AQUA
Water Infrastructure, Ecosystems and Society

IWA PUBLISHING

 AQUA — Water Infrastructure, Ecosystems and Society Vol 72 No 12, 2216 doi: 10.2166/aqua.2023.202

An angle-based leak detection method using pressure sensors in water distribution networks

Huimin Yu[a,b], Hua Zhou[c], Xiaodan Weng[c], Zhihong Long[d], Yu Shao[a,b] and Tingchao Yu[a,b,*]

[a] Zhejiang Key Laboratory of Drinking Water Safety and Distribution Technology, Zhejiang University, Hangzhou 310058, China
[b] Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing 314100, China
[c] Huadong Engineering Corporation Limited, Hangzhou 311122, China
[d] Guangzhou Water Supply Co., Ltd, Guangzhou 510600, China
*Corresponding author. E-mail: yutingchao@zju.edu.cn

ABSTRACT

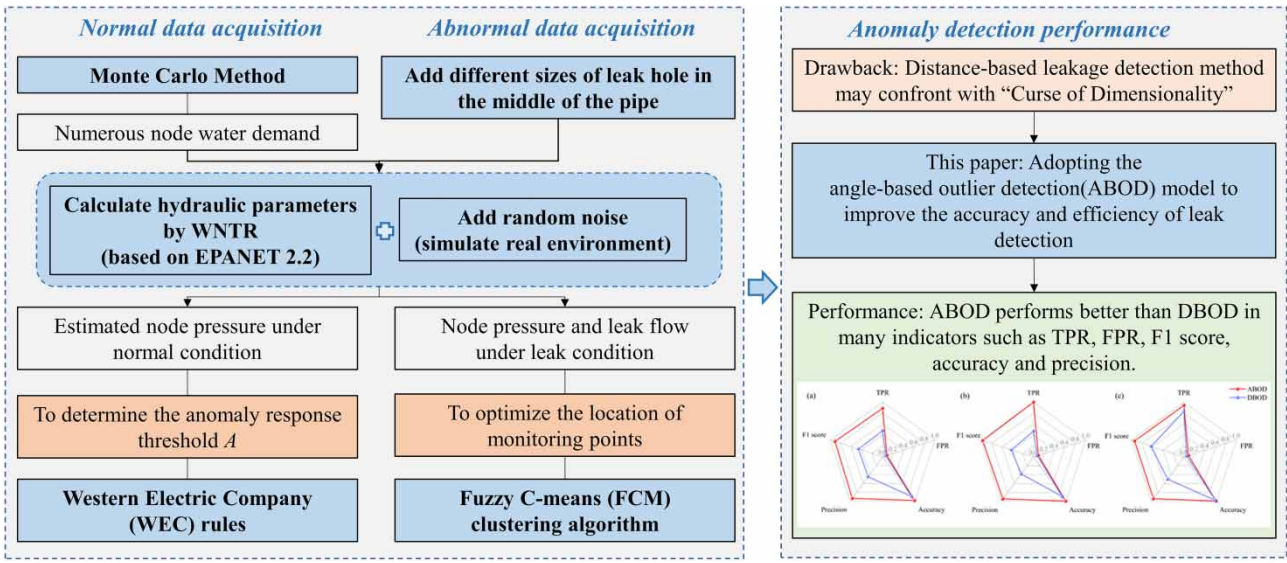
Leak detection has significant implications for the long-term stable operation of water distribution networks (WDNs). This study presented a novel leak detection method by calculating the angular variance between a pressure vector and other vectors in the database, to evaluate the presence of an anomaly in a network. The top priority for this method was to establish a reliable dataset collected from the pressure sensors, which is generated by EPANET 2.2. Numerous node water demand data in normal conditions were generated by the Monte Carlo method, and leak conditions with various leak flows were simulated by creating leak holes in the pipes. Through learning the composite normal and abnormal data in a certain proportion, the angle-based outlier detection model was employed to identify abnormal events. This angle-based method was applied in an actual WDN and the identification performance for anomalies was compared with that of previous detection methods. The results indicated that the novel method proposed in this study could significantly improve the accuracy and efficiency of leak detection compared to the threshold-based and distance-based detection methods.

Key words: angle-based algorithm, hydraulic analysis, leak detection, leak flow rate, leak simulation, water distribution networks

HIGHLIGHTS

- The proposed method for leak detection in the WDN combines the hydraulic model with outlier detection.
- Numerous normal and leak scenarios are simulated by the hydraulic model to calculate residuals.
- A clustering algorithm is used to determine the optimal locations of pressure sensors.
- The performance of the proposed method is compared with the distance-based method in the literature.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Leakages in water distribution networks (WDNs) are key issues in water resource allocation management. Not only the depletion of water resources, serious leakages have caused enormous energy dissipation for water treatment and the intrusion of external contaminants through broken pipes (Fontanazza *et al.* 2015; Beker & Kansal 2022). According to statistics, in 2020, the water loss was 7.85 billion m^3 in Chinese cities, with a comprehensive leak rate of 13.39% (CUWA 2020). Therefore, a timely and reliable detection method of anomaly makes assurance for water utilities to detect leakages more promptly, and economize more precious water resources.

Leakages can be formed on mains and service pipes (Farley & Trow 2003), which may be caused by various factors, such as the poor quality of pipes, inappropriate operation, and extreme weather (e.g., freezing weather). Therefore, it is hard to find out when and where leakages happened. To improve the efficiency of the anomaly identification method, numerous methods have been studied and developed. These methods can be generally divided into two main categories: one based on hardware and the other on software (Valizadeh *et al.* 2009).

Many technologies are based on highly specialized hardware equipment, such as leak noise correlators (Guo *et al.* 2021), leak noise loggers (Muggleton *et al.* 2006), gas injection (Hunaidi *et al.* 2000), ground penetrating radar (Hunaidi 1998), and infrared photography (Fahmy & Moselhi 2010). Despite the fact that they perform well in leak detection and location (Puust *et al.* 2010), some drawbacks they have cannot be ignored. For example, they are labor-intensive, expensive, run slowly, and may also require long interruption of pipeline operations (Romano *et al.* 2011).

