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

Feng Feng[1,2†], Yuanxun Zhang[1†], Zhenru Chen[3], Jianyuan Ni[4], 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 to assess 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 municipal 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, focusing 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 illustrate the distribution of ML models across research topics in safe drinking water supply. Finally, we discuss the challenges and outlooks for the applications of machine learning in safe drinking water supply. 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 other 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 and socioeconomic development ADDIN EN.CITE (. Public water systems are essential drinking water sources in modern society. For instance, in the United States (U.S.), over 90% of people obtain drinking water from approximately 150,000 public water systems (U.S. EPA, 2023). Public water systems provide water for human consumption through engineered water systems, including drinking water distribution systems (DWDSs) and building premise plumbing systems (Zhang and Lu, 2021b). Poor drinking water quality causes disease outbreaks and chronic diseases, 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 public health and the welfare of society (WHO, 2011).

Assessing and ensuring water quality in engineered water systems and related settings are complex (Li and Wu, 2019). Multiple ever-changing variables affect drinking water quality such as the quality of water sources, treatment processes and technologies, water pipe materials, distribution system configuration and length, natural disasters, and water stagnation in the pipes (Delpla et al., 2009; Li and Wu, 2019; Proctor et al., 2020). For instance, drinking water effluent at water utilities many be high but could deteriorate in engineered water systems because of microbial (re)growth and instruction, the formation of disinfection by-products (DBPs), pipe failures (e.g., breaks and leaks), and the detachment of heavy metals from pipes (Liu et al., 2016). Meanwhile, drinking water in engineered water systems and other settings have various hazardous agents such as harmful microbes (especially opportunistic pathogens or OPs) (Li et al., 2019; Zhang and Lu, 2021a), 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 the agents frequently fails to secure drinking water quality. For instance, increasing disinfectant residual concentrations in engineered systems suppresses microbial (re)growth but promotes the formation of DBPs (Zhang and Lu, 2021a). By contrast, reducing the dose of disinfectant residuals in engineered water systems can mitigate the DBP issue, but microbes (including pathogens) can thrive. A chlorine burn (i.e., a short conversion from chloramination and free chlorination) is an effective means to control nitrification in chloraminated engineered water systems (AWWA, 2013). However, chlorine burns significantly enhance DBP formation in water pipes, posing serious public health risks (Alexander et al., 2024; Alfredo, 2021; Allen et al., 2022).

Because of the complex nature of water quality in engineered water systems and related settings, assessing and ensuring drinking water quality using conventional means is challenging. Those 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, feasibility, 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 to enhance drinking water treatment and quality is an emerging

area of research and practice (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), the treatment processes (Li et al., 2021; Lowe et al., 2022; Ortiz-Lopez et al., 2022), and detection of quality anomaly (Dogo et al., 2019). However, a state-of-the-art, comprehensive review is missing that focuses on the applications of 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 understand and 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. We focus on engineered water systems but also cover other drinking-water-related settings such as source water and drinking water treatment processes. First, we introduce ML and common ML models. Then, we summarize recent progresses on the applications of ML in safe drinking water supply from the physical, chemical, microbiological, and temporal perspectives. Finally, we present the challenges and outlook for applying ML to ensure safe drinking waters supply. We focus on the applications of ML to access and control key factors and etiologies that affect drinking water quality and public health, such as water pipeline failures, DBPs, heavy metals, OPs, biofilms, and antimicrobial resistance genes (ARGs). This is the first review summarizing the applications

of ML in assessing and controlling water quality in municipal engineered water systems, while the applications of ML in other drinking-water-related settings are also discussed.

2. Machine learning primers

In the past decade, ML has driven significant 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), weather forecasting (Bi et al., 2023), and design-by-analogy (Jiang et al., 2021). Three key advancements in AI 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).

2.1. Categories of machine learning

On the basis of the 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 few 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 by exposing 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., 2019c; Zhou et al., 2019). Unsupervised learning is also useful in safe drinking water supply for tasks such as identifying the major factors affecting water quality and classifying data. Conversely, in safe drinking water supply, the application of semi-supervised learning is scarce, and reinforcement learning is unexplored.

Supervised learning acquires a mapping function from the input data to the corresponding output data on the basis of labeled 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 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 affecting water quality (Kazemi et al., 2023).

2.2. Workflow chart of machine learning

2.2.1 Problem definition

Defining the problem is critical in applying ML that converts 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).

2.2.2. Data collection

In any ML endeavor, the quality of data is the key to the success of the 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 developing ML models. Along with data collection, one needs to document the sources, methods, and potential biases associated with the data to ensure transparency and reproducibility.

2.2.3. 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 involves filtrating incorrect values, detecting outliers, removing duplicate entries, and converting data types. Normalization, an example of data transformation, scales data into a specific range, such as 0 to 1. Normalization ensures that each feature of the data contributes equally to model training, preventing any single attribute from disproportionately influencing the results. Organizing raw data entails structuring and optimizes the data for the specific ML algorithms to be applied.

2.2.4. Model training

The first step in model 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 model. The second step is to select an appropriate model on the basis of the nature of the problem. 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 with optimization techniques such as gradient descent.

2.2.5. 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 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 the 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).

3. Machine learning to ensure safe drinking water supply from the physical perspective

Drinking water production and distribution is critical to public health, socioeconomic growth, and urban development (Grey and Sadoff, 2007). Water demand prediction is a critical component in drinking water production. Traditional methods for estimating water demand often lead to either overestimation, resulting in high costs and waste of resources, or underestimation, resulting in water supply shortages during peak times (Donkor et al., 2014). Additionally, water systems suffer from losses because of leaks and inefficiencies with severe financial and operational consequences (Lambert et al., 1999; Lee and Schwab, 2005; Reis et al., 2023). ML models are a compelling solution to predict drinking water production and demand and identify system losses with greater precision than conventional approaches. This section examines how ML contributes to the

management of water resources, the prediction of water demand, and the assurance of sustainability and reliability in drinking water supply systems (Table 1).

3.1. Predicting water production and demand

Prediction of water production and demand is critical in safe drinking water supply. A hybrid model that combines genetic algorithms (GA) and GA artificial neural networks (GA-ANN) can predict drinking water production (Figure 1a) (Zhang et al., 2019c). The model uses temperature, COD, and electricity and chemical consumption as the inputs. The GA-ANN was trained and validated with monthly data from 45 water utilities across China. The R^2 (0.93) of GA-ANN is substantially higher than that of the ANN (0.71) when more training data are incorporated. GA not only optimizes the weights and biases to enhance prediction accuracy but also increases the tolerance to imprecision, uncertainties, and approximates in the inputs. Moreover, GA-ANN could effectively forecast fluctuations in water production for various scenarios, highlighting its feasibility in adjusting water treatment operations. To assess drinking water demand patterns, one can apply unsupervised learning algorithms (such as the hierarchical K-means algorithm) to raw time-series data of drinking water consumption (Leitão et al., 2019). In that algorithm, daily time-series demand with hourly records are the inputs within a 24-dimensional feature space to identify dense and distinctly separated temporal patterns of water demand. In contrast to directly clustering the water demand patterns, short-time water demand forecasting is more intriguing and has more practical merits in optimizing drinking water supply. A gated recurrent unit network (GRUN) predicts short-term water demand for different district metering areas for the upcoming 15 min and 24 h using a time-step of 15 min (Guo et al., 2018). The GRUN has a higher accuracy with a lower mean absolute percentage error (MAPE) between 2.06% and 2.46% than the conventional three-

dense-layered ANN (MAPE between 2.46% and 2.54%) and the seasonal autoregressive integrated moving average (SARIMA) model (MAPE between 2.57% and 2.85%) for the 15-min prediction. For the 24-h prediction, the GRUN also achieves more precise forecasts with MAPE between 4.33% and 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 Iran (Ghiassi et al., 2017). These models include a dynamic artificial neural network (DAN2), a focused time-delay neural network (FTDNN), and a KNN. Given its inherent design of adjusting dynamically to the data-driven learning, DAN2 is promising at time-series forecasting, catering to datasets characterized by evolving temporal patterns. Correspondingly, DAN2 achieves remarkable prediction accuracies (96% for daily, 99% for weekly, and 98% for monthly water demand forecasts) and outperforms FTDNN and KNN.

3.2. Monitoring pipeline integrity

Pipeline failures in engineered water systems cause significant water loss and contamination (Renwick et al., 2019). These failures can introduce harmful microbes and chemicals from the surroundings into distributed water. Addressing these issues requires precisely localizing failures in complex water networks, thoroughly assessing their impacts, and preventing future incidents. ML is a powerful tool to address these challenges, offering innovative solutions for detecting and predicting pipeline failures. This section reviews recent advancements in the application of ML to understand, detect, forecast, and mitigate pipeline failures, highlighting the role of deep learning and ensemble models in this application (Table 1).

To detect pipe burst locals, a study developed a burst location identification framework by fully-linear DenseNet (BLIFF) (Figure 1b) (Zhou et al., 2019), which 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. Using 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, ranging from 62.35% (highest probability pipe match) to 98.58% (top five pipes match), of BLIFF are two times those of the original DenseNet model. The remarkable improvement in the prediction accuracy is attributed to the linear-connection layer, which discerns global features in 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 work proposed an advanced meta-learning (AdvaML) model to predict the failure of water pipelines (Almheiri et al., 2021). AdvaML comprises an input layer with 33 neurons (mirroring the 33 input variables including pipe and climate data), four hidden layers, and an output layer that yields the failure/hazard index of a pipe. AdvaML forecasts the risk index associated with pipe failures and detects pivotal determinants of pipeline service life. Of these determinants, the number of traffic lanes and chlorine residual concentration are paramount, collectively contributing approximately 9% to the service life analysis of water pipes. Benefiting from its knowledge transfer from initial parameterization to the ultimate learning phase, AdvaML has commendable performance even with scant training data compared with cox-proportional hazards (Cox-PH), survival support vector machine (SSVM), and random survival forest (SRF).

While inherent system vulnerabilities cause pipeline failures, external factors exacerbate the issue (Fan et al., 2023). 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 predict the repercussions of calamities on water supply infrastructures (e.g., water pipelines) (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 water supply systems. 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, 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.

4. Machine learning to ensure safe drinking water supply from the chemical perspective

Numerous chemicals, such as DBPs, disinfectant residuals, and heavy metals can appear in drinking water, deteriorating water quality and affecting public health (Levallois and Villanueva, 2019; Valbonesi et al., 2021). Therefore, monitoring and controlling those chemicals are essential to ensuring drinking water quality. Conventional approaches to assess, monitor, and/or control

315 chemicals can be time-consuming, inaccurate, and costly. ML opens a promising venue for
316 monitoring and ensuring chemical drinking water quality. This section complies the applications
317 of ML to assess and control chemicals in drinking water with a focus on engineered water systems
318 (Table 2).
319
320 **4.1. Optimizing drinking water disinfection**
321 Drinking water disinfection is critical to ensuring microbial drinking water quality and
322 safeguarding public health (Zhang and Lu, 2021a). Disinfection is effective in killing pathogens,
323 impeding microbiological recontamination, and inhibiting biofilm development in drinking water
324 (Mazhar et al., 2020). Chlorine-based disinfectants, such as free chlorine (e.g., chlorine gas and
325 sodium hypochlorite), bound or combined chlorine (e.g., monochloramine), and chlorine dioxide,
326 are widely used in water treatment because of their cost-effectiveness and high efficiency (Jefri et
327 al., 2022; Zhang et al., 2018). Nonetheless, when these disinfectants interact with natural organic
328 matter (NOM) and anthropogenic compounds (such as pharmaceuticals and antimicrobials), they
329 generate DBPs such as trihalomethanes (THMs), haloacetic acids (HAAs), haloketones (HKs),
330 haloacetonitriles, halophenols, and halopropanoles (Favere et al., 2021; Xiao et al., 2023). DBPs
331 cause reproductive defects, carcer, and other serious health issues (Pandian et al., 2022; Zhou et
332 al., 2023a). Therefore, monitoring and controlling DBPs in drinking water is vital to public health
333 (He et al., 2021; Helte et al., 2023; Redondo-Hasselerharm et al., 2022). Conventional methods to
334 monitor DBPs require expensive equipment such as gas chromatography (GC) and liquid
335 chromatography (LC) combined with mass spectrometry (MS) and complicated pre-treatment
336 processes. Thus, those conventional methods are labor-intensive, costly, and time-consuming,
337 limiting the ability of water utilities to reduce DBP formation. By contrast, ML to monitor DBPs

in drinking water are accurate, efficient, inexpensive, and easy to handle (Table 2) (Balogun et al., 2021; Jia et al., 2021; Podgorski and Berg, 2022).

