Y., Zheng, Z.-X., Ai, J., Ye, T., Zhao, H.-X., Hu, C.-Y., Tang, Y.-L., Fan, J.-J., Geng, B. and Xu, B. 2023. Insight into mixed chlorine/chloramines conversion and associated

water quality variability in drinking water distribution systems. Science of The Total Environment 880, 163297.
Pandian, A.M.K., Rajamehala, M., Singh, M.V.P., Sarojini, G. and Rajamohan, N. 2022. Potential risks and approaches to reduce the toxicity of disinfection by-product–A review. Science of the Total Environment 822, 153323.
Pang, G., Cao, L., Chen, L., Lian, D. and Liu, H. 2018. Sparse Modeling-Based Sequential Ensemble Learning for Effective Outlier Detection in High-Dimensional Numeric Data. Proceedings of the AAAI Conference on Artificial Intelligence 32(1).
Park, J., Park, J.-H., Choi, J.-S., Joo, J.C., Park, K., Yoon, H.C., Park, C.Y., Lee, W.H. and Heo, T.-Y. 2020. Ensemble Model Development for the Prediction of a Disaster Index in Water Treatment Systems. Water 12(11), 3195.
Park, M., Anumol, T. and Snyder, S.A. 2015. Modeling approaches to predict removal of trace organic compounds by ozone oxidation in potable reuse applications. Environmental Science: Water Research & Technology 1(5), 699-708.
Peleato, N.M. 2022. Application of convolutional neural networks for prediction of disinfection by-products. Scientific Reports 12(1), 612.
Peleato, N.M., Legge, R.L. and Andrews, R.C. 2018. Neural networks for dimensionality reduction of fluorescence spectra and prediction of drinking water disinfection by-products. Water Research 136, 84-94.
Pifer, A.D. and Fairey, J.L. 2012. Improving on $SUVA_{254}$ using fluorescence-PARAFAC analysis and asymmetric flow-field flow fractionation for assessing disinfection byproduct formation and control. Water Research 46(9), 2927-2936.
Pinto, A.J., Schroeder, J., Lunn, M., Sloan, W. and Raskin, L. 2014. Spatial-temporal survey and occupancy-abundance modeling to predict bacterial community dynamics in the drinking water microbiome. mBio 5(3), e01135-01114.
Podgorski, J. and Berg, M. 2022. Global analysis and prediction of fluoride in groundwater. Nat Commun 13(1), 4232.
Prescott, J.F. 2017. History and Current Use of Antimicrobial Drugs in Veterinary Medicine. Microbiology Spectrum 5(6), 10.1128/microbiolspec.arba-0002-2017.
Proctor, C.R., Lee, J., Yu, D., Shah, A.D. and Whelton, A.J. 2020. Wildfire caused widespread drinking water distribution network contamination. AWWA Water Science 2(4), e1183.
Qian, K., Jiang, J., Ding, Y. and Yang, S. 2020 Deep Learning Based Anomaly Detection in Water Distribution Systems, pp. 1-6.
Ramos-Martínez, E., Herrera, M., Izquierdo, J. and Pérez-García, R. 2014. Ensemble of naïve Bayesian approaches for the study of biofilm development in drinking water distribution systems. International Journal of Computer Mathematics 91(1), 135-146.
Ramos-Martínez, E., Herrera, M., Izquierdo, J. and Pérez-García, R. 2016. A Multi-disciplinary Procedure to Ascertain Biofilm Formation in Drinking Water Pipes. International Congress on Environmental Modelling and Software 13.
Redondo-Hasselerharm, P.E., Cserbik, D., Flores, C., Farré, M.J., Sanchís, J., Alcolea, J.A., Planas, C., Caixach, J. and Villanueva, C.M. 2022. Insights to estimate exposure to regulated and non-regulated disinfection by-products in drinking water. Journal of Exposure Science & Environmental Epidemiology, 1-11.
Renwick, D.V., Heinrich, A., Weisman, R., Arvanaghi, H. and Rotert, K. 2019. Potential Public Health Impacts of Deteriorating Distribution System Infrastructure. J Am Water Works Assoc 111(2), 42-53.

Richards, C.E., Tzachor, A., Avin, S. and Fenner, R. 2023. Rewards, risks and responsible deployment of artificial intelligence in water systems. Nature Water 1(5), 422-432.
Richardson, S.D., Thruston, A.D., Caughran, T.V., Chen, P.H., Collette, T.W., Floyd, T.L., Schenck, K.M., Lykins, B.W., Sun, G.-r. and Majetich, G. 1999. Identification of New Ozone Disinfection Byproducts in Drinking Water. Environmental Science & Technology 33(19), 3368-3377.
Ritter, K.J., Carruthers, E., Carson, C.A., Ellender, R.D., Harwood, V.J., Kingsley, K., Nakatsu, C., Sadowsky, M., Shear, B., West, B., Whitlock, J.E., Wiggins, B.A. and Wilbur, J.D. 2003. Assessment of statistical methods used in library-based approaches to microbial source tracking. J Water Health 1(4), 209-223.
Roca, I., Akova, M., Baquero, F., Carlet, J., Cavaleri, M., Coenen, S., Cohen, J., Findlay, D., Gyssens, I., Heuer, O.E., Kahlmeter, G., Kruse, H., Laxminarayan, R., Liébana, E., López-Cerero, L., MacGowan, A., Martins, M., Rodríguez-Baño, J., Rolain, J.M., Segovia, C., Sigauque, B., Tacconelli, E., Wellington, E. and Vila, J. 2015. The global threat of antimicrobial resistance: science for intervention. New Microbes New Infect 6, 22-29.
Rodriguez, M.J., Sérodes, J.B. and Levallois, P. 2004. Behavior of trihalomethanes and haloacetic acids in a drinking water distribution system. Water Res 38(20), 4367-4382.
Rodriguez-Perez, J., Leigh, C., Liquet, B., Kermorvant, C., Peterson, E., Sous, D. and Mengersen, K. 2020. Detecting Technical Anomalies in High-Frequency Water-Quality Data Using Artificial Neural Networks. Environmental Science & Technology 54(21), 13719-13730.
Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A. and Chen, L.-C. 2018 MobileNetV2: Inverted Residuals and Linear Bottlenecks, pp. 4510-4520.
Sathasivan, A., Fisher, I. and Tam, T. 2008. Onset of severe nitrification in mildly nitrifying chloraminated bulk waters and its relation to biostability. Water Research 42(14), 3623-3632.
Savadatti, M.B., Dhivya, M., Meghanashree, C., Navya, M., Lokesh, Y. and Kawri, N. 2022 An Overview of Predictive Analysis based on Machine learning Techniques, pp. 1-6.
Senoro, D.B., de Jesus, K.L.M., Nolos, R.C., Lamac, M.R.L., Deseo, K.M. and Tabelin, C.B. 2022. In Situ Measurements of Domestic Water Quality and Health Risks by Elevated Concentration of Heavy Metals and Metalloids Using Monte Carlo and MLGI Methods. Toxics 10(7), 342.
Shi, Y., Babatunde, A., Bockelmann-Evans, B., Li, Q. and Zhang, L. 2020. On-going nitrification in chloraminated drinking water distribution system (DWDS) is conditioned by hydraulics and disinfection strategies. Journal of Environmental Sciences 96, 151-162.
Shi, Y., Wang, J., Wang, Q., Jia, Q., Yan, F., Luo, Z.-H. and Zhou, Y.-N. 2022. Supervised Machine Learning Algorithms for Predicting Rate Constants of Ozone Reaction with Micropollutants. Industrial & Engineering Chemistry Research 61(24), 8359-8367.
Simpson, A.M.A. and Mitch, W.A. 2022. Chlorine and ozone disinfection and disinfection byproducts in postharvest food processing facilities: A review. Critical Reviews in Environmental Science and Technology 52(11), 1825-1867.
Sincak, P., Ondo, J., Kaposztasova, D., Vircikova, M., Vranayova, Z. and Sabol, J. 2014. Artificial intelligence in public health prevention of legionelosis in drinking water systems. Int J Environ Res Public Health 11(8), 8597-8611.
Singh, D., Vardhan, M., Sahu, R., Chatterjee, D., Chauhan, P. and Liu, S. 2023. Machine-learning- and deep-learning-based streamflow prediction in a hilly catchment for future scenarios using CMIP6 GCM data. Hydrol. Earth Syst. Sci. 27(5), 1047-1075.

