Original Manuscript

SHM Structural Health Monitoring

An innovative machine learning based framework for water distribution network leakage detection and localization

Structural Health Monitoring
2021, Vol. 0(0) 1–19

DOI: 10.1177/14759217211040269
journals.sagepub.com/home/shm
SAGE

Xudong Fan and Xiong (Bill) Yu

Abstract

Leakages in the underground water distribution networks (WDNs) waste over 1 billion gallon of water annually in the US and cause significant socio-economic loss to our communities. However, detecting and localization leakage in a WDN remains a challenging technical problem despite of significant progresses in this domain. The progresses in machine learning (ML) provides new ways to identify the leakage by data-driven methods. However, in-service WDNs are short of labeled data under leaking conditions, which makes it infeasible to use common ML models. This study proposed a novel machine learning (ML)-based framework for WDN leak detection and localization. This new framework, named clustering-then-localization semi-supervised learning (CtL-SSL), uses the topological relationship of WDN and its leakage characteristics for WDN partition and sensors placement, and subsequently utilizes the monitoring data for leakage detection and leakage localization. The CtL-SSL framework is applied to two testbed WDNs and achieves 95% leakage detection accuracy and around 83% final leakage localization accuracy by use of unbalanced data with less than 10% leaking data. The developed CtL-SSL framework advances the leak detection strategy by alleviating the data requirements, guiding optimal sensor placement, and locating leakage via WDN leakage zone partition. It features excellent scalability, extensibility, and upgradeability for applications to various types of WDNs. It will provide valuable a tool in sustainable management of the WDNs.

Keywords

Water distribution network, artificial intelligence, machine learning, leakage detection, leakage localization, partition, clustering-then-localization semi-supervised learning

Introduction

Fast detection and localization of underground water pipe leakage is an important yet challenging issue in water distribution system management. Due to the deterioration of underground water pipes, a large amount of water is lost every year, mostly unnoticed. According to Sadeghioon,[1] about 3281 ML (10^6) was wasted in the UK during 2009–2011, and about 15% of supplied water was wasted annually in the US. In historical water districts, such as Cleveland, OH or Boston, MA, the percentage of water lost is significantly higher. Moreover, unnoticed water leakage can lead to serious social impacts due to traffic delay,[2] water contamination,[3] and water scarcity. Therefore, a system that provides real-time water pipe monitoring and enables fast leakage response is critical for agencies to institute preventative strategies with significant socio-economic benefits.

A significant number of studies have been conducted on water pipe leak detection. As reviewed by Chan,[4] the strategies are broadly classified into 5 categories, that is, visual observation-based, sensor/instrumentation-based, transient response based, hydraulic model-based, and data-driven based strategies. However, these strategies have encountered different limitations. For instance, the conventional sensor/instrumentation-based technologies require well-trained inspectors to conduct the inspection along the pipes with the help of different types of detection equipment including those based on the optical, acoustic, or electromagnetic sensing principles.[5–7] This method can be labor-intensive, time-consuming, and cost-prohibitive.[8–10]

Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, OH, USA

Corresponding author:
Xiong (Bill) Yu, Department of Civil Engineering, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106-7078, USA.
Email: xxy21@case.edu

Moreover, the reliability of detection are influenced by various factors including the type of leakage, size of the leakage, pipe materials, environmental conditions, and the skill level of the inspector.[11] *The transient based technology* uses transient pressure or acoustic signals associated with burst events. Such transient signals travel along the pipe at the speed of sound starting from the burst location.[12] However, the transient responses decay with distance and diminish over a short time, and therefore require sensors with high spatial and temporal resolutions, which makes it not suitable for continuous monitoring in all environments. *The hydraulic model-based approach* requires the use of the hydraulic model simulation of the water distribution network (WDN). But information such as customer water usage, pipe deterioration conditions, pipe physical information is often difficult to collect or is typically not available.[13–15] *Data driven-based leak detection*, which is based on learning from historical data with statistical or pattern recognition algorithms, is emerging.[16] Such technologies mainly depend on the available historical dataset without the requirement of collecting a comprehensive set of information of the hydraulic model. Empowered with the Internet of Things (IoT) and artificial intelligence (AI), data-driven technologies have been proven capability in knowledge discovery,[17] image processing,[18] and event forecasting, etc.[19] The development of supervisory control and data acquisition (SCADA) systems also promotes the progress in using data-driven methods for leakage detection since real-time monitoring data of water pressure and/or flow rate are available via SCADA system.[20–22]

A few data-driven methods have been developed to detect leakage in the WDNs. The previous studies typically formulate the leakage detection problem as either a supervised machine learning (ML) problem or an unsupervised ML problem. For instance, Zhou [23] developed a supervised ML method by using fully connected DensNet for leakage detection. The sensors were first assumed to be placed at different junctions determined by an optimization process. The simulated water pressure data obtained by these sensors was used to train the developed ML model and achieved promising results. For another example, Kang[5] collected the data by piezoelectric accelerometer under non-leaking condition and under leaking condition. The labeled data was trained by a Convolutional Neural Network (CNN) and a Support Vector Machine (SVM) to class leaking versus non-leaking conditions. Saade and Mustapha[23] setup a laboratory environment for leakage detection by using a network of Fiber Bragg Grating (FBG) sensor and utilizes a machine learning model for data analyses. Recent studies also explored leakage detection by clustering junctions into multiple leakage zones to enhance the final detection accuracy.[24,25] Although supervised learning approach can achieve a high leakage localization accuracy, it requires a balanced dataset, which means it needs a sufficient amount of WDN operational data under both leaking conditions and non-leaking conditions. However, as pointed by Mounce et al.,[26] datasets under leaking conditions are very scarce. Consequently, unsupervised ML model are more feasible practically. For example, Mounce et al.[27] used Artificial Neural Network to predict the water flow and water pressure one day ahead. Leakage warning was triggered if the difference between the actual data and the predicted data exceeds a threshold. However, the detection accuracy was dependent upon a stable water pressure pattern in the WDN, which however can be significantly affected by water usage behaviors. Another widely used unsupervised-learning approach is to treat the leakage as a ML-based abnormally detection problem. For example, Roya developed an autoencoder algorithm (AE) for the leakage detection[28] and validated the results on a small-scale laboratory water pipe testbed. The developed methods have only been used for leakage detection and not attempted for leakage localization. Compared to the supervised ML models, the unsupervised ML models have a promising advantage since they can work with just non-leaking data.

To further advance the data-driven approach for leakage detection and localization in water distribution network, this study explores a new ML framework that combines the advantages of both supervised ML and unsupervised ML approaches. This new framework, named clustering-then-localization semi-supervised learning (CtL-SSL), uses the topological relationship of WDN and its leakage characteristics for WDN partition, sensors placement, and subsequently utilize the monitoring data for leakage detection and leakage localization. Compared with previous studies, this framework, (1) considers the spatial relationship of the sensors in WDN portioning and sensor placement, such relationship is the cornerstone for later leakage detection and localization; (2) does not require any historical leakage data for leakage detection; (3) can be used when only limited historical leakage data is available. More specifically, the leakage detection uses unsupervised-learning algorithm to compress and decompress the normal water pressure data. This process performs poorly when the input is leakage data. The leakage localization uses supervised learning algorithm to extract the spatial relationship from the available leakage data. Only limited leakage data is required for each leakage zone with the help of proposed WDN partition process.

The organization of this article is briefly summarized as following: *Methodology* introduces the main components of the proposed CtL-SSL framework, including the considered leakage characteristics matrix, a modified *k-means* algorithm for WDN leakage zones partition, and ML models for leakage detection and leakage zone localization. *WDN operation data generation* describes the dataset used in this study. Due to the lack of real-world monitoring dataset, the dataset used in this study were generated by running holistic

hydraulic simulations using publically accepted model for WDNs. *Case study I: C-Town WDN* introduces the evaluation of the CtL-SSL framework by a widely used benchmark WDN. Finally, *Case study II: Rancho Solano Zone III WDN* gives the conclusions.

Methodology

Figure 1 illustrates the developed CtL-SSL framework for WDN leak detection and localization. It starts with a basic hydraulic model of the WDN to generate the simulated operational data with consideration of the structural, physical, topological, and hydraulic characteristics of the WDN as well as the characteristics of water user demands. It includes two stages, that is, the WDN partition stage and leakage monitoring stage. In the WDN partition stage, the WDN is partitioned into k leakage zones by using a modified *k-means* clustering algorithm that considers the junctions' leakage characteristics matrix (a matrix that describes the leakage behaviors of each junction, details are given in *Determination of the leakage characteristics matrix*) and the physical locations. The centroid of partitioned clusters also provides the optimal locations of sensor placement. In the leakage monitoring stage, a ML model was trained with non-leaking data to determine the leakage that occurs in the WDN, and another ML model was trained with available labeled leakage data to locate the exact leakage occurrence zone. Components involved in the implementation of the workflow in Figure 1 are introduced below.

Determination of the leakage characteristics matrix

In previous studies, the leakage characteristic vector is simply determined by using the difference of pressure at monitored junctions before and after leakage at a given junction. Hence, the length of the vector equals to the number of sensors m. A novel leakage characteristic matrix was proposed in this study by using the PCA and AE algorithm to find the spatial relationship among the sensors, which extract the first k principal components, with k equals to $m/2$. The conventional leakage matrix is then projected to the k principal components. The resultant leakage characteristics vector achieved dimension reduction from m to k (or by half since k is set to be m/2). The leakage characteristic matrix is subsequently used for clustering of the WDN. Details about the conventional leakage characteristic matrix and proposed leakage characteristic matrix are given below.