4.1.1. Predicting the formation of disinfection by-products from operation conditions and water quality metrics

Finding the optimal disinfectant dosages to minimize the levels of DBPs in finished water is crucial (He et al., 2021; Zhang and Lu, 2021a). Nevertheless, reaching this goal with traditional methods is time-consuming, expensive, and complex. Conversely, ML is effective in predicting the formation of DBPs, significantly reducing capital and human investment for DBP control. Common input parameters in these ML models are operational and water quality variables, such as water temperature, contract time, pH, absorbance of light at 254 nm (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, including ANN, SVM, and gene expression programming (GEP), to forecast THM formation in chlorinated river water on the basis of a 63-point dataset (Singh and Gupta, 2012). Specifically, pH, water temperature, contact time (t), bromide concentration, and DOC-normalized chlorine dose (Cl_2/DOC) are the inputs. SVM outperforms the other two models, exhibiting the highest R^2 and the lowest RMSE values. Furthermore, sensitivity analysis reveals that pH, hydraulic retention time (HRT), and water temperature are the top three contributors to DBP formation. In addition, radial basis function (RBF) based ANN

(RBF-ANN) can predict the formation of typical DBPs such as HAAs (Lin et al., 2020), THMs (Hong et al., 2020), and HKs (Deng et al., 2021) in drinking water. For instance, a study extracted 64 representative data points from the literature to predict HAA formation using pH, water temperature, DOC, UV_{254}, C_{resCl^-}, C_{Br^-}, $C_{NO_2^--N}$, and $C_{NH_4^+-N}$ as the inputs (Figure 2a) (Lin et al., 2020). RBF-ANN outperforms the linear and log-linear models by 21% and 47 % in accuracy, respectively. Therefore, RBF-ANN is promising in assessing DBP formation and optimizing disinfection. A follow-up study used 64 data points to train an RBF-ANN to predict THM formation (Hong et al., 2020). RBF-ANN achieves accuracies between 92% and 98% and regression coefficients between 0.76 and 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 trained with 63 data points of tap water predicts the formation of HK (Deng et al., 2021). Both RBF-ANN and back propagation (BP) ANN outstrip the linear and log-linear models with the RBF-ANN displaying higher accuracies in both internal and external validations. Another study applied a decision tree boost (DTB) model to predict the concentrations of THM4 and HAAs (Pan et al., 2023). The study correlated water quality parameters with mixed chlorine/chloramine species. The work then selected seven variables such as NH_2Cl, $NHCl_2$, 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 higher 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. Additionally, a study implemented multiple ML models to predict emerging DBPs in small water distribution systems across Canada by analyzing the data from eleven such networks

384 (Hu et al., 2023). The models use parameters such as water temperature, total chlorine residual,
385 DOC, turbidity, pH, conductivity, and UV_{254} to predict the concentrations of THMs, HAAs,
386 dichloroacetonitrile (DCAN), chloropicrin (CPK), and trichloropropanone (TCP). Among the
387 evaluated models, support vector regression (SVR) and Gaussian process regression (GPR) show
388 superior performance with SVR exhibiting the highest prediction accuracy ($R^2 = 0.94$) and stability
389 for DCAN and TCP, while GPR is optimal for predicting CPK ($R^2 = 0.92$).
390
391 **4.1.2. Assessing disinfection by-products using online spectroscopy**
392 Fluorescence spectroscopy is the preferred technique to monitor DBPs (Krasner et al., 2006;
393 Rodriguez et al., 2004). Fluorescence spectroscopy is sensitive in assessing the characteristics and
394 reactivity of NOM because of its minimal sample preparation requirement and short acquisition
395 time (Pifer and Fairey, 2012). However, the complex high-dimensional characteristics of
396 fluorescence spectroscopy make it difficult to predict DBP formation. The significant resources
397 and time required for DBP analysis through fluorescence spectroscopy restrict the capacity of
398 water utilities to reduce DBP formation. This situation has prompted the development of ML
399 models to assess DBP formation. Compared with traditional fluorescence spectroscopy, ML-
400 powered fluorescence spectroscopy can accurately assess DBP formation with limited resource
401 and time requirement.
402
403 Autoencoder-neural networks (AE-NN) can predict the concentrations of both THMs and HAAs
404 in river water from fluorescence spectra (Figure 2b) (Peleato et al., 2018). To manage the high
405 dimensionality of the fluorescence spectra, the researchers applied three dimension-reduction
406 techniques, including AE-NN, parallel factors analysis (PARAFAC), and PCA. Afterward, they

407 trained NN to identify fluorescence regions associated with DBP formation and to predict DBP
408 concentrations. The AE-NN model has superior predictive accuracies for THMs and HAAs,
409 achieving validation MAE values of 9.65 μg/L and 9.64 μg/L, respectively. These figures exceed
410 those of PCA, which has higher validation MAE values of 13.19 μg/L for THMs and 11.92 μg/L
411 for HAAs. Furthermore, the precision of the AE-NN model surpasses that of PARAFAC, which
412 has validation MEA values of 20.39 μg/L for THMs and 14.00 μg/L for HAAs. In addition,
413 convolutional neural networks (CNN) can predict DBP concentrations from fluorescence spectra
414 without extensive data pre-processing (Peleato, 2022). Compared with multilayer perceptron
415 (MLP) and dimensionality reduction techniques, CNN not only exhibits superior prediction
416 accuracies for THMs and HAAs but also identifies the fluorescence spectra regions highly
417 associated with DBP formation.
418
419 **4.1.3. Unraveling the formation mechanisms of disinfectant by-products**
420 ML is promising in predicting DBP formation using either water quality and operational
421 parameters or via online spectrum monitoring. However, even with the knowledge of DBP
422 concentration, controlling DBPs in drinking water remains costly and inefficient (Bond et al., 2011;
423 Rodriguez et al., 2004). An effective approach for DBP control is to remove DBP precursors and
424 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 DBP formation.
426
427 A multiple linear regression (MLR) model can predict the production of chloroform (a THM
428 compound) from organic precursors (Bond and Graham, 2017). Relying on 211 precursors from
429 22 studies, the MLR model uses 19 descriptors as the inputs and chloroform yield as the output.

The well-trained MLR model has a promising prediction accuracy with an R^2 value of 0.91 and an RMSE value of 8.93 mol/mol. Further chemical insights pinpoint that functional groups, such as hydroxyl, chlorine, and carboxyl groups, significantly affect chloroform formation. ML can also forecast the formation of HAAs from the interaction between organic precursors and free chlorine (Cordero et al., 2021). The training dataset comprises 283 organic compounds and 732 chemical descriptors as the inputs with HAA yield as the output. These organic compounds are converted into 2D and 3D chemical descriptors with their simplified molecular input line entry system (SMILE) strings used for ML compatibility. Three 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. RF 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.

4.1.4. Evaluating alternative drinking water disinfectants

Since chlorine-based disinfectants produce harmful DBPs, alternative disinfectants, such as ozone, for drinking water disinfection attract attention (Lin and Lin, 2024; Manasfi, 2021; Zhang et al., 2019a; Zhang et al., 2019b). 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 ozonated DBPs are a concern (Simpson and Mitch, 2022; Srivastav et al., 2020). Therefore, while ozone does not generate chlorinated DBPs, its use requires careful considerations for other toxic byproducts. This section summarize

how ML models have been developed and applied in the field of ozonation. ML can help control the formation of bromate and reduce micropollutants, microbes, and organic contaminants during ozonation.

Compared with MLR, ANN has two advantages 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, $C_{NH_4^+-N}$, bromide concentration (C_{Br^-}), pH, water 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 increased sample size because MLR cannot effectively process nonlinear components.

ML models based on routinely measured physical-chemical water quality parameters can predict the oxidation of micropollutants (such as pharmaceuticals and personal care products) during ozonation. For instance, RF can predict the oxidation of micropollutants during ozonation (Cha et al., 2021) (Figure 2c). That study introduced four distinct RF models, all 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 with a different resolution of FEEM data used as the unique predictors. Integrating FEEM data results in more accurate prediction of ozone exposures. The high-resolution FEEM data yield better predictions for micropollutant abatement (R^2 = 0.904; RMSE = 6.6%). However, the improvement in prediction accuracy when

using FEEM data is less substantial for predicting micropollutant abatement than for predicting the exposures of oxidants (i.e., ozone and hydroxyl radicals) during ozonation.

ML-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 R^2 values of greater than 0.75 and 0.78, respectively. Conversely, a DTB model has a higher R^2 value of greater than 0.97 (Gupta and Basant, 2016). A recent study compared several ML models including MLR, SVM, DT, RF, and deep neural network (DNN) for predicting $\log k_{O3}$ (Shi et al., 2022). Of these, RF has the highest effectiveness with a peak R^2 value of 0.91. RF has two primary benefits: robustness and a lower tendency of overfitting. On the other hand, DT has a complex structure and subsequently increases the overfitting risk. In addition, DNN, promising at recognizing nonlinear features, underperforms in predicting $\log k_{O3}$. A similar situation occurs when ML models predict the elimination of recalcitrant trace organic compounds (TOrCs) by ozonation for municipal wastewater reuse (Park et al., 2015). Specifically, ANN is susceptible to overfitting. Incorporating PCA into ANN creates a PC-ANN workflow, which addresses the overfitting issue. PCA transforms the input variables (Table 2) to linearly independent variables, thereby resolving the issue of collinearity among explanatory variables. The PC-ANN workflow (R^2 = 0.934) surpasses the standalone ANN (R^2 = 0.914) regarding predictive power.

4.2. Surveilling drinking water nitrification

Chloramine, a commonly used chlorine-based disinfectant, can maintain a higher level of residual while minimizing DBP formation (Shao et al., 2023; Shi et al., 2020). However, nitrification is a major concern in chloraminated engineered water systems (Allen et al., 2022; AWWA, 2013). During nitrification, ammonia-oxidizing microbes oxidize free ammonia to nitrite, and nitrite-oxidizing bacteria further oxidize nitrite to nitrate. Nitrification deteriorates water quality by destroying chloramine residuals, releasing free ammonia, promoting microbial (re)growth, and producing toxic nitrite and nitrate. Therefore, monitoring and controlling nitrification in engineered water systems is critical to ensuring drinking water quality and protecting public health.

ML is useful in detecting drinking water nitrification (Table 2). NB classifier, a supervised ML model, relies on biomass and microbiome datasets to detect nitrification in engineered water systems (Gomez-Alvarez and Revetta, 2020). After being trained with microbial indicators, the model has a binary classification accuracy of up to 85% with an AUC of 0.825 when distinguishing between nitrification and stable events. Nitrification can also be monitored using spectrum fingerprint since one can isolate the combined nitrate and nitrite from the total spectra. SVR was used and trained to predict the concentrations of nitrate and nitrite from nitrate/nitrite spectra at various wavelengths (Figure 2d) (Hossain et al., 2021). SVR negates the need for any chemical supplements, is easy to use, and can reach a high level of precision of up to ±0.01 mg N/L.

4.3. Monitoring and regulating heavy metals in drinking water

The detachment of heavy metals from water pipes (i.e., leaching) deteriorates drinking water quality in engineered water systems (Mays, 2000; Proctor et al., 2020). Heavy metals are toxic to human beings and significantly affect public health (Abd Elnabi et al., 2023; Fu and Xi, 2020).

ML is useful in assessing heavy metals in drinking water (Yaseen, 2021; Zhu et al., 2022). Therefore, this section summarizes the applications of ML in monitoring and regulating heavy metals in drinking water (Table 2).

4.3.1. Assessing heavy metal concentration and distribution

A continuous on-site, *in situ* system can estimate lead (Pb) concentration 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 Pb/L and an RMS prediction error of 0.71 μg Pb/L in the presence of interfering metals such as copper (Cu^{2+}), ferric (Fe^{3+}), and zinc (Zn^{2+}) ions. Other than estimating heavy metal concentration in individual samples, ML is promising in broader analytical applications. For instance, ML is useful in spatially interpolating environmental variables, significantly enhancing its performance (Li et al., 2011). This approach is valid in developing spatial interpolation maps depicting the concentrations of heavy metals such as Fe, Mn, Ni, Pb, and Zn in groundwater (i.e., a 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 heavy metal concentration and estimates health quotient indices (HQI) to offer a more refined risk assessment (Senoro et al., 2022). While integrating ML algorithms elevates the efficacy and robustness of spatial interpolation, traditional interpolation techniques are still critical 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, highlighting the enduring relevance and applicability of traditional interpolation methods (Tian et al., 2020).

4.3.2. Predicting heavy metal removal with porous materials

Adsorption is proficient in mitigating heavy metal contamination in drinking water (Joseph et al., 2019; Wołowiec et al., 2019). The adsorption of heavy metals by porous materials has highly stochastic, non-linear, and non-stationary dynamics coupled with redundancy (Bhagat et al., 2020). Therefore, ML is a preferred technique for analyzing the removal of heavy metals of porous media. Many ML techniques can enhance the precision and efficacy of predicting heavy metal adsorption dynamics by porous materials. Common predictors for those ML techniques are adsorbent dosage, operating temperature, contact time, and pH, whereas the output is the removal 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. A study used four tree-based ML models, including light gradient-boosting machine (LightGBM), XGB, gradient-boosted decision trees (GBDT), and RF, to predict the adsorption of arsenate [As(V)] by MOFs (Abdi and Mazloom, 2022). Among these models, LightGBM yields the most accurate and reliable prediction with R^2 and RMSE values of 0.996 and 2.069, respectively. The sensitivity analysis indicates that the adsorption process is adversely affected by the initial As(V) concentration and is directly influenced by the specific surface area and dosage of MOFs. ANN and symbiotic organisms search (SOS) algorithm can predict the removal of five heavy metals (Al, Cd, Co, Cu, Fe, and Pb) by two adsorbents, Chitosan and Chitosan-Montmorillonite nanocomposite (Hamidian et al., 2019). First, RBF-ANN has a higher prediction

accuracy than MLP-ANN for the two adsorbents and the five heavy metals. Second, when integrated with SOS algorithm, RBF-ANN facilitates the identification of optimal performance parameters and increases the adsorption performance than the experimental results. Finally, RBF-ANN outperforms Langmuir and Freundlich models when considering the five heavy metals and three operational parameters (pH, absorbent dose, and contact time).