Singh, K.P. and Gupta, S. 2012. Artificial intelligence based modeling for predicting the disinfection by-products in water. Chemometrics and Intelligent Laboratory Systems 114, 122-131.
Sluban, B. and Lavrač, N. 2015. Relating ensemble diversity and performance: A study in class noise detection. Neurocomputing 160, 120-131.
Smith, A., Sterba-Boatwright, B. and Mott, J. 2010. Novel application of a statistical technique, Random Forests, in a bacterial source tracking study. Water Res 44(14), 4067-4076.
Sokolova, M. and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45(4), 427-437.
Speight, V.L., Mounce, S.R. and Boxall, J.B. 2019. Identification of the causes of drinking water discolouration from machine learning analysis of historical datasets. Environmental Science: Water Research & Technology 5(4), 747-755.
Srivastav, A.L., Patel, N. and Chaudhary, V.K. 2020. Disinfection by-products in drinking water: Occurrence, toxicity and abatement. Environmental Pollution 267, 115474.
Steel, R.G.D. and Torrie, J.H. (1960) Principles and procedures of statistics : with special reference to the biological sciences, McGraw-Hill, New York.
Sudhakaran, S. and Amy, G.L. 2013. QSAR models for oxidation of organic micropollutants in water based on ozone and hydroxyl radical rate constants and their chemical classification. Water Res 47(3), 1111-1122.
Syafrudin, M., Kristanti, R.A., Yuniarto, A., Hadibarata, T., Rhee, J., Al-Onazi, W.A., Algarni, T.S., Almarri, A.H. and Al-Mohaimeed, A.M. 2021. Pesticides in drinking water—a review. International Journal of Environmental Research and Public Health 18(2), 468.
Taheran, M., Naghdi, M., Brar, S.K., Verma, M. and Surampalli, R.Y. 2018. Emerging contaminants: here today, there tomorrow! Environmental Nanotechnology, Monitoring & Management 10, 122-126.
Tian, C., Feng, C., Chen, L. and Wang, Q. 2020. Impact of water source mixture and population changes on the Al residue in megalopolitan drinking water. Water Res 186, 116335.
Tolaymat, T.M., El Badawy, A.M., Genaidy, A., Scheckel, K.G., Luxton, T.P. and Suidan, M. 2010. An evidence-based environmental perspective of manufactured silver nanoparticle in syntheses and applications: a systematic review and critical appraisal of peer-reviewed scientific papers. Science of the Total Environment 408(5), 999-1006.
U.S. EPA 2022 Information about Public Water Systems, EPA Office of Ground Water and Drinking Water, Washington, DC, U.S.A.
Umesh C. Gupta, S.C.G. 2011. Heavy Metal Toxicity in Humans and its Preventive and Control Measures. Current Nutrition & Food Science 7(4), 221-231.
Valko, M., Morris, H. and Cronin, M.T. 2005. Metals, toxicity and oxidative stress. Curr Med Chem 12(10), 1161-1208.
Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J.P., Jaderberg, M., Vezhnevets, A.S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T.L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wunsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C. and Silver, D. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature 575(7782), 350-354.
Wagner, E.D. and Plewa, M.J. 2017. CHO cell cytotoxicity and genotoxicity analyses of disinfection by-products: An updated review. J Environ Sci (China) 58, 64-76.

Walesch, S., Birkelbach, J., Jézéquel, G., Haeckl, F.P.J., Hegemann, J.D., Hesterkamp, T., Hirsch, A.K.H., Hammann, P. and Müller, R. 2023. Fighting antibiotic resistance—strategies and (pre)clinical developments to find new antibacterials. EMBO reports 24(1), e56033.
Wang, C., Yang, H., Liu, H., Zhang, X.-X. and Ma, L. 2023. Anthropogenic contributions to antibiotic resistance gene pollution in household drinking water revealed by machine-learning-based source-tracking. Water Research 246, 120682.
Weigert, M., Schmidt, U., Haase, R., Sugawara, K. and Myers, G. 2020 Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy, pp. 3655-3662, IEEE Computer Society.
WHO (2011) Guidelines for drinking-water quality, World Health Organization, Geneva.
WHO 2021. Antimicrobial resistance.
Willard, J., Jia, X., Xu, S., Steinbach, M. and Kumar, V. 2022. Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. ACM Comput. Surv. 55(4), Article 66.
Willmott, C.J. and Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research 30(1), 79-82.
Wołos, A., Koszelewski, D., Roszak, R., Szymkuć, S., Moskal, M., Ostaszewski, R., Herrera, B.T., Maier, J.M., Brezicki, G., Samuel, J., Lummiss, J.A.M., McQuade, D.T., Rogers, L. and Grzybowski, B.A. 2022. Computer-designed repurposing of chemical wastes into drugs. Nature 604(7907), 668-676.
Wołowiec, M., Komorowska-Kaufman, M., Pruss, A., Rzepa, G. and Bajda, T. 2019. Removal of Heavy Metals and Metalloids from Water Using Drinking Water Treatment Residuals as Adsorbents: A Review. Minerals 9(8), 487.
Wolpert, D.H. 1992. Stacked generalization. Neural Networks 5(2), 241-259.
Wu, J., Song, C., Dubinsky, E.A. and Stewart, J.R. 2020. Tracking Major Sources of Water Contamination Using Machine Learning. Front Microbiol 11, 616692.
Wu, Y., Jiang, P., Goh, S.G., Yu, K., Chen, Y., He, Y. and Gin, K.Y.H. 2022. Predicting Relative Risk of Antimicrobial Resistance using Machine Learning Methods. IFAC-PapersOnLine 55(10), 1266-1271.
Xiao, R., Ou, T., Ding, S., Fang, C., Xu, Z. and Chu, W. 2023. Disinfection by-products as environmental contaminants of emerging concern: a review on their occurrence, fate and removal in the urban water cycle. Critical Reviews in Environmental Science and Technology 53(1), 19-46.
Xie, Y., Sattari, K., Zhang, C. and Lin, J. 2023. Toward autonomous laboratories: Convergence of artificial intelligence and experimental automation. Progress in Materials Science 132, 101043.
Xu, X., Talbot, S. and Selvaraja, T. 2020. ParasNet: Fast parasites detection with neural networks. arXiv preprint arXiv:2002.11327.
Zainurin, S.N., Wan Ismail, W.Z., Mahamud, S.N.I., Ismail, I., Jamaludin, J., Ariffin, K.N.Z. and Wan Ahmad Kamil, W.M. 2022. Advancements in monitoring water quality based on various sensing methods: a systematic review. International Journal of Environmental Research and Public Health 19(21), 14080.
Zanoni, M.G., Majone, B. and Bellin, A. 2022. A catchment-scale model of river water quality by Machine Learning. Science of the Total Environment 838, 156377.

