Machine Learning Model and Strategy for Fast and Accurate Detection of Leaks in Water Supply Network

Xudong Fan
Case Western Reserve University
Xijin Zhang
Case Western Reserve University
Xiong (Bill) Yu (xxy21@case.edu)
Case Western Reserve University https://orcid.org/0000-0001-6879-2567

Research Article

Keywords: Water supply network, Artificial intelligence, Machine learning, Artificial Neural Network, Autoencoder Neural Network, Leak detection

Posted Date: February 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-214294/v1

Version of Record: A version of this preprint was published at Journal of Infrastructure Preservation and Resilience on April 15th, 2021. See the published version at https://doi.org/10.1186/s43065-021-00021-6.

Machine learning model and strategy for fast and accurate detection of leaks in water supply network

Xudong Fan[1], Xijin Zhang[2], Xiong (Bill) Yu[3*]

[1]Graduate Research Assistant, Department of Civil and Environmental Engineering, Case Western Reserve University, 2104 Adelbert Road, Bingham 248, Cleveland, OH, US 44106-7201, xxf121@case.edu.

[2]Graduate Research Assistant, Department of Civil and Environmental Engineering, Case Western Reserve University, 2104 Adelbert Road, Bingham 249C, Cleveland, OH, US 44106-7201, xxz677@case.edu.

[3]*Professor, Department of Civil and Environmental Engineering, Case Western Reserve University, 2104 Adelbert Road, Bingham 206, Cleveland, OH US 44106-7201, xxy21@case.edu, Corresponding Author

Abstract

The water supply network (WSN) is subjected to leaks that compromise its service to the communities, which, however, is challenging to identify with conventional approaches before the consequences surface. This study developed Machine Learning (ML) models to detect leaks in the WDN. Water pressure data under leaking versus non-leaking conditions were generated with holistic WSN simulation code EPANET considering factors such as the fluctuating user demands, data noise, and the extent of leaks, etc. The results indicate that Artificial Neural Network (ANN), a supervised ML model, can accurately classify leaking versus non-leaking conditions; it, however, requires balanced dataset under both leaking and non-leaking conditions, which is difficult for a real WSN that mostly operate under normal service condition. Autoencoder neural network (AE), an unsupervised ML model, is further developed to detect leak with unbalanced data. The results show AE ML model achieved high accuracy when leaks occur in pipes inside the sensor monitoring area, while the accuracy is compromised otherwise. This observation will provide guidelines to deploy monitoring sensors to cover the desired monitoring area.

A novel strategy is proposed based on multiple independent detection attempts to further increase the reliability of leak detection by the AE and is found to significantly reduce the probability of false alarm. The trained AE model and leak detection strategy is further tested on a testbed WSN and achieved promising results. The ML model and leak detection strategy can be readily deployed for in-service WSNs using data obtained with internet-of-things (IoTs) technologies such as smart meters.

Keywords: Water supply network, Artificial intelligence, Machine learning, Artificial Neural Network, Autoencoder Neural Network, Leak detection

1. Introduction

Water supply system provides one of the most essential service for the communities. However, due to the deterioration of the underground water pipe network, a large amount of water is lost every year, mostly unnoticed. According to (Sadeghioon, Metje et al. 2014), about 3281 megaliters (10^6) of water was wasted in the UK during 2009-2011, and about 15% of supplied drinking water was wasted each year in the US. The percentage of lost water is significantly higher in historical water districts such as Cleveland, Ohio, Boston, MA, etc. Many factors can cause leaks, such as pipe corrosion, aging, defects and inappropriate installation (Kiliç 2016). A detailed discussion is presented about the cause of water main failures by Sadia et.al (Sadiq, Rajani et al. 2004). Due to the complex underground environment, predicting underground water pipe failure remains a challenging problem. The state of practice with most agencies is to rehabilitate pipes after leaks are directly observed (Walski and Male 2000), while many small leaks remain undetected until the damages surfaced in the form of ground cavitation etc. Evidently, the agency perspective on cost does not include the socio-economic cost to the communities such as healthcare cost due to compromised water quality. Technologies to detect leakage and forecast water pipe failure will enable agencies to evolve into preventative strategies with significant socio-economic benefits.

A significant amount of efforts has been made on water pipe leak detection. Strategies can be broadly classified in five categories, i.e., vision based, sensor/instrumental based, transient response based, model based, and data-based (Chan,

Chin et al. 2018). The first two technologies require to use specialized mobile inspection equipment with optical, acoustic, or electromagnetic sensors (Gao, Shi et al. 2006; Ozevin and Yalcinkaya 2012; Kang, Park et al. 2017), which is expensive and time-consuming. For example, leak detection with acoustic signals can often be influenced by the type of leak, opening size, pipe materials and soil conditions (Butler 2000). Technology such as ground penetrating radar can detect leak around pipe but requires heavy human involvement in signal analyses (Bimpas, Amditis et al. 2010; Amran, Ismail et al. 2017; De Coster, Medina et al. 2019). The pressure or acoustic transient signals are used for pipe burst detection, since such transient signals travel along the pipe at the speed of sound starting from the burst location (Srirangarajan, Allen et al. 2013). However, the transient responses decay with distance and diminish over a short time, and therefore requires sensor with high spatial and temporal resolution. Model-based approach for leak detection has been theoretically shown to be capable of identifying leakage and localize its position. They are, however, very difficult to be implemented in real systems (Colombo, Lee et al. 2009; Mashford, De Silva et al. 2012; Adedeji, Hamam et al. 2017) due to its requirements on detailed information required for a hydraulic model such as the user demand, pipe condition, water pressure distribution, etc. Such information is difficult to collect or is typically not available. Empowered with the Internet of Things (IoT) and artificial intelligence (AI), data-driven technologies have been proven capability knowledge discovery (Liao, Chu et al. 2012), image processing (Buch, Velastin et al. 2011), and event forecasting, etc. (Lin, Hu et al. 2011). Data-driven leak detection, which is based on learning from historical data with statistical or pattern recognition algorithms, is emerging (Romano, Kapelan et al. 2012). It does not require collecting comprehensive set of information as needed for a model based approach.