5. Machine learning to ensure safe drinking water supply from the microbiological perspective

Microbial drinking after quality is essential to public health and the economic development of society (Abkar et al., 2024; Figueras and Borrego, 2010; Wen et al., 2020). The microbial community in drinking water is highly diverse and ever-changing spatiotemporally, especially in engineered water systems (Ashbolt, 2015; Jing et al., 2023; Zhou et al., 2023b). Therefore, monitoring and ensuring microbial drinking water quality is complex, time-consuming, and often ineffective with conventional assays. Microbes in drinking water encompass general heterotrophic bacteria, fecal contaminants, microbial indicators, protozoans (such as amoebae, ciliates, and slime molds), and OPs (Abkar et al., 2024; Siponen et al., 2024; Zhang et al., 2021b). Because of the complex nature of microbial drinking water quality, ML models are the preferred approaches for analyzing and ensuring microbial drinking water quality (Mahajna et al., 2022; Naloufi et al., 2021; Saboe et al., 2021; Zhu et al., 2022). This section summarizes the applications of ML in monitoring, predicting, and ensuring microbial drinking water quality (Table 3).

5.1. Surveilling and mitigating opportunistic pathogens

In municipal water, OPs are the most significant aspect of microbial drinking water quality because of their frequent occurrence, high concentration, high resistance to disinfectant residuals, and

proliferation within amoebae (Zhang and Lu, 2021b). OPs are the major disease-causing agents in drinking water and significantly affect the health of the end consumers. Therefore, closely monitoring OPs in drinking water, especially in engineered water systems, is critical to assessing drinking water quality and protecting public health. Dominant water-related OPs are *Legionella* (especially *L. pneumophila*), *Mycobacterium* (e.g., nontuberculosis mycobacteria or NTM and *M. avium* complex or MAC), *Pseudomonas aeruginosa*, *Vermamoeba vermiformis*, *Naegleria fowleri*, and *Acanthamoeba* (Donohue et al., 2019; Isaac and Sherchan, 2020; Lytle et al., 2021). Among them, *Legionella* is the most important OP. In addition, compared with conventional microbial drinking water quality indicators such as fecal coliforms and , *Legionella* is a better candidate to indicate microbial drinking water quality (Zhang and Lu, 2021b). Therefore, this section focuses on the applications of ML in monitoring and controlling *Legionella* in municipal water (Table 3).

Studies using ML to assess the risks of OPs in drinking water 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 water temperature as the inputs, that study presented a NN-based simulator relying an approximate reasoning architecture (NARA) neuro-fuzzy system to predict and simulate water tank temperature. 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 temperature with an accuracy exceeding 97%. A recent study integrated both unsupervised and supervised ML to correlate the spread of *Legionella* with environmental variables in retirement homes, health-related facilities, tourism-related buildings, and swimming-pools s from 2002 to 2019 in Italy (Brunello et al., 2022). That study used an unsupervised ML algorithm to identify the spatiotemporal distribution of atypical *Legionella*

through an ordinal regression model. The results indicate how the distribution is correlated with the types of healthcare facilities. The propagation of *Legionella* and both the nature of the facilities and broader geographical characteristics have strong correlations. Hospitals have the highest contamination cluster locations 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 Shapley values evaluates the contribution of each predictor to the final classification. The Shapley values 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.

5.2. Analyzing drinking water microbial communities

The microbial community in drinking water is complex with ever-changing structure and significant public health implications (Abkar et al., 2024; Zhang et al., 2021a). Assessing the structure and composition of the microbial community helps understand and ensure microbial drinking water quality. Conventional approaches can monitor the microbial community in drinking water, such as phenotypic/genotypic matching, molecular marking, high-throughput metagenomic sequencing, and microbial and chemical indication. However, these methods have limitations in terms of cost, time, and spatial/temporal coverage. ML, on the other hand, can overcome those

limitations and are suitable for analyzing, monitoring, and source-tracking microbial communities in municipal water (Table 3).

When using the NB theorem to estimate the distribution of microbes in water sources, one could apply either maximum posterior probability (evaluation metrics: $RMSE_c$) (Ritter et al., 2003) or direct averaging posterior probability (evaluation metrics: $RMSE_p$) (Greenberg et al., 2010). Direct averaging of the source posterior probability yields more precise source distribution estimates with $RMSE_c$ being significantly lower than $RMSE_p$. The more precise source distribution estimates are because direct estimation bypasses the information loss that typically happens when frequencies are first classified and then averaged (Greenberg et al., 2010). SourceTracker as a ML tool estimates the proportion of contaminants (Knights et al., 2011). SourceTracker employs the Gibbs sampling technique within a Bayesian framework and is more efficient than both NB- and RF-based source tracking methods (Smith et al., 2010). The superior performance of SourceTracker is because it can handle ambiguity in the source and sink distributions and can model a sink sample as a blend of various sources. SourceTracker can track the origin of bacteria in drinking water and water sources (Liu et al., 2018). For instance, a study developed six ML models (XGBoost, KNN, NB, SVM, NN, and RF) to predict microbial contamination in a watershed in California using data on land cover, weather, and hydrologic variables (Wu et al., 2020). That study used SourceTracker to generate ground-truth data for training purposes. Out of the six models, XGBoost outperforms the other models in terms of accuracy and AUC (average AUC = 0.88) when tracking the primary sources of microbial contamination in the watershed.

Unsupervised ML can unveil hidden features, trends, or patterns in bacterial communities in 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 an unsupervised ML tool uses principal coordinates analysis (PCoA) coupled beta diversity measure to analyze the differences among microbial communities (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).

5.3. Detecting drinking water parasites

Cryptosporidium and *Giardia* are protozoan parasites in municipal water with substantial public health risks by causing cryptosporidiosis and giardiasis, respectively (CDC, 2021a, 2021b). These pathogens are highly resistant to disinfectants, challenging drinking 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 ML models in monitoring *Cryptosporidium* and *Giardia* in drinking water (Table 3).

ML to detect *Cryptosporidium* and *Giardia* are robust and precise. 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 have 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 in surface water and drinking water (Ligda et al., 2020), offering a valuable tool to control waterborne diseases.

ParasNet (Xu et al., 2020) 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 time. For instance, ParasNet can reach a detection accuracy of above 95.6% with analysis speeds of up to 100 frame-per-second (fps) on embedded Jetson TX2 platform. MCellNet (Luo et al., 2021), another image classification pipeline, uses a DNN optimized from MobileNetV2 (Sandler et al., 2018) to recognize objects. 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 fast analysis of MCellNet are due to the cascading six IRBs.

An alternative statistical model uses linear discriminant function analysis (LDFA) to predict the appearance of *Cryptosporidium* and *Giardia* in drinking water (Ligda et al., 2020). That model uses microbiological, physicochemical, and meteorological parameters 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 *Cryptosporidium* and *Giardia*, respectively.

5.4. Assessing biofilm development in engineered water systems

Biofilms are complex microbial communities adhering to surfaces (Flemming and Wingender, 2010). In engineered water systems, biofilms develop on the inner wall of water pipes and pose significant risks. Drinking water biofilms harbor pathogenic and antimicrobial resistant bacteria,

corrode pipes, reduce water flow rate, deteriorate water quality, and increase the costs and complexity of water distribution (Simoes and Simões, 2013; Wingender and Flemming, 2011; Zhou et al., 2023c). Monitoring and controlling biofilm formation in engineered water systems with conventional approaches are challenging because assessing the biofilms inside of the water pipes is difficult and biofilms are protected by an extracellular matrix (Flemming et al., 2023; Karygianni et al., 2020). ML provides novel solutions for assessing and controlling drinking water biofilms (Table 3).

ML models studying biofilm development in engineered water systems use relevant physical (such as water age, flow velocity, hydraulic regime, pipe material, and pipe age) factors to assess the dynamics of biofilm development, where the heterotrophic plate count (HPC) is the output. Established ML algorithms are preferred models to study the dynamics of biofilm development, such as NB, RT, and RF (Ramos-Martínez et al., 2014, 2016). These algorithms have high prediction accuracy and provide an in-depth understanding of the impact of physical factors on biofilm development in engineered water systems. For instance, a Bagging naïve Bayesian tree (B-NBT) model proposes optimal flow velocities for different types of pipes to mitigate biofilm development (Ramos-Martínez et al., 2014). Predicted biofilm development probabilities show that, to control biofilm accumulation, water utilities need to avoid cement pipes, implement medium- or high-flow velocities in metal pipes, and sustain water ages above 0.035 in plastic pipes. The 'water age' is a synthetic index derived from the normalized HRT and the distance from the disinfection source.

Recent studies have enabled more detailed, single-cell level analyses and predictions of biofilm development in engineered water systems (Berne et al., 2018). In addition, studies have expanded ML algorithms to incorporate deep learning (Jelli et al., 2023; Weigert et al., 2020). These innovative approaches can enhance the assessment of biofilm dynamics in various settings including engineered water systems. A recent work (Jelli et al., 2023) optimized StarDist (Weigert et al., 2020), a cutting-edge CNN-based segmentation algorithm, to segment single cells in biofilms, track cell lineages, and measure single-cell growth rates (Figure 3a). First, an iterative semi-automated annotation workflow was developed to accelerate the annotation of bacterial cells in 3D images for training data. Then, a new post-processing algorithm (StarDist OPP) that reconstructs the bacterial cell shapes were developed to increase the accuracy of bacterial segmentation. The second step was to overcome the embedded limitations of the StarDist non-maximum-suppression post-processing that considers only the shape information in the voxel with the highest assigned label probability. StarDist OPP achieves unprecedented accuracy in biofilm segmentation, surpassing other algorithms under scrutiny, such as Cellpose (Stringer et al., 2021), a multi-class U-Net (Zhang et al., 2020), and BCM3D 2.0 (Zhang et al., 2022). Finally, the accurate single-cell segmentation results were used to track cell lineages and to spatiotemporally measure single-cell growth rates.

5.5. Analyzing the risks and tracking the sources of antimicrobial resistance

Antimicrobials have been extensively used since the 1920s in the medical industry, animal husbandry, consumer goods production, and other fields (Chang et al., 2015; Hutchings et al., 2019; Prescott, 2017; WHO, 2021). However, a significant portion of antimicrobials consumed by humans and animals is not metabolized but excreted, entering waterbodies. Antimicrobials in

748 waterbodies contributes to the development of antibiotic-resistant bacteria (ARB) and ARGs,
749 posing a serious threat to aquatic ecosystems and public health by causing antimicrobial resistance
750 (AMR) (Roca et al., 2015; Walesch et al., 2023). The advent of ML has introduced novel
751 methodologies that enhance our ability to assess the risks and track the sources of AMR in in
752 drinking water with unprecedented precision (Table 3).
753
754 **5.5.1. Assessing the risks of antimicrobial resistance**
755 AMR presents challenges in treatment, costs, and mortality rates compared with non-resistant
756 infections in humans and animals. Despite the efforts to assess and control AMR, challenges
757 remain because of uncertainties in data acquisition and dose-response mechanisms. To streamline
758 the estimation process and minimize labor, a study developed three ML models (LR, DT, and RF)
759 to rapidly predict the relative risks of AMR in drinking water (Wu et al., 2022). These models take
760 four land-use type factors (residential, urban, green, and agriculture) and eleven environmental
761 factors (water temperature, pH, oxidation-reduction potential, electrical conductance, resistivity,
762 total dissolved solids, salinity, pressure, DO, turbidity, and 24-h accumulated rainfall) as the inputs.
763 Given limitations in data, particularly in field data, employing classification over direct regression
764 for relative risk assessment is more robust. That study used a binary classification framework,
765 labeling relative risk scores above the median as 1 for relatively high risks and scores below the
766 median as 0 for relatively low risks. Compared with LR and DT, RF has the highest accuracy
767 (0.86), precision (1.0), recall (0.75), F1 (0.86), and AUC (0.88). Finally, the feature importance
768 analysis from RF reveals that green (areas designated for natural vegetation, parks, forests, or other
769 green spaces), DO, and pH are the top-three significant influencing factors of AMR in drinking
770 water.

771

772 **5.5.2. Tracking the sources of antimicrobial resistance**

773 Leveraging the capabilities of ML for environmental monitoring, recent studies have harnessed
774 SourceTracker, an ML tool based on a Bayesian classification algorithm (Knights et al., 2011), to
775 identify the sources of ARGs in drinking water and water sources. A study used SourceTracker to
776 identify the complex sources of ARGs and assessed their contributions to ARG pollution in a peri-
777 urban river (Chen et al., 2019). The results show that the discharge from sewage plants was the
778 largest contributor of ARGs (81.6% to 92.1%) in the river sediments. Another work used
779 SourceTracker to monitor the presence of ARGs in household drinking water and tracked their
780 origins back to anthropogenic sources, highlighting the significant impact of human activities on
781 drinking water quality (Figure 3b) (Wang et al., 2023). The data generated by SourceTracker have
782 a strong Pearson correlation (r = 0.98) with the corresponding expected proportion by artificial
783 source inputs. Source tracking analysis from that study indicates that a significant proportion of
784 ARGs (37.1%) was from anthropogenic sources, especially wastewater effluent.