Zhang, C., Brown, P.J.B. and Hu, Z. 2018. Thermodynamic properties of an emerging chemical disinfectant, peracetic acid. Science of the Total Environment 621, 948-959.
Zhang, C., Hu, Z., Li, P. and Gajaraj, S. 2016. Governing factors affecting the impacts of silver nanoparticles on wastewater treatment. Science of The Total Environment.
Zhang, C. and Lu, J. 2021a. Legionella: A Promising Supplementary Indicator of Microbial Drinking Water Quality in Municipal Engineered Water Systems. Front Environ Sci 9, 1-22.
Zhang, C. and Lu, J. 2021b. *Legionella*: a supplementary indicator of microbial water quality in municipal engineered water systems. *Frontiers in Environmental Science* 9, 684319.
Zhang, C. and Lu, J. 2021c. Optimizing disinfectant residual dosage in engineered water systems to minimize the overall health risks of opportunistic pathogens and disinfection by-products. Science of The Total Environment 770, 145356.
Zhang, C., Struewing, I., Mistry, J.H., Wahman, D.G., Pressman, J. and Lu, J. 2021. Legionella and other opportunistic pathogens in full-scale chloraminated municipal drinking water distribution systems. Water Res 205, 117571.
Zhang, Y., Gao, X., Smith, K., Inial, G., Liu, S., Conil, L.B. and Pan, B. 2019. Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. Water Research 164, 114888.
Zhong, S., Zhang, K., Bagheri, M., Burken, J.G., Gu, A., Li, B., Ma, X., Marrone, B.L., Ren, Z.J., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B.M., Xiao, X., Yu, X., Zhu, J.-J. and Zhang, H. 2021. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. Environmental Science & Technology 55(19), 12741-12754.
Zhou, Q., Bian, Z., Yang, D. and Fu, L. 2023. Stability of Drinking Water Distribution Systems and Control of Disinfection By-Products. Toxics 11(7), 606.
Zhou, X., Tang, Z., Xu, W., Meng, F., Chu, X., Xin, K. and Fu, G. 2019. Deep learning identifies accurate burst locations in water distribution networks. Water Research 166, 115058.
Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B. and Ye, L. 2022. A review of the application of machine learning in water quality evaluation. Eco-Environment & Health 1(2), 107-116.
Ziegelbauer, K., Speich, B., Mäusezahl, D., Bos, R., Keiser, J. and Utzinger, J. 2012. Effect of sanitation on soil-transmitted helminth infection: systematic review and meta-analysis. PLOS Medicine 9(1), article No. e1001162.

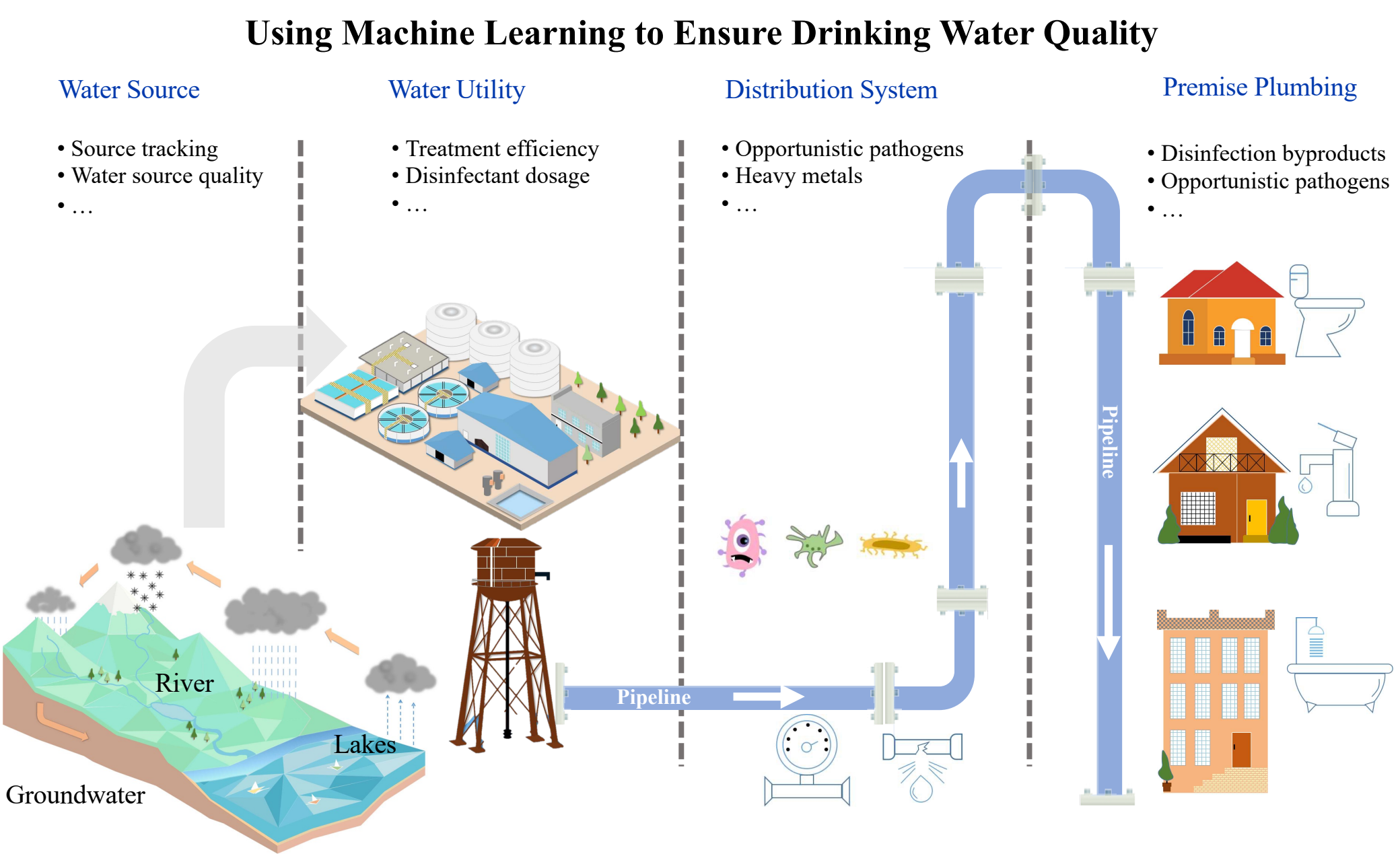

Figure 1. A stage-by-stage overview of the application of machine learning to ensure drinking water quality

