Article:
Kyritsakas, G., Boxall, J.B. and Speight, V.L. (2023) A Big Data framework for actionable information to manage drinking water quality. AQUA — Water Infrastructure, Ecosystems and Society, 72 (5). pp. 701-720. ISSN: 2709-8028

https://doi.org/10.2166/aqua.2023.218

AQUA
Water Infrastructure, Ecosystems and Society

IWA PUBLISHING

AQUA — Water Infrastructure, Ecosystems and Society Vol 00 No 0, 1 doi: 10.2166/aqua.2023.218

A Big Data framework for actionable information to manage drinking water quality

Grigorios Kyritsakas*, Joseph B. Boxall and Vanessa L. Speight

Department of Civil and Structural Engineering, University of Sheffield, Sir Frederick Mappin Building, Mappin Street, Sheffield S1 3JD, UK
*Corresponding author. E-mail: g.kyritsakas@sheffield.ac.uk

GK, 0000-0003-0945-3754; JBB, 0000-0002-4681-6895; VLS, 0000-0001-7780-7863

ABSTRACT

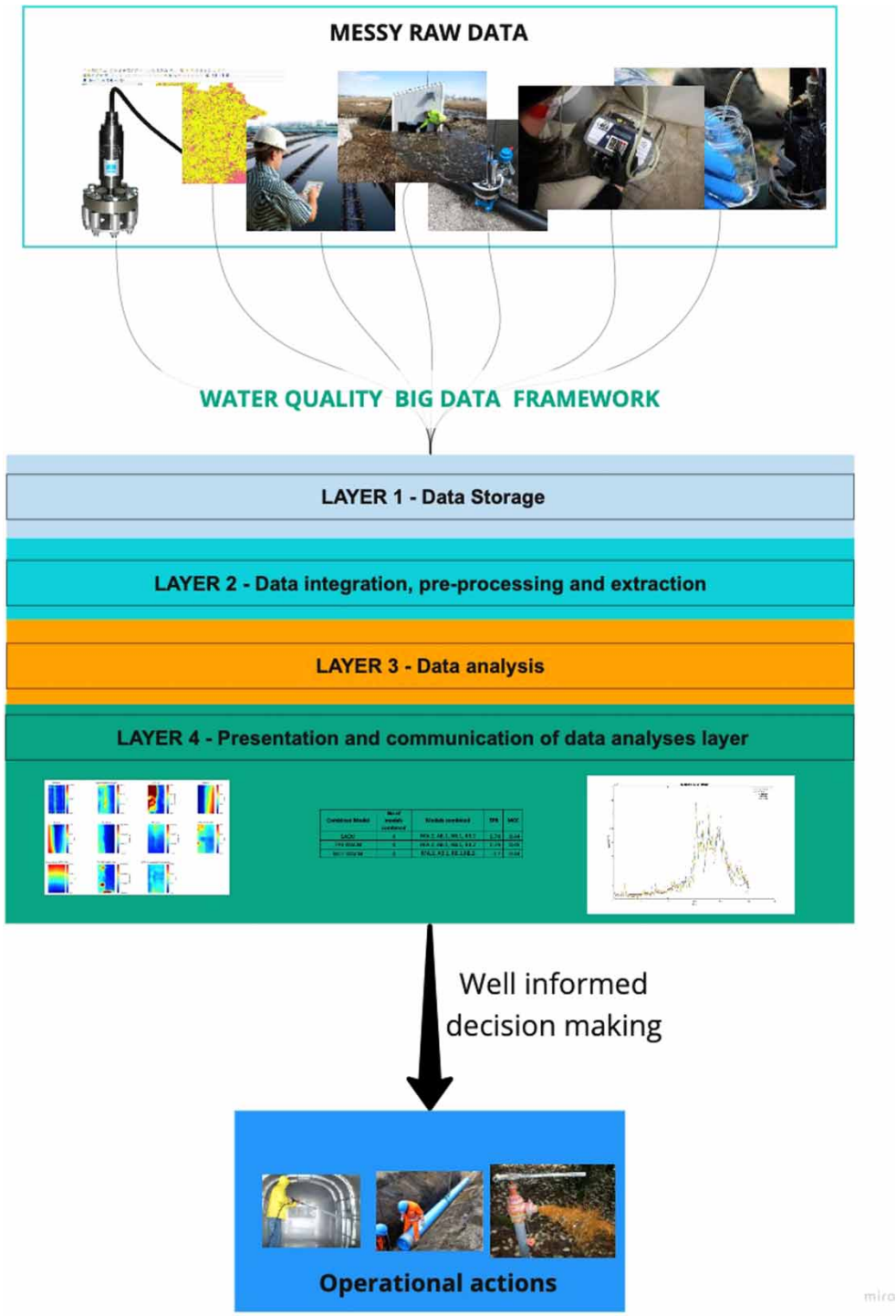
Water utilities collect vast amounts of data, but they are stored and utilised in silos. Machine learning (ML) techniques offer the potential to gain deeper insight from such data. We set out a Big Data framework that for the first time enables a structured approach to systematically progress through data storage, integration, analysis, and visualisation, with applications shown for drinking water quality. A novel process for the selection of the appropriate ML method, driven by the insight required and the available data, is presented. Case studies for a water utility supplying 5.5 million people validate the framework and provide examples of its use to derive actionable information from data to help ensure the delivery of safe drinking water.

Key words: Big Data analytics, data management, drinking water quality, machine learning, water supply systems

HIGHLIGHTS

- A four-layer Big Data framework for better water quality management is proposed.
- Framework consists of data collection, integration, analysis, and visualisation.
- Machine learning method selection tool driven by data availability is included.
- Framework yields information for interventions to manage drinking water quality.
- Two case studies demonstrate the success of the framework.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Water utilities have a duty to provide drinking water that complies with the high-quality standards set by national and international regulations. This effort requires a multi-step approach that includes the application of efficient and innovative treatment technologies in the water treatment works (WTWs) as well as the prevention of water deterioration during its

travel through the drinking water distribution system (DWDS), the proper maintenance of DWDS, and monitoring of the treated water quality from source to tap.

In the UK, water utilities monitor treatment processes using sensors to measure various water quality parameters in a regular frequency (typically every 5–15 min), with resulting data, also known as telemetry data, stored in the supervisory control and data acquisition (SCADA), or similar system. In addition, utilities take samples from different points across their DWDS, including exit points from the WTW and service reservoirs (SRs) and randomly selected consumers' taps. The water quality parameters measured in a typical DWDS monitoring programme are microbial indicators, disinfectant residual, iron, manganese, and turbidity as defined by the regulators (DWI 2016; DWQR 2019).

The typical procedure for DWDS monitoring results at UK water utilities is to archive all the data, once checked for compliance, thereby creating a large store of data in various formats. This archived data are generally not used or analysed further, with their volume increasing year after year. Analysis of these datasets, if done correctly, and when considered with the wider asset, operational or even third-party data, can provide a better understanding of the complex processes that occur inside the ageing DWDS and can be used as evidence to direct capital, operational, and maintenance activities within a water utility. Advanced data analytics, including tools that broadly fall under the umbrella of artificial intelligence (AI), offer the opportunity to unlock the potential value of otherwise ignored DWDS water quality data.

A few research studies have applied AI technologies, such as data mining (DM) or machine learning (ML), to drinking water quality problems for understanding factors contributing to water quality deterioration (Blokker *et al.* 2016; Speight *et al.* 2019), predicting future deterioration events (Meyers *et al.* 2017; Mounce *et al.* 2017; Kazemi *et al.* 2018), and optimising treatment processes at WTWs (Li *et al.* 2021). These studies demonstrate the potential that individual ML techniques could have for analysis of historical water quality data. While the focus of research is often the specific ML technique, the individual ML techniques are just one component of the 'Big Data' analytics approach required to support decision-making and inform investment choices. Water utilities who want to benefit from Big Data analytics will also need to transform the ways that they collect, store, process, and visualise data and results. This transformation, sometimes referred to as the pathway to 'Digital Water' (IWA 2019), requires holistic consideration of data issues to facilitate the Big Data applications for improving the delivery of safe drinking water.

This paper proposes a Big Data framework for water utilities, using examples drawn from DWDS water quality applications to demonstrate its application. This framework fills the gap between different individual data-driven applications and the integration and processing of the various types of raw data collected by water utilities. This framework is meant to be a guiding approach for water utilities with specific examples related to solving water quality problems in DWDS. By presenting this framework, this work aims to contribute to the 'Digital Water' transformation and to lay out the steps for water utilities to undertake on the journey to this digital revolution.

2. BACKGROUND

2.1. Big Data analytics

Big Data refers to the collection of massive amounts of data that modern digital technologies generate and/or store. The volume of data, however, is just one of the characteristics of Big Data that also includes velocity, variety, veracity, variability, and value (Gandomi & Haider 2015). Briefly, these six characteristics respectively refer to the generation and collection of extreme amounts of data, the speed that the data are generated and analysed, the complexity of the datasets as these could be composed with data in various types of formats, the reliability (in terms of quality) of the available data, the variation of data sources and data flows, and the important information that could provide once analysed. Big Data analytics is the science that includes all the processes and the tools required to uncover valuable information hidden in these massive datasets, from the data collection to the mining and the predictive methods (usually DM and ML techniques and algorithms) used for providing outputs to decision makers. Undeniably, the successful application of Big Data science requires collaboration and integration with domain expertise.