785

786 **6. Machine learning to ensure safe drinking water supply from the temporal perspective**

787 The increasing use of ML in safeguarding drinking water quality has led to the development of
788 innovative approaches to detect drinking water quality from the temporal perspective (Table 4)
789 (Zhong et al., 2021). The term 'temporal' refers to the time-related applications of ML to track,
790 predict, and mitigate contamination events in drinking water as they unfold over time. By
791 examining the temporal patterns and trends of contamination events, we can bolster the predictive
792 power and responsiveness of ML to ensure effective measures against accidental contamination in
793 drinking water. This section explores the advancements, versatility, and potential of ML in

794 revealing drinking water quality from the temporal perspective focusing on accidental drinking
795 water contamination events.
796
797 Three studies developed multiple ML approaches to detect anomalies in the drinking water quality
798 datasets from GECCO Industrial Challenges (GECCO IC) (Fehst et al., 2018; Muharemi et al.,
799 2019; Qian et al., 2020). These studies pinpointed shifts or spotted anomalies in drinking water
800 quality over time. Various parameters such as pH, redox potential, electric conductivity, turbidity,
801 and chlorine dioxide concentration are the predictors, whereas events in Boolean form are the
802 outputs. A study developed SVM, DNN, long short-term memory (LSTM), recurrent NN (RNN),
803 LR, simple NN, and linear discriminant analysis (LDA) to detect water quality anomaly in the
804 dataset from 2017 GECCO IC (Muharemi et al., 2019). SVM shows the highest performance with
805 an F1-score of 0.99 in cross-validation. Nevertheless, all the models have poor performance with
806 the unseen test dataset with a maximum F1-score of 0.36. In the other two studies focusing on the
807 dataset from GECCO IC 2018, LSTM demonstrates superior results, scoring a higher F1-score
808 than traditional models such as LR and SVM with F1-scores of 0.80 and 0.78, respectively (Fehst
809 et al., 2018; Qian et al., 2020).
810
811 The existing research, including the three key studies using the GECCO IC datasets (Fehst et al.,
812 2018; Muharemi et al., 2019; Qian et al., 2020), has made significant progress in understanding
813 anomaly in drinking water through ML. A notable trend in recent research is using real-time or
814 online applications to reflect a crucial evolution toward practical, real-world implementations.
815 Specifically, a study implemented an LSTM-based approach to detect anomalies in water quality
816 focusing on turbidity and conductivity (Rodriguez-Perez et al., 2020) (Figure 4a). 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 suitable in identifying long-term anomalies linked to gradual changes. Notably, the LSTM-based approach surpasses regression-based autoregressive integrated moving average (ARIMA) in detecting these long-term anomalies. Another study introduced an innovative stacking ensemble model designed for contamination detection (Figure 4b) (Li et al., 2022). The model uses various water quality parameters such as total chlorine, pH, electrical conductivity, water temperature, TOC, and turbidity. That model integrates multiple predictors into 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 has predictors such as ANN, SVM with a linear kernel, linear regressor, extra trees, uniform weighted KNN, and an RF meta-predictor. The ensemble demonstrates superior performance in detecting contamination compared with an ANN benchmark method, achieving higher accuracy, lower false positive rates, and improved F1-scores.

However, these models focus on single-site, one-dimensional time series data, neglecting the spatial relationships inherent in multi-site sensor data. This limitation could increase false alarm rates, particularly under conditions of high hydraulic variability. To address this issue, a follow-up study proposed a novel unsupervised, generative-adversarial-networks-based (GAN-based) multivariate method to detect multi-site contamination events (Figure 4c) (Li et al., 2023). That method effectively captures spatiotemporal patterns by transforming water quality data from single and multiple sites into superimposed images. The GAN-based model, having a generator and a discriminator, evaluates the degree of abnormality at each time step by generating anomaly scores.

The generator is trained to map historical image data to expected current images, while the discriminator differentiates between generated and actual normal images. That method is benchmarked against a multivariate unsupervised method using a minimum-volume-ellipsoid (MVE)-based event detection model (Oliker and Ostfeld, 2014). That method demonstrates superior performance in all contamination scenarios, including enhanced detection rates and reduced false alarms, particularly for sensor groups positioned at varying distances from the contamination source. Another unique ML approach can rapidly signal potential contamination risks in drinking water (Asheri Arnon et al., 2019). That approach uses an algorithm for the early detection of drinking water contamination against an unpredictable stochastic background. By extracting key features from the spectrophotometric characteristics of water, the algorithm can effectively identify contamination using a unique affinity measure (Asheri-Arnon et al., 2018). The measure compares the absorbance spectra of different water sources, thereby amplifying the feature dissimilarity between portable and contaminated water, followed by processing via SVM and post-processing. That chain of data processing generates a reliable early warning system 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 to obtain minimal false positives.

7. Machine learning model distribution in safe drinking water supply

We provide a macroscopic visual illustration to elucidate the distribution of ML models across research topics in safe drinking water supply (Figure 5). To facilitate a clear and 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 ML models are the top two frequently implemented in safe drinking water supply. NN-based models have significant applications in managing the production and demand of drinking water and accessing and controlling DBPs. 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 NN-based models. Water management and DBP assessment often involve multifaceted, nonlinear, and high-dimensional data that demand robust modeling (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, NN-based models 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 NN-based models in accommodating changing inputs makes them promising in assessing the dynamic nature of drinking water quality. The wide applications of NN-based models in safe drinking water supply are due to this harmonious fit between the challenges posed by these fields and the advantages of these models.

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). The suboptimal performances do not undermine the value of regression-based models. However, their linear or predefined non-linear structures may limit their effectiveness, especially when compared with the adaptive and intricate abilities of NN-based models.

The superior performance of NN-based models is widely acknowledged (Goodfellow et al., 2016). These general strengths become pertinent when NN-based models are applied to the complexities of drinking water research. First, unlike regression-based models which are limited by their linear or defined non-linear structures, NN-based models capture intricate, non-linear associations. Second, the mutable architecture of NN-based models allows them to modify their framework during training, optimizing alignment with the inherent data distribution. Lastly, given abundant data, NN-based models excel in discerning subtle data patterns because of their proficiency in processing high-dimensional input attributes, whereas regression-based models may have tendencies of underfitting. This proficiency of NN-based models is further enhanced by the use of techniques such as grid search for hyperparameter optimization, particularly crucial in fine-tuning the performance of NN-based models 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 (LeCun et al., 1989; Lecun et al., 1998). Therefore, we list the CNN-based models out of the broader NN category (Figure 5). The practical implications of CNN-based models are evident in drinking water research: They can interpret 2D fluorescence spectra and predict the formation of DBPs during 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 drinking 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 offers 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, and/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-based 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 5). These models are more deterministic and often rooted in foundational principles and established, theoretical, and/or empirical 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 (Li and Wu, 2019). 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).

8. Challenges and outlooks

While ML has made significant progress 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 explored. 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 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, supply, and distribution. This knowledge gap impedes the effective deployment of ML in enhancing safety drinking water supply. Addressing this dichotomy is

beneficial and essential, necessitating educational and collaborative efforts to build a shared understanding and to develop interdisciplinary skillsets.

Further complicating the matter is the absence of standardized toolkits tailored to safe drinking water supply. 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 is affected by numerous spatiotemporal factors, 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 safe drinking water supply.

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 global challenges is critical to 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 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 tool 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, microbiological, and temporal 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, assess AMR risks, and study microbial communities in municipal water, especially in engineered water systems. In addition, ML is a useful tool in assessing drinking water quality from the temporal perspective, especially in detecting accidental

drinking water contamination. Taken together, ML is feasible in assessing and ensuring drinking water quality with a great potential to mitigate the global water crisis.

Acknowledgment

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

Declaration of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abd Elnabi, M.K.; Elkaliny, N.E.; Elyazied, M.M.; Azab, S.H.; Elkhalifa, S.A.; Elmasry, S.; Mouhamed, M.S.; Shalamesh, E.M.; Alhorieny, N.A.; Abd Elaty, A.E. **2023**. Toxicity of heavy metals and recent advances in their removal: a review. *Toxics* 11(7), 580.

Abdi, J.; Mazloom, G. **2022**. Machine learning approaches for predicting arsenic adsorption from water using porous metal-organic frameworks. *Scientific Reports* 12, 16458.

Abkar, L.; Moghaddam, H.S.; Fowler, S.J. **2024**. Microbial ecology of drinking water from source to tap. *Science of the Total Environment* 908, 168077.

Adeyemo, F.E.; Singh, G.; Reddy, P.; Bux, F.; 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.; 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.; Elshafie, A. **2019**. Machine learning methods for better water quality prediction. *Journal of Hydrology* 578, 124084.

Alexander, M.T.; Woodruff, P.; Mistry, J.H.; Buse, H.Y.; Lytle, D.A.; Pressman, J.G.; Wahman, D.G. **2024**. Evaluation of distribution system water quality during a free chlorine conversion [Manuscript submitted for publication].

Alfredo, K. **2021**. The "Burn": water quality and microbiological impacts related to limited free chlorine disinfection periods in a chloramine system. *Water Research* 197, 117044.

Aliashrafi, A.; Zhang, Y.; Groenewegen, H.; 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.

Allen, J.M.; Plewa, M.J.; Wagner, E.D.; Wei, X.; Bokenkamp, K.; Hur, K.; Jia, A.; Liberatore, H.K.; Lee, C.-F.T.; Shirkhani, R. **2022**. Feel the burn: disinfection byproduct formation and cytotoxicity during chlorine burn events. *Environmental Science & Technology* 56(12), 8245–8254.

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.; Consortium, I.D. **2023**. Discovery of drug-omics associations in type 2 diabetes with generative deep-learning models. *Nature Biotechnolgy* 41(3), 399-408.

Almalioglu, Y.; Turan, M.; Trigoni, N.; Markham, A. **2022**. Deep learning-based robust positioning for all-weather autonomous driving. *Nature Machine Intelligence* 4(9), 749-760.

Almheiri, Z.; Meguid, M.; Zayed, T. **2021**. Failure modeling of water distribution pipelines using meta-learning algorithms. *Water Research* 205, 117680.

Ashbolt, N.J. **2015**. Microbial contamination of drinking water and human health from community water systems. *Current Environmental Health Reports* 2, 95-106.

Asheri-Arnon, T.; Ezra, S.; Fishbain, B. **2018**. Contamination Detection of Water with Varying Routine Backgrounds by UV-Spectrophotometry. *Journal of Water Resources Planning and Management* 144(9), 04018056.

Asheri Arnon, T.; Ezra, S.; 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.

Ates, N.; Civelekoglu, G.; Kaplan-Bekaroglu, S.S. **2022**. Management strategies for minimising DBPs formation in drinking water systems. In: Bahadir, M. and Haarstrick, A. (Eds.) *Water and Wastewater Management: Global Problems and Measures*. pp. 67-82. Springer International Publishing, Cham.

Avni, N.; Fishbain, B.; Shamir, U. **2015**. Water consumption patterns as a basis for water demand modeling. *Water Resources Research* 51(10), 8165-8181.

AWWA **2013**. *Nitrification Prevention and Control in Drinking Water*, 2nd Edition. American Water Works Association, Denver, Colorado, U.S.A.

Bagriacik, A.; Davidson, R.A.; Hughes, M.W.; Bradley, B.A.; 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.; 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.; 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.; 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.; 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.; Ostfeld, A. **2023**. Coupling machine learning and agent-based modeling to characterize contamination sources in water distribution systems for changing demand regimes. In: *World Environmental and Water Resources Congress 2023*. pp. 10.1061/9780784484852.9780784484082.

Berne, C.; Ellison, C.K.; Ducret, A.; Brun, Y.V. **2018**. Bacterial adhesion at the single-cell level. *Nature Reviews Microbiology* 16(10), 616-627.

Bhagat, S.K.; Tung, T.M.; 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.; 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.; Jefferson, B. **2011**. Treatment of disinfection by-product precursors. *Environmental Technology* 32(1-2), 1-25.