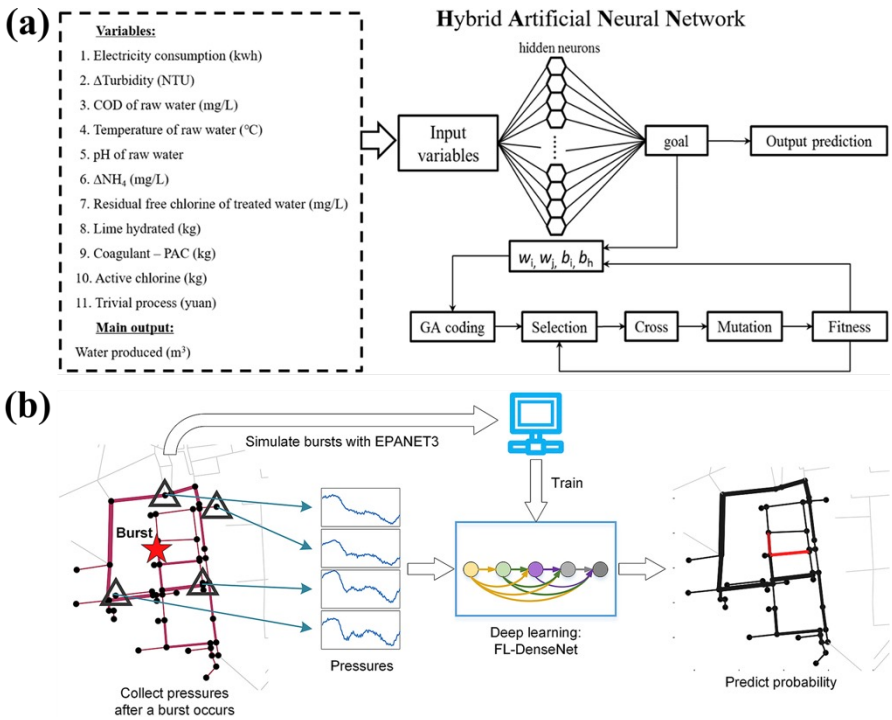

Figure 2. (a) Input and output variables used for modeling and the proposed hybrid artificial neural network framework. Reproduced with permission from Zhang et al., 2019. Copyright 2019 Elsevier. (b) Schematic of fully-linear DenseNet (BLIFF) model for accurate identification of burst locations in EWS networks. Reproduced with permission from Zhou et al., 2019. Copyright 2019 Elsevier.

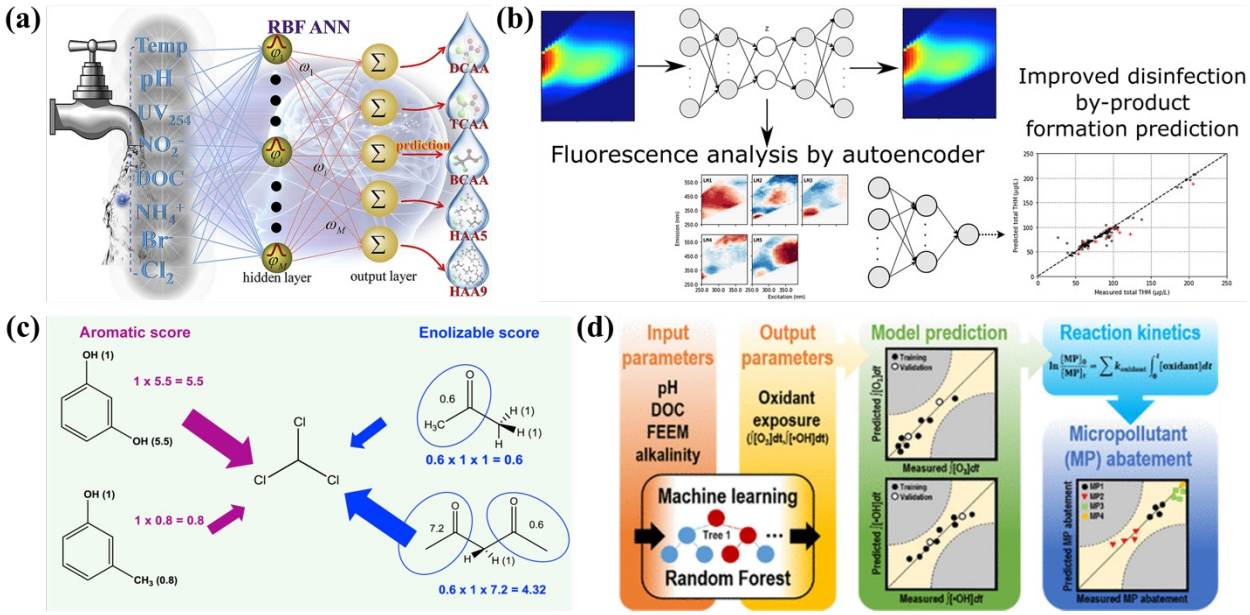

Figure 3. (a) Schematic of radial basis function (RBF) artificial neural network (ANN) model for prediction of disinfection by-products (DBPs). Reproduced with permission from Lin et al., 2020, Copyright 2020 Elsevier. (b) Schematic of parallel factor analysis (PARAFAC) model for prediction of DBPs. Reproduced with permission from Peleato et al., 2018, Copyright 2018 Elsevier. (c) Schematic of multiple linear regression (MLR) model for prediction of DBPs from organic precursors. Reproduced with permission from Bond and Graham 2017, Copyright 2017 Elsevier. (d) Schematic of random forest (RF) model for prediction of micropollutant abatement. Reproduced with permission from Cha et al., 2021, Copyright 2021 ACS.

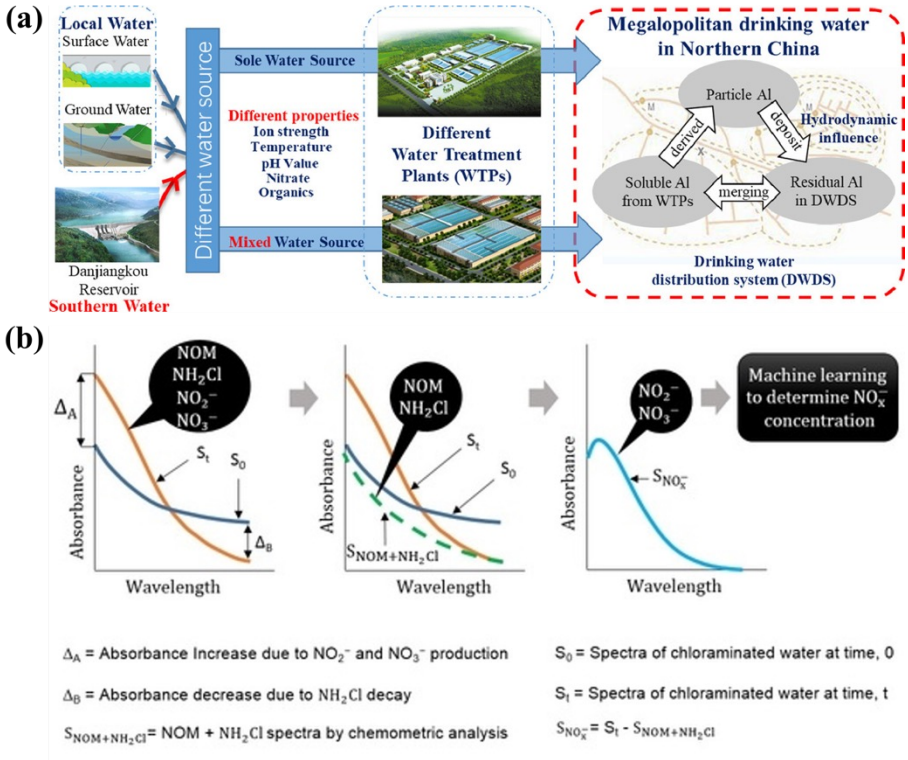

Figure 4. (a) The impact of water source mixture and population changes on the aluminum (Al) residue. Reproduced with permission from Tian et al., 2020, Copyright 2020 Elsevier. (b) Prediction of nitrate and nitrite concentrations over support vector regression (SVR) model. Reproduced with permission from Hossain et al., 2021, Copyright 2021 MDPI.