2.2. Machine learning

ML is the area of AI that develops the algorithms used to optimise future performance or understand patterns by learning from existing data or past experiences (Alpaydin 2014). ML algorithms are the most common tools that Big Data analytics use for the identification of patterns in data and predictions of future trends. There are two main categories of ML algorithms: supervised and unsupervised. Supervised learning algorithms are trained on data that have been labelled as input or output

with the aim to predict future outputs based on new unseen inputs. Supervised ML predictive modelling is further split into regression and classification depending on the type of required outputs. In classification, for the given inputs, a specific category or class is specified as the output (e.g., above or below a threshold) and a specific class is required as predictive output. In regression, the ML algorithms are trained to predict future continuous output values when given new unseen inputs. Unsupervised learning is typically applied for data exploration as it uses unlabelled data as inputs to generate clusters of different groups, uncovering hidden structures in the datasets and identifying correlations between the various parameters of the analysis.

ML is gaining traction in water-related applications, with recent studies having developed and applied algorithms for topics including leak detection in pipes (Mounce *et al.* 2010; Romano *et al.* 2014; Carreño-Alvarado *et al.* 2017), water demand forecasting (Herrera *et al.* 2010; Xenochristou *et al.* 2021), wastewater treatment plant operations (Dairi *et al.* 2019; Mamandipoor *et al.* 2020; Xu *et al.* 2021), sewer overflow predictions (Mounce *et al.* 2014; Rosin *et al.* 2022), prediction of chlorine decay at consumers taps (Gibbs *et al.* 2006), prediction of indicator microorganisms in drinking water supply (Mohammed *et al.* 2017), and prediction of water quality events in DWDS using sensors (Vries *et al.* 2016; Fellini *et al.* 2018; Garcia *et al.* 2020). The aforementioned studies generally cover a single application but collectively demonstrate the potential for ML techniques to provide value to water utility operations. However, Speight *et al.* (2019) and Mounce *et al.* (2017) report challenges in applying ML techniques in the collection of the data and the processes required to construct a dataset suitable for analysis.

The need for a well-founded question or a more exact articulation of the insight sought, to ensure that the Big Data exercise is well-directed and leads to consequential new understanding is evident when exploring and differentiating past research. These observations reinforce the need to include consideration of the insight sought along with data collection, storage, and organisation; a Big Data framework that encompasses these and guides the selection of the ML algorithm is essential to create lasting value for water utility applications.

2.3. Big Data analytics frameworks

The complexities in Big Data applications vary from one organisation to another, depending on the type of collected data and the knowledge that needs to be derived from the datasets. In many scientific domains, discussions over holistic structures, also known as Big Data analytics frameworks, have begun to emerge. Chandarana & Vijayalakshmi (2014) documented the challenges that organisations face using the different types of data that they collect and the requirements for the development of frameworks to organise and analyse these data. As part of this work, the authors emphasised the importance of Big Data analysis for deriving valuable information and making better decisions and gave example areas such as health care and intelligence where Big Data frameworks could be beneficial.

Most frameworks proposed in the literature comprise a series of rules, in the form of layers, to (1) address the specific data storage and data integration complexities; (2) apply the proper ML, DM, or other data analysis methods depending on the desired outputs; and (3) visualise the outputs (Abdullah *et al.* 2018; Zekić-Sušac *et al.* 2020; Ahmed *et al.* 2021). For example, Osman (2019) proposed a 3-layer framework for smart cities applications that includes the platform layer which specifies the operating systems and communication protocols for collecting the various types of data, the security layer which specifies the protocols for controlling access to the data and the protocols for data integration, and the data processing layer which specifies the data pre-processing, data analytics, and management of the analytics model. The author also included a discussion of principles required for successful implementation, including the integration of static and real-time data as well as standardisation of data acquisition. Zacarias *et al.* (2018) suggested a 4-layer framework for the manufacturing sector that includes data storage and integration, consideration of the available IT resources for data pre-processing and analysis, selection of the appropriate ML algorithms for the data analytics, and a dashboard for the visualisation of the different solutions.

Within the water sector, there were two published studies found that discuss Big Data analytics frameworks. One examined the benefits that water utilities could gain in the reduction of chemicals in their wastewater using Big Data tools and ML techniques within their datasets (Romero *et al.* 2017). This study described the current situation in the water sector, referred to applicable ML tools, and provided two different examples where the incorporation of Big Data tools could strengthen the existing approaches. However, the authors did not develop a specific Big Data framework or ML technique selection process. The second study proposed a 5-layer framework for improving urban domestic wastewater treatment and reducing environmental pollution, consisting of a data perception layer, data transmission layer, data storage layer, data analysis and application layer, and user interface layer (Du *et al.* 2019). The authors analysed the volume and type of information that

would be required for the application of such data-driven approaches and emphasised the importance of collecting all the necessary data from the wastewater treatment works and networks to support the Big Data framework implementation. The degree of data proposed would require a significant transformation of monitoring practices in wastewater networks compared with the typical level today and the application of the application-specific framework with a smaller dataset or with non-sensor data was not demonstrated. Neither of these two studies refer to the selection of specific ML techniques and the criteria required for the application of those techniques.

This paper proposes a comprehensive Big Data framework that is driven by the insight sought, addresses data collection, storage, and management aspects, integrates ML technique selection, and includes visualisation and communication of outputs. Importantly, the criteria for ML selection are based upon the desired drinking water quality investigation and the existing data that are available for analysis, and the integration of data science and water engineering is essential for this.

3. PROPOSED BIG DATA FRAMEWORK

The data that water utilities collect do not compare with the amounts of data that some other sectors collect every day. In addition, their collection and storage systems are often siloed and slow, thus the value of these data is rarely explored beyond its direct individual purpose (often regulatory reporting). Thus, at present, water utility data do not comply with the Big Data definition. However, water utilities' aim is to follow the digital revolution like the other sectors (energy telecommunications, etc.). Recognising the need for a holistic approach to the management of water quality data in DWDS for data-driven applications, we propose a Big Data framework consisting of four layers: (1) data storage; (2) data connection and integration; (3) data analysis; and (4) presentation and communication of data analyses outcomes (Figure 1). Importantly, the involvement of different types of expertise for each layer is noted to emphasise that the development of a Big Data framework is not solely within the domain of computer scientists and IT specialists but rather requires collaboration across a number of water utility teams. In this section, the purpose and main principles required for the implementation of each layer are described based upon its application to water quality in DWDS.

3.1. The layers of the proposed framework

3.1.1. Layer 1: data storage layer

The data storage layer includes the storage of various types of databases through the use of data warehouse or cloud storage technology. The data storage software and system specifications will differ from one water utility to another, but the key capability for all such systems is the ability to store all available types of data including structured, unstructured, and asset information. In addition, it is very important that the format of the stored datasets makes them easily accessible for current and future use and supports linkages across databases with the use of unique IDs for every asset or sample. Therefore, the main principle required for this layer is the standardisation of data deposition, which sounds straightforward but is not trivial.

The data that water utilities collect may be grouped into two categories: the data regarding their assets that are static or changing infrequently over time, and time series data such as water quality and system operations or control data. Within the time series data, three further subcategories of data that are collected may be usefully defined: (1) telemetry data from permanent sensor installations at the WTW and key DWDS assets which are transmitted to a central data repository (typically SCADA), (2) discrete water quality (grab) samples from the WTWs and SRs outlets and sparsely from consumers' taps, and (3) other time series monitoring data from temporary sensors in the DWDS, installed for water quality investigations, research, and similar purposes, that are often not stored in the central data repository but rather in a separate application. In addition to internal data collected by the water utilities, Big Data applications may also require data from external sources, including parameters such as rainfall and air temperature.

Supplementary Appendix Table S1 summarises the key parameters required to support analytics for different types of water utility data, with a focus on water quality. The specifics of the parameters will differ by water utility, but baseline information will be required for assets and samples.

To gain the greatest value from Big Data techniques, linkages between datasets including aspects of physical connectivity are important to be included. For example, for a given water quality sample from the DWDS, it is ideal to be able to identify the pipes, water treatment works, and other relevant infrastructure supplying the sample location.

Regarding the quality of the available data sources (accounting for errors in measurement, sensor errors, missing values, GIS errors, etc.), this is partially examined in this work and partially related to the regulations. More specifically, GIS errors and static data are checked in this layer to follow the standardisation process presented in Supplementary Appendix