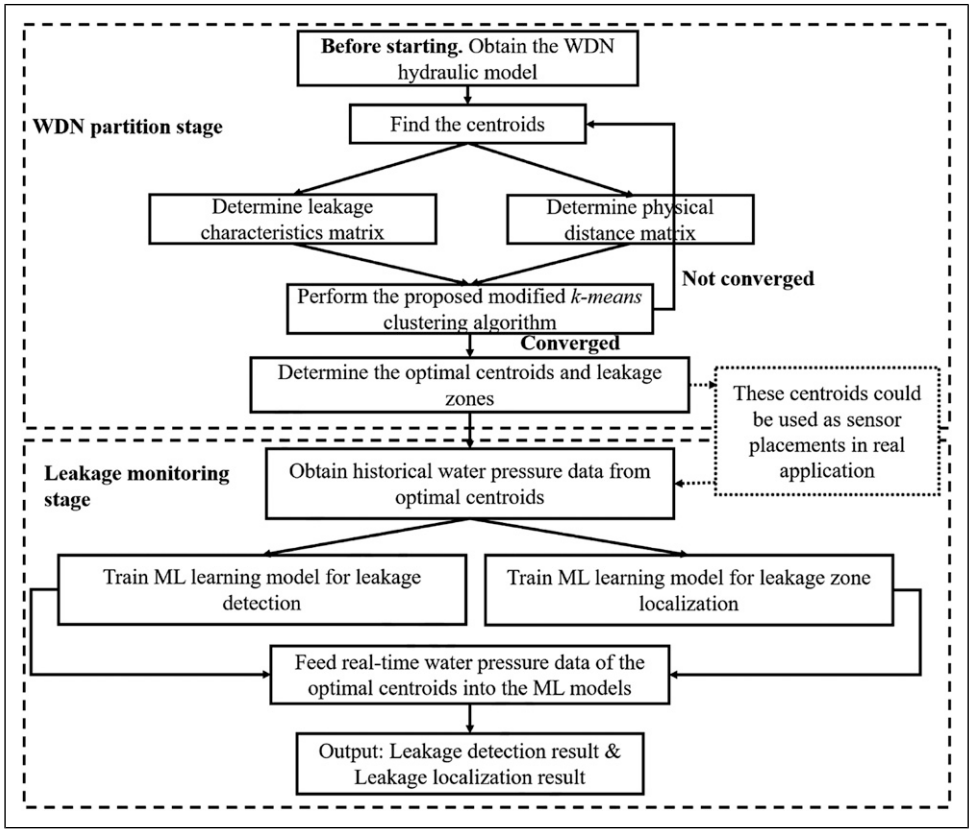

Figure 1. Flowchart of the proposed cluster-then-localization semi-supervised learning algorithm for water distribution networks leak detection and localization.

Conventional Leakage Characteristics Matrix

Zhang[24] and Chen[25] defined the leakage characteristics matrix using the change of monitored water pressure due to a given leakage occurring at each junction compared with non-leaking conditions. Table 1 illustrates the calculation of the leakage characteristics matrix, where each row is the leakage characteristics vector corresponding to the junction.

The leakage characteristics matrix defined in Table 1 does not consider the internal relationships among the monitoring sensors. Previous studies have proven that such internal relationship is sensitive to leakage occurrence and therefore can be used for leakage detection and localization.[26–28] Hereby, to further extract the underlying relationships between the junctions, this study proposed two new leakage characteristics extracted by unsupervised-learning algorithms, that is, the Principal Component Analysis (PCA) and Autoencoder neural network (AE). Details of the PCA-based and AE-based leakage characteristics matrix are described in the following.

PCA-based leakage characteristics matrix

PCA is an unsupervised ML model that is often used for the dimensional reduction of data samples.[29] The process to calculate PCA-based leakage characteristics is illustrated in Figure 2. It is assumed m monitoring sensors are installed in the WDN for data collection. A non-leaking dataset contains t samples, each includes a vector of data by m monitoring sensors, is first used to train the PCA model to obtain the first k principal directions (step 1 in Figure 2). There is no requirement for the number of training samples. However, the more the samples, the better the PCA model in finding the relationships among the monitoring sensors. Then, the leakage matrix $[p_{ij}]$ is transformed by using the PCA model (step 2 in Figure 2). The leakage matrix, $[p_{ij}]$, here is defined as a matrix consisting of the monitored water pressure vector at the m monitoring sensors when a leakage happens to each of the n junctions in turn. That is, the i^{th} row of the leakage matrix is a vector consisting of the water pressure by the m sensors when the leak occurs at the i^{th} junction. By feeding the leakage matrix, $[p_{ij}]$, to the trained PCA model, the output matrix $[d]$ is named as the PCA-based leakage

Table 1. Leakage characteristics matrix, p_i^j, based on water pressure change for a water distribution networks with n junctions and m monitoring sensors (i is the sensor *No.*, j is the junction *No.*).

Pressure change	Pressure sensor 1	Pressure sensor 2	…	Pressure sensor m
Junction 1	$p_1^{\text{nonleak}} - p_1^{\text{jun 1}}$	$p_2^{\text{nonleak}} - p_2^{\text{jun 1}}$	…	$p_m^{\text{nonleak}} - p_m^{\text{jun 1}}$
Junction 2	$p_1^{\text{nonleak}} - p_1^{\text{jun2}}$	$p_2^{\text{nonleak}} - p_2^{\text{jun2}}$	…	$p_m^{\text{nonleak}} - p_m^{\text{jun2}}$
…	…	…	…	…
junction n	$p_1^{\text{nonleak}} - p_1^{\text{jun } n}$	$p_2^{\text{nonleak}} - p_2^{\text{jun } n}$	—	$p_m^{\text{nonleak}} - p_m^{\text{jun } n}$

Note: p_i^{nonleak} is the water pressure measured by sensor i under non-leaking conditions. $p_i^{jun\ j}$ is the water pressure of sensor i when leaking occurs at junction j. i is the index for sensors which ranges from 1 to m, j is the index for junctions which ranges from 1 to n.

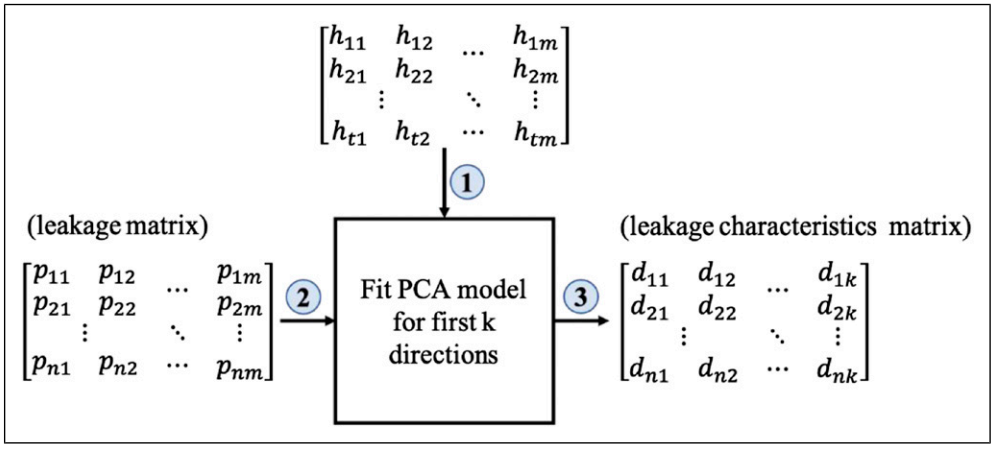

Figure 2. Illustration of PCA-based leakage characteristics matrix (Note: h is the training dataset that consists of non-leaking dataset, p is the leakage matrix containing vectors of the monitored water pressure when leakage happens at each junction, d is the PCA-based leakage characteristics matrix, m is the number of monitoring sensors, t is the number of samples for training PCA model, n is the number of junctions, and k equals $m/2$).

characteristics matrix (step 3 in Figure 2). Due to dimension reduction by the PCA, for each of the n junctions, its leakage characteristics are represented by a projected vector with k elements. The dimension of principal components k is set to be around $m/2$, as this study found this well captures the relationship among the monitored junctions.

AE-based leakage characteristics matrix

AE neural network is an unsupervised-learning algorithm based on deep neural network. Unlike the PCA method, which is an orthogonal linear transformation, AE neural network can extract non-linear relationships among data samples. A common architecture of AE network is shown in Figure 3. For each sample with m features, the AE network encodes the original dataset into a compressed dimensional space and then decode it to the original dimension. By minimizing the difference between output and input, the neural network is forced to learn the features of the samples and their relationships.

AE-based leakage characteristics matrix is defined as illustrated in Figure 3. First, an AE neural network is built with a middle layer consisting of k neurons (Figure 3). The AE model is pre-trained with the monitoring dataset from the WDN under non-leaking situations (this step is not shown in Figure 3). This pre-trained AE neural network learned the internal relationship among the monitored junctions under non-leaking conditions. Then, the leakage matrix $[p_{ij}]$ (which is the same as the PCA process) is fed into the AE neural network (see Figure 3 step 1). The output of the middle layer of the AE neural network consists of the leakage characteristics vector corresponding to each junction. Matrix $[d]$, which is dimension reduced from m to k compared with leakage matrix $[p_{ij}]$, is defined as the AE-based leakage characteristics matrix. Similar to the PCA-based leakage characteristics matrix, the reduced dimension k is set as $m/2$.

Both the PCA- and AE-based leakage characteristics matrix are derived from the projected leaking data matrix by PCA or AE models which are pre-trained with non-leaking dataset only. This process effectively utilized these ML models for the feature extraction. Meanwhile, as a byproduct of the feature extraction process, the utilized characteristics matrix is only half-the-size of the monitored data size and conventional leakage characteristics matrix. Such data size reduction could potentially increase the computing and data storing efficiency as more data are collected by the sensors.

WDN partition stage: modified k-means clustering algorithm

K-means clustering algorithm clusters data based on their Euclidian distance. The standard *k-means* algorithm has been used in previous studies for the WDN zone partition to reduce the degree of freedoms in leakage detection, based on the conventional leakage characteristics matrix.[24,25] The partition aims to improve leakage detection and localization accuracy.

We hereby noted that this current WDN partition procedure has a few limitations. First, the standard *k-means* algorithm requires the number of sensors and their placements to be predefined. That is, the monitoring data collection schema must be set in advance and the standard algorithm cannot consider the influence of choosing different sensor placements during the clustering process. For example,

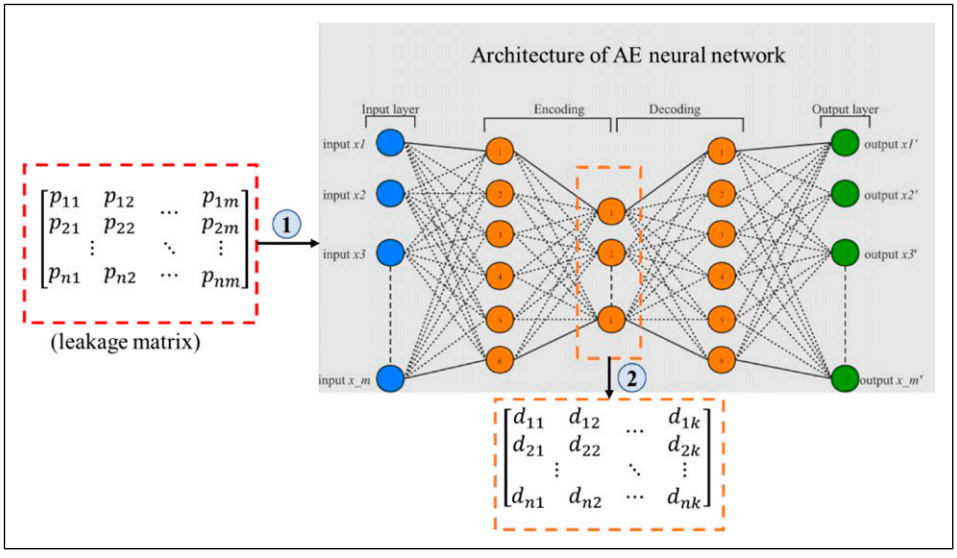

Figure 3. Illustration of AE and AE-based leakage characteristics matrix (Note: the *p, d, m, n, k* are the same as in Figure 2).
Note: AE: autoencoder algorithm.