With the widespread applications of advanced meter infrastructure, data loggers, and sensors, it is imperative to integrate the hydraulic model with data processing methods for higher quality and efficiency of water management (Wu & He 2021). Methods based on software are to actively analyze the signals measured in WDNs to detect leakages. The obvious signals of wireless sensors installed in WDNs are flow and pressure. Methods based on software can be roughly categorized into transient-based approaches, model-based approaches, and data-driven approaches. Transient-based approaches focus on analyzing information about the presence of leakages from transient pressure signals measured within a WDN (Liggett & Chen 1994; Christodoulou *et al.* 2017). Experimental results demonstrate that even a small leak point can be detected by this method (Mpesha *et al.* 2001; Covas *et al.* 2005). However, restricted by technologies and costs, current transient-based methods were not specifically authenticated in actual networks (Li *et al.* 2015).

Benefitting from technical improvements in modeling software and the popularity of supervisory control and data acquisition (SCADA) systems, model-based approaches have improved over the last two decades (Kim *et al.* 2010; Yu *et al.* 2021). These methods detect leakages by comparing flow or pressure data by simulating the hydraulic conditions of WDNs. Compared with other methods, establishing a hydraulic model with certain monitoring information to detect leakages

can be more economical and simpler in the application. However, these models need to be updated whenever the topology changes and necessitate more sophisticated equations for calibration to have a good performance.

Given limitations in the aforementioned approaches, with the development of computing capabilities, data-driven approaches have been intensively studied since the beginning of the 21st century. The data-driven approaches can be classified as supervised and unsupervised approaches including prediction, classification, and clustering (Zaman *et al.* 2020). The purpose of the prediction stage is to generate estimated data under normal conditions of networks. Mounce *et al.* (2010) introduced artificial neural networks for leak detection by analyzing the similarity of abnormal pressure or flow variation, which can model any function without specific parameters involved to handle complex hydraulic problems (Mounce *et al.* 2002; Romano *et al.* 2014). The classification stage is then adopted to compare the residuals between predicted values and actual measurements to evaluate abnormal events (Mounce *et al.* 2011; Bakker *et al.* 2014). These methods make it difficult to learn complex features, so Zhou *et al.* (2019) proposed a novel burst location identification framework by fully linear DenseNet, which supersedes the convolutional layers in DenseNet by linear connections. In order to reduce the misjudgments in the prediction process, the clustering-based approaches were studied to detect leakages according to the similarities of monitoring data (Wu *et al.* 2016), which are implemented on the district metering area (DMA) level. However, due to the large scale of the networks in many cities, distance-based anomaly detection algorithms may cause 'Curse of Dimensionality', leading to the deterioration of performance.

This paper aims to improve the accuracy and efficiency of leak detection, and the method used in this paper is calculating the angle variance for each candidate data vector (containing pressure values from all sensors at the same time). The current objectives of the research include four parts: (1) importing the sensitivity matrix to the fuzzy c-means (FCM) algorithm, and the cluster centers were calculated to determine installation positions of sensors, (2) simulating normal operating conditions by the Monte Carlo method and collecting various abnormal events by the water network tool for resilience (WNTR), (3) evaluating the performance of the proposed method and other outlier detection methods in large leakages, and (4) comparing the accuracy of angle-based method with the distance-based method in minor leak scenarios.

2. METHODS

The proposed method is a novel data-driven leak detection in the WDN. In the following research contents, the steps for normal and leak scenario simulations will be given priority. Then, the FCM clustering algorithm was used to cluster nodes according to the node pressure values under completely burst conditions, and the clustering centers were selected as the placement of the monitoring points. Finally, the angle-based outlier detection (ABOD) model was adopted to identify abnormal events in the WDN accurately by learning from a certain amount of unlabeled normal and abnormal data. A schematic of the leak detection method is shown in Figure 1, with a detailed explanation in the subsequent sections.

2.1. Leak scenarios generation

There are many methods of simulating leakages such as the emitter method (Giustolisi *et al.* 2008), the artificial reservoir method (Ang & Jowitt 2006), and the additional node demand method (Shao *et al.* 2019). In this paper, the simulating method is adding a leak hole in the middle of the pipe and analyzing the hydraulic operating parameters by the WNTR. WNTR is an open-source Python package designed to help water utilities simulate and analyze the resilience of WDN, which is compatible with EPANET 2.00.12 and EPANET 2.2.

The leak model adopted by the WNTR simulator was established by a general form of the equation proposed by Crowl & Louvar (2001) where the flow rate of fluid through the leak hole is expressed as

$$q_{\mathrm{l}} = C_d A\sqrt{2gh}, \quad \text{when } \alpha = 0.5 \tag{1}$$

where q_{l} is the leak demand (m^3/s), C_{d} is the discharge coefficient (unitless), A is the area of the hole (m^2), α is an exponent related to characteristics of the leak (unitless), g is the acceleration of gravity (m/s^2), and h is the gauge head (m). The discharge coefficient C_d and leak exponent α were set to 0.75 and 0.5 in this paper (Lambert 2001; Greyvenstein & van Zyl 2007).

According to Equation (1), it is essential to confirm the diameter of the leak hole, so that the area of the hole can be calculated. Then, adding leak flow to each pipe and calculating the hydraulic parameters one after another, the node pressure values of every leak event are generated to calculate the node pressure sensitivity matrix.