With the development of supervisory control and data acquisition (SCADA) systems, real-time monitoring data of water pressure and/or flow rate are available and can be collected for the leak detection and localization (Stoianov, Nachman et al. 2007; Kim, Sharma et al. 2010; Chim, Yiu et al. 2011). Other data such as monitored acoustic sensor data was found significantly affected by the environmental noise and limited transmission distance (Loth, Morris et al. 2004; Srirangarajan, Iqbal et al. 2010). The monitored water pressure data of districted metering areas (DMA) can be trained with a state of art ML

model to detect possible leak by used of the historical data, which is combined with traditional methods such as vision-based or instrument-based inspection to pinpoint leak location. Different ML algorithms have been developed for leaks detection, for example the ANNs and a fully-line DensNet (Mounce, Day et al. 2002; Mounce, Boxall et al. 2010; Zhou, Tang et al. 2019). The leaks or bursts were also detected through comparing the predicted water demand or the water pressure at nodes versus the actual demand or pressure (Ye and Fenner 2011; Bakker, Vreeburg et al. 2013).

The existing ML algorithms for leak detection generally include classification model, prediction-classification model, and statistical model (Wu and Liu 2017), each featuring advantages and limitations. The classification model uses supervised learning model and requires large datasets under both normal conditions and leaking conditions. In the practice, data under leaking conditions is generally less commonly available. The prediction-classification method belongs to unsupervised ML method and can be trained just with data under normal service conditions. Machine learning algorithms based on artificial neural network and autoencoder neural network were developed and evaluated based on simulation data and achieved reasonable accuracy (Tao, Huang et al. 2013; Pal and Kant 2019; Zhou, Tang et al. 2019). The accuracy of statistical ML model is dependent upon the uncertainty levels.

To address this important issue, this paper explored data driven machine learning (ML) model for leak monitoring of a desired water area in the WSN. The performance of a data-driven approach is highly dependent upon the availability of historical data. Since data in an actual water distribution network is not widely available at this time, simulated data with industry certified hydraulic model EPANET is used to generate data used in this study. The Artificial Neural Network (ANN), a supervised ML algorithm, and Autoencoder neural network (AE), an unsupervised ML models, were developed to detect leak in a water supply network from sensor data serving a District Meter Area (DMA). Both models, which required balanced or unbalanced datasets respectively, were found to achieve satisfactory results. Strategy to improve detection accuracy is further developed by multiple independent attempts of detection. The article is organized as following: an introduction of the background and methodologies including the theoretical basis of the hydraulic model for water pipe network, background on ML models including artificial

intelligence neural network (ANN) and autoencoder neural network (AE), and data generation scheme under both normal and leaking conditions. This is followed by the case studies of two hypothetic water distribution network simulated by EPANET. The performance of ANN, AE and a postprocessing framework are then described and analyzed by using the first water distribution network. The performance of AE model is future validated on a larger network with complex water usage scenario, i.e. C-Town water distribution network. Finally, the conclusions are provided to summarize the major discoveries of this paper.

2. Theoretical Background

2.1 Hydraulic model for water pipe network

A hydraulic model is commonly used to compute the hydraulic parameters such as water pressure or water head and flow rate for the design of a water distribution network. The governing hydraulic equations describe the conservation of mass and conservation of energy considering the topological characteristics of a water pipe network. The hydraulic model allows to account for the water usage behaviors (described as water demand fluctuations at the service nodes) and events such as leakages on the network performance. While hydraulic model is regarded as sufficiently accurate for water network planning purpose (Wu and Liu 2017), there are uncertainties of the model prediction results due to fluctuating water demands, deteriorating pipe conditions, etc. A calibrated hydraulic model serves as the basis for model-based leak detection. Given it is sufficiently reliable, hydraulic model can be utilized to generate holistic artificial datasets for the development and validation of ML-based leak detection algorithms. As a general note, using holistic artificially generated data is a common strategy in the development of ML technologies when data is not available due to practice constraint. The key equations used for the hydraulic computations are introduced in following.

Equation (1) of the hydraulic model describes the conversation of mass at a pipe node, which prescribes that under no leak condition the inflow of water to a pipe node must be equal to the outflow of water. The outflow of the water including the demand or use of water at that node as well as water flowing from this node to other nodes.

$$\sum_{p \in P_n} q_{p,n} - D_n^{act} = 0 \qquad n \in N \tag{1}$$

where P_n is the set of pipes connected to the node n, $q_{p,n}$ is the flow rate of water into node n from pipe p (m^3/s), D_n^{act} is the actual water demand at node n (m^3/s), and N is the set of all nodes in the pipe network. $q_{p,n}$ is positive when water is flowing into node n from pipe p, otherwise, it is negative.

Equation (2) of the hydraulic model describes the conversation of energy. For water pipe network, the total energy is typically referred as the total water head, which includes components describing the kinetic energy (kinetic water head), hydraulic potential energy (pressure head), and gravitational potential energy (elevation head), i.e.,

$$h_A = \frac{u_A^2}{2g} + \frac{p_A}{\gamma_w} + z_A = h_B + H_L = \frac{u_B^2}{2g} + \frac{p_B}{\gamma_w} + z_B + H_L \tag{2}$$

where h is the total water head, u is the water velocity at each node, and z is the altitude of each node. H_L is the energy loss between node A and node B.