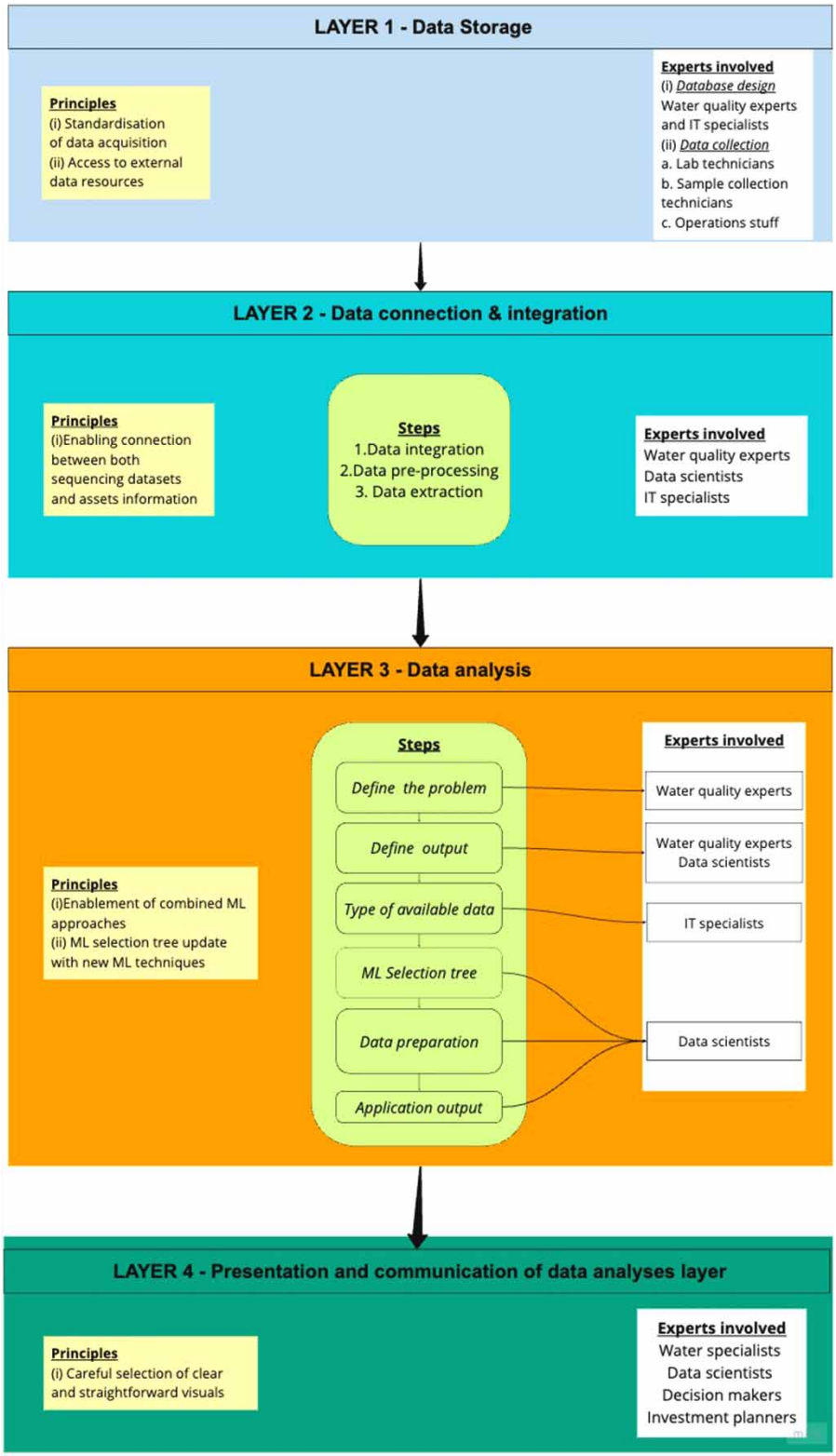


Figure 1 | The proposed Big Data framework labelled for data analysis applications in DWDS water quality.

Table S1. The quality of the telemetry and any other time series data are checked in the next layer using specific denoising and cleaning techniques. Finally, for the grab samples data quality, there are specifications and regulations that clearly describe the steps that should be followed from the sample collection up to the parameters' measurement in the laboratories (Standing Committee of Analysts 2002).

3.1.2. Layer 2: data connection and integration layer

This layer addresses the challenge of combining the various types of data and extracting the necessary parameters from storage in the previous layer data. The spatial connectivity between data elements is an important feature of DWDS data and harnessing this information in the Big Data analysis yields much more significant insight than considering water quality parameters alone.

This layer includes the production of a dataset fit for analysis by integrating across and between the different types of data. The integration component links the various types of data to each other and with their associated DWDS assets. For example, linking each water quality sample to its local pipe and service area asset hierarchy (district metered area, water operations zone, pressure zone, WTW, water source, etc.) is required to fully understand the route that the water follows to enable identification of the causes of deterioration. Data integration can be a complicated task. However, if good standardisation principles are followed in the data storage layer, the process can be facilitated using references and indices like water operations zone names or geospatial coordinates.

In the data pre-processing component, the integrated raw data are further cleaned to remove outliers, bad quality data, data with missing or not measured variables, or chronological periods that should not be included in the given analysis. The pre-processing may also require certain decisions to be made regarding the raw data, such as determining the way to connect WTWs' dataset with the dataset of the customers taps that they serve. Filling in missing values and removing unwanted parameters from the dataset are also included in this pre-processing. As regards the time series data, a further cleaning maybe required in this layer for denoising and cleaning the dataset.

Once pre-processing has been completed, the final component of this layer is the extraction of the clean dataset in the appropriate, accessible format, and ready for further analysis in the next layer.

3.1.3. Layer 3: data analysis layer

This data analysis layer is where analysis is performed using ML techniques for the creation of new knowledge from the available data. A critical component of this layer is the selection of the appropriate ML technique, which is dependent on the water quality question to be addressed, the type of output desired, and the type and quantity of the available data. We propose a six-step ML selection and implementation process as follows:

1. *Define the water quality problem:* The first step requires water quality experts to specify a task or a water quality question that has the potential to be addressed using data analytics solutions.
2. *Define the type of required output:* Once the problem is defined, the water quality experts should specify the goal of the investigation and desired type of output. For this framework, ML outputs have been categorised into four types: (1) prediction of future class (classification output – e.g., prediction of a water quality failure); (2) prediction of future behaviour (regression output – e.g., predicting the future values of certain parameters); (3) grouping of unlabelled data (clustering – e.g., splitting large datasets into groups based on various criteria); and (4) identifying relationships between parameters (correlation – e.g., identifying parameters that influence water quality deterioration). This definition should not be constrained by the available data initially, although it may need to be adapted through an iterative process with Step 3 to reflect the practicalities of the actual data.
3. *Type of available data:* This step connects this data analysis layer with the previous data integration layer. Here, the final extracted dataset is reviewed to determine the type, format, and most important quantity of data available. For example, continuous water quality monitoring data from the DWDS typically results in a dataset that is spatially and chronologically sparse and covers only a few water quality parameters. Given that the quantity of available data and the number of included parameters influence the performance of some ML techniques, this review of available data is important for ML technique selection. For example, artificial neural networks (ANNs) are generally not applicable to small and sparse datasets with a significant number of missing values (Ennett *et al.* 2001). In some cases, some initial exploratory analyses will be required to determine if the available data are sufficient to address the given water quality question and iterative reconsideration of Step 2 will be required.

4. *ML technique selection:* Building upon the previous three steps, the appropriate ML technique is selected in this step, facilitated by the use of *a ML technique selection tree* (2, further detail below). Some ML techniques cannot handle missing data and it is not always possible to infill the missing values, or the target of interest may be a rare event and the technique must be carefully selected to address this. Monitoring sample datasets are temporally and spatially sparse and, moreover, from each sample taken, not all the water quality parameters are measured. The ML methods selected for this framework are presented in detail in the next section.
5. *Data preparation:* Once the ML technique has been selected, this step includes any final changes to the data format required for the selected ML technique.
6. *Application output:* In this step, the selected ML technique/techniques are trained using the available data and tested to check their performance on unseen data. Once the simulations are finished, the outputs are specified. These could include images, values, or tables. The outputs are then reviewed to ensure that the ML technique has produced effective results using performance metrics techniques and multiple evaluation tests (k-fold techniques, further investigation on unseen data, sensitivity analysis, etc.)

3.1.4. ML selection tree

For a defined water quality problem (Figure 2, Box A), a path in the tree is followed considering the available data (Figure 2, Box B) and the desired output (Figure 2, Box C), all of which are considered in Steps 1 through 3 of the ML selection and implementation process.

The ML technique selection then proceeds by considering an additional factor that needs to be specified (Figure 2, Box D). This factor, termed 'interpretability', has been defined as the ability of the techniques to offer a transparent explanation of how

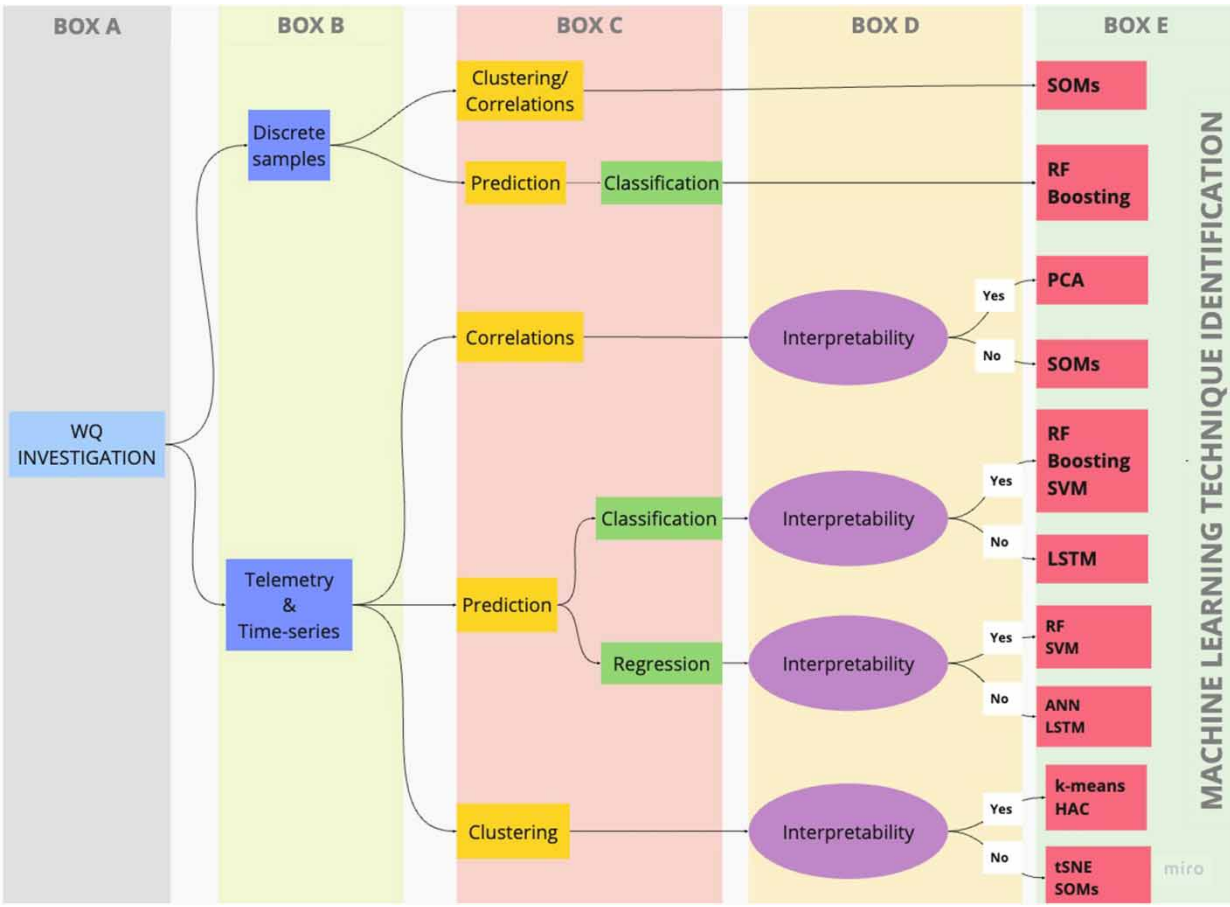


Figure 2 | Machine learning method selection tree.