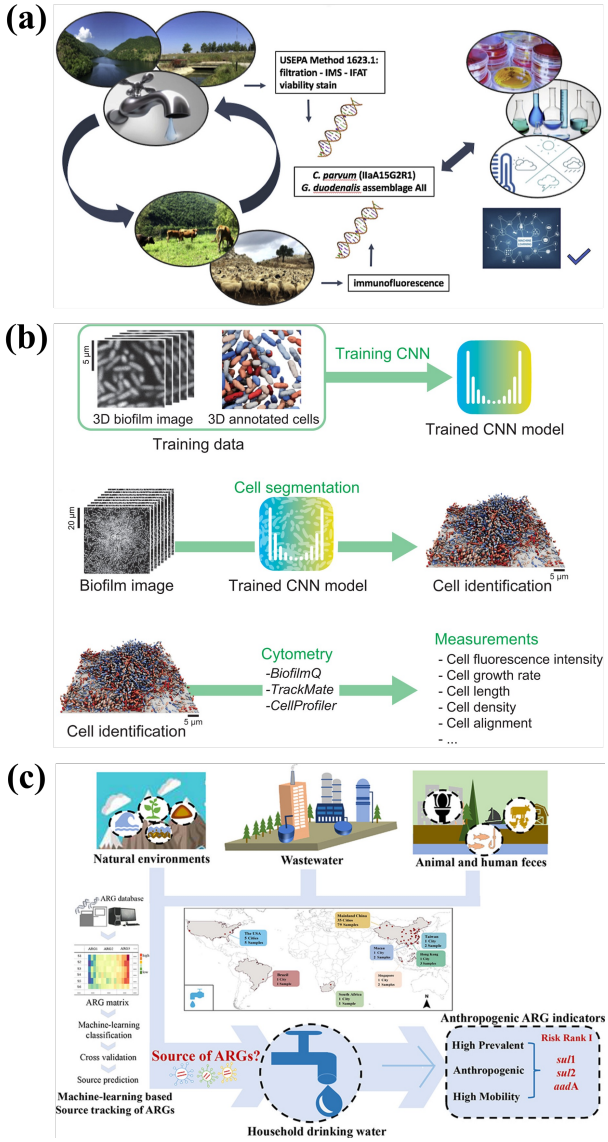

Figure 5. (a) The application of machine learning to predict contamination of *Cryptosporidium* and *Giardia* in surface water and drinking water. Reproduced with permission from Ligda et al., 2020, Copyright 2020 Elsevier. (b) Deep-learning-based workflow for single-cell measurements in three-dimensional biofilms. Reproduced with permission from Jelli et al., 2023, Copyright 2023 Elsevier. (c) SourceTracker was performed to investigate the pollution sources of antimicrobial resistance genes (ARGs) in household drinking water. Reproduced with permission from Wang et al., 2023, Copyright 2023 Elsevier.

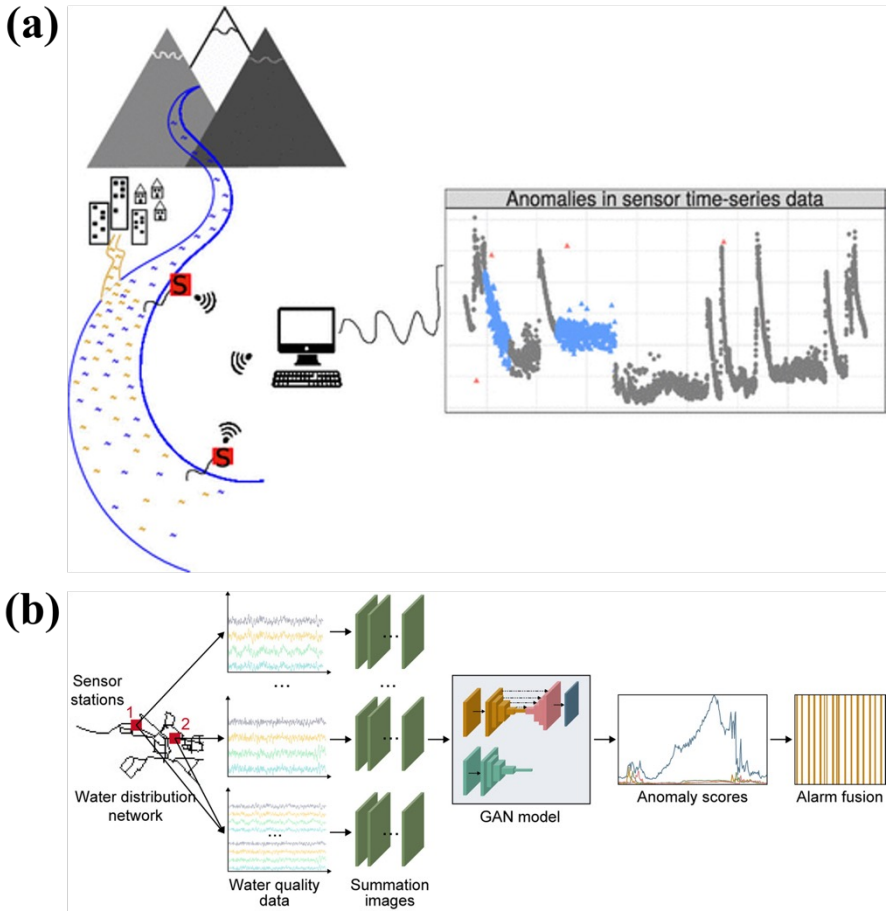

Figure 6. (a) Detection of technical anomalies in water quality using artificial neural network (ANN) model. Reproduced with permission from Rodriguez-Perez et al., 2020, Copyright 2020 ACS. (b) Detection of contamination events using generative adversarial network (GAN) model. Reproduced with permission from Li et al., 2023, Copyright 2023 Elsevier.