There are two major mechanisms for the energy lose in a pipe flow (Twort, Ratnayaka et al. 2000), i.e., the distributed energy loss and localized energy loss. The distributed energy loss along the pipe due to hydraulic resistance is mainly determined by the velocity of the flow V, the internal diameter of the pipe d, the length of the pipe L, and the roughness of the pipe wall, which is described by the Hazen-Williams formula (Liou 1998), i.e., Equation (3).

$$H(m) = \left(\frac{6.78L}{d^{1.165}} \right) (V/C)^{1.85} \tag{3}$$

where C is the roughness coefficient of pipe wall.

The localized energy loss is due to turbulence associated with change of flow conditions (such as flow speed, direction, or flow area etc.), which is determined by the topology of water distribution network connections.

An important phenomenon in a water supply network is the water usage or demand. Two major types of models are generally used for water demand at pipe nodes, i.e., demand-driven model and pressure-driven model. A comparison of both models is

described in (Braun, Piller et al. 2017). A pressure-driven water demand model is used in this study to consider the effects of losing pressure due to change of water demand or leaks.

$$D = \begin{cases} 0 & p \le P_0 \\ D_f \left(\dfrac{p \quad P_0}{P_f \quad P_0} \right)^{\frac{1}{2}} & P_0 \le p \le P_f \\ D_f & p \ge P_f \end{cases} \tag{4}$$

where D is the demand at a particular node, D_f is the desired demand(m^3/s), p is the water pressure, P_f is the pressure above which the desired demand D_f should be met, P_0 is the water pressure below which no water will be supplied at the node.

The leaking is modeled as a special type of water demand in this study. The demand due to a leaking scenario is related to the size of the leak and is described in Equation (5) (Crowl and Louvar 2001).

$$d_{leak} = C_d A p \sqrt{\frac{2}{}} \tag{5}$$

where d_{leak} is the equivalent water demand due to leak (m^3/s), C_d is the discharge coefficient, with a default value 0.75, A is the area of leak, p is the internal water pressure, the exponential is the discharge coefficient, which is 0.5 for steel pipe, and is the water density.

The model is implemented in EPANET, a certified hydraulic model for water supply network (WSN).

2.2 Artificial Neural Networks (ANN)

Artificial Neural Network (ANN) is a supervised machine learning (ML) model. The architecture of ANN includes interconnected neurons in the input layers, hidden layers, and output layer. The number of layers and the number of neuron in each layer determine its overall performance (Abokifa, Haddad et al. 2018). Increasing the number of neurons and hidden layers can improve the ability of ANN model to describe complex nonlinear relationships. It, however, also increase the computational demand and potentially lead to overfitting. An optimal ANN architecture for this study is determined by an optimization

process, which leads to an ANN model with one input layer, three hidden layers, and one output layer, as shown in Fig 1.

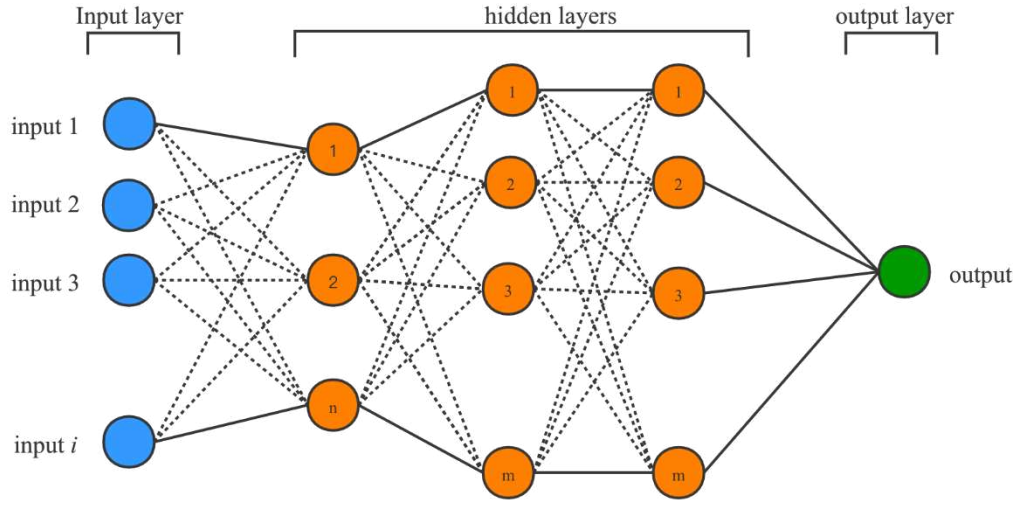


Fig 1. Schematic of ANN architecture

The input layer consists of ***I*** neurons, which are corresponding to the number of input features. The hidden layers provides the capability to model the complex non-linear relationships which are fine tuned with the training data. The output layer consists of one neuron which is used to classify the output as leaking or not leaking condition.

The hidden layers includes fully connected neurons, the output of each neuron is written as follows.

$$y_k = f(\sum_{r=1}^{I} x_{r,k} \quad _{r,k} + b) \tag{6}$$

where y_k is the output of each neuron at the hidden layer, $x_{r,k}$is the output of the last layer, for the first layer of neural network, $x_{r,k}$is the sample data $_i$. $_{r,k}$ is the weight of that neuron and b is the bias of that neuron. The weight and bias are trained with the training datasets by the back-propagation algorithm. $f(.)$ is the activation function used to increase the nonlinear property during the propagation. In this study, the 'ReLu' function is used as the activation function of the hidden layer, i.e, Equation (7).

$$f(x) = \max(0, x) \tag{7}$$

The output of the last hidden layer is then transferred into the neurons in the output layer, whose actions is written as below.

$$y_z = g(y_k \quad + b) \tag{8}$$

where and y_z is the output of the output layer, y_k is the output of the neurons in the last hidden layer, and b are the weight and bias. $g(.)$ is the tangent sigmoid transfer function defined as

$$g(x) = \frac{2}{1+e^{2x}} \quad 1 \tag{9}$$

The ANN model learns the relationship between the output and input by a training process to classify the observed data into leaking and non-leaking situations. More detailed mathematical information about ANN can be found at (Jain, Mao et al. 1996).