outputs were calculated. Techniques with high interpretability, also known as 'white boxes', offer a way to clearly demonstrate the logic behind outputs and indicate the contributions of various parameters to decisions. Techniques with low interpretability, also known as 'black boxes', provide outputs without any explanatory elements. Different water quality questions might necessitate the selection of techniques with higher or lower interpretability. For example, while seeking a prediction of future iron failures in a DWDS, deriving an understanding of the parameters that influence the prediction could be as useful as the prediction itself.

Following the selection tree process, the final part identifies the appropriate ML technique (Figure 2, Box E) based upon the requirements for that specific water quality problem. Table 1 summarises such methods covering those with examples of successfully tested ML techniques. New ML methods are continually emerging from research; hence, the likes of Table 1 require regular updating. The methods investigated for this framework are those with demonstrated applications in the water sector. Some other ML methods that are used in natural and water systems presented in a review paper by Huang *et al.* (2021) could also be included in this tree. However, further investigation over their potential is required.

Neural networks and deep learning applications are applied to discover knowledge from large and complicated datasets (Lecun *et al.* 2015). Therefore, ANNs and LSTM techniques were both excluded from applications where discrete sample data are used because these datasets are spatially and temporally sparse with many missing values. Similarly, predictions based on regression techniques are not recommended with discrete samples due to their spatially and temporally sparse nature.

Table 1 | Summary of ML methods investigated

Method	Type	Output	Interpretable	Notes/Comments	Example References
k-means	Unsupervised	Clustering	Yes	Not suitable for datasets with missing values	Maimon & Rokach (2006)
Hierarchical Agglomerative Clustering (HAC)	Unsupervised	Clustering	Yes	Not suitable for datasets with missing values	Perruchet (1983)
Principal Component Analysis (PCA)	Unsupervised	Clustering/Linear correlations/ Dimensionality reduction	Yes	Not suitable for datasets with missing values	Jolliffe (2002)
Self-Organising Maps (SOMs)	Unsupervised	Clustering/Non-linear correlations	No	Good for datasets with missing values	Kohonen (1990)
Random Forest (RF)	Supervised	Classification/Regression	Yes	Ensemble decision trees with equal contribution to the final decision	Breiman (2001)
Boosting Trees (Boosting)	Supervised	Classification	Yes	Ensemble decision trees with weighted contribution to the final decision	Dietterich (2000)
Support Vector Machines (SVM)	Supervised	Classification/Regression	Yes	Method that splits data with hyperlanes	Kadyrova & Pavlova (2014)
Artificial Neural Networks (ANNs)	Supervised	Regression	No	Not suitable for datasets with missing values	Mounce *et al.* (2014)
t-Distributed Stochastic Neighbour Embedding (tSNE)	Unsupervised	Dimensionality reduction/clustering	No	Ignores input rows with missing values so not suitable for many discrete sample datasets	Maaten & Hinton (2008)
Long Short-Term Memory (LSTM)	Supervised	Regression/Classification	No	Deep learning method requires a large amount of data	Hochreiter & Schmidhuber (1997)

3.1.5. Layer 4: presentation and communication of data analyses outcomes

This layer overlaps with the application output step of the data analysis layer, taking the ML model outputs and presenting them using graphs, tables, and images to facilitate understanding and interpretation of the results by decision-makers. While the visual formatting is important, it is also critical that the most important and relevant results be carefully selected to clearly explain the ML outputs. Presentation of all outputs created in the ML analysis may create confusion for non-technical stakeholders but editing outputs for clarity must be balanced with providing sufficient evidence for interventions. Well-presented results provide sufficient and correct information to utility staff to make informed decisions for proactive interventions in the DWDS.

4. APPLICATION OF THE BIG DATA FRAMEWORK

To validate and evidence the value of the proposed framework, its application with a water utility located in the north of the UK is presented. This water utility serves more than 5.5 million people via 250 WTWs and greater than 50,000 kilometres of pipes, with approximately 1,100 SRs. Two case studies are presented covering different aspects of water quality performance in SRs. Specifically, the insights sought were evaluating the factors related to bacteriological activity in SRs, and the prediction of low chlorine concentration events in the SR outlets.

4.1. Example 1: factors related to increased bacteriological activity in SRs

4.1.1. Layer 1: data storage

The data storage layer of the Big Data framework comprised the water utility's in-house data management system with manual integration of external sources like weather data. The in-house data included discrete water quality samples taken from the outlets of SRs and WTWs. For each sample, various parameters were measured including bacteriological indicator parameters such as heterotrophic plate counts (HPCs) at 22 °C as well as flow cytometry total cell counts (FC_TCCs) and intact cell counts (FC_ICCs), disinfectant residual parameters, and other physical and chemical parameters. The data were stored in a parameter per line way, so that the number of lines required for each sample is equal to the number of parameters measured from that sample. No other information apart from the location that the sample was taken (WTW name, etc.) was given. In addition, databases that included asset information such as estimated retention time (water age) within each SR, type of secondary disinfection at each WTW (chlorine/chloramine), and connections between SRs and their source WTWs were stored in different files in their storage system. Finally, daily and hourly precipitation data were retrieved from the relevant Met Office weather stations (Met Office 2021).

4.1.2. Layer 2: data connection and integration

The investigation period was the period between January 2012 and May 2020, thus, discrete water quality samples taken from the outlets of SRs and WTWs in that period were collected. Initial cleaning of the dataset was made so that the exported dataset had changed each format from the original one, described in the previous layer, to one where each line represented a different sample and each column a different parameter measured, and every empty cell represented a non-measured parameter in that sample. Furthermore, pre-processing was made that included the generation of links between the water quality data, the connections between the SRs and the WTWs, and the disinfection type for each SR and WTW. The links between water quality data and the corresponding physical asset were created based on spatial, naming, and/or connectivity data. The connection between the SRs data and the water quality data of the WTWs that fed them was achieved by calculating the monthly average values for each parameter at WTWs and then by linking these data to each corresponding SR. Finally, the linkage between each SR and each WTW and their corresponding weather station was made using the spatial distance between them.

Overall, the additional parameters that were created and integrated with the main SRs' discrete samples dataset were as follows: (1) the age of water exiting an SR (hours) as the sum of the retention time of the given SR plus the retention time of the SRs that the water passed through upstream of the given SR (AgeofWaterLeavingSR); (2) the time (days) between the last reported SR cleaning date and the sample date (DaysFromCleaningDay), with negative values indicating samples that were taken before the last cleaning; (3) the monthly average total organic carbon in the WTWs (TOC_WTW_AVE); (4) the monthly average temperature of water exiting the WTWs (Temperature_WTW_AVE); (5) the monthly average flow cytometry total cell counts exiting the WTWs (FC_TCC_WTW_AVE); and (6) the daily average precipitation per month near the WTW (WTW_AverageDailyPrecipitation).

4.1.3. Layer 3: data analysis

The six steps within the ML selection and implementation process were performed for this water quality investigation as follows:

1. *Define the water quality problem*

The aim of this investigation is to understand the factors related to increased bacteriological activity in the SRs.

2. *Define the type of required output*

The required output in this investigation is correlations between parameters, with bacteriological parameters as the outcome parameters of interest. Numerical predictions are not required to understand these correlations.

3. *Type of available data*

As described above, the available data stems from the data integration layer. The outcome parameters of interest (bacteriological measurements HPC, FC_TCC, and FC_ICC) are only available as discrete samples taken from the outlet of the SRs. Telemetry data on water quality, as well as calculated daily average water quality from the WTW outlet, are also available. Weather data are available as time series data and calculated as daily average data. Physical data on the WTWs and SRs is also available.

4. *ML technique selection*

Given that this problem has an outcome characterised by discrete sample data and that the required output is correlation/clusters, the ML technique selection tree (Figure 2) directs towards SOMs as the ML method.

5. *Data preparation*

For the SOM application, the SR water quality data were prepared so that each row represents a discrete sample and each column is a different measured parameter from that sample, including the average monthly values of the water quality parameters at the WTW outlets feeding the sample location and the average daily precipitation per month in the given SR.

6. *Application output*

Two SOMs were selected from the analysis (Figures 3 and 4). These outputs consider many of the factors that have been shown to influence bacteriological water quality in the literature. The SOM analysis produces output planes for each parameter that visualise clusters of similar data by colour (low is blue, high is red) based on the range within the dataset, which in this investigation was set to colour-code based on the 5th and 95th percentile for each parameter without excluding any data. The SOM Toolbox 2.1 for MATLAB was used for all analyses (Helsinki University of Technology 2015).