Zhang[24] used Zheng's[30] algorithm to optimize sensor placements before initializing the WDN partitioning via the *k-means* clustering process. Second, the previous WDN partitioning (i.e., [27,28]) only considered the leakage characteristics. It did not consider the physical distance among the monitoring junctions. The consequence is the junctions clustered into the same WDN zone can be geographically scattered on the WDN.

To overcome these limitations, a modified *k*-means clustering algorithm is developed in this study for WDN partition. Compared to the standard *k*-means WDN clustering which only considers the leakage characteristics of junctions, the new algorithm also considers the shortest physical path distance between the junctions over the WDN. The pseudocode of the proposed *k-means* algorithm is shown in the following Table 2.

It is noted that in step 3.1, the leakage characteristics matrix can be obtained by using different definitions, that is, conventional leakage characteristic matrix based on pressure change or feature extraction with ML algorithms. Although PCA and AE models are used in this study, other ML models can also be integrated into this framework, such as the Mahalanobis classification system (MCS).[31] In step 3.2, the physical distance between pairs of junctions is obtained by using Dijkstra's[32] shortest pathfinding algorithm. Other shortest path algorithms could also be considered when dealing with different types of graphs, such as Floyd[33]–Warshall algorithm. This step guarantees the clustered junctions are concentrated based on their network path distance. Both of the pair distance matrices are normalized by dividing their largest value. Therefore, the range of these distances is from 0 to 1. In step 5, the represented distance between junctions is defined as the unweighted average of physical distance and leakage characteristics distance. The different ratios between the weight of leakage characteristics distance and physical distance will be discussed in *Hybrid approach for leakage detection and leakage zone localization*. In Step 6, the process of centroid redistribution of each cluster requires the re-acquire of the leakage characteristics matrix with the new set of centroids. Also, in step 6, unlike the standard *k-means* which used the mean value of each cluster as its centroids, the

Table 2. Modified k-means cluster algorithm for WDN partition.

Algorithm: Modified *k-means* algorithm for WDN leakage zone partition

Step 1: Initialize parameters: set the number of cluster *k*, tolerance, maximum iteration number

Step 2: Randomly select *k* junctions from the WDN as the first group of centroids

Step 3: Data preparation

Step 3.1: Prepare the leakage characteristics matrix

I.a) For the conventional leakage characteristics matrix, use Table 1

I.b) For PCA- or AE-based leakage characteristics matrix, follow Figures 2 and 3 respectively

II. Normalize the leakage characteristics matrix by dividing its maximum value

Step 3.2: Calculate the WDN physical pair distance matrix by computing the shortest path between all junctions. Standardize the matrix by dividing its maximum value

Step 4: Calculate the total Euclid distance between junctions

Step 4.1: Calculate the junctions' Euclid distance matrix measured by the junction leakage characteristics matrix pairs, $L_{leakage}$

Step 4.2: Calculate the component of Euclid distance matrix measured by the physical distance between junctions, $L_{physical}$

Step 5: Assign each junction $(v \in J)$ to its nearest clusters based on the total Euclid distancedefined as

$$L_{v,c_i} = (L^{leakage}_{(v,c_i)} + L^{physical}_{(v,c_i)})/2$$

$$v \in cluster_i,\ if\ L_{v,\,c_i} \le L_{v,\,c_l},\ \forall l \in \{1, 2, 3, \ldots, k\}$$

where J is the set of all junctions, L_{v,c_i} is the represent distance between junction i and centroid c_i, c_i is the centroid of cluster i

Step 6: Centroids redistribution

Step 6.1: For $cluster_i$, set junction v_k as the new centroid. Replace the original centroid c_i and determine the new group of centroids $(v_k, c_2, c_3, \ldots, c_k)$

Step 6.2: Recalculate the leakage characteristics matrix in **step 3.1** with the new group of centroids

Step 6.3: Recalculate the distance from to v_k all other junctions in $cluster_i$

Step 6.4: For every junction in $cluster_i$, repeat **step 6.1** to **6.3** to find the junction with the minimum total distance as the new centroid for $cluster_i$, i.e,

$$c_i^{new} = v_k,\ if \sum_{m=1}^{M} L_{v_k,v_m}\ is\ minimum,$$

where c_i^{new} is the new centroid for cluster i, M is all junctions in cluster i

Step 6.5: Repeat **step 6.1** to **6.4** for all clusters. Until the centroid distribution is stabilized

Step 7: Determine the sum of the distance of all clusters to their corresponding centroid from **step 6**. Repeat from **step 3** to **step 7** until the following relationship is satisfied

$$abs(\text{sum}(L_{v,c^{new}}) - \text{sum}(lL_{v,\,c})) < \text{tolerance}$$

$$or\ (\text{iteration} > \text{number of iteration})$$

where $l_{v,\,c}$ is the distance of each junction to its corresponding centroid, c is the old set of centroids, c^{new} is the new set of centroids

optimal junctions (that minimize distance within the cluster) is set as the new centroids so that centroids remain on the junctions in the WDN. The sensors are recommended to be placed at the centroid to maximize the value of data acquisition. Therefore, the influence of sensor placement is also considered during the WDN partition process.

Leakage monitoring stage: leakage detection and leakage zone localization

The WDN partition stage clusters the WDN into partition zones based on leakage characteristics and physical connectivity. The monitoring sensors are recommended to be installed at the centroids of the partition zones to achieve the best value of monitoring data. With the partition zones and sensor data, stage 2 implements algorithms for leakage detection and localization using the sensor monitoring data.

Leakage detection

Two unsupervised ML models, PCA and AE models, are used for leakage detection. Both PCA and AE models are capable of extracting the most important features from the training dataset. Testing data are projected or decomposed into the dimension-reduced feature space; and from the projected components, the original data can be reconstructed with small errors. However, abnormal data that carries unknown features will lead to large reconstruction errors from the pre-trained ML models. This allows abnormal events such as leakages to be detected. The advantage of the proposed models is that they can be trained with unbalanced data (normal non-leaking data in the case of WDN) to detect abnormal conditions.

The ML model training process for leakage detection is illustrated in part of Figure 4(a)). First, the unsupervised ML model is trained with a dataset under normal non-leaking conditions, as shown in step 1. In the testing stage, the dataset which contains non-leaking data (labeled by N) and limited number of labeled leaking data (labeled by J_i) is fed into the trained ML model (step 2). Using the reconstruction capability of the unsupervised ML model (step 3 and step 4), the input data is reconstructed (step 5). A reconstruction error θ for each sample is computed based on distance measure $\theta = p' - p$. Since the ML model is trained with non-leaking dataset only, among the testing dataset,

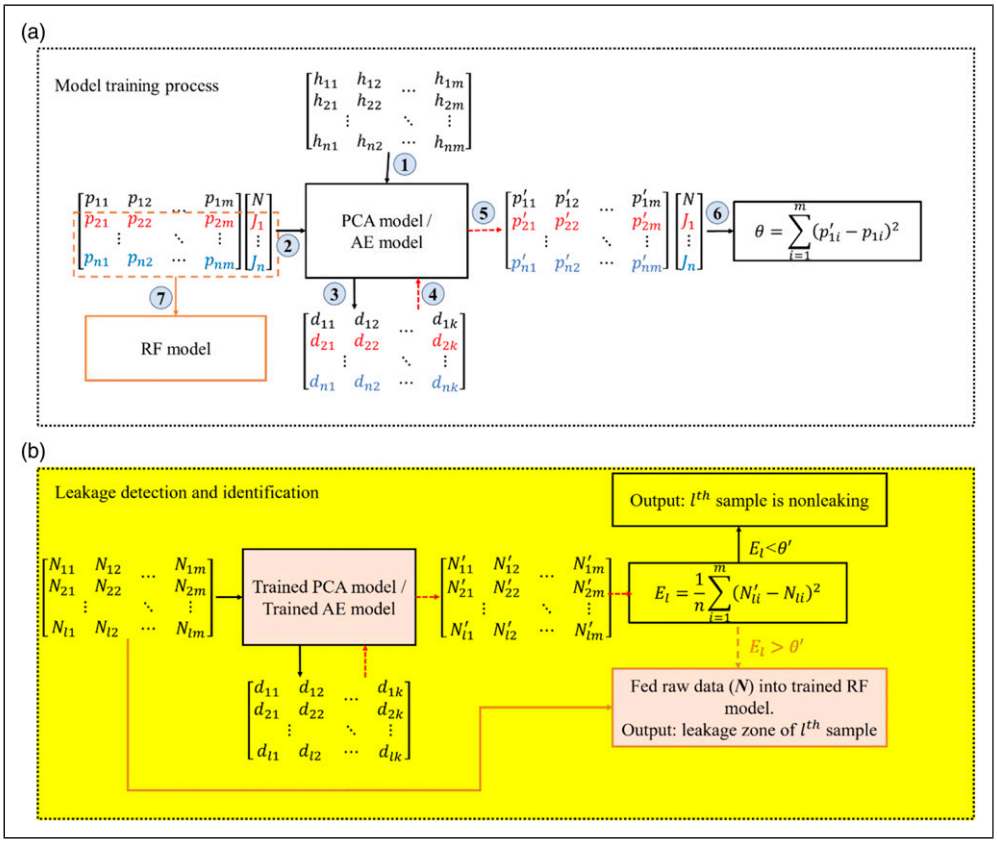

Figure 4. Illustration of the ML models for leakage detection and localization: (a) procedures for training ML models (note: h is the historical data matrix, p is the dataset with labels. p' is the reconstructed matrix of p. θ is the reconstruction error); (b) procedures for model applications in leakage detection and localization (note: N is the new observed dataset without labels. N' is the corresponding reconstructed matrix. E_l is the reconstruction error of sample l. θ' is the reconstruction error threshold).

non-leaking samples will have a small reconstruction error while leaking samples will have a larger reconstruction error. A threshold of reconstruction error can be used to differentiate leaking versus non-leaking dataset. This threshold can be obtained by only using non-leaking samples based on the characteristics of reconstruction errors, or further fine-tuned by labeled leaking samples.

Leakage localization

The leakage localization is defined as a classification problem, that is, the leakage conditions are classified into different WDN partition zones. There are various types of ML models for classification problems, such as the Artificial Neural Network, Support Vector Machine, Decision Tree, Random Forest (RF). In this study, Random Forest (RF) is used because (1) RF is an efficient classification algorithm, and (2) it only needs a very few hyperparameters to be tuned. These help with the efficiency and consistency during the evaluation process. It is noted that the other types of ML-based classifiers can also be used for leakage localization. The RF is trained with leaking samples with leakage zone labels (step 7).

With the trained models for leakage detection (PCA and AE) and model for leakage localization (RF), for each operational dataset, the leak is detected based on reconstruction error larger than the threshold θ. If a leak is detected, the data will be fed into the RF classifier for leakage zone localization. The detection and localization process for real-time monitored data is illustrated in Figure 4(b).