2.3 Autoencoder neural (AE) networks

Autoencoder neural network is an unsupervised ML model. It is based on a special type of neural network that is trained to reconstruct its input, so the output $(y_1, y_2, y_3, ..., y_n)$ would contain the same information as its input $(x_1, x_2, x_3, ..., x_n)$. To reduce the reconstruction error, the network is forced to learn the hidden patterns between the input data. An innovative strategy is proposed in this study to detect the leaking situation by autoencoder neural network based on its reconstruction error. The reconstruction error is characterized by the mean square error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{10}$$

where MSE is the mean square error or reconstruction error, n is the dimension of the input vector x, x_i is the sample data and y_i is the predicted sample data.

A typical architecture of the autoencoder neural (AE) network is shown in Fig 2. The training process of the AE network involves firstly compresses the input vector $\boldsymbol{x}$ into a small dimension, which is called the encoding process. Then the model will reconstruct the compressed data into its original space, which is called the decoding process. By reducing the error between output and input, the weights and bias of the neurons in the neural network are adjusted to learn the relationship among the input data.

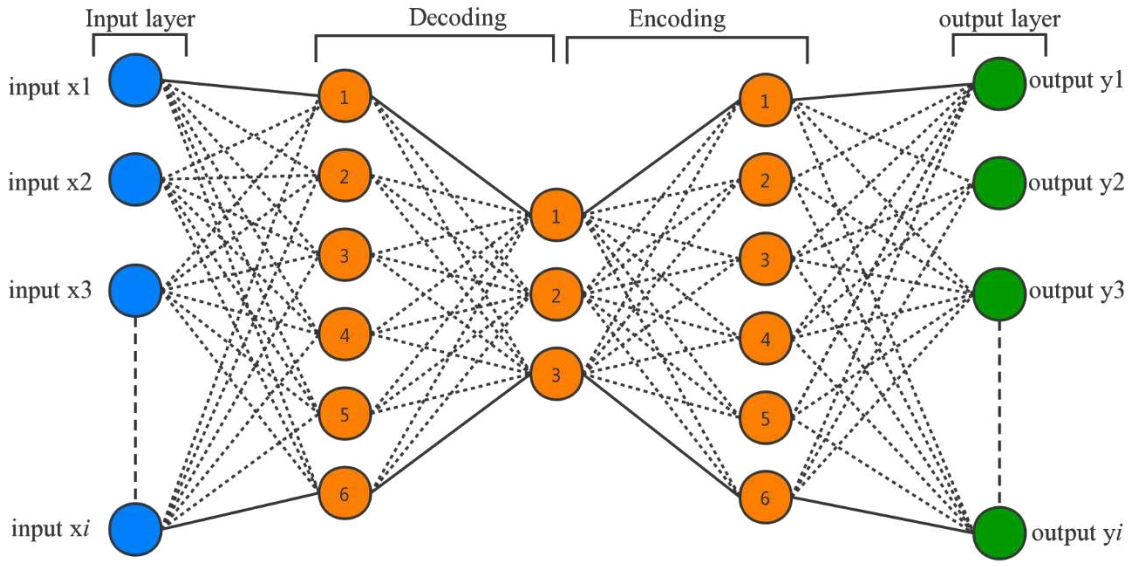


Fig 2. Schematic of architecture of an autoencoder neural network (the numbers of neurons in the decoding layers and encoding layers are conceptual)

For a model trained by dataset of only non-leak condition, a large reconstruction error occurs it is inputted with data of leaking condition because the relationship described by the trained AE neural network is not valid under such condition. By setting a threshold in the construction error, the AE model can classify if a set of data corresponds to a leaking situation or a non-leaking situation.

2.4 Generation of monitoring data in the water supply network under normal and leaking conditions

Machine learning (ML) model requires a sufficient amount of data for training and validation. When short of real-world data, it is a common approach to develop and evaluate ML algorithms based on simulated data (Prasad, Park et al. 2004; Housh and Ohar 2018; Taormina, Galelli et al. 2018). A python package Water Network Tool for Resilience (WNTR) is used to build the water supply network and solve the hydraulic equations (which is shown in section 2) for water flow in the pipeline system (Klise, Murray et al. 2018). WNTR is an open-source python package for hydraulic simulations of water pipe system based on EPANET. This package was adopted to run iterative simulations with a

combination of a set of random parameters that describes the fluctuation in water demand, data noise, and leaking conditions.

With the WNRT package, WSN operation data (i.e., pressure, flow rates, etc.) with fluctuating water demands, leaking or non-leaking conditions can be generated. The randomness of the WSN model, including water usage fluctuations, pipe conditions and noise, are considered in data generation. *The water usage* here is an overall estimated during a period. For example, the baseline water usage at each node was chosen from a uniform distribution in the range of 0.008 to 0.012 *L/s* according to Funk et.al (Funk and DeOreo 2011). According to (Pal and Kant 2019), the real time water usage may fluctuate from 0.3 times to 1.3 times that of the baseline usage depending on the different time in a day. *The pipe conditions* are described with the dimensionless roughness coefficient. The roughness coefficient of each pipe was selected from a uniform distribution from 100 to 300. *Different levels of noise* were also added to account for the uncertainty of WSN such as the water usage or sensor error at a certain time interval and is described with a Gaussian noise $N(0, \sigma)$. Similar data generation framework was used by Zhou et.al (Zhou, Tang et al. 2019) for a different purpose. The detailed data generation process is introduced in Case study I.