1099 Bond, T.; Templeton, M.R.; Graham, N. **2012**. Precursors of nitrogenous disinfection by-products
1100 in drinking water—A critical review and analysis. *Journal of Hazardous Materials* 235-236,
1101 1-16.
1102 Bond, T.; Graham, N. **2017**. Predicting chloroform production from organic precursors. *Water*
1103 *Research* 124, 167-176.
1104 Bradley, A.P. **1997**. The use of the area under the ROC curve in the evaluation of machine learning
1105 algorithms. *Pattern Recognition* 30(7), 1145-1159.
1106 Breiman, L. **2001**. Random Forests. *Machine Learning* 45(1), 5-32.
1107 Brunello, A.; Civilini, M.; De Martin, S.; Felice, A.; Franchi, M.; Iacumin, L.; Saccomanno, N.;
1108 Vitacolonna, N. **2022**. Machine learning-assisted environmental surveillance of Legionella: A
1109 retrospective observational study in Friuli-Venezia Giulia region of Italy in the period 2002–
1110 2019. *Informatics in Medicine Unlocked* 28, 100803.
1111 Bruno, A.; Sandionigi, A.; Bernasconi, M.; Panio, A.; Labra, M.; Casiraghi, M. **2018**. Changes in
1112 the drinking water microbiome: Effects of water treatments along the flow of two drinking
1113 water treatment plants in a urbanized area, Milan (Italy). *Frontiers in Microbiology* 9, 2557.
1114 CDC **2021a**. Parasites - *Cryptosporidium* (also known as "Crypto").
1115 CDC **2021b**. Parasites - *Giardia*.
1116 Cha, D.; Park, S.; Kim, M.S.; Kim, T.; Hong, S.W.; Cho, K.H.; Lee, C. **2021**. Prediction of oxidant
1117 exposures and micropollutant abatement during ozonation using a machine learning method.
1118 *Environmental Science & Technology* 55(1), 709-718.
1119 Chang, Q.; Wang, W.; Regev-Yochay, G.; Lipsitch, M.; Hanage, W.P. **2015**. Antibiotics in
1120 agriculture and the risk to human health: how worried should we be? *Evolutionary Applications*
1121 8(3), 240-247.
1122 Chen, H.; Bai, X.; Li, Y.; Jing, L.; Chen, R.; Teng, Y. **2019**. Source identification of antibiotic
1123 resistance genes in a peri-urban river using novel crAssphage marker genes and metagenomic
1124 signatures. *Water Research* 167, 115098.
1125 Chowdhury, S.; Mazumder, M.A.J.; Al-Attas, O.; Husain, T. **2016**. Heavy metals in drinking water:
1126 occurrences, implications, and future needs in developing countries. *Science of the Total*
1127 *Environment* 569-570, 476-488.
1128 Cordero, J.A.; He, K.; Janya, K.; Echigo, S.; Itoh, S. **2021**. Predicting formation of haloacetic acids
1129 by chlorination of organic compounds using machine-learning-assisted quantitative structure-
1130 activity relationships. *Journal of Hazardous Materials* 408, 124466.
1131 Craun, G.F.; Brunkard, J.M.; Yoder, J.S.; Roberts, V.A.; Carpenter, J.; Wade, T.; Calderon, R.L.;
1132 Roberts, J.M.; Beach, M.J.; Roy, S.L. **2010**. Causes of outbreaks associated with drinking water
1133 in the United States from 1971 to 2006. *Clinical Microbiology Reviews* 23(3), 507-528.
1134 Daniel, I.; Abhijith, G.R.; Kadinski, L.; Ostfeld, A.; Cominola, A. **2023**. A machine learning-based
1135 surrogate model for coupled hydraulic and water quality simulation in water distribution
1136 networks. In: *World Environmental and Water Resources Congress 2023*. pp.
1137 10.1061/9780784484852.9780784484077.
1138 De Jesus, K.L.M.; Senoro, D.B.; Dela Cruz, J.C.; Chan, E.B. **2021**. A hybrid neural network-
1139 particle swarm optimization informed spatial interpolation technique for groundwater quality
1140 mapping in a small island province of the Philippines. *Toxics* 9(11), 273.
1141 Delpla, I.; Jung, A.-V.; Baures, E.; Clement, M.; Thomas, O. **2009**. Impacts of climate change on
1142 surface water quality in relation to drinking water production. *Environment International* 35(8),
1143 1225-1233.

Deng, Y.; Zhou, X.; Shen, J.; Xiao, G.; Hong, H.; Lin, H.; Wu, F.; 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.; 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.

Donkor, E.A.; Mazzuchi, T.A.; Soyer, R.; Roberson, J.A. **2014**. Urban water demand forecasting: Review of methods and models. *Journal of Water Resources Planning and Management* 140(2), 146-159.

Donohue, M.; King, D.; Pfaller, S.; 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.

Ewuzie, U.; Bolade, O.P.; Egbedina, A.O. **2022**. Application of deep learning and machine learning methods in water quality modeling and prediction: a Review. In: Marques, G. and Ighalo, J.O. (Eds.) *Current Trends and Advances in Computer-Aided Intelligent Environmental Data Engineering*. pp. 185-218. Elsevier Inc., San Diego, California, U.S.A.

Fan, X.; Zhang, X.; Yu, A.; Speitel, M.; 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.; 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.; Boon, N. **2021**. Safeguarding the microbial water quality from source to tap. *NPJ Clean Water* 4, 28.

Feng, S.; Sun, H.; Yan, X.; Zhu, H.; Zou, Z.; Shen, S.; Liu, H.X. **2023**. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature* 615(7953), 620-627.

Figueras, M.J.; Borrego, J.J. **2010**. New perspectives in monitoring drinking water microbial quality. *International Journal of Environmental Research and Public Health* 7(12), 4179-4202.

Flemming, H.-C.; Wingender, J. **2010**. The biofilm matrix. *Nature Reviews Microbiology* 8(9), 623-633.

Flemming, H.-C.; van Hullebusch, E.D.; Neu, T.R.; Nielsen, P.H.; Seviour, T.; Stoodley, P.; Wingender, J.; Wuertz, S. **2023**. The biofilm matrix: multitasking in a shared space. *Nature Reviews Microbiology* 21(2), 70-86.

Friedman, J.H. **2001**. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* 29(5), 1189-1232.

Fu, Z.; Xi, S. **2020**. The effects of heavy metals on human metabolism. *Toxicology mechanisms and methods* 30(3), 167-176.

Ghiassi, M.; Fa'al, F.; 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.

Ghobadi, F.; 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.; Kumar, M. **2018**. Occurrence and fate of emerging contaminants in water environment: a review. *Groundwater for Sustainable Development* 6, 169-180.

Gomez-Alvarez, V.; Revetta, R.P. **2020**. Monitoring of nitrification in chloraminated drinking water distribution systems with microbiome bioindicators using supervised machine learning. *Frontiers in Microbiology* 11, 2254.
Gong, J.; Guo, X.; Yan, X.; Hu, C. **2023**. Review of urban drinking water contamination source identification methods. *Energies* 16(2), 705.
Gonzalez, S.; Lopez-Roldan, R.; 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.; Courville, A. **2016**. *Deep Learning*. MIT Press.
Greenberg, J.; Price, B.; Ware, A. **2010**. Alternative estimate of source distribution in microbial source tracking using posterior probabilities. *Water Research* 44(8), 2629-2637.
Grey, D.; 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.; 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.; 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.
Hamidian, A.H.; Esfandeh, S.; Zhang, Y.; 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.; Friedman, J.H. **2009**. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition. Springer, New York.
He, L.; Bai, L.; Dionysiou, D.D.; Wei, Z.; Spinney, R.; Chu, C.; Lin, Z.; 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.; 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* 115(12), 1597-1604.
Henrique Alves Ribeiro, V.; 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* 23(1), 447-474.
Hong, H.; Zhang, Z.; Guo, A.; Shen, L.; Sun, H.; Liang, Y.; Wu, F.; 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.; Hewa, G.A. **2021**. Development of an optical method to monitor nitrification in drinking water. *Sensors* 21(22), 7525.
Hu, G.; Mian, H.R.; Mohammadiun, S.; Rodriguez, M.J.; Hewage, K.; 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.; He, Q. **2021**. Machine learning in natural and engineered water systems. *Water Research* 205, 117666.

1234 Huang, Y.; Li, T.; Zheng, S.; Fan, L.; Su, L.; Zhao, Y.; Xie, H.B.; Li, C. **2020**. QSAR modeling for
1235 the ozonation of diverse organic compounds in water. *Science of the Total Environment* 715,
1236 136816.
1237 Hutchings, M.I.; Truman, A.W.; Wilkinson, B. **2019**. Antibiotics: past, present and future. *Current*
1238 *Opinion in Microbiology* 51, 72-80.
1239 Isaac, T.S.; Sherchan, S.P. **2020**. Molecular detection of opportunistic premise plumbing pathogens
1240 in rural Louisiana's drinking water distribution system. *Environmental Research* 181, 108847.
1241 Jefri, U.H.N.M.; Khan, A.; Lim, Y.C.; Lee, K.S.; Liew, K.B.; Kassab, Y.W.; Choo, C.-Y.; Al-Worafi,
1242 Y.M.; Ming, L.C.; Kalusalingam, A. **2022**. A systematic review on chlorine dioxide as a
1243 disinfectant. *Journal of Medicine and Life* 15(3), 313.
1244 Jelli, E.; Ohmura, T.; Netter, N.; Abt, M.; Jiménez-Siebert, E.; Neuhaus, K.; Rode, D.K.H.; Nadell,
1245 C.D.; Drescher, K. **2023**. Single-cell segmentation in bacterial biofilms with an optimized deep
1246 learning method enables tracking of cell lineages and measurements of growth rates. *Molecular*
1247 *Microbiology* 119(6), 659-676.
1248 Jia, X.; O'Connor, D.; Shi, Z.; Hou, D. **2021**. VIRS based detection in combination with machine
1249 learning for mapping soil pollution. *Environ Pollut* 268(Pt A), 115845.
1250 Jiang, S.; Hu, J.; Wood, K.L.; Luo, J. **2021**. Data-driven design-by-analogy: state-of-the-art and
1251 future directions. *Journal of Mechanical Design* 144(2), 020801.
1252 Jing, Z.; Lu, Z.; Zhao, Z.; Cao, W.; Wang, W.; Ke, Y.; Wang, X.; Sun, W. **2023**. Molecular
1253 ecological networks reveal the spatial-temporal variation of microbial communities in drinking
1254 water distribution systems. *Journal of Environmental Sciences* 124, 176-186.
1255 Joseph, L.; Jun, B.-M.; Flora, J.R.V.; Park, C.M.; Yoon, Y. **2019**. Removal of heavy metals from
1256 water sources in the developing world using low-cost materials: A review. *Chemosphere* 229,
1257 142-159.
1258 Karygianni, L.; Ren, Z.; Koo, H.; Thurnheer, T. **2020**. Biofilm matrixome: extracellular
1259 components in structured microbial communities. *Trends in Microbiology* 28(8), 668-681.
1260 Kazemi, E.; Kyritsakas, G.; Husband, S.; Flavell, K.; Speight, V.; Boxall, J. **2023**. Predicting iron
1261 exceedance risk in drinking water distribution systems using machine learning. *IOP*
1262 *Conference Series: Earth and Environmental Science* 1136(1), 012047.
1263 Khattak, A.; Bukhsh, R.; Aslam, S.; Yafoz, A.; Alghushairy, O.; Alsini, R. **2022**. A hybrid deep
1264 learning-based model for detection of electricity losses using big data in power systems.
1265 *Sustainability* 14(20), 13627.
1266 Kirstein, I.V.; Hensel, F.; Gomiero, A.; Iordachescu, L.; Vianello, A.; Wittgren, H.B.; Vollertsen, J.
1267 **2021**. Drinking plastics?–Quantification and qualification of microplastics in drinking water
1268 distribution systems by µFTIR and Py-GCMS. *Water Research* 188, 116519.
1269 Knights, D.; Kuczynski, J.; Charlson, E.S.; Zaneveld, J.; Mozer, M.C.; Collman, R.G.; Bushman,
1270 F.D.; Knight, R.; Kelley, S.T. **2011**. Bayesian community-wide culture-independent microbial
1271 source tracking. *Nature Methods* 8(9), 761-763.
1272 Krasner, S.W.; Weinberg, H.S.; Richardson, S.D.; Pastor, S.J.; Chinn, R.; Sclimenti, M.J.; Onstad,
1273 G.D.; Thruston, A.D. **2006**. Occurrence of a New Generation of Disinfection Byproducts.
1274 *Environmental Science & Technology* 40(23), 7175-7185.
1275 Krasner, S.W.; Mitch, W.A.; McCurry, D.L.; Hanigan, D.; Westerhoff, P. **2013**. Formation,
1276 precursors, control, and occurrence of nitrosamines in drinking water: a review. *Water*
1277 *Research* 47(13), 4433-4450.

Lambert, A.; Brown, T.G.; Takizawa, M.; Weimer, D. **1999**. A review of performance indicators for real losses from water supply systems. *Journal of Water Supply: Research and Technology—AQUA* 48(6), 227-237.
LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. **1989**. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* 1(4), 541-551.
Lecun, Y.; Bottou, L.; Bengio, Y.; 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.; 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, E.J.; Schwab, K.J. **2005**. Deficiencies in drinking water distribution systems in developing countries. *Journal of Water and Health* 3(2), 109-127.
Lee, J.; Kim, E.-S.; Roh, B.-S.; Eom, S.-W.; 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.; Croue, J.P. **2004**. Modeling of bromate formation by ozonation of surface waters in drinking water treatment. *Water Research* 38(8), 2185-2195.
Leitão, J.; Simões, N.; Sá Marques, J.A.; Gil, P.; Ribeiro, B.; 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.; 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.; Daniell, J.J. **2011**. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software* 26(12), 1647-1659.
Li, L.; Rong, S.; Wang, R.; Yu, S. **2021**. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. *Chemical Engineering Journal* 405, 126673.
Li, P.; 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.; Ren, L. **2017**. Microbial communities shaped by treatment processes in a drinking water treatment plant and their contribution and threat to drinking water safety. *Frontiers in Microbiology* 8, 2465.
Li, R.A.; McDonald, J.A.; Sathasivan, A.; Khan, S.J. **2019**. Disinfectant residual stability leading to disinfectant decay and by-product formation in drinking water distribution systems: a systematic review. *Water Research* 153, 335-348.
Li, Z.; Zhang, C.; Liu, H.; Zhang, C.; Zhao, M.; Gong, Q.; Fu, G. **2022**. Developing stacking ensemble models for multivariate contamination detection in water distribution systems. *Science of the Total Environment* 828, 154284.
Li, Z.; Liu, H.; Zhang, C.; Fu, G. **2023**. 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.
Ligda, P.; Claerebout, E.; Kostopoulou, D.; Zdragas, A.; Casaert, S.; Robertson, L.J.; 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. *Environmental Pollution* 264, 114766.