The first SOM (Figure 3) investigates the effect of disinfectant residual type and concentration, along with retention time in the SR and temperature, on the bacteriological indicator parameters of HPC at 22 °C (HPC_22), flow cytometry total cell counts (FC_TCCs), and intact cell counts (FC_ICCs). Both free (FreeCl) and total (TotalCl) chlorine are plotted, with the type of disinfectant (chlorine or chloramine) used for a post-clustering labelled plot (right-hand side of Figure 3).

This SOM shows a large cluster of high HPC values (left half of the plane) with correlations to high and medium TCCs and high HPCs. The high HPC cluster also has a tendency to correlate with higher age of water exiting the SRs, correlates strongly with low free chlorine, and somewhat with elevated temperature. The labelled map, which is developed post-clustering analysis to assign categorical parameters that best match the members of each cell, shows a very strong correlation between high HPC values and chloraminated systems. High HPCs corresponded to the entire range of total chlorine concentrations and therefore indicate that bacteriological activity is less strongly associated with loss of disinfectant residual than with the type of disinfectant.

A cluster with increased numbers of ICCs (top centre of the plane) is correlated with high age of water exiting the SRs, high temperature, and low free and total chlorine in both chloraminated and chlorinated SRs. Interestingly, this analysis shows no clear correlation between ICCs and TCCs.

The impact of SR cleaning on bacteriological activity is somewhat less clear from this analysis. There is a cluster of low HPC values (lower left of the plane) within chloraminated systems that correlates with a higher number of days after cleaning. Considering that recently cleaned SRs show as light blue (horizontal band in the middle of the plane), there seem to be slightly lower HPC values corresponding to this cleaning, but not exclusively so.

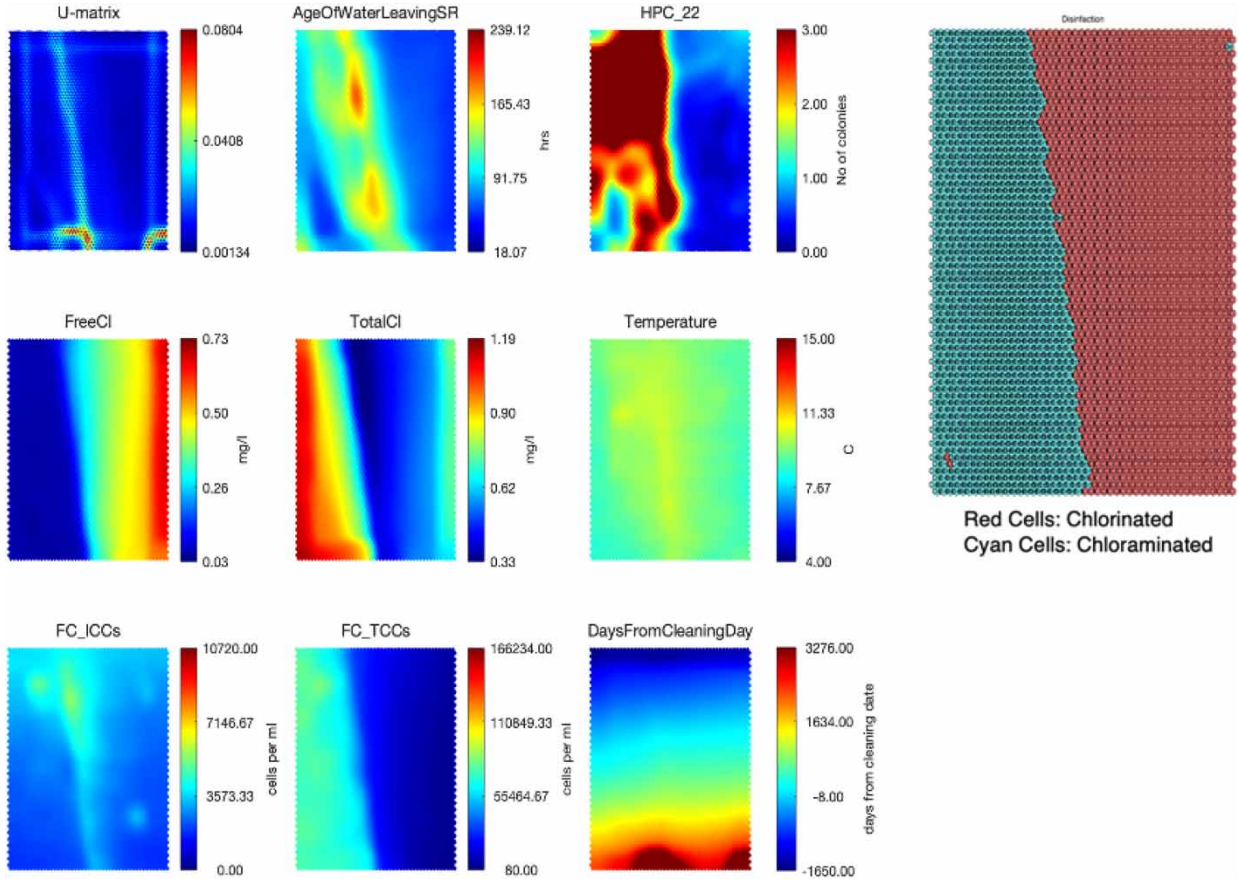


Figure 3 | SOM showing the impact of disinfectant residual type and concentration on bacteriological activity in SRs, including secondary disinfection labelled map. Please refer to the online version of this paper to see this figure in colour: https://doi.org/10.2166/aqua.2023.218.

The second SOM in this analysis (Figure 4) investigated the impact of key WTW water quality parameters (TOC and FC_TCCs) in addition to disinfectant type and concentration, water age in the SRs, and precipitation.

4.1.4. Layer 4: presentation and communication of data analyses outcomes

In this layer, the application outputs were presented in the form of colour-coded output planes (Figures 3 and 4). These outputs allow for the visual observation of correlations across multiple parameters and could be understood by a variety of water utility stakeholders. The outputs provided a good indication of which factors influence bacteriological activity in the SRs, although evidence in this format does not provide numerical values or estimates. As such, correlation analyses like these can answer general questions like the one posed in Step 1 of the ML selection and implementation process for this case study but cannot address questions that require a numerical output like a score or ranking.

4.2. Example 2: predicting low chlorine concentration events in the SRs

In this example application of the framework, the aim was to classify the SRs into either high-risk or low-risk categories based on a prediction of monthly low chlorine events by the ML models. The monthly temporal scale was selected as a prediction horizon as, on average, 2–4 monitoring samples per SR per month are collected for regulatory compliance (DWQR 2019). A low chlorine event was defined as a sample where the chlorine concentration was measured below 0.3 mg/l to allow a small margin above the allowable minimum free chlorine concentration at customers taps of 0.2 mg/l. The dataset used for this example was the same as the one used for Example 1 and therefore Layers 1 and 2 were already complete.

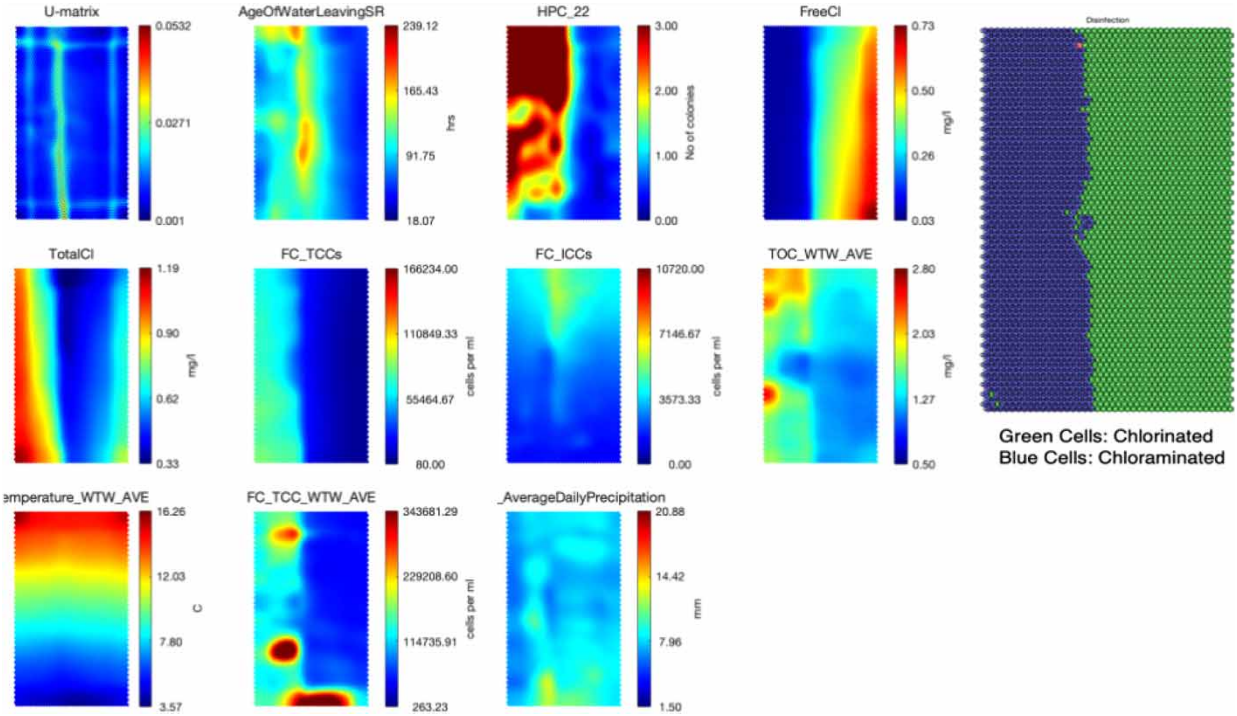


Figure 4 | SOM showing the additional impact of key WTW parameters on bacteriological activity SRs, including secondary disinfection labelled map. Please refer to the online version of this paper to see this figure in colour: https://doi.org/10.2166/aqua.2023.218.