3. Challenges in leak detection and strategy to overcome the challenges

There are two primarily challenges in implementing a data-driven model for leak detection of WSN:

1) *Unbalanced data*. Many existing leak detection methods, Support Vector Machine (SVM), Convolutional Neural Network –Support Vector Machine (CNN-SVM), ANN, etc., treat leak detection as a classification problem (Mounce, Day et al. 2002; Caputo and Pelagagge 2003; Kang, Park et al. 2017). These methods require a sufficient large leaking dataset as well as non-leaking datasets. However, as pointed out by Mounce et.al (Mounce, Mounce et al. 2010), there are relatively limited leaking situations compared with non-leaking situation. Let alone to require data of leaking occurring at each pipe of a water distribution system. This leads to overly unbalanced datasets, i.e., much more non-leaking data than leaking data. A major challenge of a data driven leak detection approach is how

to use little or no data collected under leaking condition to train ML model for leak detection.

2) *Uncertainty of user demands*. The water pressure pattern in the service water pipe network may be unstable due to the fluctuation of water use behaviors (Wu and Liu 2017). The water use in the water supply network is strongly affected by the user behaviors and can show strong fluctuations. The prediction-classification method has been developed for this purpose (Wu and Liu 2017). For example, methods by (Abokifa, Haddad et al. 2018; Chan, Chin et al. 2018) predicted the water pressure at the next time step by different methods. Baker et.al used an adaptive forecasting model to predict the short term water demand of a DMA(Bakker, Vreeburg et al. 2013). Leaking alert is triggered if the difference between the actual water pressure and predicted water pressure exceeded a threshold. Water flow data at midnight when user water demand is small has been used to identify the leaks, under the assumption that baseline flow at midnight is most likely due to leaks (Mazzolani, Berardi et al. 2017) when fluctuation in user demand is small. However, most of these detection methods require a consistent or predictable trend of water demand, otherwise it will trigger false alarms. The spatial relationship of multiple nodes in the water distribution network can be used to mitigate false alarms in leak detection. For example, Zhou et.al (Zhou, Tang et al. 2019) used a full linear DenseNet neural network with the spatial information of multiple sensors in a water distribution network for leak detection. However, such spatial information was only used in supervised learning which requires sufficient amount of data under leaking conditions and therefore present challenges as stated in item 1).

To address these two challenges, a novel leak detection method is proposed in this paper based on both the spatial and temporal information. The spatial pattern among a group of nodes is used in leak detection and identify leak conditions. The temporal information is used to further improve the detection accuracy. The advantage of the new leak detection model is that it can be trained with data under non-leaking conditions only.

4. Case studies of ML models for leak detection in water supply network

Two water supply networks (WSN) were analyzed to illustrate the proposed data generation framework and develop new leaking detection methods. The first network is a

relatively small WSN that has been widely used as a standard testbed, which was chosen to illustrate the data generation process, development and validation of proposed leak detection methods. The second network is a large size WSN containing 5 district meter areas (DMA), multiple water sources (7 tanks and 1 reservoir) and 6 different control rules (such as valve controls and pump controls). The second WSN network was used to demonstrate the performance of the developed leak detection method under more complex conditions.

4.1 Case Study – I: Rancho Solano Zone III Water Distribution System

Rancho Solano Water Network located in the city of Fairfield, California. This network is published by ASCE task committee on a research database for water distribution systems (Hernadez, Hoagland et al. 2016). The graph of this water supply network is shown in Fig 3. There are 112 nodes in total, including one reservoir as the source of water, and 126 pipes. The elevations of the nodes in this pipe network range from 90 meters to 120 meters and the length of the pipes range from 90 meters to 130 meters.

The water distribution network and basic water demand at each service node are shown in Fig. 3. The basic demand of each node is chosen from a uniform distribution of 0.008 to 0.012 *L/s*. The real demand at each node is generated by adding a random Gaussian distribution with variance $\sigma = 0.01L/s$. 11 demand ratios from 0.3 to 1.3 are considered during the data generation with the hydraulic model for the WSN. The monitoring sensors are assumed to be deployed in the area shown in the red circle area, i.e., the water pressure data of the nodes which are located in the red circle in Fig. 3 is used for leak detection. Fig 3 also shows some key nodes and pipes that are analyzed in the study.

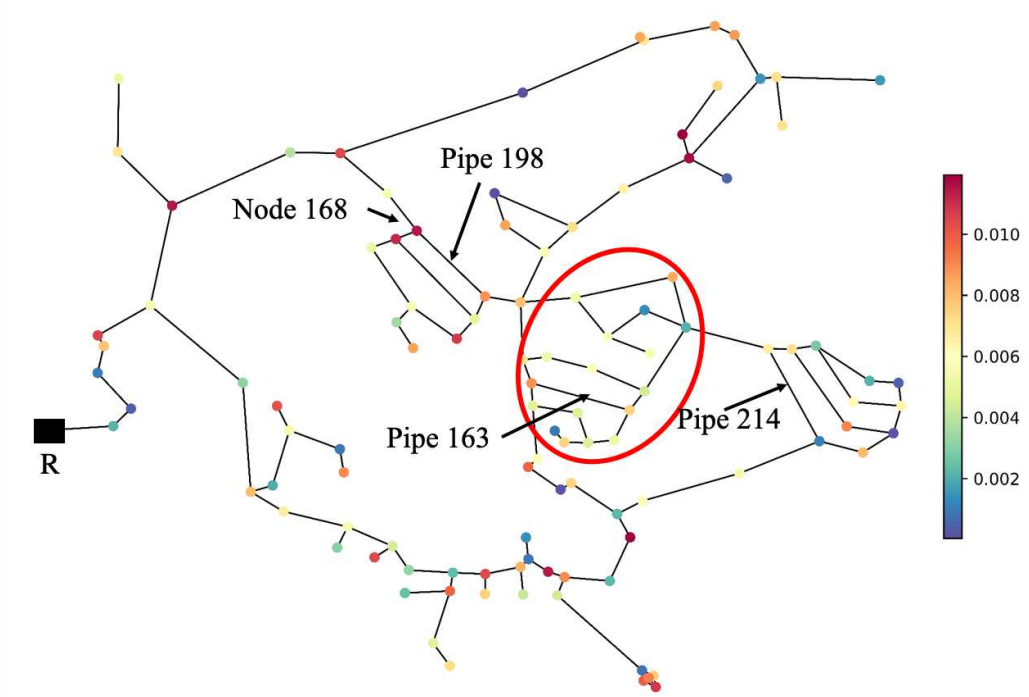


Fig. 3 Illustration of the water supply network and water demand at each node (color code corresponds to basic water demand in the unit of *L/s*)

The pressure-driven demand model, which relates the water discharge to the water pressure head at the node (i.e., Eq. 4), is used in the hydraulic analysis of the WSN. The lower bound of the pressure head at the node is set as 5 meters and the upper bound as 30 meters. An example of the relationship between demand/discharge and pressures head at a node with base demand of 0.02 m^3/s is shown in Fig 4.