Lin, H.; Dai, Q.; Zheng, L.; Hong, H.; Deng, W.; Wu, F. **2020**. Radial basis function artificial neural network able to accurately predict disinfection by-product levels in tap water: Taking haloacetic acids as a case study. *Chemosphere* 248, 125999.

Lin, H.H.-H.; Lin, A.Y.-C. **2024**. Peracetic acid as an alternative disinfectant for micropollutants degradation and disinfection byproducts control in outdoor swimming pools. *Journal of Hazardous Materials* 464, 132988.

Ling, F.; Whitaker, R.; LeChevallier, M.W.; Liu, W.T. **2018**. Drinking water microbiome assembly induced by water stagnation. *The ISME Journal* 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.; 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 Research* 138, 86-96.

Liu, S.; Gunawan, C.; Barraud, N.; Rice, S.; Harry, E.; Amal, R. **2016**. Understanding, monitoring and controlling biofilm growth in drinking water distribution systems. *Environmental Science & Technology* 50(17), 8954-8976.

Lowe, M.; Qin, R.; 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.; Knight, R. **2011**. UniFrac: an effective distance metric for microbial community comparison. *The ISME Journal* 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.; 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.; 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.; Euverink, G.J.W.; Keesman, K.J.; Jayawardhana, B. **2022**. Clean and safe drinking water systems via metagenomics data and artificial intelligence: state-of-the-art and future perspective. *Frontiers in Microbiology* 13, 832452.

Manasfi, T. **2021**. Ozonation in drinking water treatment: an overview of general and practical aspects, mechanisms, kinetics, and byproduct formation. *Comprehensive Analytical Chemistry* 92, 85-116.

Mao, Y.; Wang, X.; Yang, H.; Wang, H.; 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*, 1st Edition. 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.; Vambol, V. **2020**. Chlorination disinfection by-products in municipal drinking water – A review. *Journal of Cleaner Production* 273, 123159.

Mounce, S.R.; Mounce, R.B.; 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.; 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.; 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.

Naloufi, M.; Lucas, F.S.; Souihi, S.; Servais, P.; Janne, A.; Wanderley Matos De Abreu, T. **2021**. Evaluating the performance of machine learning approaches to predict the microbial quality of surface waters and to optimize the sampling effort. *Water* 13(18), 2457.

Narita, K.; Matsui, Y.; Matsushita, T.; 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.; 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, 9513909.

Oliker, N.; 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.; 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.; 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.; 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.; Liu, H. **2018**. Sparse modeling-based sequential ensemble learning for effective outlier detection in high-dimensional numeric data. *Thirty-Second AAAI Conference on Artificial Intelligence* 32(1), 10.1609/aaai.v1632i1601.11692.

Park, J.; Park, J.-H.; Choi, J.-S.; Joo, J.C.; Park, K.; Yoon, H.C.; Park, C.Y.; Lee, W.H.; 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.; 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.; Legge, R.L.; 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.

Peleato, N.M. **2022**. Application of convolutional neural networks for prediction of disinfection by-products. *Scientific Reports* 12(1), 612.

Pifer, A.D.; 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.; Raskin, L. **2014**. Spatial-temporal survey and occupancy-abundance modeling to predict bacterial community dynamics in the drinking water microbiome. *MBio* 5(3), e01135-14.

Podgorski, J.; Berg, M. **2022**. Global analysis and prediction of fluoride in groundwater. *Nature Communications* 13, 4232.

Prescott, J.F. **2017**. History and current use of antimicrobial drugs in veterinary medicine. *Microbiology Spectrum* 5(6), ARBA-0002-2017.
Proctor, C.R.; Lee, J.; Yu, D.; Shah, A.D.; Whelton, A.J. **2020**. Wildfire caused widespread drinking water distribution network contamination. *AWWA Water Science* 2(4), e1183.
Ramos-Martínez, E.; Herrera, M.; Izquierdo, J.; 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.; Pérez-García, R. **2016**. A multi-disciplinary procedure to ascertain biofilm formation in drinking water pipes. *International Congress on Environmental Modelling and Software* 3, 619-626.
Redondo-Hasselerharm, P.E.; Cserbik, D.; Flores, C.; Farré, M.J.; Sanchís, J.; Alcolea, J.A.; Planas, C.; Caixach, J.; 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* 34, 23-33.
Reis, A.L.; Lopes, M.A.; Andrade-Campos, A.; Antunes, C.H. **2023**. A review of operational control strategies in water supply systems for energy and cost efficiency. *Renewable and Sustainable Energy Reviews* 175, 113140.
Renwick, D.V.; Heinrich, A.; Weisman, R.; Arvanaghi, H.; Rotert, K. **2019**. Potential public health impacts of deteriorating distribution system infrastructure. *Journal AWWA* 111(2), 42-53.
Richards, C.E.; Tzachor, A.; Avin, S.; 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.; 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.; Wilbur, J.D. **2003**. Assessment of statistical methods used in library-based approaches to microbial source tracking. *Journal of Water and 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.; Vila, J. **2015**. The global threat of antimicrobial resistance: science for intervention. *New Microbes New Infect* 6, 22-29.
Rodriguez-Perez, J.; Leigh, C.; Liquet, B.; Kermorvant, C.; Peterson, E.; Sous, D.; Mengersen, K. **2020**. Detecting technical anomalies in high-frequency water-quality data using artificial neural networks. *Environmental Science & Technology* 54(21), 13719-13730.
Rodriguez, M.J.; Sérodes, J.B.; Levallois, P. **2004**. Behavior of trihalomethanes and haloacetic acids in a drinking water distribution system. *Water Research* 38(20), 4367-4382.
Saboe, D.; Ghasemi, H.; Gao, M.M.; Samardzic, M.; Hristovski, K.D.; Boscovic, D.; Burge, S.R.; Burge, R.G.; Hoffman, D.A. **2021**. Real-time monitoring and prediction of water quality parameters and algae concentrations using microbial potentiometric sensor signals and machine learning tools. *Science of the Total Environment* 764, 142876.
Senoro, D.B.; de Jesus, K.L.M.; Nolos, R.C.; Lamac, M.R.L.; Deseo, K.M.; Tabelin, C.B. **2022**. In situ measurements of domestic water 1uality and health risks by elevated concentration of heavy metals and metalloids using Monte Carlo and MLGI methods. *Toxics* 10(7), 342.

Shao, B.; Shen, L.; Liu, Z.; Tang, L.; Tan, X.; Wang, D.; Zeng, W.; Wu, T.; Pan, Y.; Zhang, X. **2023**. Disinfection byproducts formation from emerging organic micropollutants during chlorine-based disinfection processes. *Chemical Engineering Journal* 455, 140476.

Shi, Y.; Babatunde, A.; Bockelmann-Evans, B.; Li, Q.; 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.; 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.

Simoes, L.C.; Simões, M. **2013**. Biofilms in drinking water: problems and solutions. *RSC advances* 3(8), 2520-2533.

Simpson, A.M.A.; 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.; Sabol, J. **2014**. Artificial intelligence in public health prevention of legionelosis in drinking water systems. *International Journal of Environmental Research and Public Health* 11(8), 8597-8611.

Singh, D.; Vardhan, M.; Sahu, R.; Chatterjee, D.; Chauhan, P.; Liu, S. **2023**. Machine-learning- and deep-learning-based streamflow prediction in a hilly catchment for future scenarios using CMIP6 GCM data. *Hydrology and Earth System Sciences* 27(5), 1047-1075.

Singh, K.P.; Gupta, S. **2012**. Artificial intelligence based modeling for predicting the disinfection by-products in water. *Chemometrics and Intelligent Laboratory Systems* 114, 122-131.

Siponen, S.; Jayaprakash, B.; Hokajärvi, A.-M.; Gomez-Alvarez, V.; Inkinen, J.; Ryzhikov, I.; Räsänen, P.; Ikonen, J.; Pursiainen, A.; Kauppinen, A. **2024**. Composition of active bacterial communities and presence of opportunistic pathogens in disinfected and non-disinfected drinking water distribution systems in Finland. *Water Research* 248, 120858.

Sluban, B.; Lavrač, N. **2015**. Relating ensemble diversity and performance: A study in class noise detection. *Neurocomputing* 160, 120-131.

Smith, A.; Sterba-Boatwright, B.; Mott, J. **2010**. Novel application of a statistical technique, Random Forests, in a bacterial source tracking study. *Water Research* 44(14), 4067-4076.

Sokolova, M.; 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.; 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.; Chaudhary, V.K. **2020**. Disinfection by-products in drinking water: Occurrence, toxicity and abatement. *Environmental Pollution* 267, 115474.

Steel, R.G.D.; Torrie, J.H. **1960**. *Principles and procedures of statistics: with special reference to the biological sciences*. McGraw-Hill, New York.

Stringer, C.; Wang, T.; Michaelos, M.; Pachitariu, M. **2021**. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods* 18(1), 100-106.

Sudhakaran, S.; 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 Research* 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.; 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.; Surampalli, R.Y. **2018**. Emerging contaminants: here today, there tomorrow! *Environmental nanotechnology, monitoring & management* 10, 122-126.

Tian, C.; Feng, C.; Chen, L.; Wang, Q. **2020**. Impact of water source mixture and population changes on the Al residue in megalopolitan drinking water. *Water Research* 186, 116335.

U.S. EPA **2023** (accessed 2024). *Information about Public Water Systems*. https://www.epa.gov/dwreginfo/information-about-public-water-systems.

Valbonesi, P.; Profita, M.; Vasumini, I.; Fabbri, E. **2021**. Contaminants of emerging concern in drinking water: Quality assessment by combining chemical and biological analysis. *Science of the Total Environment* 758, 143624.

Walesch, S.; Birkelbach, J.; Jézéquel, G.; Haeckl, F.P.J.; Hegemann, J.D.; Hesterkamp, T.; Hirsch, A.K.H.; Hammann, P.; 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.; 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.

Wen, X.; Chen, F.; Lin, Y.; Zhu, H.; Yuan, F.; Kuang, D.; Jia, Z.; Yuan, Z. **2020**. Microbial indicators and their use for monitoring drinking water quality—A review. *Sustainability* 12(6), 2249.

WHO **2011**. *Guidelines for Drinking-Water Quality*, 4th Edition. World Health Organization, Geneva.

WHO **2021**. Antimicrobial resistance.

Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V. **2022**. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys* 55(4), 66.

Willmott, C.J.; 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.

Wingender, J.; Flemming, H.C. **2011**. Biofilms in drinking water and their role as reservoir for pathogens. *International Journal of Hygiene and Environmental Health* 214(6), 417-423.

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.; 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.; 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.; Stewart, J.R. **2020**. Tracking major sources of water contamination using machine learning. *Frontiers in Microbiology* 11, 616692.

Wu, Y.; Jiang, P.; Goh, S.G.; Yu, K.; Chen, Y.; He, Y.; Gin, K.Y.H. **2022**. Predicting relative risk of antimicrobial resistance using machine learning lethods. *IFAC-PapersOnLine* 55(10), 1266-1271.

Xiao, R.; Ou, T.; Ding, S.; Fang, C.; Xu, Z.; 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.; Lin, J. **2023**. Toward autonomous laboratories: Convergence of artificial intelligence and experimental automation. *Progress in Materials Science* 132, 101043.

Xu, X.; Talbot, S.; Selvaraja, T. **2020**. ParasNet: Fast parasites detection with neural networks. *arXiv:2002.11327*, 10.48550/arXiv.42002.11327.

Yaseen, Z.M. **2021**. An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. *Chemosphere* 277, 130126.

Zainurin, S.N.; Wan Ismail, W.Z.; Mahamud, S.N.I.; Ismail, I.; Jamaludin, J.; Ariffin, K.N.Z.; 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.; 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.; Hu, Z. **2018**. Thermodynamic properties of an emerging chemical disinfectant, peracetic acid. *Science of the Total Environment* 621, 948-959.

Zhang, C.; Brown, P.J.B.; Hu, Z. **2019a**. Higher functionality of bacterial plasmid DNA in water after peracetic acid disinfection compared with chlorination. *Science of the Total Environment* 685, 419-427.

Zhang, C.; Brown, P.J.B.; Miles, R.J.; White, T.A.; Grant, D.G.; Stalla, D.; Hu, Z. **2019b**. Inhibition of regrowth of planktonic and biofilm bacteria after peracetic acid disinfection. *Water Research* 149, 640-649.

Zhang, C.; Lu, J. **2021a**. 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.; Lu, J. **2021b**. *Legionella*: a supplementary indicator of microbial water quality in municipal engineered water systems. *Frontiers in Environmental Science* 9, 684319.