WDN operation data generation

The developed method for WDN leakage detection and localization can be readily applied to operational WDN. However, the monitoring data of in-service WDN is scarce. A hydraulic simulator of WDN is therefore utilized to generate a dataset to evaluate the developed framework. Simulation-based data generation is commonly used to develop ML models to overcome the limitations of physical data. For example, Tao[34] evaluated an artificial immune network with the dataset generated by water pipeline hydraulic simulation code EPANET. Similar works have also been done by many studies.[4,35–38] In this study, a python package WNTR is utilized to build the hydraulic model for WDN. The package implements the hydraulic model and solver of EPANET 2.2, which is an industrial hydraulic standard.[39] It is also capable of performing Monte Carlo simulations of WDN operations under different scenarios.[40]

By default, the hydraulic simulator considers the user node could always get designed water demand (d) even when the water pressure at that node is 0. However, due to the leakages, the supplied water demand (d^*) could be less than the designed water demand (d) when the water pressure is low. Herein, a pressure-dependent water model is used to consider the influence of water pressure on the water supply at each junction, which is assumed to follow Wagner's formulas[41] as shown in equation (1)

$$d^* = f(p) = \begin{cases} 0, & p \le P_0 \\ d\left(\dfrac{p - P_0}{P_f - P_0}\right)^{1/\eta}, & P_0 < p \le P_f \\ d, & p > P_f \end{cases} \quad (1)$$

where p is the water pressure at the junction, d is the designed water demand, d^* is the supplied water at different water pressure. P_0 is the minimum water pressure, P_f is the required water pressure to meet the designed water demand. η is the pressure exponent which is set as 2 in this study. The values of P_0 and P_f are set as 2 and 30 m, respectively, based on recommendation byReference 42.

The equation by Crowl and Louvar[43] is used as the leaking model. The model assumes there is a turbulent flow of water as leak occurs. The mass flow rate of the leakage is expressed by equation (2)

$$d_{\text{leak}} = C_d A \sqrt{2gh} \quad (2)$$

where $d_{\text{leak}}$is the leaking demand which depends on the water pressure. C_d is the discharge coefficient which is set as 0.75 in this study. A is the leaking area in the unit of m^2, h is the water head with unit of m, g is the gravity acceleration (m/s^2). To emulate the uncertainty of leakage size, a randomly generated value of the leaking area A is used in simulating different leaking scenarios.

The following procedures are used to generate dataset under normal (non-leaking) conditions and leaking conditions:

Data generation for normal operation scenario of the WDN:

1. *Define water pipe network*: Build the WDN pipe network with the corresponding pipeline geometry and material properties following EPANET data input format.
2. *Set the water demands at WDN junctions*: Each junction on the WDN has a design water demand. A Gaussian fluctuation is added to the design water demand as the total design water demand to consider the variations in user needs, that is

$$\text{Demand}_i = \left| D^i_{\text{base}} + N\left(0, \sigma_i^2\right) \right| \quad (3)$$

where D^i_{base} is the baseline design water demand at junction i which is defined in the original pipe network. A Gaussian term is added to consider the water usage fluctuations.

3. *Data generation*: With the defined water pipe network, the hydraulic model for the WDN is solved with proper hydraulic boundary conditions with the WNTR solver package. The results include all hydraulic information such as water pressure or flow rate at any location in the WDN.
   a. The results of water pressure at selected monitoring nodes under different WDN operational conditions are obtained.
   b. Gaussian noise $N(0, \gamma)$ is added to the water pressure data to mimic the noise in the monitoring data (due to sensor performance or other random factors).
   c. Store the data.

Steps 2 to 3 are repeated to generate data under different water demand conditions.

Data generation for WDN under leaking scenario:

Similar procedures are used to generate a dataset for WDN under the leaking scenario (steps 1–3). Except for the effects of leakages are considered in step 3 before solving the hydraulic model for the WDN. Leaking is assumed to occur at each junction, which is convenient for clustering purposes.[24,25] The leakage size is assumed to be randomly set between 0.05 m to 0.1 m in this study, which leads to $0.0012m^2$ to $0.0078\ m^2$ leaking size. These, however, can be easily changed to more complex leaking conditions.

Case study I: C-Town WDN

C-Town water distribution network is a WDN that was used for calibration competition in Battle of the Water Calibration Networks (BWCN) [49]. The topology of this WDN is shown in Figure 5. The WDN has 388 junctions that are connected by 429 pipes. The water source includes 1 reservoir and 7 water tanks. The water is powered by 11 pumps and controlled by 4 different kinds of valves. Each pump and valve have its functionality which will perform differently under different water usage situations. The WDN hydraulic model includes the junction locations, pipe lengths, pipe diameters, pipe roughness, water demands, user patterns, and working rules of pumps, valves, tanks, etc. The original data of the WDN shared by ASCE include hydraulic simulation at 168-time steps. The hydraulic model for the C-Town WDN was made public after the competition. The model can be downloaded from the ASCE Library, which also includes the EPANET input format file. The details of the network can be found in the original article.

Although the parameters of the C-Town WDN provided by the original article are deterministic values, the uncertainties of the WDN are considered in this study by adding randomness to the parameters. For example, a Gaussian distributed random value (equation (3)) was added to the

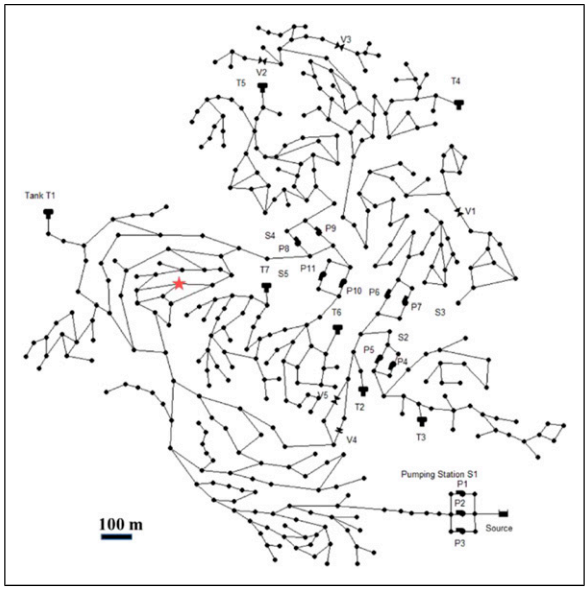

Figure 5. Topology of C-Town water distribution network (T: tank, S: meter district, P: pumps, V: valves, red star: Junction J370).

water demand of each junction to represent the uncertainties of water demand, the standard deviation is 10% of the junction's designed water demand. The leakage size of each leakage scenario was chosen from uniform distribution of 0.05 m–0.1 m. Besides, to consider the sensor noise, a random error of Gaussian distribution is also added to the water pressure data, which has a 0 mean value and 0.1 m standard deviation. Using the EPANET model for the C-Town WDN with proper hydraulic boundary conditions, simulations are conducted on the WDN under different operational conditions (i.e., the operational rules of the pumps and valves, water demand, and leakage occurrence). From these, the hydraulic data (i.e., water head and flow rate) at any location in the WDN can be obtained.

Figure 6 compares the total water head at junction "J370" (noted by a red star in Figure 5) with and without a nearby leakage, which clearly shows that leakage affects the hydraulic conditions in the WDN. Here, the water head is defined as total water head including the summation of the pressure head and junction's elevation head. It is noted that the water pressure under leaking conditions is sometimes higher and sometimes lower than that under normal conditions, possibly due to fluctuation in the WDN operational status. These make leaking detection and localization to be a challenging task.

C-town WDN partition results

The C-town WDN is partitioned following the procedures described in *WDN partition stage: modified k-means clustering algorithm*. Datasets of C-town WDN were

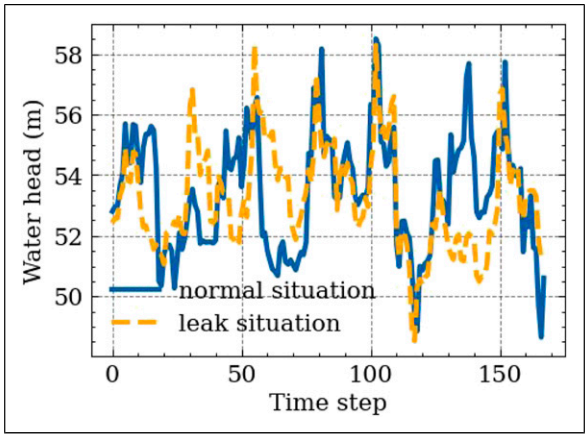

Figure 6. Example of water head fluctuation under normal or leaking conditions at Junction 370.

generated using the python package WNTR for both non-leaking conditions and leaking conditions via the procedures described in *WDN operation data generation*.

To calculate the leakage characteristic matrix, without loss of generality, a fixed leakage size of 0.05 m was assumed in the simulation, since the subsequent data normalization will take away the effects of leakage size. A leakage matrix is essential to obtain the leakage characteristics matrix for WDN partitioning. The leakage matrix is used to obtain the influence of leakage at different junctions on the monitoring locations. A fixed leakage size of 0.05 m was used to build the leakage matrix for the simplify consideration. This leakage size is selected based on the lower bound of leakage size. The effects of selected leakage size are minor since the leakage matrix can be normalized to determine the leakage characteristics matrix.

To evaluate the relative performance of the proposed partitioning method, the testbed C-town WDN were partitioned into different numbers of leakage zones based on 5 different partitioning methods, which utilizes different data feature versus Euclid distance measures. These comparative approaches are described as following.

1. Standard k-means clustering using the conventional leakage characteristics matrix (Table 1), which was used in previous studies such as Zhang.[24] This is named conventional partition in this article.
2. Modified *k*-means clustering using conventional leakage characteristics matrix and physical distance matrix (named as modified partition).
3. Modified *k*-means clustering using PCA extracted leakage characteristic matrix and physical distance matrix (named as PCA-based partition).
4. Modified *k*-means clustering using AE-based leakage characteristic matrix and physical distance matrix (named as *AE-based partition*).
5. Modified *k*-means clustering only using the physical distance matrix (named as *graph distance-based partition*).

The results of WDN partition into different numbers of clusters (6, 10, and 14) by the five different partition procedures are shown in Figure 7. Junctions of the cluster are represented with different colors, with the centroid of each cluster indicated with a rectangle symbol with the same color as its cluster.