4.2.1. Layer 3: data analysis

The six steps within the ML selection and implementation process were performed for this water quality investigation as follows:

1. *Define the water quality problem*

The aim of this investigation is to classify SRs as high or low risk by predicting their low free chlorine events in the upcoming month.

2. *Define the type of required output*

The required output is the classification of each SR into low and high-risk categories for the upcoming month based on low chlorine events.

3. *Type of available data*

The available data for the investigation are the water quality parameters from SR and WTW outlets during the period between January 2012 and December 2019 (complete years of data required, final 5 months from Example 1 not utilised). The outcome parameter of interest (FreeCl) is only available as discrete samples taken from the outlet of the SRs. Telemetry data on water quality as well as calculated daily average water quality from the WTW outlet are also available. Weather data are available as time series data and calculated as daily average data. Physical data on the WTWs and SRs are also available.

4. *ML technique selection*

Following the ML method selection tree in Figure 2, two types of ML techniques are suitable for this investigation, random forest and boosting trees. The SRs dataset is heavily unbalanced, meaning that most of the available data for training the ML model belong to the non-event, low-risk class. Therefore, RusBoost, a technique that combines random under sampling (of the non-event data) with the boosting tree algorithm, was selected for this analysis (Seiffert *et al.* 2008).

5. *Data preparation*

The initial dataset contained water quality data taken from all SRs and WTWs on different days. Therefore, this dataset required a final transformation to a monthly scale for analysis. Monthly averaged values per parameter per SR and the chlorine standard deviation per month per SR were calculated. Given that the results of the previous investigation revealed different behaviour in chlorinated and chloraminated SRs, all chloraminated SRs were excluded from this analysis. The historical classification of low-risk or high-risk was calculated for each SR for every month of the dataset. Within a given month, high-risk SRs were those that had one or more low chlorine events for that month (chlorine measured value below 0.3 mg/l in at least one discrete sample in that month). Low-risk SRs had no low chlorine events in the given month.

6. *Application output*

a. ML Model training

Two options for the ML model were tested during training. The first option used the water age exiting the SRs, the average daily precipitation per SR and the average monthly values of 15 water quality parameters per SR per month (Model RB.1 in Table 2). The second focused on the free chlorine parameters. More specifically, in the second option, the average monthly free chlorine, the monthly free chlorine standard deviation, and the average monthly WTW free chlorine along with water age exiting the SRs, and average water temperature comprising Model RB.2 (Table 2). The two latter parameters were included as input variables because both indicated a high correlation with chlorine in the previous case study.

During the training period, the ML model was used to predict the class (high-risk/low-risk) for each SR in the following month and this prediction was paired with the historical class for model training. For example, using the January 2012 water quality data as inputs, the historical SRs classification for February 2012 was produced as output (Figure 5). The ML models were developed in MATLAB (2019b) utilising 1,000 weak learners for each model and tested for their performance based on their predictions for August 2019. A simple schematic of the model training and testing is presented in Figure 5.

The accuracy of the models was evaluated by using a true positive rate as calculated with Equation (1) and the Matthews correlation coefficient (MCC) (Baldi *et al.* 2000) as calculated with Equation (2), as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{1}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{2}$$

where TP is the number of true positive predictions, FN is the number of false negative predictions, TN is the number of true negative predictions, and FP is the number of false positive predictions. True positive rate is used for quantifying the proportion of correctly predicted positives (events) over all the actual events. The MCC is used for evaluating the overall

Table 2 | Summary of the performance metrics for the two ML model options tested in Example 2

Algorithm	Model	Parameters	Most important predictor	TPR	MCC
RUSBoost	RB.1	Monthly free chlorine average, monthly free chlorine standard deviation, average monthly WTW total chlorine, monthly average HPCs @23C, monthly average HPCs @37C, monthly average ICCs, monthly average TCCs, average water temperature, water age exiting the SRs, average daily precipitation per month per SR, average monthly WTW free chlorine, average monthly WTW total chlorine, average monthly WTW TCCs, average monthly ICCs, average monthly WTW water temperature, average monthly WTW TOC, average monthly WTW pH	Monthly free chlorine average	0.71	0.44
	RB.2	Monthly free chlorine average, monthly free chlorine standard deviation, average monthly WTW free chlorine, water age exiting the SRs, average water temperature	Monthly free chlorine average	0.72	0.44

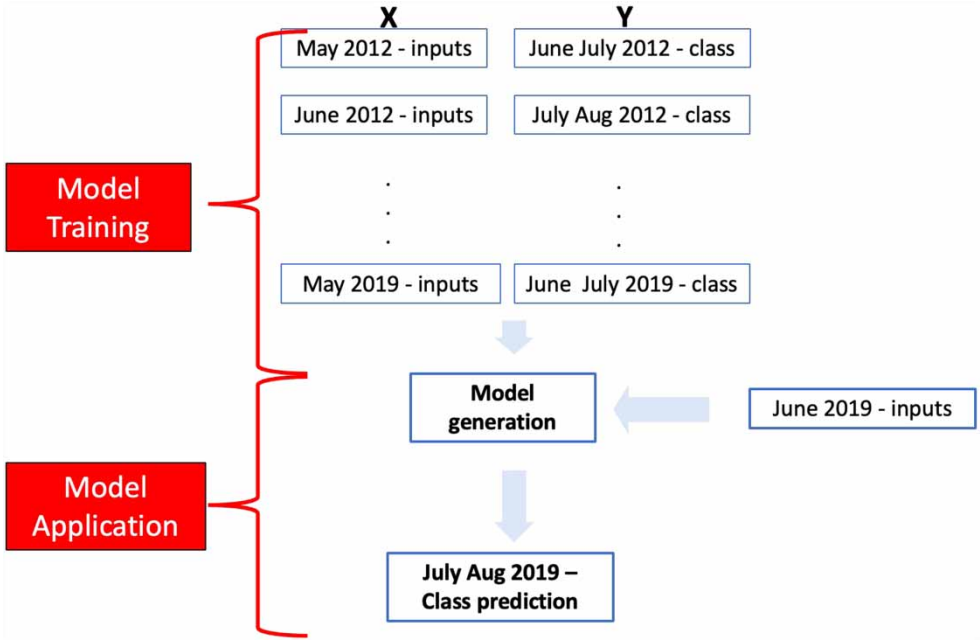


Figure 5 | Monthly predictive model schematic for SR class prediction for August 2019.

performance of the model and has a range between −1 and +1, where a result of −1 indicates a poor predictive capability for the model and values close to +1 indicate good predictive capability.

b. ML Model results

The MCC results for the two ML model options (Table 2) show that both RB.1 and RB.2 performed well, especially in keeping a balance between correctly predicting more high-risk SRs (TPR = 0.71 and TPR = 0.72, respectively) and creating less false positives (MCC = 0.44 for both). Overall, RB.2 can be considered the best model given its slightly higher true positive rate.

4.2.2. Layer 4: presentation and communication of data analyses outcomes

In this example, the outputs presented to the utility decision-makers included a list of SRs and their predicted risk categories for a given month. Given that the boosting tree algorithm is a white-box model, part or all the decision trees that contributed to the final predictions were exported to illustrate which water quality parameters contributed the most to the model predictions (e.g., the most important predictor in Table 2). This valuable additional information, associated with the Interpretability factor for the selected ML method, allows the utility to understand not only the risk classification but also the factors that lead to this risk.

5. DISCUSSION

5.1. Actionable information

The two example case studies delivered a variety of actionable information. In the first example, the water quality question was to identify factors that influence the bacteriological activity inside the SRs. The SOM outputs provided the evidence to support actions including: (1) closer investigation of chloraminated SRs and maintenance of disinfectant residual, especially when the temperature increases; (2) control and optimisation of the retention time in both the chlorinated and chloraminated SRs; (3) management and reduction of TOC exiting the WTWs; (4) improvements in the WTWs operation to respond to sudden changes in the quality of the raw water due to increased precipitation. These case study outputs agree with findings in different water quality research works. In their review paper, Prest *et al.* (2016) indicated that water age and temperature as key parameters that influence the bacteriological activity in the DWDS. Moreover, TOC was indicated as one of the key factors that govern bacteriological regrowth in the DWDS (LeChevallier *et al.* 1991; Fish *et al.* 2016). Additional studies may be

required to fully explore the suggested by this case study actions, including collection of data that were not available for the SOM analysis, but the Big Data approach has provided focus and clarity for such additional studies that would not otherwise have been possible.

In the second case study example, a prediction model for high-risk SRs based on their low chlorine events each month was created. The model was able to predict up to 72% of the low chlorine events for the investigation month, which is a high degree of accuracy that allows targeted interventions to take place. Furthermore, the fact that low chlorine events can be accurately predicted using mainly chlorine-related parameters is important, demonstrating that monitoring of supplemental parameters (e.g., other bacteriological indicators) would not be warranted to address this particular water quality question. The inclusion of the age of water and the water temperature as input parameters in the predictive model (RB.2 model) contributed to the improved results. This output reinforces the argument that low chlorine is in general related to high values of those two parameters as presented in the previous case study and also agrees with the work of Kerneïs *et al.* (1995) regarding the factors that influence chlorine residual in the DWDS.