Zhang, C.; Qin, K.; Struewing, I.; Buse, H.; Santo Domingo, J.; Lytle, D.; Lu, J. **2021a**. The bacterial community diversity of bathroom hot tap water was significantly lower than that of cold tap and shower water. *Frontiers in Microbiology* 12, 625324.

Zhang, C.; Struewing, I.; Mistry, J.H.; Wahman, D.G.; Pressman, J.; Lu, J. **2021b**. *Legionella* and other opportunistic pathogens in full-scale chloraminated municipal drinking water distribution systems. *Water Research* 205, 117571.

Zhang, J.; Wang, Y.; Donarski, E.D.; Toma, T.T.; Miles, M.T.; Acton, S.T.; Gahlmann, A. **2022**. BCM3D 2.0: accurate segmentation of single bacterial cells in dense biofilms using computationally generated intermediate image representations. *npj Biofilms and Microbiomes* 8(1), 99.

Zhang, M.; Zhang, J.; Wang, Y.; Wang, J.; Achimovich, A.M.; Acton, S.T.; Gahlmann, A. **2020**. Non-invasive single-cell morphometry in living bacterial biofilms. *Nature Communications* 11(1), 6151.

Zhang, Y.; Gao, X.; Smith, K.; Inial, G.; Liu, S.; Conil, L.B.; Pan, B. **2019c**. 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.; 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.; Fu, L. **2023a**. Stability of drinking water distribution systems and control of disinfection by-products. *Toxics* 11(7), 606.

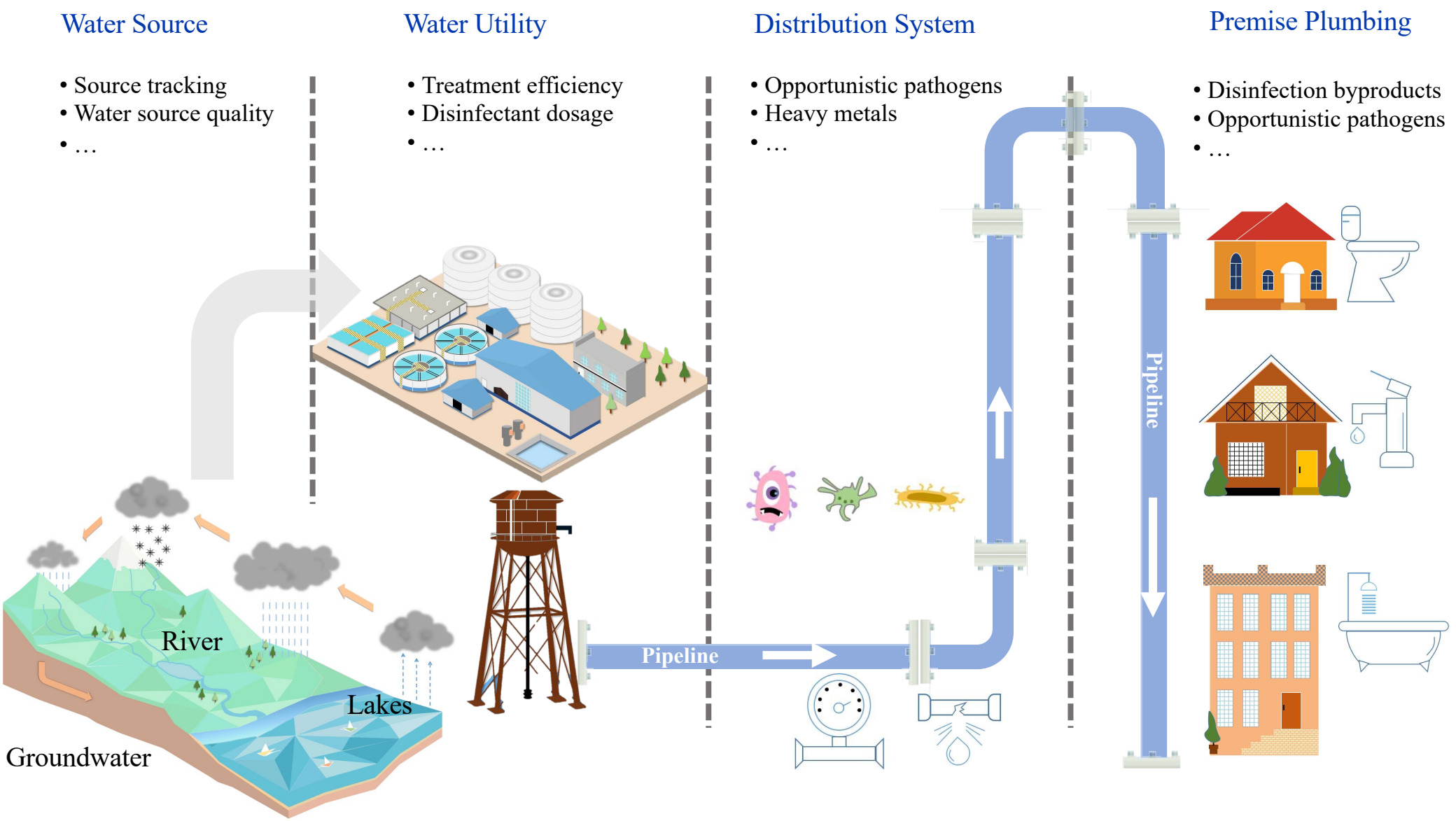
Zhou, Q.; Huang, J.; Guo, K.; Lou, Y.; Wang, H.; Zhou, R.; Tang, J.; Hou, P. **2023b**. Spatiotemporal distribution of opportunistic pathogens and microbial community in centralized rural drinking water: One year survey in China. *Environmental Research* 218, 115045.

Zhou, X.; Tang, Z.; Xu, W.; Meng, F.; Chu, X.; Xin, K.; Fu, G. **2019**. Deep learning identifies accurate burst locations in water distribution networks. *Water Research* 166, 115058.

Zhou, Z.; Zhong, D.; Zhang, Z.; Ma, W.; Chen, J.; Zhuang, M.; Li, F.; Zhang, J.; Zhu, Y.; Su, P. **2023c**. Biofilm on the pipeline wall is an important transmission route of resistome in drinking water distribution system. *Environmental Pollution* 335, 122311.

Zhu, M.; Wang, J.; Yang, X.; Zhang, Y.; Zhang, L.; Ren, H.; Wu, B.; Ye, L. **2022**. A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health* 1(2), 107-116.

Using Machine Learning to Ensure Drinking Water Quality



Graphical abstract

Table 1 Machine learning to ensure safe drinking water supply from the physical perspective

Topic	Task	Model	Inputs	Outputs	Metrics (Selected)	Performances (Selected)	Reference
Predicting 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 prediction	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
Monitoring pipeline integrity	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 Machine learning to ensure safe drinking water supply from the chemical perspective

Topic	Task	Model	Inputs	Outputs	Metrics (Selected)	Performances (Selected)	Reference
Optimizing drinking water disinfection	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
	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; 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;

Table 2 Machine learning to ensure safe drinking water supply from the chemical perspective

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 Machine learning to ensure safe drinking water supply from the chemical perspective (*cont.*)

Topic	Task	Model	Inputs	Outputs	Metrics (Selected)	Performances (Selected)	Reference
Optimizing drinking water disinfection	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
Nitrification surveillance	Nitrification episodes classification	NB	16S rRNA profiling	Nitrification episodes: stable or failure	AUC	0.83	Gomez-Alvarez and Revetta, 2020
	Estimate NOx concentrations	SVR	NOx absorbances at various wavelengths	C_{NO_x}	RMSE and R^2	RMSE < 0.04	Hossain et al., 2021
Heavy metal monitoring regulation	Pb ions concentration detection	SVR	S_{11}	Pb concentration	RMS	0.71	Oh et al., 2021
	Spatial concentration mapping of heavy metal	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 of Al residue	Kriging interpolation	Spatial and temporal data	Temporal-spatial distribution of residual Al	-	-	Tian et al., 2020
	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 and Mazloom, 2022
	Heavy metal 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

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 Machine learning to ensure safe drinking water supply from the chemical perspective (*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; NB, naïve Bayes; AUC, area under the curve; NOx, nitrite and nitrate; 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.

Table 3 Machine learning to ensure safe drinking water supply from the microbiological perspective

Topic	Task	Model	Inputs	Outputs	Metrics (Selected)	Performances (Selected)	Reference
Surveilling and mitigating 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 higher-risk, serogroup and contamination levels prediction of *Legionella* spread	SaTScan, XGB, LR, and SVM	Survey, spatial and meteorological info., and risk level to *Legionella*;	Higher-risk level clusters; serogroup of a sample and the contamination level	Accuracy and F1-score	XGBoost > SVM > LR	Brunello et al., 2022
Analyzing drinking water microbial communities	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

NARA, Neural network designed on approximate reasoning architecture; *Q*, flow rate; 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 3 Machine learning to ensure safe drinking water supply from the microbiological perspective (*cont.*)

Topic	Task	Model	Inputs	Outputs	Metrics (Selected)	Performances (Selected)	Reference
Parasite detection	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
Biofilm development assessment	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
Risk analysis and source tracking of antimicrobial resistance	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 pollution 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

LDFA, linear discriminant function analysis; RT, regression trees; HPC , heterotrophic plate count; OSA, over-segmentation abundances; IoU, intersection-over-union; 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; ARG, Antimicrobial resistance genes

Table 4 Machine learning to ensure safe drinking water supply from the temporal perspective

Topic	Task	Model	Inputs	Outputs	Metrics (Selected)	Performances (Selected)	Reference
Anomalies and contamination events detection	Anomalies 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.

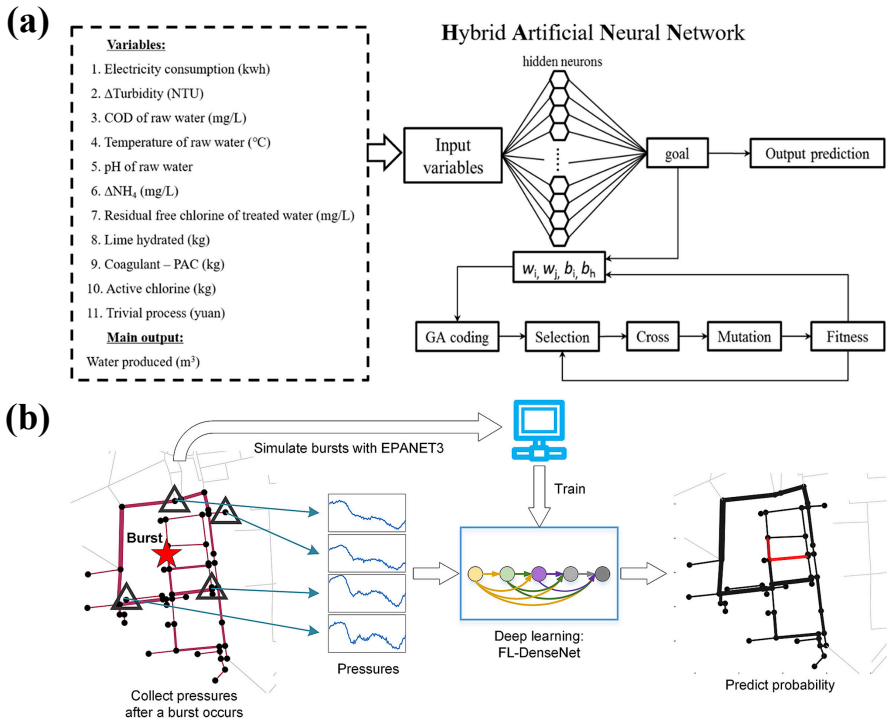


Figure 1. (a) Input and output variables used for modeling and the proposed hybrid artificial neural network framework for prediction of drinking water production. Reproduced with permission from Zhang et al., 2019. Copyright 2019 Elsevier. (b) Schematic of fully-linear DenseNet (BLIFF) model for accurate identification of burst locations in EWS networks. Reproduced with permission from Zhou et al., 2019 (CC BY 4.0).