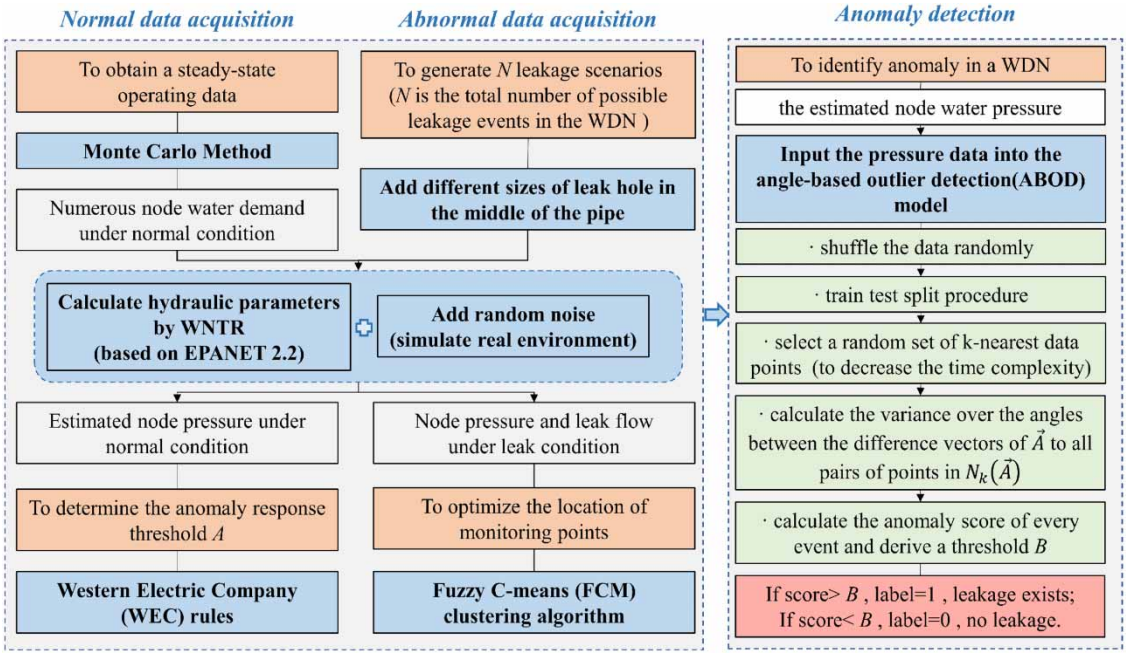


Figure 1 | Diagrammatic representation of the leak detection method.

The quantity and quality of data have an important effect on the performance of the leak detection method. However, leak events in realistic WDN are not frequent, and it is hard to collect the recording of leaks in actual pipe networks. Therefore, synthetic leak data were generated by hydraulic model simulation in this paper. These data represent most of the leak events.

In this paper, the degree of the leaks depends on the diameter of the leak hole. The minimum leak diameter $d^l_{\min}$ indicates the minimum extent of leakage to trigger its associate sensor. Then, set the diameter of the leaking pipe as the maximum leak diameter $d^l_{\max}$, if there is no negative node pressure node existing and at least one sensor responds, define the current diameter as the maximum leak diameter. Otherwise, reduce the leak diameter by step $0.005d$ until the condition is satisfied. The maximum extent of leakage was redefined as the current leak diameter.

2.2. Pressure sensor deployment

A reasonable sensor deployment method contributes to the detection of leakages in WDNs. As the variation of node pressure reflects the sensitivity of each node to different leak events, the sensitivity analysis is used to classify all the nodes into different clusters. By splitting the pipe at the midpoint and adding different sizes of the leak hole to the pipe, various leak scenarios are simulated by the WNTR. The pressure data of all nodes are normalized as the vector of the pressure sensitivity matrix:

$$\Delta P' = mP_0 - P_l \tag{2}$$

where $\Delta P'$ is the pressure sensitivity matrix between normal conditions and the leakages, P_0 is the normal node pressure with n-dimensional column vectors, n is the number of nodes in the network, P_l is the node pressure under various leaking scenarios, and m is the total number of pipes with potential for leakages.

By normalizing all nodes' pressure under different leaking conditions, a node pressure sensitivity analysis matrix as matrix formalism is shown in the following:

$$\Delta P = \begin{bmatrix} \Delta P'_{11} \\ \vdots \\ \Delta P'_{nm} \end{bmatrix} = \begin{bmatrix} \Delta p_{11} & \cdots & \Delta p_{1m} \\ \vdots & \ddots & \vdots \\ \Delta p_{n1} & \cdots & \Delta p_{nm} \end{bmatrix} \tag{3}$$

where n is the number of nodes where pressure sensors can be installed, m is the number of pipes in the network, and Δp_{nm} is the pressure difference of node n between normal conditions and pipe m leaking.

To detect leak events accurately according to the monitoring pressure data, the spatial correlation of data must be optimized. Therefore, the FCM clustering algorithm is adopted to classify the nodes by calculating the node pressure sensitivity matrix (Equation (3)). According to the number of clusters and sensitivity matrix, the clustering centers are calculated by the algorithm, which represents the installation sites of the pressure sensors. The priority of this method is dividing the dataset X into k categories and setting the cluster center matrix V. Equation (4) is the objective function of the FCM algorithm. The function sums the pairwise difference of every data value and cluster center (Askari 2021).

$$J = \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^{m} \parallel x_j - v_i \parallel^2, \qquad \sum_{k=1}^{c} u_{kj} = 1 \tag{4}$$

where J is the objective function, n is the number of nodes, $c \in [2,n]$ is the total number of clusters, v_i is the i-th cluster center, x_j is the j-th data, u_{ij} is the membership value of the j-th data in the i-th cluster, and m is the weighting index ($m > 1$) that affects the clustering results.

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{c} u_{ij}^{m}}, \quad 1 \le i \le c \tag{5}$$

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{\parallel x_j - v_i \parallel^2}{\parallel x_j - v_k \parallel^2}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{6}$$

The FCM algorithm starts with the initialized membership matrix $U = \text{rand}\,(c, n) = [u_{ij}]$. Cluster center v_i and membership degree u_{ij} are updated until the convergence of the algorithm. It should be pointed out that the maximum number of iterations τ should be set as the parameter of the algorithm. A convergence condition $\delta(\delta > 0)$ is an indicator for the end of the loop, and the program continues until the objective function value is less than δ. When the convergence condition is satisfied, the clustering process is over, and the output is the result of the calculation.

The FCM clustering algorithm was applied to ascertain the installation sites of pressure sensors. By normalizing the node pressure, the data point corresponding to the samples with the shortest Euler distance of node j at cluster i was determined as the site of sensors.

2.3. Leak event detection

A leak event in the WDN is one of the modalities of abnormal events, which can be reflected in monitoring data, such as outliers in monitoring pressure or flow. Anomaly detection is devoted to identifying the outlier points from the dataset accurately and efficiently. The dominant anomaly detection algorithms are three types: distance-based anomaly detection algorithm, density-based local anomaly detection algorithm, and statistical-based anomaly detection algorithm. These methods are mainly based on statistical theory and use Euclidean distance as an anomaly evaluation criterion.

There are dozens of monitoring sensors in large WDN, which will generate high-dimensional monitoring data at the same time. However, as the dimensionality increases, the distance between points is concentrated to a certain level, which means the nearest neighbor based on distance is close to the farthest neighbors (Beyer *et al.* 1999). Therefore, it is hard to identify the anomaly in large WDN which causes the potential risk to the stability of the WDN.

Different from the methods mentioned above, ABOD is an effective method for the detection of outliers in high-dimensional datasets, which evaluates the degree of outliers in terms of the variance of the angles (VOA) between the target object and the other objects. Some researchers studied the performances of distance-based algorithm and angle-based algorithm in several high-dimensional datasets and testified that ABOD is more stable with increasing data dimensionality (Ye *et al.* 2014).