As shown in Figure 7, subgraph 1 indicate that conventional partition using the traditional *k-means algorithm* without the consideration of the graph distance between junctions, the junctions in clusters are scattered and intermingled. The scattering increases with the increasing number of partition zones. From subgraph 2, the modified partition based on modified *k-means* clustering algorithm effectively reduces the scattering since it considers the graph distance of junctions in addition to the conventional leakage characteristics. This leads to a much smaller number of isolated junctions. Comparison of subgraph 3 and 4 vs 2 shows that PCA-based partition and AE-based partition further reduce the scattering based on raw leakage characteristic matrix. PCA-based partition and AE-based partition achieved comparable results. It is noted that the AE-based partition of the C-town WDN took more than 10 h while the PCA-based partition required only a few minutes. The main reasons include the AE method needs more time for training and requires more iteration times to get convergence. A potential solution is using regularization methods for AE-based partition.[44]

C-town leakage Detection

The leakage detection is demonstrated on the clustering result when the C-town WDN is partitioned into 10 clusters by PCA-based partition. The monitoring sensors are assumed to be installed at the centroid of each cluster and the corresponding data are used (shown in Figure 7 (3) by rectangles) for leakage detection. PCA and AE models are used for leakage detection.

With the data generation procedures outlined in *WDN operation data generation*, the dataset with 1000 non-leaking samples under different operation conditions of the WDN is generated. The non-leaking data are randomly split into a subset of 700 and 300 samples. Then, 300 leaking samples were generated by setting a random leakage size at a randomly picked junction. The subset of 700 non-leaking samples is used as the training dataset. The subset of 300 non-leaking samples together with the 300 leaking samples is used as the testing dataset.

The ML-based leakage detectors (AE model or PCA model) are first individually trained with the training dataset (700 non-leaking samples). With the trained ML models, the

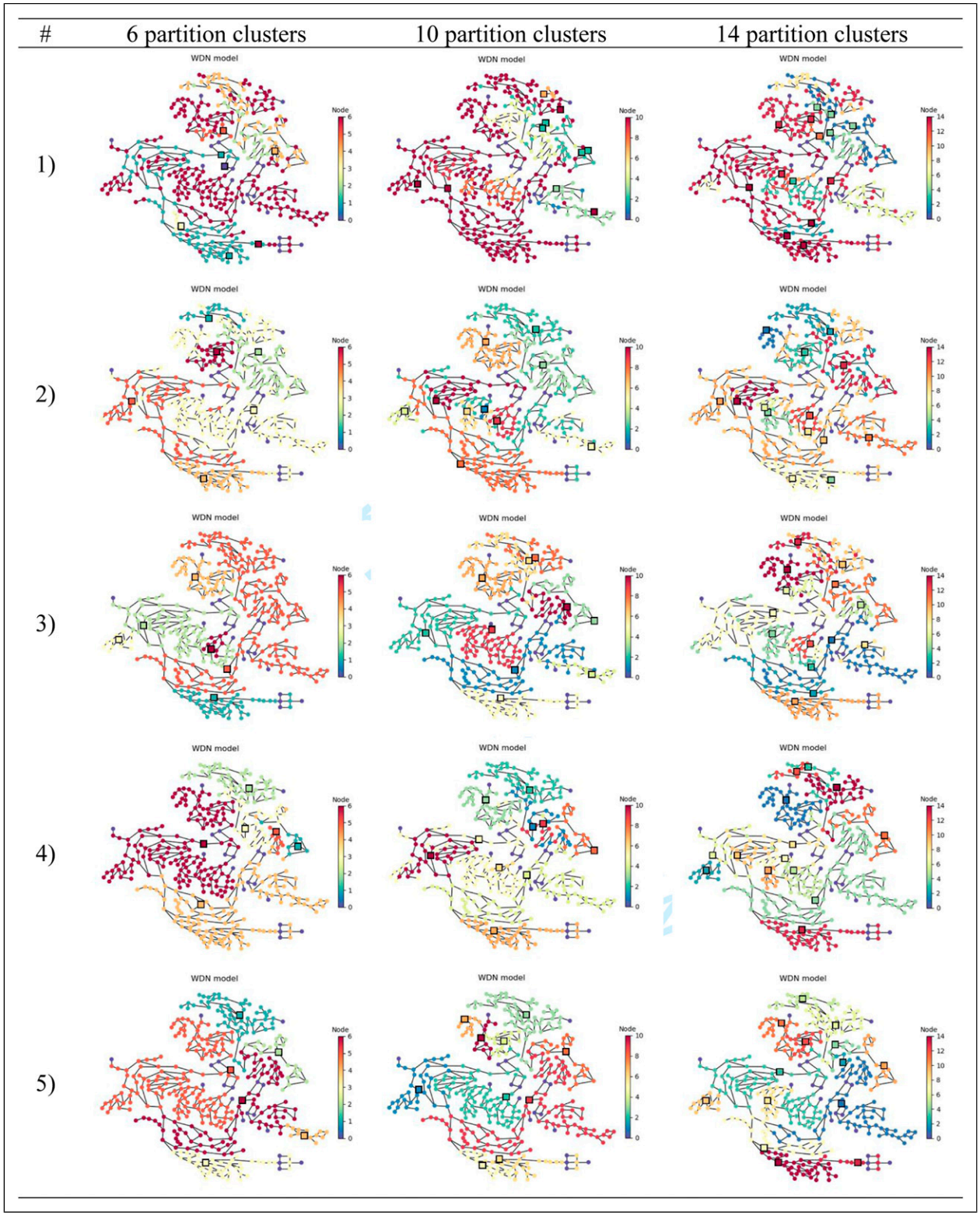

Figure 7. Results of water distribution networks leakage zone partition with five different approaches (Note: class 1 to *k* are the leakage zone IDs. Class 0 is the pump/tank/reservoir nodes. Black rectangular denotes the pressure sensor location).

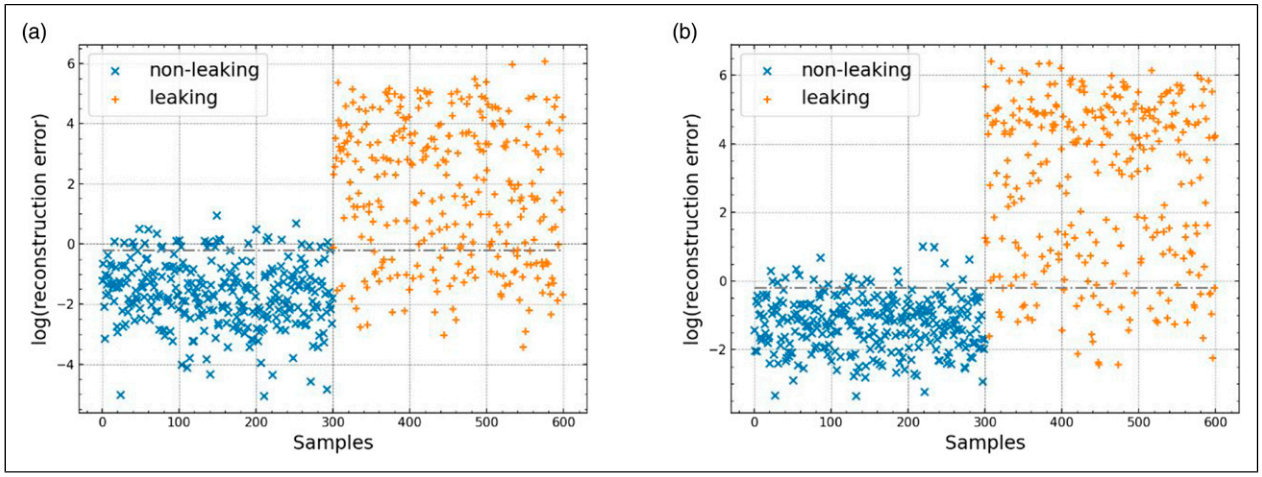

Figure 8. Reconstruction error of non-leaking versus leaking samples: (a) AE detector; (b) PCA detector.
Note: AE: autoencoder algorithm.

testing datasets are fed as inputs. The reconstruction errors of the input data by the AE detector and PCA detector are shown in Figure 8. The models feature larger reconstruction error with leaking samples than non-leaking samples, which is the basis for differentiating leaking versus non-leaking cases. A threshold can be set to achieve the best leaking detection performance. For practical implementation, this threshold can be empirically set initially based on statistical analyses of the reconstruction error distribution with monitored data under non-leaking condition. For example, the maximum value or third quantile value can be used. It can be fine-tuned when more leaking data become available.

If the reconstruction error of a non-leaking sample is smaller than the threshold or a leaking sample is larger than the threshold, this sample will be recognized as correctly classified. Misclassifications happen with the set threshold, that is, leaking samples are classified as non-leaking, or non-leaking samples are classified as leaking. The leakage detection accuracy is assessed by the number of correctly classified cases over the total number of testing samples.

The leakage detection performance of PCA and AE detectors when the WDN is partitioned into different numbers of leakage zones is summarized in Figure 9. Since monitoring data are assumed to be at the centroid of each partition, the dataset for each number of partitions has to be regenerated for each case using the data generation process described in *WDN operation data generation*. The size of training dataset and testing dataset are kept the same throughout. To overcome the uncertainties during the data generation, the average accuracy of 5 cross-validations is reported for each case. As seen in Figure 9, the leakage detection accuracy steadily increased with the increasing number of WDN partitions. It is understandable since the more sensors deployed in the WDN, the more water leakage scenarios would be covered. The results also show that the AE leak detector overperforms the PCA detector by about 5%.

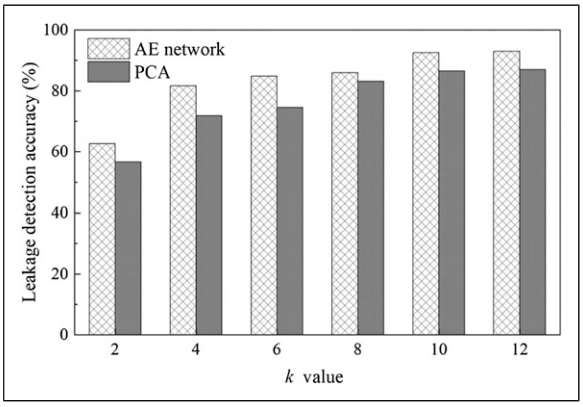

Figure 9. Leakage detection accuracy with two different unsupervised ML models under different number of partitions (k value).

C-town leakage zone localization

Besides detecting leakage, localizing the leakage is also important for retrofit actions on the WDN. Conventional supervised ML classifier requires training dataset must include data of leaking occurring at each WDN junction. In practice, however, leakage only appears at limited locations, which makes it infeasible to well train a supervised ML model. With the partition of WDN, the leakage localization problem is defined as a semi-supervised classification problem. The Random Forest (RF) model is chosen for leakage zone localization following the procedures outlined in Figure 4(b).

The leakage localization performance of using different partition methods is first evaluated in WDN by partitioning the WDN into 6 leakage zones. A small portion of total junctions (assumed as 10% in this study which can be changed to other assumptions without loss of generality) are assumed to have experienced leakage. The leaking junctions are assumed to be evenly distributed among the 6 partition zones. Based on this assumption, leaking data samples are generated by assuming a leakage of random size occurring at one of these selected junctions. For each of the 6 partition zones, 400 leaking data samples are generated. Therefore, the total training dataset includes 2400 data samples, with their respective labels of partition zones they belong to.