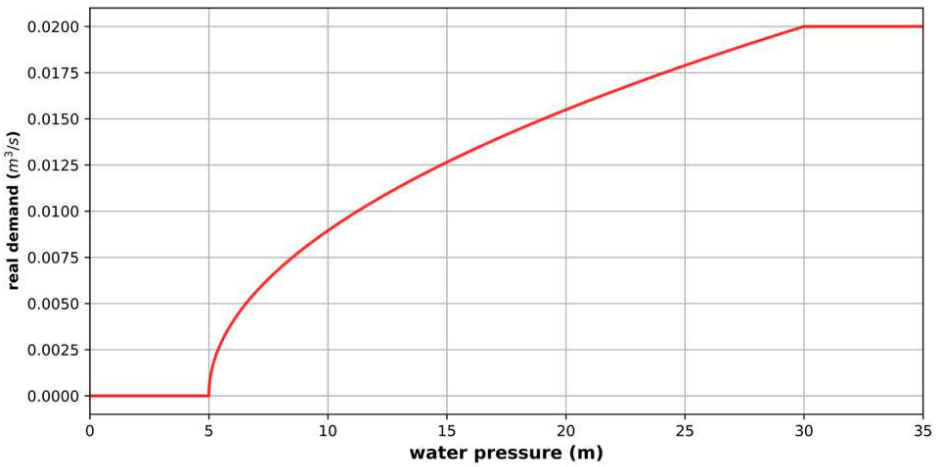


Fig. 4 Example of the relationship between water demand and pressure head at a node with base demand of 0.02 m^3/s

The overall data generation procedures of balanced dataset with the hydraulic model for the WSN are briefly summarized below. Similar amount of data are produced for the monitoring area in the WSN under both leaking and non-leaking conditions.

Dataset for Non-Leaking conditions:

1. *Define Water Pipe Network*: Construct the pipe network according to Rancho Solano Water Network (Fig. 3) using EPANET data input format.
2. *Assign Water Pipe Conditions*: Assign the pipe roughness of pipe with a random number from uniform distribution $U(100, 300)$. The length of each pipe is already defined in the original water pipe network
3. *Set the Baseline Demand and Actual Demand at User Nodes*: the baseline demand of each node is randomly selected from a uniform distribution $U(0.008, 0.012) L/s$, following Funk et al (Funk and DeOreo 2011). The actual demand at each node is set by considering both the base demand and the demand uncertainty, i.e.,

   $$Demand = deman\,ratio * D_{base} + |N(0,\ \sigma)|. \quad (11)$$

   where D_{base} is the predefined base demand. The $demand\ ratio$ is set from 0.3 to 1.3 to account for the fluctuation in water usage demand during a day or between different days. Gaussian noise $N(0, \sigma)$ considers the uncertainty due to the water usage fluctuation.
4. *Data Generation*: solve the hydraulic model of the WSN with WNRT using the EPANET built-in module and record the water pressure at selected monitoring nodes (i.e, the nodes inside the red circle in Figure 3).
5. *Data generation for different water demand situations*: step 2 to step 4 are repeated for each water demand scenery. 200 rounds of simulations were conducted for each scenery to generate sufficient amount of data under different water demand situations,

Dataset for Leaking Conditions:

1. Similar procedures as for non-leaking conditions are followed to build the water network, assign pipe roughness and water use demands (Step 1-3 for non-leaking conditions).

2. *Leak Scenario*: Set pipe i as the leaking pipe. By default, the leaking position is located at the middle of the pipe, which, however, can be easily changed for more complex scenario.
3. *Data generation*: solve the hydraulic simulator with WNRT using the EPANET built-in module and record the water pressure at selected monitoring nodes (inside the red circle in Figure 3).
4. *Data generation for different water demand situations*: Repeat steps 2-4 200 times for each pipe at demanding level similar as what is done for non-leak conditions.
5. Repeat the above step for each pipe leaking scenario.

The water pressure data under different scenarios were generated via the processes described. The non-leaking situation and leaking situation at each pipe contain 2200 cases respectively (11 different water demand levels with 200 rounds of simulations). It is noted that the model-generated data can be easily replaced with real-world data when measurement data is available.

The code for data generation is published in this link for the sake of open source[1]. Overall, the water pressure is affected by the average demand at the node, fluctuations in water demand, and if leak occurs. To illustrate the characteristics of water pressure data, the water pressure of node '168' under a few demand ratios are shown in Fig 5. As can be seen, the water pressure can be highly influenced by the demand levels. There are significant differences in the water pressure between demand ratios of 0.3 versus 1.2. A higher water pressure corresponds to a lower water demand. It should be noted that water pressure is also affected by leak. For example, Figure 6 shows the water pressure at node '168' is similar for intact water pipe at high average water demand ratio of 1.0 versus leaking pipe (pipe 198) with low average water demand ratio of 0.3. Such overlap in the influence on water pressure by water demand and leaking makes it difficult to detect leak from data from a single node. Machine Learning model, however, allows to extract features from the spatial pattern in the pressure data at multiple nodes and therefore allows to differentiate leaking versus non-leaking conditions.