The ABOD method used in this paper is to calculate the VOA of different pressure vectors, and each vector consists of pressure data from all the sensors at a certain time, which is expressed as the angle-based outlier factor (ABOF) value. Figure 2 shows the intuition of the identification of outlier based on ABOD, which comprises a set of data points to form a cluster. For

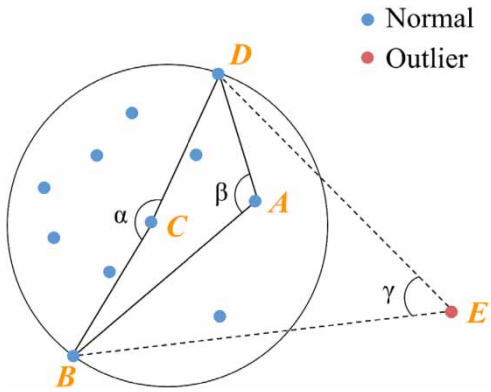


Figure 2 | An intuition of identifying the outlier from data points based on the ABOD algorithm.

each point O, it has the distance-weighted angle variance as its outlier score, and the angle $\angle AOB$ for each pair of points A, B ($A \neq O$ and $B \neq O$) is calculated and compared with the outlier score. Note that if it is a normal data point (e.g., A and C), within a cluster, the angles between different vectors and other data point pairs vary greatly. If a point is located at the border of a cluster (e.g., B and D), the variation of angle is smaller. As for an outlier point (e.g., E), the spectrum of angles for the data point is substantially narrow, and other points are located merely in certain directions (Kriegel *et al.* 2010), which means the point positioned outside of the cluster is an outlier.

The definition of ABOF is as follows:

Given a database $\mathcal{D} \subseteq R^d$, a point $\vec{A} \in \mathcal{D}$, and a norm $||.|| : R^d \to R_0^+$, where $R_0^+ = \{x | x \in R \wedge x \geq 0\}$. The scalar product is denoted by $\langle ., . \rangle : R^d \times R^d \to R$. For two points $\vec{B}, \vec{C} \in \mathcal{D}$, $\overrightarrow{BC}$ denotes the difference vector $\vec{C} - \vec{B}$. $\mathcal{N}_k(\vec{A}) \subseteq \mathcal{D}$ denotes the set of the k nearest neighbors of $\vec{A}$. The approximate $\mathrm{ABOF}_k(\vec{A})$ is the variance over the angles between the difference vectors of $\vec{A}$ to all pairs of points in $\mathcal{N}_k(\vec{A})$ weighted by the distance of the points:

$$\mathrm{ABOF}_k(\vec{A}) = \mathrm{VAR}_{\vec{B},\vec{C} \in \mathcal{N}_k(\vec{A})} \left(\frac{\langle \overline{AB}, \overline{AC} \rangle}{\| \overline{AB} \|^2 \cdot \| \overline{AC} \|^2} \right) \tag{7}$$

The further a data point is away from the clusters, the smaller the variance of angles of a point is, and the smaller the ABOF. Therefore, the ABOD calculates the ABOF for each data point and outputs the list of points in the dataset according to the ascending order of ABOF.

3. CASE STUDY AND RESULTS

3.1. Network description and dataset establishment

The pipe network of the case study is an actual one obtained from the literature (Lin 2013), as shown in Figure 3. It is comprised of 78 pipes, 4 reservoirs, and 49 demand nodes. The ability of water supply and average node demand were 5,992 and 122 L/s, respectively. By comparing the effectiveness of different quantities of monitoring sites in the literature, seven sensors showed the best performance (Wu *et al.* 2022). Therefore, the number of sensors was adopted in the case study. To detect abnormal events more effectively, the FCM algorithm was used to determine the installation sites of sensors and sensitive areas of each sensor. The core step is importing the sensitivity matrix into the model, and calculating cluster centers by the FCM algorithm. According to the results of clustering, the settlements of pressure sensors are 3, 9, 26, 27, 29, 41, and 44, which are represented by colored triangles in Figure 3.

The methodology of detecting an anomaly in the WDN relies on the data acquired from a SCADA system. As mentioned in Section 2.1, the construction of the SCADA system contained two parts: (1) the establishment of a pressure database under normal operating conditions and (2) the establishment of a pressure database under different extents of leakage. The construction of the database considered the uncertainty of modeling error and measurement error, which reflects on the parameters as random noise.

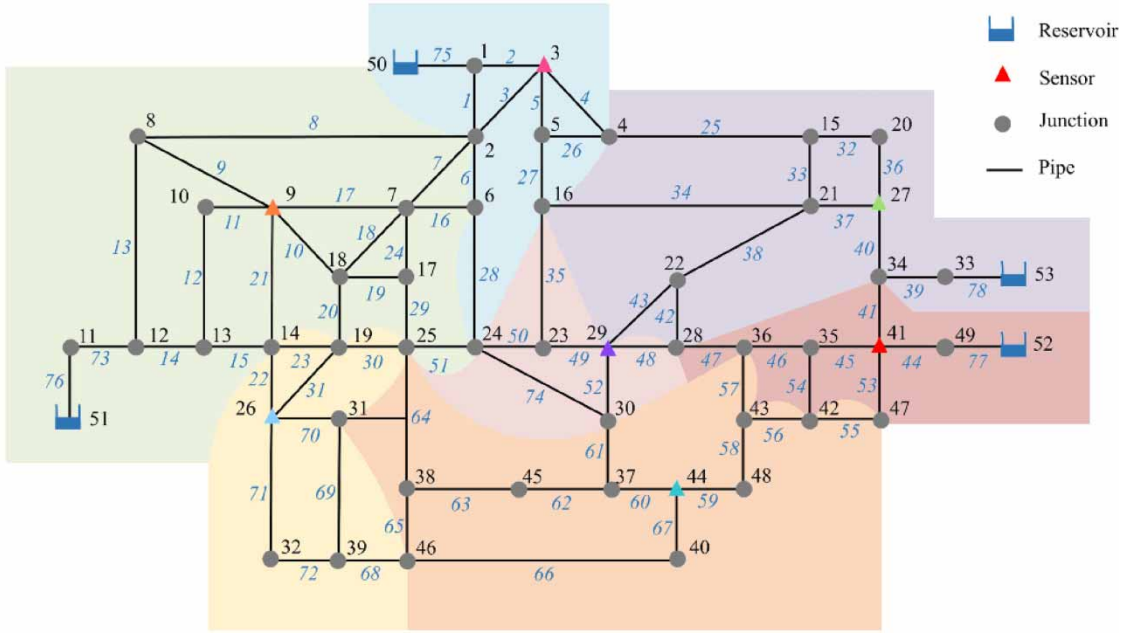


Figure 3 | Network layout for the case study.