Part of the appeal of Big Data analytics is the ability to answer increasingly complex questions and make predictions for the future. Drawing upon the case study examples, basic data analysis could map the SRs with low chlorine measurements each month, perhaps identifying geographical areas with clusters of high-risk SRs. But such an analysis cannot identify underlying factors that contribute to the low chlorine events and cannot predict which SRs may have problems next month. It is this deeper understanding through iteration of the third layer that was key in obtaining the actionable information derived.

The predictive model developed in the second case study is completely data-driven without any use of the hydraulic model. Thus, it has the potential to be transferred in any other part of the DWDS (investigation on DMAs or WTWs, etc.) where enough data exist. This is one of its main advantages in comparison to the deterministic/numerical models. Another advantage of this model is that it does not require any parameter calibration as the traditional deterministic models do. Moreover, for this case study, creating a deterministic model that predicts low chlorine in SRs' outlet is practically impossible as the mechanisms that work inside them are not well known, characterised or quantified.

For the water quality examples used, changes occur due to complex physical, chemical, and biological reactions and interactions occurring inside water infrastructure. It was, therefore, important that the Big Data analytics investigation was directed by domain knowledge, the posing of the question and understanding sought that drives the third layer. This is crucial to ensure actionable information results. This finding is an underlying principle of the framework, whatever the application. It requires collaboration between experts in different water utility departments and with data scientists, in all the layers of the framework.

5.2. Framework

The Big Data framework, presented in this paper with applications to drinking water quality, emphasises the necessary steps to unlock the power of ML and advanced data analytics for water utilities. The framework systemises a process to ensure that actionable information is derived by unlocking the potential of previously siloed data. Importantly, a selection tree process to identify the best ML techniques driven from the insight required and the data available is central. This is based on the knowledge and illustrates that there is much more effort required for successful Big Data applications than coding a given ML algorithm.

Standardising data acquisition and storage, organising the data to facilitate analysis including generating links between different datasets, understanding the difference between available data types, and selecting the most appropriate data-driven techniques are all necessary steps to deliver actionable information and supporting evidence to inform operation and management decisions. Implementation of standards that guarantee the collection of good quality data and the organisation of the stored data are often under-resourced tasks at water utilities yet have been shown to be core elements of this framework. As the presented examples demonstrate, the absence of a proper standardisation of the data collection and storage and the absence of access to the external datasets (precipitation data), for that particular water utility, made the process of collecting the data extremely complicated. In addition, the computational effort required for the data integration in Layer 2 was significant. This indicates the importance of a proper standardisation of the data acquisition to reduce the time spent and justifies the importance of that principle in the proposed framework. The data-specific nature of current systems within the water industry as well as the lack of historical collaboration between the relevant areas of expertise perhaps explain why there have been so few Big Data frameworks proposed for the water sector before now.

One of the most challenging but vital aspects of Layer 2 is the generation and association of links between different data. Linkages between different data sources, such as between asset information and measured water quality parameters, are critical for ML analyses yet are not often performed. For example, analysis of water quality sample data without consideration of the water treatment works supplying a given point in a network often falls short in answering the questions of interest. This is a key area where data scientists and water domain experts must work closely to understand what is possible and to ensure that appropriate associations are made. Unique ID and geocoding data are often useful here, but these should be supplemented by checking secondary data. For example, linking a pipe repair record to GIS data can be done based on location data, with a check made using pipe material data that is also frequently contained in both datasets.

The initial investment of time and effort for data collection, storage, and integration (Layers 1 and 2) is often greater than what is needed for data analysis (Layer 3). However, once created, Layers 1 and 2 can then be used to support a multitude of analyses repeating and iterating Layers 3 and 4 without the need to revisit Layers 1 and 2. This was shown for the two case study examples presented. The return on initial investment in Layers 1 and 2 can be further multiplied many times if automatic updating of datasets can be incorporated. The need for investment in Layers 1 and 2 is great but the benefits will be felt across the water industry when the analytical power and decision-making support of Layers 3 and 4 is unlocked.

Once Layers 1 and 2 are complete, the question, and opportunity, becomes to consider what new and actionable information is needed and how to extract it. Studies in the literature which explored specific algorithm(s) with application(s) to a few water quality parameters have paved the way to greater understanding of the potential in Big Data analytics, but the water industry lacked an understanding of how best to apply these techniques and which ones work best in which situations. The ML method selection tree proposes a novel, problem-driven approach to this, enabling a wide range of investigations and opening up the possibilities for taking Big Data analytics to the next level of application across the water industry.

This framework does not examine the sufficiency of the discrete water quality (grab) samples collected by water utilities to answer questions of relevance. The number of samples that water utilities are required to collect are set by the national Regulators like the Drinking Water Inspectorate (DWI) for England and Wales and the Drinking Water Quality Regulator (DWQR) for Scotland. The authors believe that further discussions between the Regulators, the water utilities and researchers are required that surpass the objectives that this work aims to address. This collaboration between these various experts should focus on answering the question ‘how much data are enough to answer the water quality-related problems’.

6. CONCLUSIONS

A Big Data framework to enable water utilities to robustly and efficiently apply data-driven methods to derive new understanding from complex, traditionally siloed data is presented. The proposed framework is based on a four-layer approach:

1. The data storage layer includes a system to categorise and sort data.
2. The data integration layer, where the importance of associating across and between data is emphasised, irrespective of types, formats, and sources of data.
3. The data analysis layer is systemised as a six-stage process that is driven from precise articulation of the decisions that are to be informed. These steps are (1) definition of the problem; (2) definition of the type of required output; (3) type of the available data; (4) selection of the ML learning technique; (5) data preparation; and (6) application output. A selection tree is used to inform the selection of ML technique based on three criteria: the available data (discrete water quality samples or time series), the required output, and the need for interpretability of the process for producing the outputs.
4. The presentation and communication of data analyses outcomes, where careful selection from the huge number of outputs generated is essential to present a logic effective and evidence-based narrative.

The need for and integration of roles across water engineering and data science are set out across the framework. Layers 1 and 2 are often complex and time-consuming. However, once comprehensively accomplished they readily enable a multitude of different explorations of the different data to derive different and deeper understanding by repeating and iterating Layers 3 and 4.

Case study examples evidence the application of the framework for drinking water quality. The examples demonstrate the derivation of new understanding, such as the association between disinfection residual type and concentration and age of water exiting a service reservoir combining to correlate with increased two-day plate counts, and for the prediction of low chlorine events at the outlet of service reservoirs. This understanding readily informs operational decisions, such as managing

disinfection residual dose and prioritisation of maintenance activities. Overall, the framework is demonstrated to provide robust data-driven understanding and evidence to inform vital water utility operational and maintenance decisions.

ACKNOWLEDGEMENTS

This work was funded by the EPSRC Centre for Doctoral Training in Engineering for the Water Sector (STREAM IDC, EP/L015412/1) and Scottish Water. The authors gratefully acknowledge Claire Thom and other staff at Scottish Water for their input and assistance.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abdullah, M. F., Zaki, M., Amin, M., Mohamad, M. F. & Marini, M. 2018 N-HyDAA – Big Data analytics for Malaysia climate change knowledge management. **13**, 1–7.

Ahmed, I., Ahmad, M., Jeon, G. & Piccialli, F. 2021 A framework for pandemic prediction using Big Data analytics. *Big Data Research* **25**. https://doi.org/10.1016/j.bdr.2021.100190.

Alpaydin, E. 2014 *Introduction to Machine Learning*, 3rd edn. The MIT Press, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. & Nielsen, H. 2000 Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16** (5), 412–424. https://doi.org/10.1093/bioinformatics/16.5.412.

Blokker, E. J., Furnass, W., Machell, J., Mounce, S., Schaap, P. & Boxall, J. 2016 Relating water quality and age in drinking water distribution systems using self-organising maps. *Environments* **3** (2), 10. https://doi.org/10.3390/environments3020010.

Breiman, L. 2001 Random forests. *Machine Learning* **45**, 5–32. https://doi.org/10.3390/rs10060911.

Carreño-Alvarado, E. P., Reynoso-meza, G., Montalvo, I. & Izquierdo, J. 2017 A comparison of machine learning classifiers for leak detection and isolation in urban networks. In: *Congress on Numerical Methods in Engineering*, 3–5 July 2017, Valencia, Spain. Available from: https://www.researchgate.net/publication/318275002_A_comparison_of_machine_learning_classifiers_for_leak_detection_and_isolation_in_urban_networks.

Chandarana, P. & Vijayalakshmi, M. 2014 Big data analytics frameworks. In: *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications, CSCITA 2014*, pp. 430–434. https://doi.org/10.1109/CSCITA.2014.6839299.

Dairi, A., Cheng, T., Harrou, F., Sun, Y. & Leiknes, T. O. 2019 Deep learning approach for sustainable WWTP operation: a case study on data-driven influent conditions monitoring. *Sustainable Cities and Society* **50**, 101670. https://doi.org/10.1016/j.scs.2019.101670.

Dietterich, T. G. 2000 Ensemble methods in machine learning. In: *Multiple Classifier Systems (MCS 2000). Lecture Notes in Computer Science, 1857 LNCS*. pp. 1–15. https://doi.org/10.1007/3-540-45014-9_1.

Du, J., Kuang, B. & Yang, Y. 2019 A data-driven framework for smart urban domestic wastewater: a sustainability perspective. *Advances in Civil Engineering* **2019**. https://doi.org/10.1155/2019/6530626.

DWI 2016 *The Water Supply (Water Quality) Regulations 2016, Statutory Instruments (England and Wales) No.614*. Available from: http://www.legislation.gov.uk/uksi/2016/614/made/data.pdf.

DWQR 2019 *Drinking Water Quality in Scotland 2018: Public Water Supply*.