For the testing data, 200 leaking data samples are randomly generated assuming leakage of random size occurring at randomly picked junctions in the remaining 90% junctions. Altogether, the testing dataset includes 1200 leaking samples.

A confusion matrix is often used to evaluate the classification performance, which is also used for leakage localization performance in this study. A typical confusion matrix contains four prediction terminologies: True Positive, True Negative, True Positive, and False Negative. For a multi-classification confusion matrix with a structure like Figure 10, the True Positive indicate the number of correctly predicted samples, the rest rows of each class are False Positive, and the rest columns are False Negative. The accuracy, recall value, and precision matrices are used for evaluation since this is a balanced testing dataset. For a better understanding, the equations to compute each matrix are shown in equations (5)–(7)

$$\text{acc} = \frac{\sum_{i=1}^{k} M_{ii}}{\text{sum}(M)} \tag{5}$$

$$\text{Recall}_i = \frac{M_{ii}}{\sum_{j=1}^{k} M_{ij}} \tag{6}$$

$$\text{Precision}_i = \frac{M_{ii}}{\sum_{j=1}^{k} \sum_{j=1}^{k} M_{ji}} \tag{7}$$

where M is the confusion matrix, i is the i^{th} class, and k is the total number of classes. M_{ij} indicates the i^{th} row and j^{th} column.

The RF leakage zone detection is implemented on WDN using different partition methods, that is, AE-based partition, PCA-based partition, Modified Partition, and Graph distance-based partition. The confusion matrices of the final leakage localization results are shown in Figure 10. From the comparison, the RF-based leakage localization using PCA-based partition achieved the highest accuracy of 91% (Figure 10(a)). This is followed by 88% accuracy using

(a) PCA-based partition

Actual Class \ Predicted class	1	2	3	4	5	6	Recall
1	176	11	13	0	0	0	0.88
2	4	184	12	0	0	0	0.92
3	16	4	158	6	6	10	0.79
4	0	0	2	198	0	0	0.99
5	0	4	9	0	187	0	0.94
6	6	2	8	0	0	178	0.92
Precision	0.87	0.90	0.78	0.97	0.97	0.90	0.91

(b) AE-based partition

Actual Class \ Predicted class	1	2	3	4	5	6	Recall
1	172	2	19	0	0	7	0.86
2	0	171	0	15	9	5	0.86
3	25	0	160	0	0	5	0.84
4	0	10	1	187	0	2	0.94
5	11	13	0	0	176	0	0.88
6	11	6	5	0	0	178	0.89
Precision	0.78	0.85	0.86	0.93	0.95	0.90	0.88

(c) Modified partition

Actual Class \ Predicted class	1	2	3	4	5	6	Recall
1	150	3	14	7	26	0	0.75
2	0	189	1	0	10	0	0.95
3	0	0	193	6	1	0	0.97
4	0	0	103	95	2	0	0.48
5	22	21	24	4	129	0	0.65
6	0	0	11	10	2	177	0.89
Precision	0.87	0.89	0.56	0.78	0.76	1.00	0.78

(d) Graph distance-based partition

Actual Class \ Predicted class	1	2	3	4	5	6	Recall
1	121	5	0	37	4	33	0.61
2	0	181	0	6	9	4	0.91
3	43	0	153	3	0	1	0.77
4	0	18	0	139	21	22	0.70
5	0	16	0	11	171	2	0.86
6	6	0	0	118	3	72	0.36
Precision	0.71	0.82	1.00	0.44	0.82	0.54	0.70

Figure 10. Accuracy of leakage zone localization using different WDN partitioning strategies: (a) PCA-based partition; (b) AE-based partition; (c) modified partition; (d) graph distance-based partition.
Note: WDN: water distribution networks; AE: autoencoder algorithm.

AE-based partition (Figure 10(b)), 78% accuracy using modified partition (Figure 10(c)), 70% accuracy using graph distance-based partition (Figure 10(d))).

The effects of the number of WDN partitions on the leakage zone localization accuracy are further evaluated and summarized in Figure 11(a)). 20 trials are conducted for each case to eliminate the randomness and the average accuracy is plotted. The accuracy in leakage localization by RF consistently achieved top 2 performance by using AE-based or PCA-based partitions. The accuracy of localization is worst when only considers the physical distance of junctions with no consideration of the leakage characteristics. It is noted that leakage zone localization using AE-based partition started to overperform that using the PCA-based partition for a larger number of partitions (i.e., 12). This might be attributed to that AE-based partition is more capable of identifying complex relationships from the data.

Figure 11(b) shows the influence of different percentages of junctions with leakage data on the accuracy of leakage zone localization, which is performed by partitioning the WDN into 10 leakage zones via different partition methods. The results show with leaking data available at more junctions, the leakage localization accuracy improves. The results also showed that the RF-based leakage zone localization achieved the best performance with PCA-based partition.

Hybrid approach for leakage detection and leakage zone localization

The analyses so far indicate that the AE model achieved higher accuracy than the PCA model for leak detection, possibly due to its ability to extract non-linear relationships among the input features. Meanwhile, the RF-based leakage localization achieved the highest accuracy when using PCA-based WDN partition, possibly because the PCA-based information extraction is easier to be learned by RF. PCA also features higher computing efficiency. Therefore, a hybrid framework is proposed that combines the use of PCA-based partition, AE-based leakage detection, and RF-based leakage localization.

In practice, resource constraints might prevent sensors to be installed at the most optimal junctions. To consider such issue, analyses are conducted under the scenery where sensors are assumed to be "randomly placed" in the WDN and the leakage zones are clustered using these sensors as the centroids. The accuracy of leakage detection and localization under "optimal sensor placement" and "random sensor placement" are determined using the C-Town WDN testbed. The final results are summarized in Figure 12. The results of random sensor placement are the mean values of 10 different random sensor deployment scenarios.

As seen in Figure 12, deployment at non-ideal locations slightly compromises the accuracy of the proposed leakage detection and localization method. However, it still achieved an overall accuracy between 70% and 80%. The overall performance is regarded satisfactory. This is vindication of the accuracy and robustness of the developed method. It is also observed from Figure 12(b) that the differences in the detection accuracy under optimal versus non-optimal sensor locations diminishes with the increasing number of leakage zones. With more leakage zones partitions, more sensors are placed in the WDN. Placing sensors at optimal locations become less important for the developed leakage detection framework.

Initially, the proposed framework is illustrated by defining distance L_{v,c_i} as the average distance of leakage characteristics distance and graph shortest path distance. The sensitive study about the different penalty weights of the leakage characteristics distance and graph distance is conducted here based on the hybrid partition and detection framework. In detail, the distance is redefined as equation (4). The target leakage zone number (k) is set as 10 for the illustration purpose.

$$L_{v,c_i} = w_1 * L^{\text{leakage}}_{(v,c_i)} + w_2 * L^{\text{physical}}_{(v,c_i)} \tag{4}$$

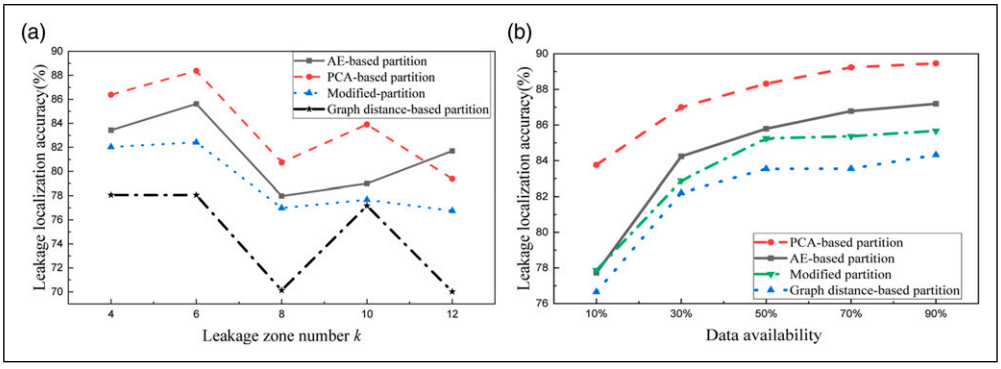

Figure 11. Leakage localization accuracy with RF model under (a) different numbers of partition zones, and (b) percentage of junctions with leakage data available (water distribution networks with 10 partitions).

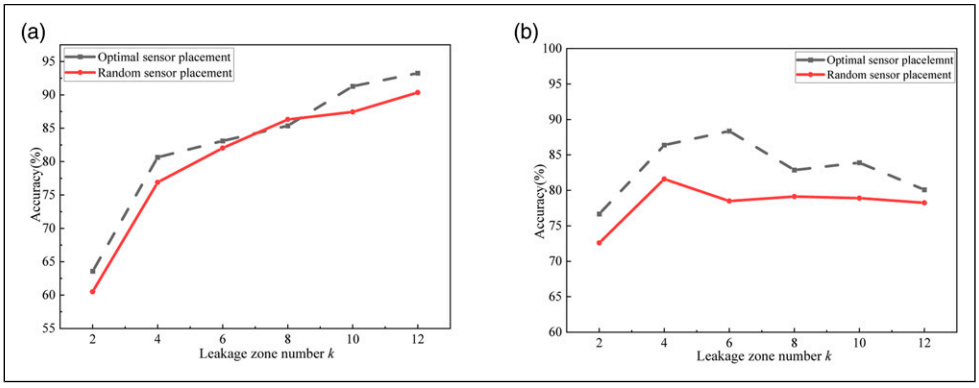

Figure 12. Comparison the influence of sensor placements on the performance of the developed CtL-SSL framework: (a) leakage detection accuracy; (b) and leakage localization accuracy.

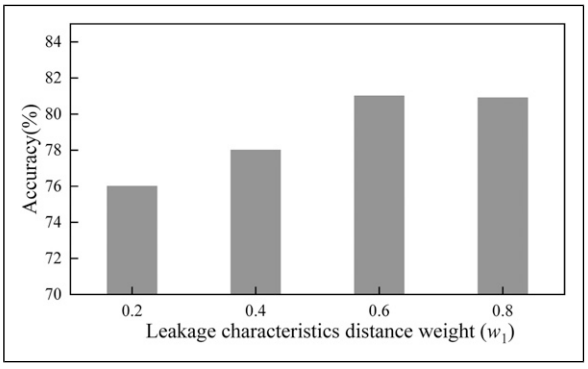

Figure 13. The influence of weights assigned to the leakage characteristics distance during Water distribution networks clustering (stage 1) on the final leakage localization accuracy (stage 2).

s.t. $w_1 + w_2 = 1$

where w_1 is the weight assigned to the leakage characteristics distance and w_2 is the weight assigned to the physical distance.