For modeling error, the Monte Carlo method was used to simulate and generate the water demand data with 20% floating of each node as random noise (Kapelan *et al.* 2005), and the corresponding pressure value was calculated by WNTR. The sample size of the Monte Carlo simulation was 10,000 times to ensure stable operating conditions. And random noise N (0, 0.2 m) was added to the pressure value as a measurement error.

For the traditional leak detection method, the corresponding threshold value of each node is commonly adopted. As for the normal condition, the pressure values follow a normal distribution. According to the Western Electric Corporation rules, the threshold in this paper was (Cheng *et al.* 2020) defined as 2 times the standard deviation of normal pressure data that is suggested in the literature. If the leak event occurs in the WDN, the pressure will fluctuate suddenly. Especially, hydraulic operating parameters vary greatly during the minimum hour demand period (which is always around midnight) (Qi *et al.* 2018). With the increase in leak flow, the pressure value will be lower than the threshold, which will trigger the alarm.

To establish a relatively complete database of leakages, various leaking events were simulated by WNTR. Different leak degrees range from each pipe. In generating samples, the minimal leak extent was set the size of the leak hole as 0.005 times the pipe diameter d^l, which is defined as the minimum leak diameter $d^l_{\min}$. Then, the diameter of the leak hole was increased at the step of 0.005 d^l until the pipe completely burst. After several iterations of hydraulic calculation, the total number of 12,742 possible leak events in this network are contained in the leak database. And the spectrum of leak flow for entire leak scenarios is shown in Figure 4.

In this paper, the leak events were divided into two classes: one is the slight leakage that pressure drop does not exceed the threshold, and the other is the large leakage that pressure sensors have responded. The large class was divided into five levels based on leak flow, in which V level indicates that the leak extent just exceeds the threshold value and triggers an alarm, and I level indicates that the leakage is the most severe. The leak degree classification results are shown in Table 1. For the slight leak class, the extent of leaks was further divided into three categories according to the leak flow, and the range of leak flow is 10–50, 50–100, and 100–175 L/s.

3.2. Detection performance

For qualitative analysis of detection results, the confusion matrix is commonly used to evaluate binary classification problems. The matrix contains the information of actual labels and predicted labels by the outlier detection model, which is depicted in Figure 5. Evaluation metrics based on confusion matrix are also important to evaluate the performance of outlier detection algorithm for detecting new data vectors, such as TPR, FPR, Accuracy, Precision, and F1 score. Notably, the F1 score is a

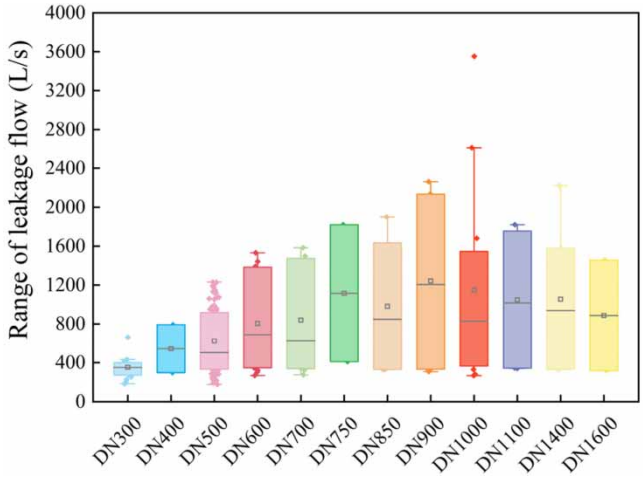


Figure 4 | Range of leak flow in different leak scenarios.

Table 1 | Classification of large leak events

Leak flow (L/s)	Leak level
176–800	V
801–1,500	IV
1,501–2,200	III
2,201–2,900	II
2,901–3,352	I

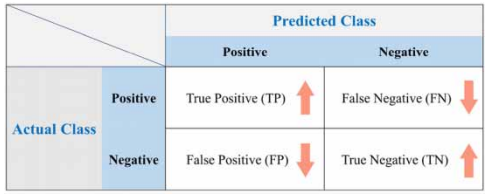


Figure 5 | Schematic diagram of the confusion matrix.

weighted average of class recall and precision, reflecting the ability of the model to recognize positive and negative samples.

$$\text{TPR} = \text{Class recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{9}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{10}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{F1} = \frac{2 \times \text{Class recall} \times \text{Precision}}{\text{Class recall} + \text{Precision}} \tag{12}$$

To evaluate the model performance more comprehensively, the area under the receiver operating characteristic (ROC-AUC) curve was adopted. ROC is a composite indicator reflecting continuous variables of sensitivity and effect, and the

ROC curve is based on a series of different dichotomies (boundary value or threshold determination), where the x-coordinate is TPR and the y-coordinate is FPR. The area under the ROC curve represents the value of the AUC. In other words, the larger value of AUC, the better the performance of the model.

To verify the accuracy and effectiveness of the ABOD for anomaly detection, four common unsupervised outlier detection algorithms were selected for comparison, which are Empirical-Cumulative-distribution-based Outlier Detection (ECOD), principal component analysis (PCA), Isolation Forest (IForest) and local outliers factor (LOF). Based on the normal pressure value and leak pressure value, the detection model was used to identify an anomaly in the WDN. In this part, 6,300 normal data and 700 abnormal data were selected randomly as the samples of the model and divided into training data and testing data according to the proportion of 8:2. In this part, various leak levels that can trigger the pressure sensors alarm were analyzed by outlier detection models, and the results of AUC and precision are shown in Tables 2 and 3.

From Tables 2 and 3, the AUC and precision of the ABOD model and LOF model are higher than any other models in the anomaly detection for each leak level. However, the LOF model does not perform well in training data, which has a potential effect on the detection results of other datasets. To select the most robust algorithm, the performance of all algorithms was analyzed under the condition of a complete pipe burst. As shown in Figure 6, the LOF model is not as stable as the ABOD model in identifying complete bursting events. Therefore, it is suggested that the ABOD model was adopted in anomaly detection.