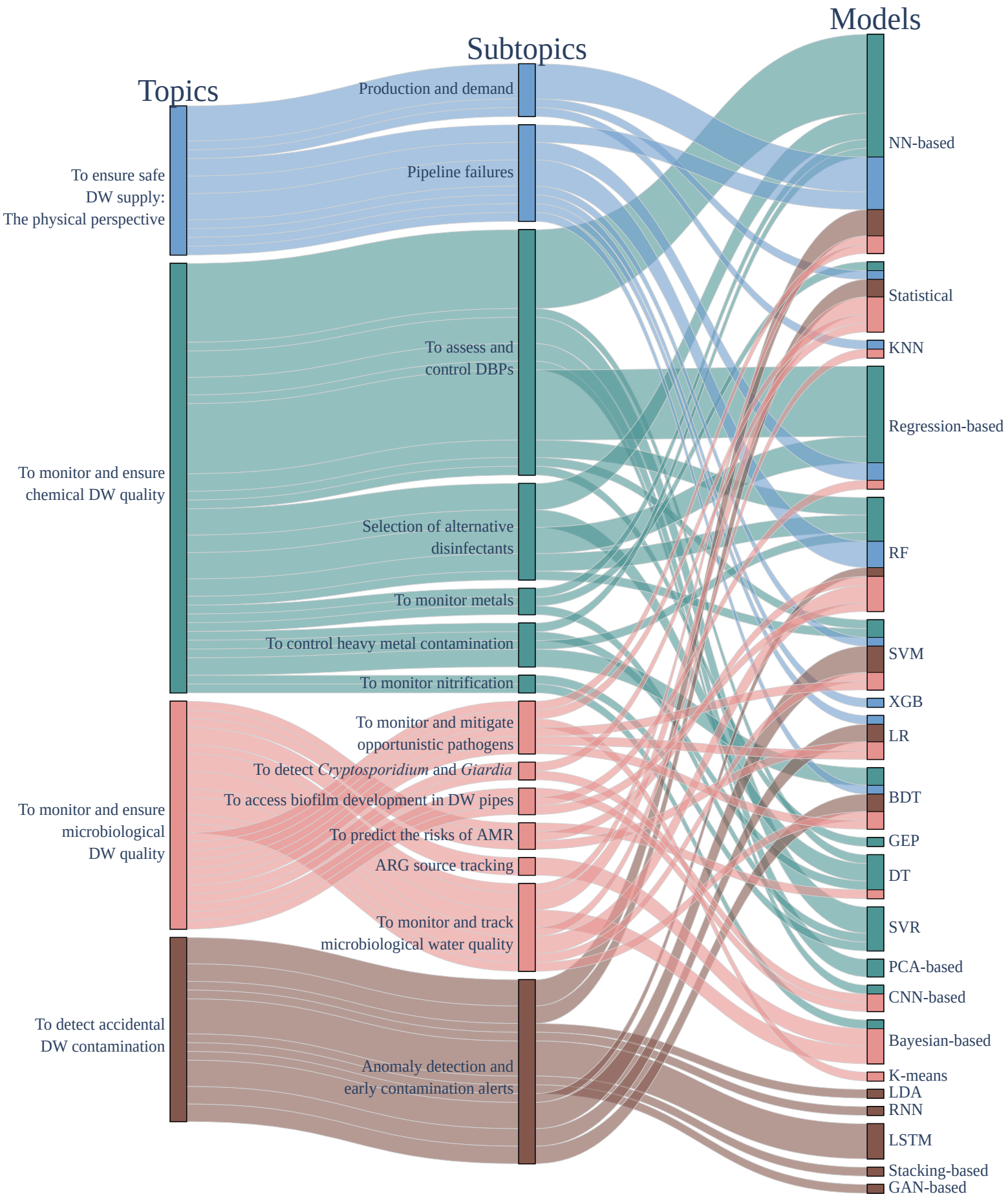

Figure 7. Distribution of machine learning models across drinking water distribution system research topics

Abbreviations: NN, Neural network; KNN, K-nearest neighbor; RF, random forest; SVM, support vector machine; XGB, extreme gradient boosting; LR, logistic regression; BDT, boosting decision tree; GEP, gene expression programming; DT, decision tree; SVR, support vector regression; PCA, principal component analysis; CNN, convolutional neutral network; LDA, linear discriminant analysis; RNN, recurrent neural network; LSTM, long short-term memory; GAN, generative adversarial network; DW, drinking water.

Table 1 The applications of machine learning to ensure safe drinking water supply from the physical perspective

Topic	Task	Model	Inputs	Outputs	Metrics (Selected)	Performances (Selected)	Reference
Drinking water production and demand	Water production prediction	GA-ANN and ML-ANN	T, COD, and operational parameters	Water production of DWTPs	MSE, R^2, MAPE	GA-ANN R^2=0.93 > ML-ANN	Zhang et al., 2019
	Short-term water demand prediction	GRUN, ANN, and SARIMA	Historical water demand data	15-min and 24-h prediction of water demand	MAE, MAPE, RMSE, NSE	GRUN > ANN and SARIMA	Guo et al., 2018
	Water demand forecasting	DAN2, FTDNN, and KNN	Daily water production and monthly water consumption	Daily, weekly, and monthly water demands	MAPE, accuracy, R^2, MSE, and SSE	DAN2 accuracies: 96% to 98%	Ghiassi et al., 2017
Drinking water pipeline failures	Pipe burst localization	FL-DenseNet	Pressure measurements	Burst occurring likelihood per pipe	Accuracy	62.35% to 98.58%	Zhou et al., 2019
	Pipe failure prediction	AdvaML, Cox-pH, SRF, and SSVM	Pipe data and climate data	Failure/Hazard Index	C-index	AdvaML $\geq$ 0.8 > Cox-pH, SRF, and SSVM	Almheiri et al., 2021
	Disaster index prediction on WTS	RF and XGB	Facility specification and operational data	Disaster index	RMSE and R^2	XGB R^2 = 0.86 > RF	Park et al., 2020
	Earthquake damage prediction	RR, LR, BRT, and RF	Earthquake-related variables and pipe attributes	Binary classification of damage status	TE, TEP, RMSE, MAE, MASE, MPSE, SN, SP, TSS, and AUC	BRT > RR, LR, and RF in overall performance	Bagriacik et al., 2018

GA-ANN, Artificial neural network with genetic algorithm; ML-ANN, multi-layered artificial neural network; T, temperature; COD, chemical oxygen demand; DWTPs, drinking water treatment plants; MSE, mean squared error; R^2, coefficient of determination; MAPE, mean absolute percentage error; GRUN, gated recurrent unit network; SARIMA, seasonal autoregressive integrated moving average; MAE, mean absolute error; RMSE, root-mean square error; NSE, Nash-Sutcliffe model efficiency; DAN2, dynamic artificial neural network; FTDNN, focused time-delay neural network; KNN, K-nearest neighbor; SSE, summing the squared differences; FL-DenseNet, fully-linear DenseNet; AdvaML, advanced meta-learning; Cox-pH, cox-proportional hazards; SRF, random survival forest; SSVM, survival support vector machine; C-index, concordance index; WTS, water treatment system; RF, random forest; XGB, extreme gradient boosting (XGBoost); RR, repair rate; LR, logistic regression; BRT, boosted regression trees; TE, error in total count; TEP, percentage error in total count; MASE, median absolute suburb error; MPSE, Median percentage suburb error; SN, sensitivity; SP, specificity; TSS, true skill statistics; AUC, area under the receiver operating characteristic (ROC) curve.

Table 2 The applications of machine learning to monitor and ensure chemical drinking water quality

Topic	Task	Model	Inputs	Outputs	Metrics (Selected)	Performances (Selected)	Reference
To assess and control DBPs	DBPs formation prediction	ANN, SVM, and GEP	pH, T, C_{Br^-}, $C_{Cl_2/DOC}$, t	C_{THMs}	MSE, RMSE and R^2	SVM > ANN and GEP	Singh and Gupta 2012
		Linear/log linear, and RBF-ANN	pH, T, UV_{254}, C_{DOC}, C_{Br^-}, $C_{residual\_cl}$, $C_{NO_2^-N}$, and $C_{NH_4^+-N}$	C_{HAAs}	Accuracy, AAE	RBF-ANN > linear/log linear	Lin et al., 2020
		Linear/log linear, and RBF-ANN	pH, T, UV_{254}, C_{DOC}, C_{Br^-}, $C_{residual\_free\_cl}$, $C_{NO_2^-N}$, and $C_{NH_4^+-N}$	C_{THMs}	Accuracy and r_p	RBF ANN > linear/log linear	Hong et al., 2020
		Linear/log linear BP-ANN, and RBF-ANN		C_{HKs}	R^2	RBF ANN:0.799 > BP ANN and linear/log linear	Deng et al., 2021
		DTB	C_{NH_2Cl}, C_{NHCl_2} + OC, pH, TDN, $C_{NO_2^-N}$, TOC, and $C_{NH_4^+-N}$	C_{THM4} and C_{HAAs}	R^2 and MSE	C_{THM4}: R^2 = 0.56 C_{HAAs}: R^2 = 0.65	Pan et al., 2023
		MLR, NN, RF, GPR and SVR	T, $C_{residual\_cl}$, DOC, Turb, pH, Leit, and UV_{254}	C_{THMs}, C_{HAAs}, C_{DCAN}, C_{CPK}, and C_{TCP}	MSE	SVR, GPR > NN > RF > MLR	Hu et al., 2023
	Spectroscopic detection of DBPs	AE-NN, AE, PCA, and PARAFAC	Fluorescence spectra	C_{THMs} and C_{HAAs}	MAE, MSE	AE-NN > AE > PCA > PARAFAC	Peleato et al., 2018
		MLP, CNN, PARAFAC-MLP, PCA-MLP, and 3-way PLS		C_{THMs}, C_{HAAs}, and C_{TCMs}		CNN > MLP, PARAFAC-MLP, PCA-MLP, and 3-way PLS	Peleato 2022
		SVR-(linear, polynomial, RBF, or sigmoid)	UV-Vis spectra	C_{NH_2Cl}	R^2 and RMSE	SVR-RBF > SVR-Polynomial > SVR-linear and SVR-Sigmoid	Hossien et al., 2020
	DBPs formation mechanism analysis	MLR	Chemical descriptors	THM yield	R^2 and RMSE	R^2 = 0.91	Bond and Graham 2017
		RF, SVR-RBF, SVR-linear, MLP, and MLR	Chemical descriptors	HAAs formation potential		RF > SVR-RBF, SVR-linear, MLP and MLR	Cordero et al., 2021