[1] https://github.com/herewego321/Random-WDN-data-generation

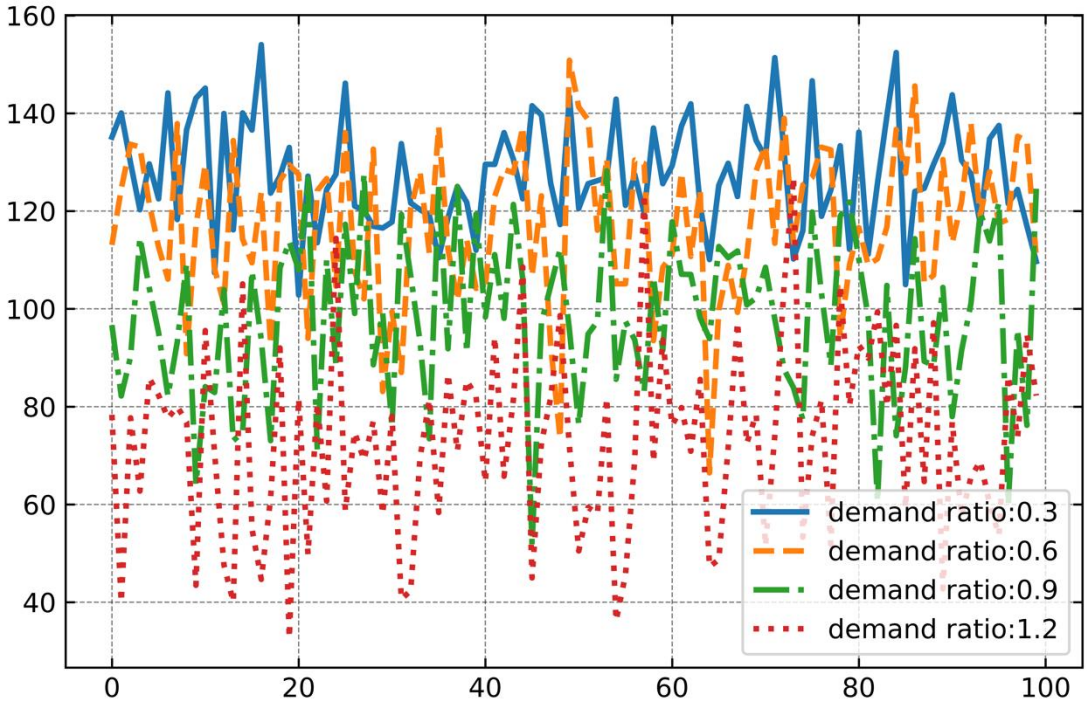


Fig. 5 Example of water head at node '168' under different water demands (note: demand ratio is defined as the average water demand to the predefined baseline demand at the node (Eq. 11); the fluctuation is due to fluctuations in the real demand).

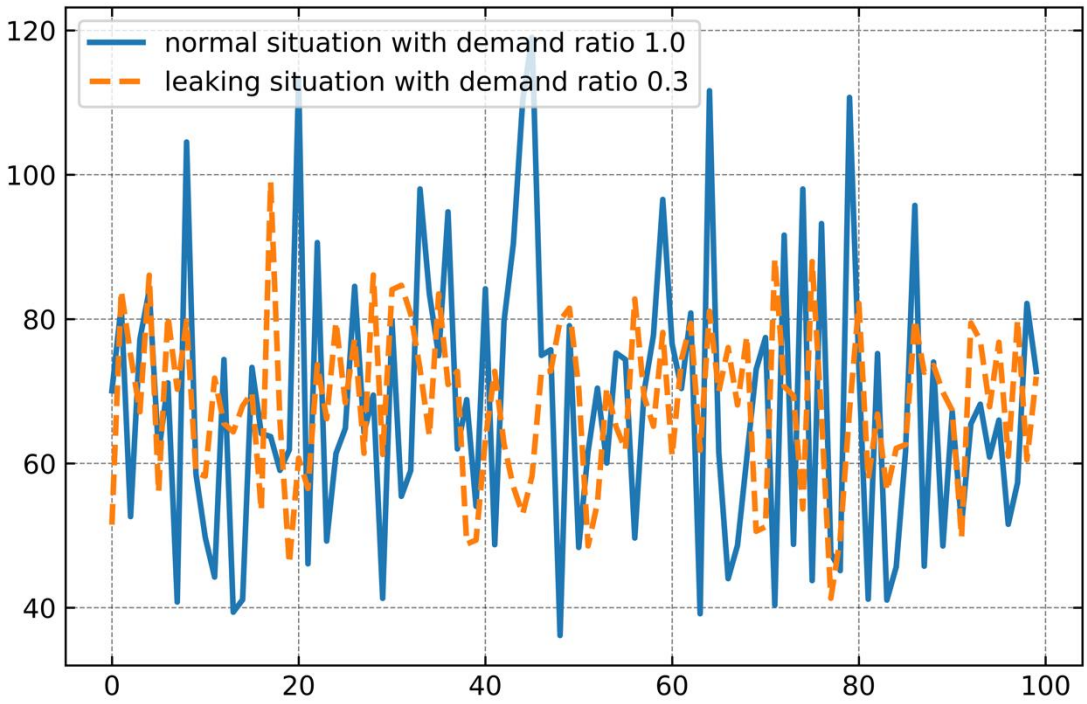


Fig. 6 Illustration of water head at node '168': the average water head is similar for non-leaking condition with high water demand versus leaking condition with low water demand

Artificial Neural Network (ANN) model for leak detection

An ANN model is developed to detect leaks. The water pressure data at a group of monitoring nodes is used for this purpose. Unlike the existing approach of using time series analyses, the water pressure data are used by ANN model to find the spatial relationship among data at monitoring nodes at a given time. The ANN model was built and trained with TensorFlow in python environment. An optimal ANN architecture for this study is determined by an optimization process, which leads to an ANN model with one input layer, three hidden layers, and one output layer. The input layer contains 11 neurons corresponding to the 11 monitored nodes. The first hidden layer contains 128 neurons and then the remaining two hidden layers contain 258 neurons respectively. The output layer contains 1 neuron, which is a categorical data indicating leaking versus non-leaking condition.