3.3. Comparison with other leak identification methods

In this section, the performance of the proposed ABOD is further compared with distance-based outlier detection (DBOD) in order to comprehensively evaluate the applicability of the two methods in the medium-scale network. The core of DBOD is to calculate the vector's local density ρ_i and its distance ξ_i from points with a higher density of each point according to the Euclidean distances d_{ij} between vectors, and then select the outlier that has a low local density and a long distance. This method created by Rodriguez & Laio (2014) has been applied to burst detection in a small WDN (Wu *et al.* 2016).

Leak events with a leak flow of less than 175 L/s are selected, which means these events are too minor to trigger the monitoring sensors (according to the threshold-based method), so that it could be ignored and developed into a serious burst event. The first step of the establishment of dataset is selecting different degrees of leak scenarios and dividing these data into three categories according to the leak flow. Then, 100 abnormal data and 900 normal data were integrated, and 80% of the data

Table 2 | The AUC of several algorithms for detecting different levels of leak events

	V		IV		III		II		I	
Algorithm	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**
ECOD	0.942	0.930	0.942	0.921	0.942	0.922	0.942	0.936	0.942	0.927
PCA	0.979	0.999	0.979	1.000	0.979	1.000	0.979	1.000	0.979	0.999
IForest	0.982	0.980	0.983	0.984	0.987	0.984	0.980	0.991	0.984	0.988
LOF	0.689	1.000	0.689	1.000	0.689	1.000	0.689	1.000	0.689	1.000
ABOD	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000

Table 3 | The precision of several algorithms for detecting different levels of leak events

	V		IV		III		II		I	
Algorithm	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**
ECOD	0.547	0.390	0.547	0.320	0.547	0.327	0.547	0.410	0.547	0.633
PCA	0.871	0.957	0.871	0.977	0.871	0.983	0.871	0.993	0.871	0.982
IForest	0.813	0.777	0.834	0.853	0.876	0.883	0.800	0.897	0.833	0.880
LOF	0.346	1.000	0.346	1.000	0.346	1.000	0.346	1.000	0.346	1.000
ABOD	0.987	1.000	0.987	1.000	0.987	1.000	0.987	1.000	0.987	1.000

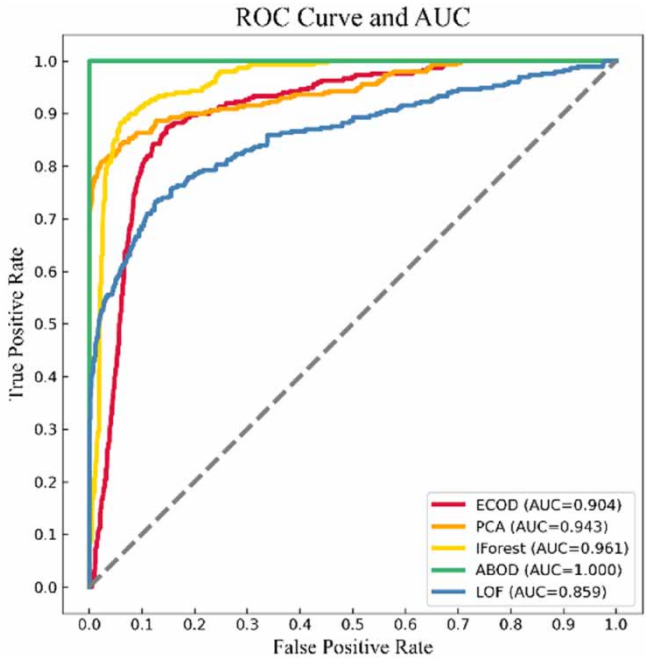


Figure 6 | Performance of five algorithms when the pipe completely burst.

were randomly selected as training data and the remaining data as a testing set. It should be noted that the information includes seven sensors monitoring data and labels about binary classification, but an unsupervised anomaly detection model was not provided with no leak or leak labels for training, which were only used to verify the accuracy of the testing set. By the calculation of the unsupervised anomaly detection algorithm, the performances of ABOD and DBOD responding to different levels of leak events were presented in Table 4.

In terms of comparative performance, the ABOD model achieves high accuracy in different scenarios, which are 98.5, 99.5, and 99.45%. The accuracy of the DBOD model in each leak scenario is 89.5, 89.0, and 98.0%, which is lower than the first model. The comparison diagram of the two detection results is shown in Figure 7.

The category axis of the radar map represents different performance evaluation indexes, and the numerical axis describes the model results of the response. Figure 8(a) and 8(b) shows two methods have similarly low false alarm rates in two types of slight leak scenarios, while ABOD performs better than DBOD in other indicators such as TPR, F1 score, and precision. For leak scenarios with leak flow between 100 and 175 L/s, DBOD does not make great progress in leak identification, and there is still a large gap between the performance of DBOD and ABOD, as shown in Figure 8(c).

The advantage of applying the ABOD algorithm for leak detection is more accurate and effective than the DBOD algorithm and overcomes the obstacle that the supervised algorithm needs a complete label dataset for accurate identification. In

Table 4 | Detection results for slight leak events

		ABOD		DBOD	
Leak flow (L/s)		**PNL**	**PL**	**PNL**	**PL**
10–50	ANL	182	1	170	11
	AL	2	15	10	9
50–100	ANL	182	1	170	13
	AL	0	17	9	8
100–175	ANL	182	1	178	10
	AL	1	16	2	10

Note: ANL, actual non-leak; AL, actual leak; PL, predicted leak; PNL, predicted non-leak.

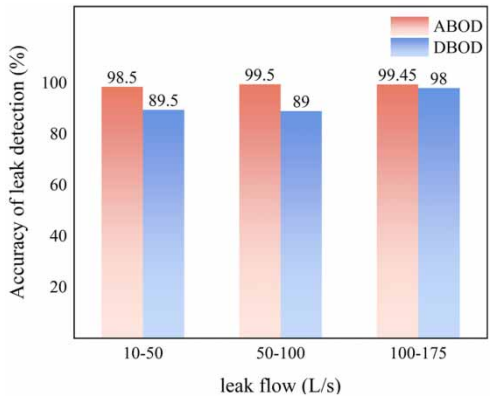


Figure 7 | Accuracy results of leak detection.