DBPs, Disinfection by-products; GEP, gene expression programming; C_{Br^-}, Br concentration; $C_{Cl_2/DOC}$, dissolved organic carbon normalized chlorine dose; t, contact time; C_{THMs}, trihalomethane concentration; linear/log Linear, linear/log linear regression models; RBF-ANN, radial basis function ANN; UV_{254}, ultraviolet absorbance at 254 nm; C_{DOC}, dissolved organic carbon concentration; $C_{residual\_cl}$, residue chlorine concentration; $C_{NO_2^-N}$, nitrite concentration; $C_{NH_4^+-N}$, ammonia concentration; C_{HAAs}, haloacetic acids concentration; AAE, average absolute error; $C_{residual\_free\_cl}$, residual free chlorine concentration; r_p, regression coefficients; BP-ANN, back propagation ANN; C_{HKs}, haloketones concentration; DTB: decision tree boost; C_{NH_2Cl}, monochloramine concentration; C_{NHCl_2} + OC, dichloramine and organic chloramines concentration; TDN, total dissolved nitrogen; TOC, total organic carbon;

Table 2 The applications of machine learning to monitor and ensure chemical drinking water quality

MLR, multiple linear regression; GPR, Gaussian process regression; SVR, support vector regression; Turb, turbidity; Leit, electric conductivity of the water; C_{DCAN}, dichloroacetonitrile concentration; C_{CPK}, chloropicrin concentration; C_{TCP}, trichloropropanone concentration; AE-NN, autoencoder-neural network; PCA, principal component analysis; PARAFAC, parallel factors analysis; MLP, multi-layer perceptron network; CNN, convolutional neutral network; 3-way PLS, 3-way partial least squares; C_{TCMs} , trichloromethane concentration; C_{NH_2Cl}, monochloramine concentration; MLP, multilayer perceptron.

Table 2 The applications of machine learning to monitor and ensure chemical drinking water quality (*cont.*)

Topic	Task	Model	Inputs	Outputs	Metrics (Selected)	Performances (Selected)	Reference
To select alternative disinfectants	Prediction of bromate formation by ozonation	MLR and ANN	C_τ, pH, C_{Br^-}, T, UV, DOC, Alk, and $C_{NH_4^+-N}$	$C_{BrO_3^-}$	R^2	ANN = 0.98 > MLR	Legube et al., 2004
	Prediction of MP/organic contaminant abatement during ozonation	RF	pH, Alk, DOC, and FEEM	Oxidant exposures	R^2 and RMSE	R^2 : 0.904	Cha et al., 2021
		MLR, SVM, DT, RF, and DNN	NI_a, E_{LUMO}, E_{HOMO}, and $C_{Ene\_val,min}$	log k_{O_3}	R^2 , MSE, MAE and Q_{ext^2}	RF: R^2 = 0.9113	Shi et al., 2022
		DTB and SDT	k_{O_3} model: AMR, minHBa, n_X, and MDEC-24 ; k_{SO_4} model: AMR, SssO, and mean$I$	k_{O_3} and k_{SO_4}	R^2 and RMSE	DTB: R^2 > 0.97	Gupta and Basant 2016
	Estimation of the TOrCs removal	MLR, ANN, and PC-ANN	C_{O_3}, TOC, $k_{O_{3,TOrC}}$ and $k._{OH,TOrC}$	TOrCs removal	R^2 and RMSE	PC-ANN: R^2 = 0.934	Park et al., 2015
To monitor HM	Pb ions concentration detection	SVR	S_{11}	Pb concentration	RMS	0.71	Oh et al., 2021
	Spatial HM concentration mapping	MLGI (NN - PSO + EBK)	Geographical coordinates	Spatial concentration maps	MSE and r	$r \approx 1.0$	De Jesus et al., 2021
	Temporal-spatial map generating	Kriging interpolation	Spatial and temporal data	Temporal-spatial distribution of residual Al	-	-	Tian et al., 2020
To control HM contamination	As adsorption removal prediction	LightGBM, XGB, GBDT, and RF	adsorbent dosage, t, C_{As_init}, pH, T, A_{MOFs}, and N_{anions}	Adsorptive removal of As(V)	AAPRE, RMSE and R^2	LightGBM > XGBoost > GBDT > RF	Abdi et al., 2022
	HM removal prediction	MLP-ANN and RBF-ANN	adsorbent dosage, τ, and pH_{init}	Al, Cd, Co, Cu, Fe, and Pb ions removal efficiency	MSE and R^2	RBF-ANN > MLP-ANN	Hamidian et al., 2019
To monitor nitrification	Nitrification episodes classification	NB	16S rRNA profiling	Nitrification episodes: stable or failure	AUC	0.83	Gomez-Alvarez et al., 2020
	Estimate NOx concentrations	SVR	NOx absorbances at various wavelengths	C_{NO_x}	RMSE and R^2	RMSE < 0.04	Hossain et al., 2021

C_τ, Disinfectant concentration and contact time product; Alk, alkalinity; $C_{BrO_3^-}$, bromate concentration; FEEM, fluorescence excitation−emission matrix; DT, decision tree; DNN, deep neutral network; NI_a , norm descriptors; E_{LUMO} and E_{HOMO}, energy of the lowest unoccupied molecular orbital and energy of the highest occupied molecular orbital; $C_{Ene\_val,min}$, minimum valence shell orbital energy on carbon atom; Q_{ext^2} , external validation parameter; SDT, single decision tree; k_{O_3} and k_{SO_4} , the rate constants for the reactions of O_3 and $SO_4^{\cdot-}$ respectively; AMR, antimicrobial resistance; minHBa, minimum E-states for (strong) hydrogen bond acceptors; n_X , number of halogen atoms; MDEC-24, molecular distance edge between all secondary and quaternary carbons; SssO, sum of atom-type E-state: -O-; mean$I$, mean intrinsic state values I.