Sensitivity analyses are conducted on the effects of weights assigned to the leakage characteristic distance. The final leakage localization accuracy when using different values of w_1 is shown in Figure 13 and the corresponding WDN partition results are shown in Figure 14. The results indicate the leakage localization framework achieved higher than 76% when 20% weight is assigned to the leakage characteristic distance (80% weight by the physical distance). The performance improved with increasing weight of the leakage characteristics. However, about 60% weight, the performance improvement becomes insignificant. Meanwhile, a more scattered leakage zone partition result is observed when the leakage characteristics are given higher weights, as demonstrated in Figure 14. These observations indicate that an optimal weight exist that achieves balanced consideration of the leakage localization accuracy and the leakage zone scattering degree.

Case study II: Rancho Solano Zone III WDN

To further illustrate the proposed leakage partition, detection, and localization framework, another WDN, Rancho Solano Zone III WDN is used as the second independent testbed. This testbed is located in Fairfield, California. The information about this WDN is published by ASCE task committee on a research database for water distribution systems and are open to download from the database of Kentucky University [50]. The graph of this water supply network is shown in Figure 15. There are 112 nodes in total, including one reservoir and one water treatment plant as the source of water, and 126 pipes. The elevations of the nodes in this pipe network range from 90 m to 120 m and the length of the pipes range from 90 m to 130 m. The original data of water demand and water supply conditions for this WDN are used in this study.

The same uncertainties of water demand and sensor noise that are considered in the Case study I are used again in this testbed. The leakage uncertainty range is set as $U(0.01, 0.05)$ since a too large leakage size could directly drainage all the water in this WDN. A 40 time steps water pressure record of Junction "F010" is shown in Figure 16 to illustrate the influence of a nearby leakage.

Rancho Solano Zone III WDN partition results

The proposed hybrid framework in *Hybrid approach for leakage detection and leakage zone localization* is applied on the Rancho Solano Zone III to illustrate the effect of WDN partition results. The considered leakage zone numbers are 2, 4, and 6 in this study. When using an equivalent penalty weight of the leakage characteristics and physical distance, the final partition results when considering different numbers (k) of partition zones are shown in Figure 17.

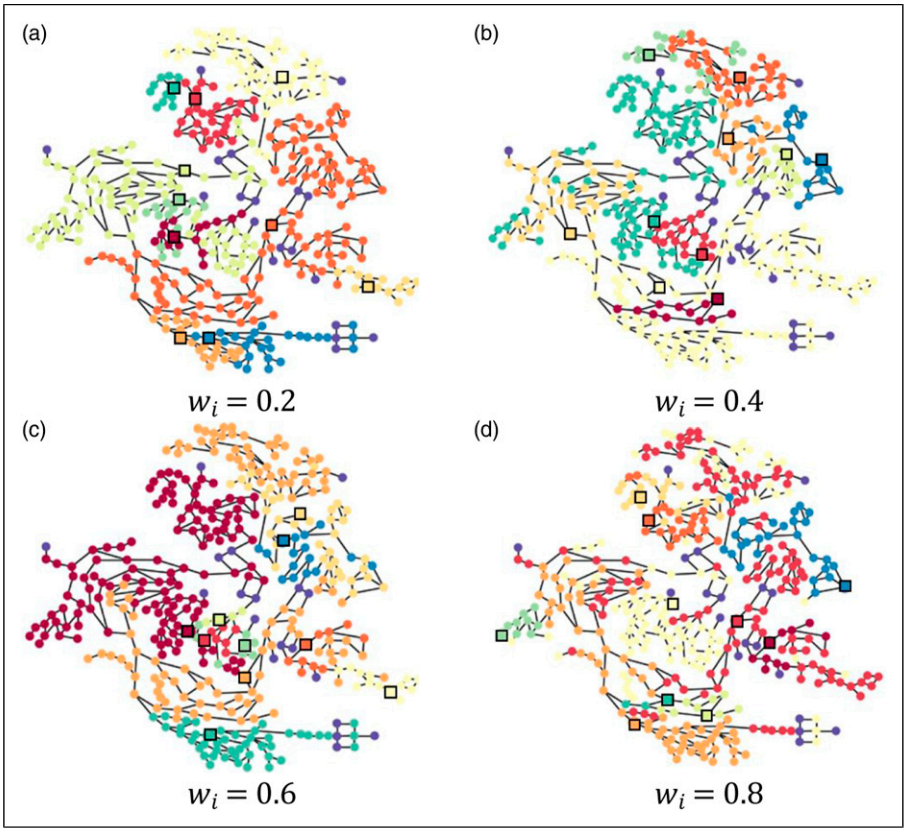

Figure 14. The water distribution networks partition results based on different weights assigned to the leakage characteristics: (a) $w_i = 0.2$; (b) $w_i = 0.4$; (c) $w_i = 0.6$; (d)$w_i = 0.8$.

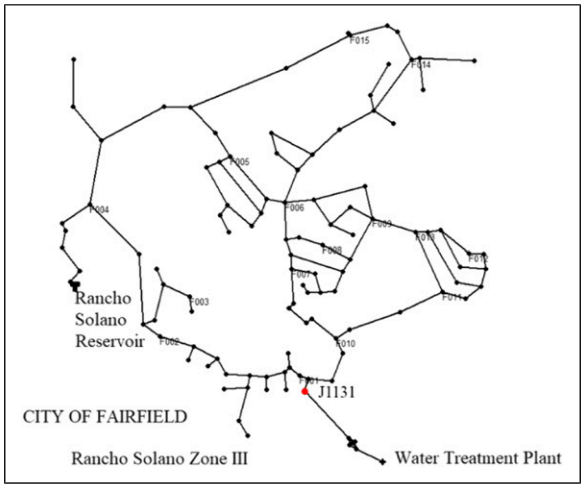

Figure 15. Water distribution networks graph structural of Rancho Solano Zone III (Red node denotes junction "J1131").

As can be seen from the results, the partitioned results are reasonably balanced and concentrated. Hence, the maintenance team can easily narrow down the inspection area after a leakage is detected.

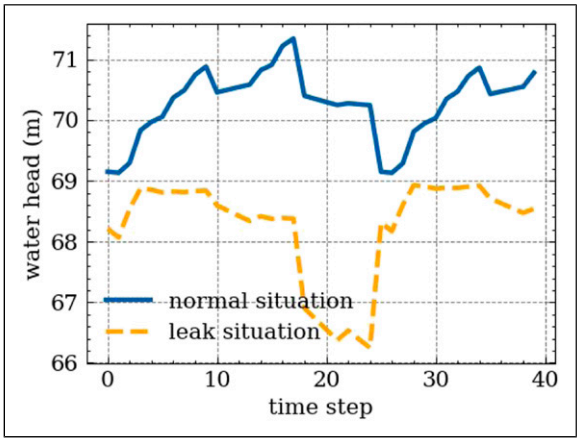

Figure 16. Example of water head fluctuation under normal or leaking conditions at Junction "1131."

Rancho Solano Zone III leakage detection and localization results

Similar to case study I, only non-leaking monitored water pressure data is used for leakage detection and only 10% of junctions of each leakage zone are assumed to have leakage

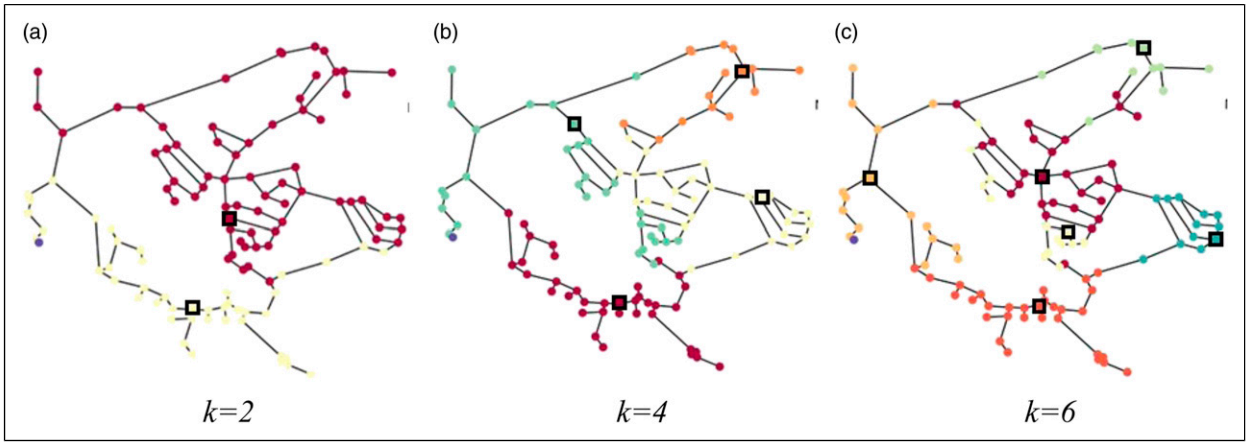

Figure 17. Water distribution networks leakage zone partition results when considering different values of *k* (where the black rectangular denotes the pressure sensors location): (a) *k*=2; (b) *k*=4; (c) *k*=6.

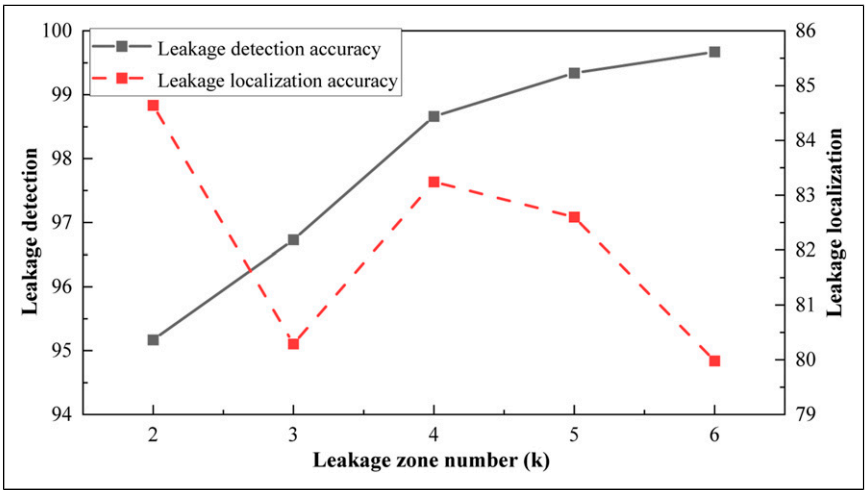

Figure 18. Final leakage detection and localization accuracy.

experience. Hence, the recorded water pressure under available leakage experiences can be used for the leakage localization model training. A similar preparing process for the training dataset and testing dataset in case study I is also used here. Figure 18 shows the final leakage detection accuracy and leakage localization accuracy when the WDN is portioned into different numbers of leakage zones. As can be seen from the results, the leakage detection accuracy increases with the increasing number of leakage zones which is also the number of placed sensors. It is understandable since the more sensors the higher chances that one of them could be impacted by the leakage. For a small-scale WDN like Rancho Solano Zone III, the water pressure of each junction is sensitive to any leakage situation, so the leakage detection achieved about 95% even with two sensors. On the other hand, the leakage zone localization accuracy fluctuates from 80% to 86% when partitioning the WDN into a different number of leakage zones. The overall accuracy is still acceptable.