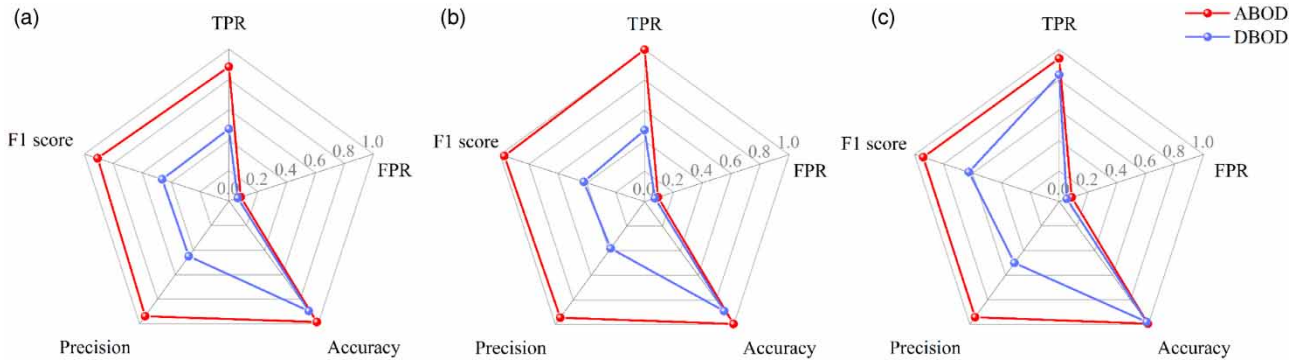


Figure 8 | Performance indicators of two methods under different leak scenarios. (a) Leak flow is 10–50 L/s, (b) leak flow is 50–100 L/s, and (c) leak flow is 100–175 L/s.

addition, benefitted from the ability of the model to process unbalanced data, it can be used even in the pipe network with fewer leak events. The recognition accuracy can be improved by monitoring the results of the model in real time and continuously training the model with new data.

4. CONCLUSION

This paper proposed a novel angle-based leak detection using pressure sensors and investigated it in a middle-scale WDN. The method adopted in this paper needs to be based on scientific sensor deployment, so the FCM clustering algorithm was adopted to determine the placement of pressure sensors. To evaluate leak detection, the leak scenarios were divided into two major categories according to the leak flow. The following conclusions were drawn:

1. For large leaks, the comprehensive indicators of the ABOD method are better than other common outlier detection algorithms, which mainly contain TPR, FPR, F1 score, accuracy, precision, and AUC. It is noteworthy that ABOD can detect all the leak scenarios correctly.
2. For minor leaks, ABOD was further compared with the distance-based leak detection method adopted in the literature. The leak events were randomly selected from the database, and the number is determined to be 200 according to the abnormal ratio of 0.2. The results demonstrate that the ABOD performed better than the DBOD in terms of accuracy and other evaluation indicators.
3. Considering the limitations, future work will be conducted on the actual network and real data for analysis to ensure scientificity and feasibility. This model will further analyze the time series to identify the moment of leakage and predict the extent of the leakage. The ABOD model presented in this study will be further improved to achieve better performance.

ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Program of China (2022YFF0606905), National Natural Science Foundation of China (52070167), and Zhejiang Provincial Natural Science Foundation of China (LHY22E080003).

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

Ang, W. K. & Jowitt, P. W. 2006 Solution for water distribution systems under pressure-deficient conditions. *Journal of Water Resources Planning and Management-ASCE* **132**, 175–182. doi:10.1061/(Asce)0733-9496(2006)132:3(175).

Askari, S. 2021 Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development. *Expert Systems with Applications* **165**. doi:10.1016/j.eswa.2020.113856.

Bakker, M., Vreeburg, J., Van De Roer, M. & Rietveld, L. 2014 Heuristic burst detection method using flow and pressure measurements. *Journal of Hydroinformatics* **16**, 1194–1209. doi:10.2166/hydro.2014.120.

Beker, B. A. & Kansal, M. L. 2022 Fuzzy logic-based integrated performance evaluation of a water distribution network. *Aqua-Water Infrastructure Ecosystems and Society* **71**, 490–506.

Beyer, K., Goldstein, J., Ramakrishnan, R. & Shaft, U. 1999 When is 'nearest neighbor' meaningful? In *Database Theory – ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7*, Springer, pp. 217–235.

Cheng, W., Chen, Y. & Xu, G. 2020 Optimizing sensor placement and quantity for pipe burst detection in a water distribution network. *Journal of Water Resources Planning and Management* **146**. doi:10.1061/(Asce)Wr.1943-5452.0001298.

Christodoulou, S. E., Kourti, E. & Agathokleous, A. 2017 Waterloss detection in water distribution networks using wavelet change-point detection. *Water Resources Management* **31**, 979–994. doi:10.1007/s11269-016-1558-5.

Covas, D., Ramos, H. & de Almeida, A. B. 2005 Standing wave difference method for leak detection in pipeline systems. *Journal of Hydraulic Engineering-Asce* **131**, 1106–1116. doi:10.1061/(Asce)0733-9429(2005)131:12(1106).

Crowl, D. A. & Louvar, J. F. 2001 *Chemical Process Safety: Fundamentals with Applications*. Pearson Education, London, UK. doi:10.1002/prs.12086.

CUWA. 2020 *Urban Water Statistics Yearbook 2020*. China Statistics Press, Beijing.

Fahmy, M. & Moselhi, O. 2010 Automated detection and location of leaks in water mains using infrared photography. *Journal of Performance of Constructed Facilities* **24**, 242–248. doi:10.1061/(ASCE)CF.1943-5509.0000094.

Farley, M. & Trow, S. 2003 *Losses in Water Distribution Networks: A Practitioner's Guide to Assessment, Monitoring and Control*, Vol. 4. IWA Publishing, London.

Fontanazza, C. M., Notaro, V., Puleo, V., Nicolosi, P. & Freni, G. 2015 Contaminant intrusion through leaks in water distribution system: experimental analysis. In *Computing and Control for the Water Industry (CCWI2015) – Sharing the Best Practice in Water Management* , Vol. 119 (Ulanicki, B., Kapelan, Z. & Boxall, J., eds.). University of Exeter, Leicester, UK, pp. 426–433.

Giustolisi, O., Savic, D. & Kapelan, Z. 2008 Pressure-driven demand and leakage simulation for water distribution networks. *Journal of Hydraulic Engineering* **134**, 626–635. doi:10.1061/(ASCE)0733-9429(2008)134:5(626).

Greyvenstein, B. & van Zyl, J. E. 2007 An experimental investigation into the pressure-leakage relationship of some failed water pipes. *Aqua-Water Infrastructure Ecosystems and Society* **56**, 117–124. doi:10.2166/aqua.2007.065.

Guo, G. C., Liu, S. M., Jia, D. L., Wang, S. H. & Wu, X. 2021 Simulation of a leak's growth process in water distribution systems based on growth functions. *Aqua-Water Infrastructure Ecosystems and Society* **70**, 521–536. doi:10.2166/aqua.2021.021.