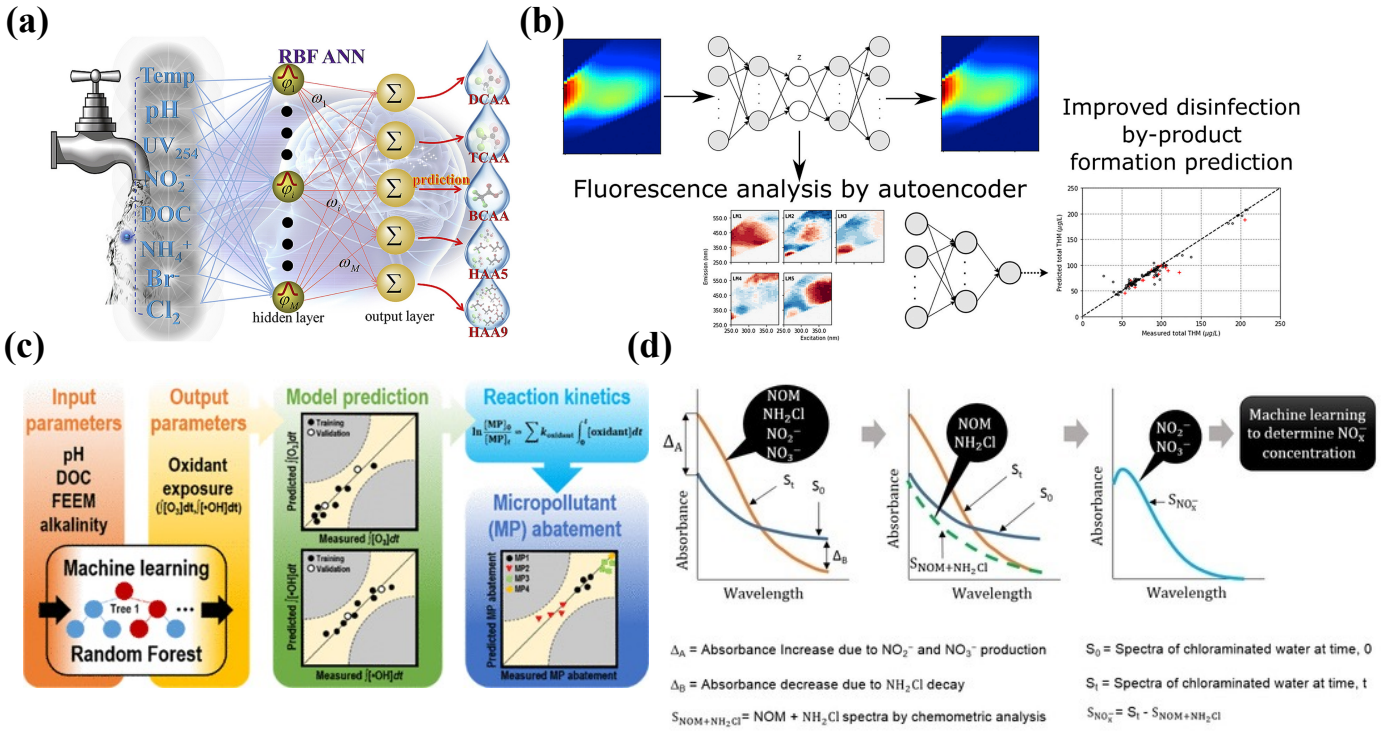


Figure 2. (a) Schematic of radial basis function (RBF) artificial neural network (ANN) model for prediction of disinfection by-products (DBPs). Reproduced with permission from Lin et al., 2020, Copyright 2020 Elsevier. (b) Schematic of autoencoder model for prediction of DBPs. Reproduced with permission from Peleato et al., 2018, Copyright 2018 Elsevier. (c) Schematic of random forest (RF) model for prediction of micropollutant abatement. Reproduced with permission from Cha et al., 2021, Copyright 2021 American Chemical Society. (d) Prediction of nitrate and nitrite concentrations over support vector regression (SVR) model. Reproduced with permission from Hossain et al., 2021 (CC BY 4.0).

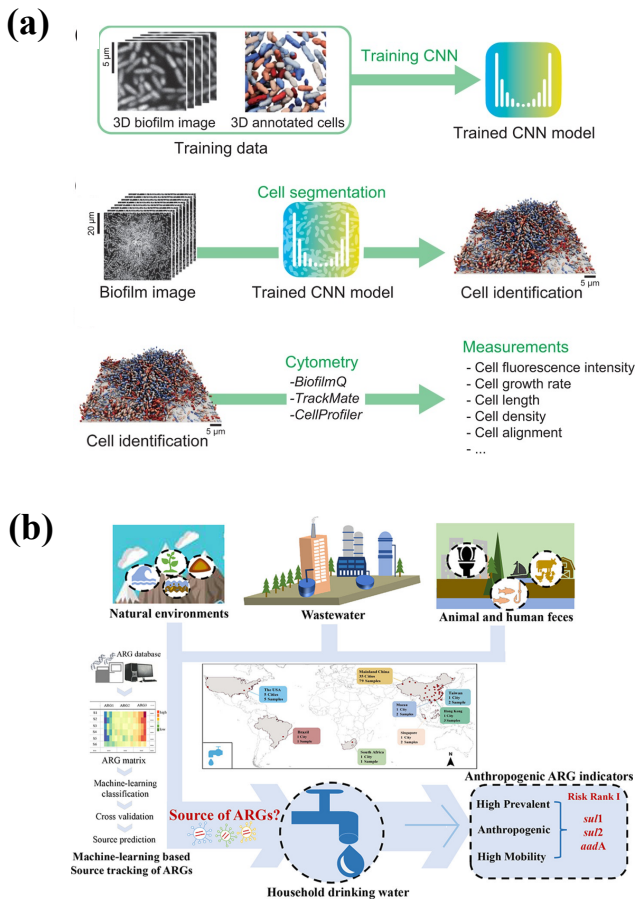


Figure 3. (a) Deep-learning-based workflow for single-cell measurements in three-dimensional biofilms. Reproduced with permission from Jelli et al., 2023, Copyright 2023 Elsevier. (b) SourceTracker was performed to investigate the pollution sources of antimicrobial resistance genes (ARGs) in household drinking water. Reproduced with permission from Wang et al., 2023, Copyright 2023 Elsevier.

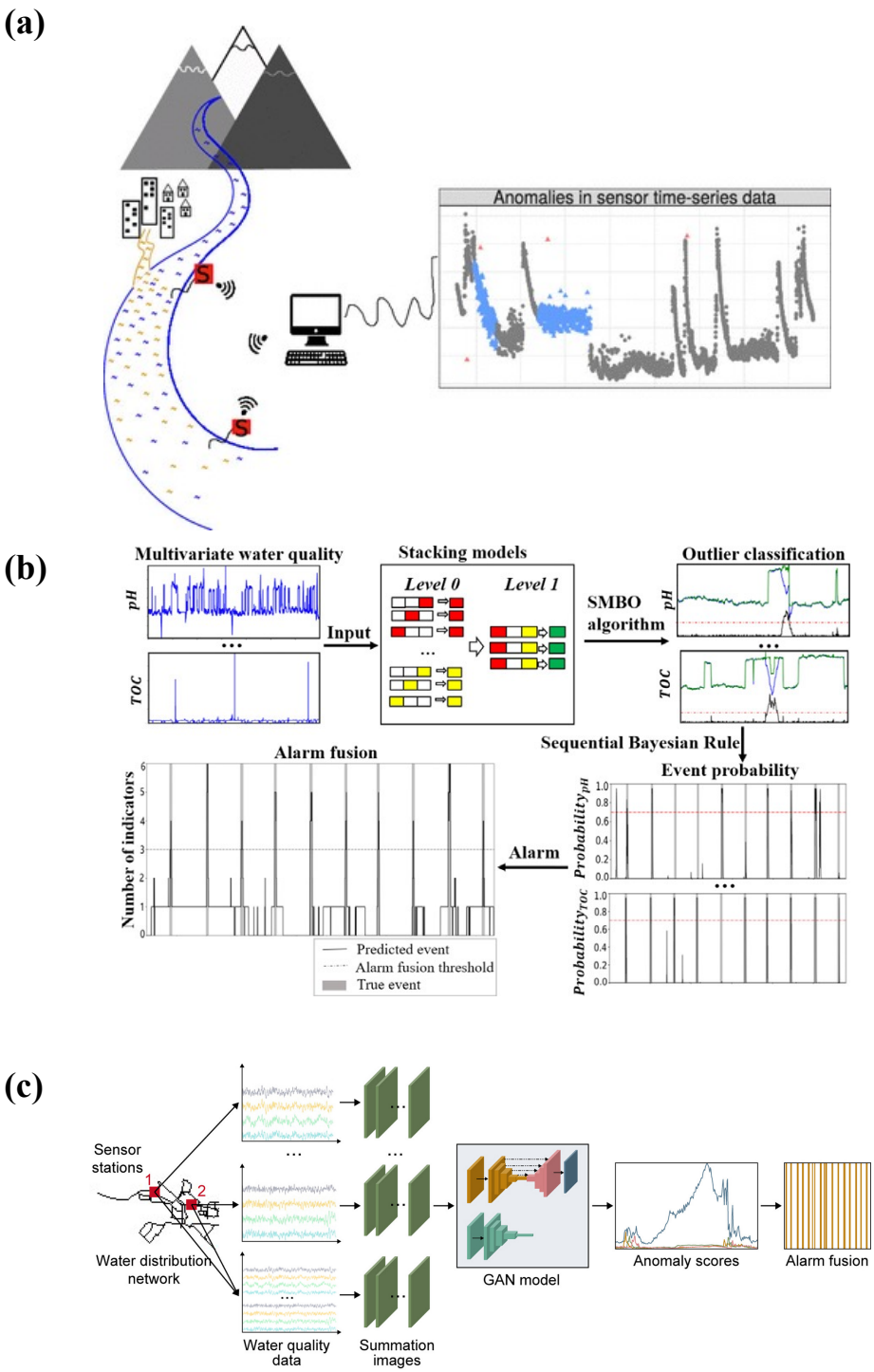


Figure 4. (a) Detection of technical anomalies in water quality using artificial neural network (ANN) model. Reproduced with permission from Rodriguez-Perez et al., 2020, Copyright 2020 American Chemical Society. (b) A stacking ensemble model for contamination event detection using multiple water quality parameters. Reproduced with permission from Li et al., 2022, Copyright 2022 Elsevier. (c) Detection of contamination events using generative adversarial network (GAN) model. Reproduced with permission from Li et al., 2023, Copyright 2023 Elsevier.

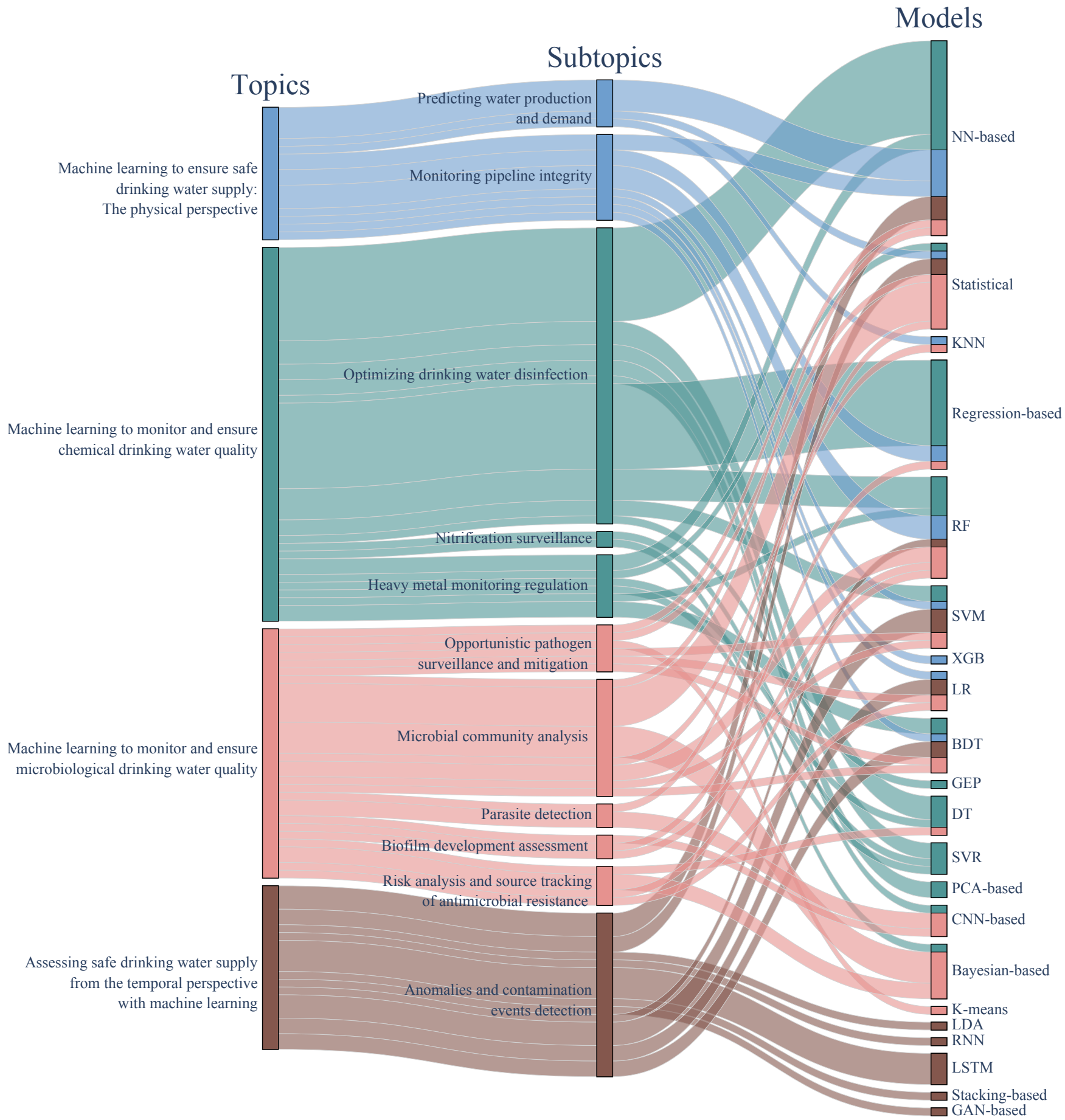


Figure 5. Machine learning model distribution across drinking water supply research topics

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.