Table 2 The applications of machine learning to monitor and ensure chemical drinking water quality (*cont.*)

TOrCs, trace organic compounds; PC-ANN, principal component ANN; C_{O_3}, applied ozone dose; $k_{O_{3,TOrC}}$ and $k._{OH,TOrC}$, rate constants of O_3 and ˙OH of TOrCs; HM, heavy metal; S_{11}, reflection coefficient; MLGI (NN-PSO+EBK), machine learning and geostatistical interpolation (neural network with the particle swarm optimization and empirical Bayesian kriging); r, Pearson's correlation coefficient; LightGBM, light gradient-boosting machine; GBDT, gradient boosting decision tree; t, contact time; C_{As_init} , initial arsenic concentration; A_{MOFs}, metal–organic frameworks surface area; N_{anions} , presence of anions; AAPRE, average absolute percent relative error; $\mathrm{pH_{init}}$, initial pH; NB, naïve Bayes; AUC, area under the curve; NOx, nitrite and nitrate.

Table 3 The applications of machine learning to monitor and ensure microbiological drinking water quality

Topic	Task	Model	Inputs	Outputs	Metrics (Selected)	Performances (Selected)	Reference
To monitor and mitigate opportunistic pathogens	To simulate conditions for preventing legionleosis outbreak	NARA	*Q* and T	T profile of the water tank	Accuracy	>97%	Sincak et al., 2014
	Bacterium clustering	K-means	16S rRNA profiling	Clusters of bacteria	-	-	Moodley and Haar 2019
	Spatio-temporal clustering of high-risk; serogroup and contamination levels prediction	SaTScan, XGB, LR, and SVM	Survey, spatial and meteorological info., and risk level to *Legionella*;	High-risk level clusters; serogroup of a sample and the contamination level	Accuracy and F1-score	XGBoost > SVM > LR	Brunello et al., 2022
To detect *Cryptosporidium* and *Giardia*	Image classification of *Cryptosporidium* and *Giardia* morphology	CNN	Cell level scattering image	Classification of *Cryptosporidium*, *Giardia,* or others	Accuracy	Accuracy: 95.6% for *Cryptosporidium* and 99.5% for *Giardia*	Xu et al., 2020
				Multiple classification or binary classification	Accuracy, precision, recall, and F1-score	Accuracy > 99.6%	Luo et al., 2021
	Cryptosporidium and *Giardia* contamination intensity prediction	LDFA	Microbiological, physicochemical, and meteorological parameters	(oo)cyst concentrations of *Cryptosporidium* and *Giardia*	Accuracy	Accuracy: 75% for *Cryptosporidium* and 69% for *Giardia*	Ligda et al., 2020
to access biofilm development in drinking water pipes	Biofilm development analysis	RT and RF	System physical and hydraulic characteristics, sampling and incubation, and physico-chemical of water	HPC	*R*	RF: 0.898	Ramos-Martínez et al., 2016
	Single-cell segmentation in 3D biofilms	StarDist OPP (CNN-based)	3D biofilm image	Cell identification	Precision and OSA	OSA = 3% Precision depends on IoU threshold	Jelli et al., 2023

NARA, Neural network designed on approximate reasoning architecture; *Q*, flow rate; LDFA, linear discriminant function analysis; RT, regression trees; HPC , heterotrophic plate count; OSA, over-segmentation abundances; IoU, intersection-over-union.

Table 3 The applications of machine learning to monitor and ensure microbiological drinking water quality (*cont.*)

Topic	Task	Model	Inputs	Outputs	Metrics (Selected)	Performances (Selected)	Reference
To predict risk of AMR and track sources of ARGs	Relative risk of AMR prediction	LR, DT, and RF	T, pH, ORP, EC, ρ, TDS, Sal, P, DO, Turb, and 24h rainfall	Relative risk score	Accuracy, precision, recall, F1-score and AUC	RF: AUC = 0.88 > DT, LR	Wu et al., 2022
	ARG source tracking	Bayesian-based	Metagenomic signatures of ARGs and microbial taxa	Relative contributions of ARGs	-	-	Chen et al., 2019
			Broad-spectrum ARG profiles	Proportion of pollution sources of AGGs	r	r = 0.98	Wang et al., 2023
To monitor and track microbiological water quality	Water source tracking	Bayesian-based	rep-PCR and ARA	Source membership	-	RMSEp < RMSEc	Ritter et al., 2003
			ARA	Source distribution	RMSE		Greenberg et al., 2010
			Bacterial 16S ribosomal RNA gene sequences	Source proportion	R^2	≥ 0.8	Knights et al., 2011
		RF	ARA	Source classification	ARCC	82.3%	Smith et al., 2010
	Microbial contamination prediction	XGB, KNN, NB, SVM, NN and RF	Weather, hydrologic and land cover data	Source classification	Accuracy and AUC	XGBoost > RF > KNN > NN > SVM > NB	Wu et al., 2020
	Hidden features of bacterial communities unveiling	Alpha and Beta diversity analyses	Sequencing data of the bacterial community	Clustering properties of bacterial community	Unweighted UniFrac score	-	Pinto et al., 2014
		UniFrac	-	-	Unweighted/weighted UniFrac score	-	Lozupone et al., 2011; Bruno et al., 2018; Ling et al., 2018; Li et al., 2017

ORP, oxidation-reduction potential; EC, electrical conductance; ρ, resistivity; TDS, total dissolved solids; Sal, salinity; P, pressure; DO, dissolved oxygen; 24h rainfall, 24h accumulated rainfall; rep-PCR, repetitive element polymerase chain reaction; ARA, antibiotic resistance analysis; RMSEp, RMSE for posterior probability averaging estimator; RMSEc, RMSE for classification method estimator; ARCC, average rates of correct classification; NN, neural network.

Table 4 The applications of machine learning to detect accidental drinking water contamination

Topic	Task	Model	Inputs	Outputs	Metrics (Selected)	Performances (Selected)	Reference
To detect anomalies and contamination events in DW	Anomaly event detection	LR, LDA, SVM, ANN, DNN, RNN, and LSTM	T, C_{ClO_2}, pH, Redox, Leit, Turb, and Q	Event (Boolean)	F1-score	SVM: F1-score = 0.36	Muharemi et al., 2019
		LSTM				LSTM: F1-score = 0.80	Fehst et al., 2018
		LR, RF, XGB, xgbDART, and LSTM				LSTM: F1-score = 0.78	Qian et al., 2020
		LSTM and ARIMA	Turb and Leit		b.Acc, F1-score, and MCC	LSTM > ARIMA	Rodriguez-Perez et al., 2020
		Stacking-based and ANN	Cl_2, pH, Leit, T, TOC, and Turb		F1-score, R^2, and MSE	Stacking > ANN	Li et al., 2022
		GAN-based and MVE-based			FAR, F1-score, and EDR	GAN > MVE	Li et al., 2023
	Contamination event detection	SVM	Cl_2, EC, pH, T, TOC, and Turb	Three-class-event classification	Accuracy and EDR	Accuracy: 0.83-0.97	Oliker and Ostfeld, 2014
	DW classification: potable vs. contaminated	SVM	UV-absorbance readings	Contamination event	Confusion matrix	False alarm: 0.19	Asheri Arnon et al., 2019

LDA, linear discriminant analysis; RNN, recurrent neural network; LSTM, long short-term memory; C_{ClO_2}, chlorine dioxide concentration; Redox, redox potential; xgbDART, extreme gradient boosting with dropouts meet multiple additive regression trees; ARIMA, auto-regressive integrated moving average; b.Acc, balanced accuracy; MCC, Matthews correlation coefficient; Cl_2, total chlorine; GANs, generative adversarial networks; MVE, minimum volume ellipsoid; FAR, false alarm rate; EDR, event detection rate.