Hunaidi, O. 1998 Ground-penetrating radar for detection of leaks in buried plastic water distribution pipes. In *Proceedings of the Seventh International Conference on Ground Penetrating Radar*, Lawrence, KA, 27–30 May 1998. IEEE. New York, NY.

Hunaidi, O., Chu, W., Wang, A. & Guan, W. 2000 Detecting leaks in plastic pipes. *Journal American Water Works Association* **92**, 82–94. doi:10.1002/j.1551-8833.2000.tb08819.x.

Kapelan, Z. S., Savic, D. A. & Walters, G. A. 2005 Multiobjective design of water distribution systems under uncertainty. *Water Resources Research* **41**. doi:10.1029/2004wr003787.

Kim, J.-H., Sharma, G., Boudriga, N. & Iyengar, S. S. 2010 SPAMMS: A sensor-based pipeline autonomous monitoring and maintenance system. In *Second International Conference on Communication Systems and Networks*, Bangalore, India, 5-9 January 2010. IEEE, New York, NY, pp. 1–10.

Kriegel, H.-P., Kröger, P. & Zimek, A. 2010 Outlier detection techniques. *Tutorial at KDD* **10**, 1–76.

Lambert, A. 2001 What do we know about pressure-leakage relationships in distribution systems. In *Proc. IWA Systems Approach to Leakage Control and Water Distribution System Management*, Brno, Czech Republic, 16–18 Ma7 2001.

Li, R., Huang, H. D., Xin, K. L. & Tao, T. 2015 A review of methods for burst/leakage detection and location in water distribution systems. *Water Science and Technology-Water Supply* **15**, 429–441. doi:10.2166/ws.2014.131.

Liggett, J. A. & Chen, L. C. 1994 Inverse transient analysis in pipe networks. *Journal of Hydraulic Engineering-ASCE* **120**, 934–955. doi:10.1061/(Asce)0733-9429(1994)120:8(934).

Lin, Y. 2013 *Design of Real-Time Monitoring System for Municipal Water Supply Network and Optimization of Monitoring Points*. South China University of Technology, Guangzhou, China.

Mounce, S. R., Day, A. J., Wood, A. S., Khan, A., Widdop, P. D. & Machell, J. 2002 A neural network approach to burst detection. *Water Sci Technol* **45**, 237–246. doi:10.2166/wst.2002.0595.

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**, 309–318. doi:10.1061/(Asce)Wr.1943-5452.0000030.

Mounce, S. R., Mounce, R. B. & Boxall, J. B. 2011 Novelty detection for time series data analysis in water distribution systems using support vector machines. *Journal of Hydroinformatics* **13**, 672–686. doi:10.2166/hydro.2010.144.

Mpesha, W., Gassman, S. L. & Chaudhry, M. H. 2001 Leak detection in pipes by frequency response method. *Journal of Hydraulic Engineering-Asce* **127**, 134–147. doi:10.1061/(Asce)0733-9429(2001)127:2(134).

Muggleton, J. M., Brennan, M. J., Pinnington, R. J. & Gao, Y. 2006 A novel sensor for measuring the acoustic pressure in buried plastic water pipes. *Journal of Sound and Vibration* **295**, 1085–1098. doi:10.1016/j.jsv.2006.01.032.

Puust, R., Kapelan, Z., Savic, D. & Koppel, T. 2010 A review of methods for leakage management in pipe networks. *Urban Water Journal* **7**, 25–45. doi:10.1080/15730621003610878.

Qi, Z. X., Zheng, F. F., Guo, D. L., Zhang, T. Q., Shao, Y., Yu, T. C., Zhang, K. J. & Maier, H. R. 2018 A comprehensive framework to evaluate hydraulic and water quality impacts of pipe breaks on water distribution systems. *Water Resources Research* **54**, 8174–8195. doi:10.1029/2018wr022736.

Rodriguez, A. & Laio, A. 2014 Machine learning. Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496. doi:10.1126/science.1242072.

Romano, M., Kapelan, Z. & Savić, D. 2011 Burst detection and location in water distribution systems. In *World Environmental and Water Resources Congress 2011: Bearing Knowledge for Sustainability*, Palm Springs, CA, 22 May 2011. ASCE, Reston, VA.

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

Shao, Y., Li, X., Zhang, T., Chu, S. & Liu, X. 2019 Time-series-based leakage detection using multiple pressure sensors in water distribution systems. *Sensors (Basel)* **19**. doi:10.3390/s19143070.

Valizadeh, S., Moshiri, B. & Salahshoor, K. 2009 Leak detection in transportation pipelines using feature extraction and KNN classification. In *ASCE Pipelines Specialty Conference 2009*, San Diego, CA, USA, pp. 580–589.

Wu, Z. Y. & He, Y. K. 2021 Time series data decomposition-based anomaly detection and evaluation framework for operational management of smart water grid. *Journal of Water Resources Planning and Management* **147**. doi:10.1061/(Asce)Wr.1943-5452.0001433.

Wu, Y., Liu, S., Wu, X., Liu, Y. & Guan, Y. 2016 Burst detection in district metering areas using a data driven clustering algorithm. *Water Research* **100**, 28–37. doi:10.1016/j.watres.2016.05.016.

Wu, J. J., Ma, D. H. & Wang, W. 2022 Leakage identification in water distribution networks based on XGBoost algorithm. *Journal of Water Resources Planning and Management* **148**. doi:10.1061/(Asce)Wr.1943-5452.0001523.

Ye, H., Kitagawa, H. & Xiao, J. 2014 Continuous angle-based outlier detection on high-dimensional data streams. In *Proceedings of the 19th International Database Engineering & Applications Symposium on – IDEAS '15*, Yokohama, Japan, 12–15 July 2014, pp. 162–167.

Yu, J., Zhang, L., Chen, J. Y., Xiao, Y., Hou, D. B., Huang, P. J., Zhang, G. X. & Zhang, H. J. 2021 An integrated bottom-up approach for leak detection in water distribution networks based on assessing parameters of water balance model. *Water* **13**. doi:10.3390/w13060867.

Zaman, D., Tiwari, M. K., Gupta, A. K. & Sen, D. 2020 A review of leakage detection strategies for pressurised pipeline in steady-state. *Engineering Failure Analysis* **109**, 104264.

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. doi:10.1016/j.watres.2019.115058.

First received 3 May 2023; accepted in revised form 14 September 2023. Available online 9 November 2023