Ennett, C. M., Frize, M. & Robin Walker, C. 2001 Influence of missing values on artificial neural network performance. *Studies in Health Technology and Informatics* **84**, 449–453. https://doi.org/10.3233/978-1-60750-928-8-449.

Fellini, S., Vesipa, R., Boano, F. & Ridolfi, L. 2018 Real-time measurement fault detection and remote-control in a mountain water supply system. In *Proceeding of 13th International Conference on Hydroinformatics*, Palermo, Italy.

Fish, K. E., Mark Osborn, A. & Boxall, J. 2016 Characterising and understanding the impact of microbial biofilms and the extracellular polymeric substance (EPS) matrix in drinking water distribution systems. *Environmental Science: Water Research & Technology* **2** (4), 614–630. https://doi.org/10.1039/C6EW00039H.

Gandomi, A. & Haider, M. 2015 Beyond the hype: Big Data concepts, methods, and analytics. *International Journal of Information Management* **35** (2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Garcia, D., Puig, V. & Quevedo, J. 2020 Prognosis of water quality sensors using advanced data analytics: application to the Barcelona drinking water network. *Sensors (Switzerland)* **20** (5). https://doi.org/10.3390/s20051342.

Gibbs, M. S., Morgan, N., Maier, H. R., Dandy, G. C., Nixon, J. B. & Holmes, M. 2006 Investigation into the relationship between chlorine decay and water distribution parameters using data driven methods. *Mathematical and Computer Modelling* **44** (5–6), 485–498. https://doi.org/10.1016/j.mcm.2006.01.007.

Helsinki University of Technology 2015 *SOM Toolbox (for MATLAB)*. Available from: https://github.com/ilarinieminen/SOM-Toolbox.

Herrera, M., Torgo, L., Izquierdo, J. & Pérez-García, R. 2010 Predictive models for forecasting hourly urban water demand. *Journal of Hydrology* **387** (1–2), 141–150. https://doi.org/10.1016/j.jhydrol.2010.04.005.

Hochreiter, S. & Schmidhuber, J. 1997 Long short-term memory. *Neural Computation* **9** (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Huang, R., Ma, C., Ma, J., Huangfu, X. & He, Q. 2021 Machine learning in natural and engineered water systems. *Water Research*. https://doi.org/10.1016/j.watres.2021.117666.

IWA 2019 *Digital_Water: Industry Leaders Chart the Transformation Journey*. IWA Publications, London.

Jolliffe, I. T. 2002 *Principal Component Analysis*, 2nd edn. Springer US, New York. https://doi.org/10.1007/BF01884351.

Kadyrova, N. O. & Pavlova, L. V. 2014 Statistical analysis of Big Data: an approach based on support vector machines for classification and regression problems. *Biophysics* **59** (3), 364–373. https://doi.org/10.1134/S0006350914030105.

Kazemi, E., Mounce, S., Husband, S. & Boxall, J. 2018 Predicting turbidity in water distribution trunk mains using nonlinear autoregressive exogenous artificial neural networks. In: *Proceeding of 13th International Conference on Hydroinformatics*, Palermo, Italy.

Kerneïs, A., Nakache, F., Deguin, A. & Feinberg, M. 1995 The effects of water residence time on the biological quality in a distribution network. *Water Research* **29** (7), 1719–1727. https://doi.org/10.1016/0043-1354(94)00323-Y.

Kohonen, T. 1990 The self-organizing map. *Proceedings of the IEEE* **78**, 1464–1480. https://doi.org/10.1109/5.58325.

LeChevallier, M. W., Schulz, W. & Lee, R. G. 1991 Bacterial nutrients in drinking water. *Applied and Environmental Microbiology* **57** (3), 857–862.

Lecun, Y., Bengio, Y. & Hinton, G. 2015 Deep learning. *Nature* **521** (7553), 436–444. https://doi.org/10.1038/nature14539.

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**. https://doi.org/10.1016/j.cej.2020.126673.

Maaten, L. v. d. & Hinton, G. 2008 Visualizing data using T-SNE Laurens. *Journal of Machine Learning Research* **9**, 2579–2605. https://doi.org/10.1007/s10479-011-0841-3.

Maimon, O. & Rokach, L. 2006 *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Researchers and Practitioners*, 1st edn. Springer US, Tel-Aviv.

Mamandipoor, B., Majd, M., Sheikhalishahi, S., Modena, C. & Osmani, V. 2020 Monitoring and detecting faults in wastewater treatment plants using deep learning. *Environmental Monitoring and Assessment* **192** (2), 1–12. https://doi.org/10.1007/s10661-020-8064-1.

Met Office 2021 *MIDAS Open: UK Daily Rainfall Data, V202007*. https://doi.org/10.5285/ec9e894089434b03bd9532d7b343ec4b. Centre for Environmental Data Analysis.

Meyers, G., Kapelan, Z. & Keedwell, E. 2017 Short-term forecasting of turbidity in trunk main networks. *Water Research* **124**, 67–76. https://doi.org/10.1016/j.watres.2017.07.035.

Mohammed, H., Hameed, I. A. & Seidu, R. 2017 Random forest tree for predicting fecal indicator organisms in drinking water supply. In: *Proceedings of 4th International Conference on Behavioral, Economic, and Socio-Cultural Computing, BESC 2017*. https://doi.org/10.1109/BESC.2017.8256398.

Mounce, S. R., Boxall, J. B. & Machell, J. 2010 Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows. *Journal of Water Resources Planning and Management* **136** (3), 309–318. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000030.

Mounce, S. R., Shepherd, W., Sailor, G., Shucksmith, J. & Saul, A. J. 2014 Predicting combined sewer overflows chamber depth using artificial neural networks with rainfall radar data. *Water Science and Technology* **69** (6), 1326–1333. https://doi.org/10.2166/wst.2014.024.

Mounce, S. R., Ellis, K., Edwards, J. M., Speight, V. L., Jakomis, N. & Boxall, J. B. 2017 Ensemble decision tree models using RUSBoost for estimating risk of iron failure in drinking water distribution systems. *Water Resources Management* **31** (5), 1575–1589. https://doi.org/10.1007/s11269-017-1595-8.

Osman, A. M. S. 2019 A novel Big Data analytics framework for smart cities. *Future Generation Computer Systems* **91**, 620–633. https://doi.org/10.1016/j.future.2018.06.046.

Perruchet, C. 1983 Constrained agglomerative hierarchical classification. *Pattern Recognition* **16** (2), 213–217. https://doi.org/10.1016/0031-3203(83)90024-9.

Prest, E. I., Hammes, F., van Loosdrecht, M. C. M. & Vrouwenvelder, J. S. 2016 Biological stability of drinking water: controlling factors, methods, and challenges. *Frontiers in Microbiology* **7**, 1–24. https://doi.org/10.3389/fmicb.2016.00045.

Romano, M., Kapelan, Z. & Savić, D. A. 2014 Automated detection of pipe bursts and other events in water distribution systems. *Journal of Water Resources Planning and Management* **140** (4), 457–467. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000339.

Romero, J. M. P., Hallett, S. H. & Jude, S. 2017 Leveraging Big Data tools and technologies: addressing the challenges of the water quality sector. *Sustainability (Switzerland)* **9** (12). https://doi.org/10.3390/su9122160.

Rosin, T. R., Kapelan, Z., Keedwell, E. & Romano, M. 2022 Near real-time detection of blockages in the proximity of combined sewer overflows using evolutionary ANNs and statistical process control. *Journal of Hydroinformatics* **24** (2), 259–273. https://doi.org/10.2166/hydro.2022.036.

Seiffert, C., Khoshgoftaar, T. M., Van Hukse, J. & Napolitano, A. 2008 RUSBoost: improving classification performance when training data is Skewed.Pdf. In: *19th International Conference on Pattern Recognition (ICPR 2008)*, 8–11 December 2008, Tampa, Florida, USA. https://doi.org/10.1109/ICPR.2008.4761297.

Speight, V., Mounce, S. & Boxall, J. B. 2019 Identification of the causes of drinking water discolouration from machine learning analysis of historical datasets. *Environmental Science: Water Research and Technology* **5** (4), 747–755. https://doi.org/10.1039/c8ew00733k.

Standing Committee of Analysts 2002 *The Microbiology of Drinking Water (2002). Part 1 - Water Quality and Public Health Methods for the Examination of Waters and Associated Materials*.

Vries, D., Van Den Akker, B., Vonk, E., De Jong, W. & Van Summeren, J. 2016 Application of machine learning techniques to predict anomalies in water supply networks. *Water Science and Technology: Water Supply* **16** (6), 1528–1535. https://doi.org/10.2166/ws.2016.062.

Xenochristou, M., Hutton, C., Hofman, J. & Kapelan, Z. 2021 Short-term forecasting of household water demand in the UK using an interpretable machine learning approach. *Journal of Water Resources Planning and Management* **147** (4). https://doi.org/10.1061/(asce)wr.1943-5452.0001325.

Xu, R., Cao, J., Fang, F., Feng, Q., Yang, E. & Luo, J. 2021 Integrated data-driven strategy to optimize the processes configuration for full-scale wastewater treatment plant predesign. *Science of the Total Environment* **785**, 147356. https://doi.org/10.1016/j.scitotenv.2021.147356.

Zacarias, V., Gabriel, A., Reimann, P. & Mitschang, B. 2018 A framework to guide the selection and configuration of machine-learning-based data analytics solutions in manufacturing. *Procedia CIRP* **72**, 153–158. https://doi.org/10.1016/j.procir.2018.03.215.

Zekić-Sušac, M., Mitrović, S. & Has, A. 2020 Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2020.102074.

First received 9 December 2022; accepted in revised form 8 April 2023. Available online 26 April 2023