Conclusions

A novel CtL-SSL framework is developed for WDN leakage management in this study. The framework includes WDN leakage zone partition, leakage detection, and leakage zone localization. The WDN partition is based on the leaking behaviors of the WDN junctions. New leakage characteristics are defined based on features extracted from non-leaking data with unsupervised ML models such as PCA or AE. Improved k-means method is proposed for WDN partition, which considers the graph distance between junctions and the leakage characteristics. Sensors are recommended to be installed at the centroid junction of each partition to acquire monitoring data. With the monitoring

data, unsupervised ML models are developed for leakage detection based on threshold criteria of reconstruction errors. This allows leakages to be detected with unbalanced dataset that contains non-leaking samples only. With the leakage zone partition of the WDN, the leakage zone localization is defined as a ML-based classification problem using partition zone numbering, which is achieved with a small percentage of leaking data.

The results indicate the new partition algorithm (stage 1) achieves less intermingling of junctions from different partitions compared with the conventional partition method. The leakage detection and localization stage (stage 2) also gained promising performance even with leakage data over only a small portion of junctions. The proposed framework achieved around 95% accuracy in leakage detection and 83% leakage localization accuracy in both case studies with less than 10% of junctions' leakage data.

The proposed CtL-SSL framework can be easily used on different WDNs and updated with more powerful models in the future, which increases its extensibility and upgradeability. The final performance may vary when the number of leakage zones and scales of WDNs are different. Determining the optimal number of leakage zones for different types of WDN is still a problem worth future investigation. In practice, an optimal number of leakage zones should not only consider the final detection and localization accuracy but also factors such as budget limitation, expected leakage zone resolution, social-economic impact, and so on. Moreover, the proposed method is developed and validated by use of data generated by use of hydraulic model for WDNs. While it is common to use holistic simulation data for development and validation of machine learning models, further validation with data from real-world in service WDN are highly recommended.

Acknowledgments

The authors would like to thank the Cleveland Water Department, especially Mr. Alex Margevicius and his team for providing guidance to the team during the study. The research is partially supported by the US National Science Foundation grant No. 1638320.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is partially supported by the US National Science Foundation.

ORCID iD

Xiong (Bill) Yu https://orcid.org/0000-0001-6879-2567

References

1. Sadeghioon A, Metje N, Chapman D, et al. SmartPipes: smart wireless sensor networks for leak detection in water pipelines. *J sensor Actuator Networks* 2014; 3(1): 64–78.
2. Berardi L, Giustolisi O, Kapelan Z, et al. Development of pipe deterioration models for water distribution systems using EPR. *J Hydroinformatics* 2008; 10(2): 113–126.
3. Fontanazza CM, Notaro V, Puleo V, et al. Contaminant intrusion through leaks in water distribution system: experimental analysis. *Proced Eng* 2015; 119: 426–433.
4. Chan TK, Chin CS and Zhong X. Review of current technologies and proposed intelligent methodologies for water distributed network leakage detection. *IEEE Access* 2018; 6: 78846–78867.
5. Kang J., Park Y, Lee J, et al. Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems. *IEEE Trans Ind Electro* 2017; 65(5): 4279–4289.
6. Ozevin D and Yalcinkaya H. Reliable monitoring of leak in gas pipelines using acoustic emission method. In: Proc. Civil struct. Health monit. Workshop on Civil Structural Health Monitoring (CSHM-4), Berlin, Germany, November 6–8, CSHM, 2012.
7. Gao J, Shi B, Zhang W, et al. Monitoring the stress of the post-tensioning cable using fiber optic distributed strain sensor. *Measurement* 2006; 39(5): 420–428.
8. Amran TST, Ismail MP, Ahmad MR, et al. Detection of underground water distribution piping system and leakages using ground penetrating radar (GPR). In: AIP conference proceedings, Selangor, Malaysia, 8–10 August 2016, AIP Publishing LLC, 2017.
9. Bimpas M, Amditis A and Uzunoglu N. Detection of water leaks in supply pipes using continuous wave sensor operating at 2.45GHz. *J Appl Geophys* 2010; 70(3): 226–236.
10. De Coster A, Pérez Medina JL, Nottebaere M, et al. Towards an improvement of GPR-based detection of pipes and leaks in water distribution networks. *J Appl Geophys* 2019; 162: 138–151.
11. Butler D. *Leakage Detection and Management: A Comprehensive Guide to Technology and Practice in the Water Supply Industry*. Nashville, TN: Palmer Environmental, 2000.
12. Srirangarajan S, Allen M, Preis A, et al. Wavelet-based burst event detection and localization in water distribution systems. *J Signal Process Syst* 2013; 72(1): 1–16.
13. Mashford J, De Silva D, Burn S, et al. Leak detection in simulated water pipe networks using SVM. *Appl Artif Intelligence* 2012; 26(5): 429–444.
14. Colombo AF, Lee P and Karney BW. A selective literature review of transient-based leak detection methods. *J hydro-environment Res* 2009; 2(4): 212–227.

15. Adedeji KB, Hamam Y, Abe BT, et al. Towards achieving a reliable leakage detection and localization algorithm for application in water piping networks: An overview. *IEEE Access* 2017; 5: 20272–20285.
16. Romano M, Kapelan Z and Savić DA. Automated detection of pipe bursts and other events in water distribution systems. *J Water Resour Plann Manag* 2012; 140(4): 457–467.
17. Liao S-H, Chu P-H and Hsiao P-Y. Data mining techniques and applications - A decade review from 2000 to 2011. *Expert Syst Appl* 2012; 39(12): 11303–11311.
18. Buch N, Velastin SA and Orwell J. A review of computer vision techniques for the analysis of urban traffic. *IEEE Trans Intell Transportation Syst* 2011; 12(3): 920–939.
19. Lin W-Y, Hu Y-H and Tsai C-F. Machine learning in financial crisis prediction: a survey. *IEEE Trans Syst Man, Cybernetics, C (Applications Reviews)* 2011; 42(4): 421–436.
20. Chim TW, Yiu SM, Hui LCK, et al. SPECS: Secure and privacy enhancing communications schemes for VANETs. *Ad Hoc Networks* 2011; 9(2): 189–203.
21. Kim J-H, Sharma G, Boudriga N, et al. SPAMMS: A sensor-based pipeline autonomous monitoring and maintenance system2010. In: Second international conference on communication systems and networks (COMSNETS 2010), Bangalore, India, 5–9 January, 2010, IEEE, 2010.
22. Stoianov I, Nachman L, Madden S, et al. PIPENETa wireless sensor network for pipeline monitoring. In: Proceedings of the 6th international conference on Information processing in sensor networks, Cambridge, MA, USA, 12 November 2007, 2007.
23. Saade M and Mustapha S. Assessment of the structural conditions in steel pipeline under various operational conditions - A machine learning approach. *Measurement* 2020; 166: 108262.
24. Zhang Q, Wu ZY, Zhao M, et al. Leakage zone identification in large-scale water distribution systems using multiclass support vector machines. *J Water Resour Plann Manag* 2016; 142(11): 04016042.
25. Chen J, Feng X and Xiao S. An iterative method for leakage zone identification in water distribution networks based on machine learning. *Struct Health Monit* 2020; 20: 1938–1956.
26. Romano M, Kapelan Z and Savić DA. Evolutionary algorithm and expectation maximization strategies for improved detection of pipe bursts and other events in water distribution systems. *J Water Resour Plann Manag* 2014; 140(5): 572–584.
27. Wu Y, Liu S, Wu X, et al. Burst detection in district metering areas using a data driven clustering algorithm. *Water Res* 2016; 100: 28–37.
28. Kadri A, Abu-Dayya A, Trinchero D, et al. Autonomous sensing for leakage detection in underground water pipelines 2011. In: Fifth international conference on sensing technology, Palmerston North, New Zealand, 28 November–1 December 2011, IEEE, 2011.
29. Wold S, Esbensen K and Geladi P. Principal component analysis. *Chemometrics Intell Lab Syst* 1987; 2(1–3): 37–52.
30. Zheng YW and Yuan S. Optimizing pressure logger placement for leakage detection and model calibration. In: WDSA 2012: 14th water distribution systems analysis conference, Adelaide, South Australia, 24–27 September 2012. Engineers Australia, 2012
31. Cheng L., Yaghoubi V, Paepegem WV, et al. Quality inspection of complex-shaped metal parts by vibrations and an integrated Mahalanobis classification system. *Struct Health Monit* 2020; doi: 10.1177/1475921720979707.
32. Dijkstra EW. A note on two problems in connexion with graphs. *Numerische mathematik* 1959; 1(1): 269–271.
33. Floyd RW. Algorithm 97: shortest path. *Commun ACM* 1962; 5(6): 345.
34. Tao T, Haidong H, Fei L, et al. Burst detection using an artificial immune network in water-distribution systems. *J Water Resour Plann Manag* 2013; 140(10): 04014027.
35. Bakker M, Vreeburg JHG., van Schagen KM, et al. A fully adaptive forecasting model for short-term drinking water demand. *Environ Model Softw* 2013; 48: 141–151.
36. Ye G and Fenner RA. Kalman filtering of hydraulic measurements for burst detection in water distribution systems. *J pipeline Syst Eng Pract* 2011; 2(1): 14–22.
37. Abokifa AA, Haddad K, Lo C, et al. Real-time identification of cyber-physical attacks on water distribution systems via machine learning–based anomaly detection techniques. *J Water Resour Plann Manag* 2018; 145(1): 04018089.
38. Mazzolani G, Berardi L, Laucelli D, et al. Estimating leakages in water distribution networks based only on inlet flow data. *J Water Resour Plann Manag* 2017; 143(6): 04017014.
39. Rossman L, Woo H, Tryby M, et al. *EPANET 2.2 User Manual*. U.S. Environmental Protection Agency, 2020. EPA/600/R-20/133.
40. Klise K, Moriarty ML, Bynum R, et al. *Water Network Tool for Resilience (WNTR) User Manual*. Albuquerque, NM (United States): Sandia National Lab.(SNL-NM), 2020.
41. Wagner JM, Shamir U and Marks DH. Water distribution reliability: simulation methods. *J Water Resour Plann Manag* 1988; 114(3): 276–294.
42. He P, Tao T, Xin K, et al. Modelling water distribution systems with deficient pressure: an improved iterative methodology. *Water Resour Manag* 2016; 30(2): 593–606.
43. Crowl DA and Louvar JF. *Chemical Process Safety: Fundamentals with Applications*. Pearson Education, 2001.
44. Kukačka J, Golkov V and Cremers D. *Regularization for Deep Learning: A Taxonomy*, 2017. arXiv preprint arXiv: 1710.10686.