Overall, the ANN model is used to classify leaking and non-leaking conditions of the monitoring area as a binary classification problem. As a supervised learning model, ANN requires the dataset to be labeled prior to training. Datasets of no leaking situation as well as data corresponding to leak in one pipe located inside the monitoring area are labeled. The dataset includes 1200 sets of data from non-leaking conditions and 1200 dataset under leaking conditions. These samples were randomly chosen from the 2200 dataset generated under non-leaking and leaking conditions happened inside the monitoring area. Standardization of the dataset is conducted to reduce the computing time and avoid potential overfitting. Each row of the dataset is transformed to a normal distribution with zero mean and unit variance. The benefits of data standardization is described by (Shanker, Hu et al. 1996). The 2400 labelled dataset is then randomly split into independent training data and testing data with a ratio of 7:3 (i.e., 1680 set of training data and 720 set of test data). The training dataset is used to train the ANN model. The independent testing data is used to validate the model results. The loss value of training and validation processes are shown in Fig 7. The loss value is the mean square error of predicted result versus

actual result. As can be seen, both loss values of training dataset and validation dataset decrease to small values during the learning process, which means the ANN model is able to uncover the relationship among data for classification of leak versus non-leak conditions.

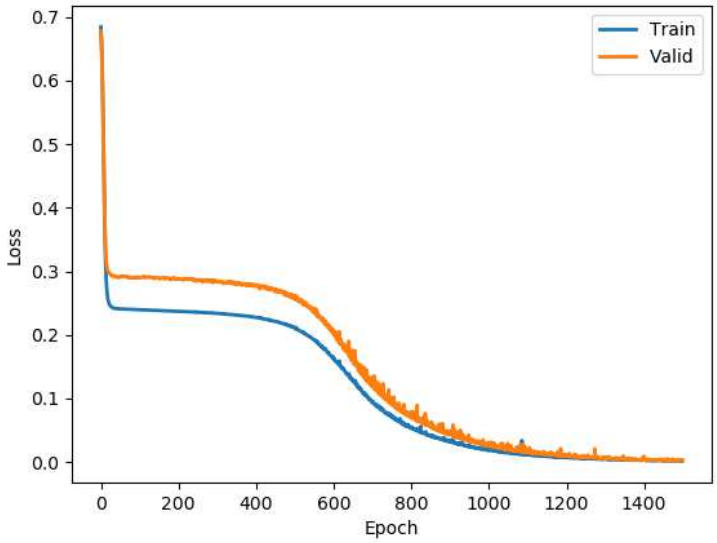


Fig. 7 Loss values during the ANN model training process

The final classification result by the trained ANN model using the testing data is shown in Fig 8 as described with the confusion matrix. There were 370 non-leaking cases and 350 leaking cases in the testing dataset, both are classified with 100% accuracy and with no misclassification. The results imply that a) the relationship of water pressure among a group of nodes is different under leaking and non-leaking scenarios; and b) the ANN model is trained to extract this relationship and to accurately classify leaking versus non-leaking conditions.

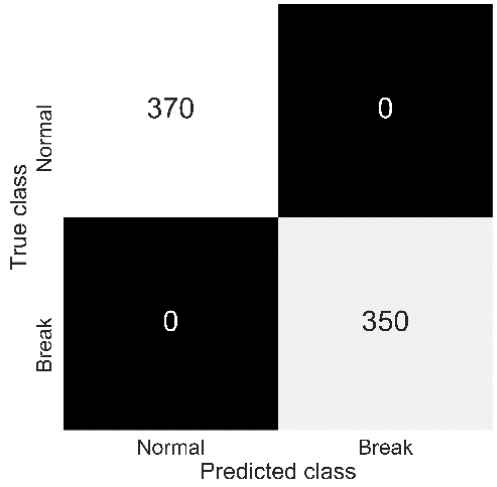


Fig 8 The confusion matrix of the classification result of leaking and non-leaking cases by the trained ANN model (achieved 100% accurate with no misclassification)

Autoencoder Neural (AE) Network Model for Leak Detection

The ANN model achieved excellent performance by utilizing the water pressure data at multiple nodes. However, as a supervised ML model, ANN model requires balanced data, i.e., similar amount of data under both normal and leaking conditions. However, in the reality, the available data is typically unbalanced. i.e., there might be only limited amount of data under leaking conditions compared with data under non-leaking conditions. Besides, labeling the dataset to leak or non-leak conditions, i.e., such as the method used by (Zhou, Tang et al. 2019) may be extremely difficult under real situation since the leaks might not be detected until their effects surface.

A variation of ANN model, the autoencoder neural (AE) network, is developed for leak detection to resolve the challenge of unbalanced data. As an unsupervised ML model, the AE model features unique advantages to work with unbalanced data. In this study, a AE model with 5 layers was built. As shown in Fig 2, the first and last layer contains 11 neurons which are corresponding to the 11 monitored nodes. The second and third layer encodes the input data from 11 nodes to a lower-dimensional space, while the fourth and fifth layer decodes the data from this lower-dimensional space back to 11 nodes. The hidden layer of the AE model contains 3 neurons. The training dataset for the AE model includes 1200 datasets randomly chosen from the 2200 generated dataset under the normal situation. Each sample in the dataset includes the water pressure information at the 11 selected monitoring nodes. 70% of the 1200 normal non-leaking dataset is standardized and used for training. The rest 30% of the normal non-leaking dataset is used for validation.

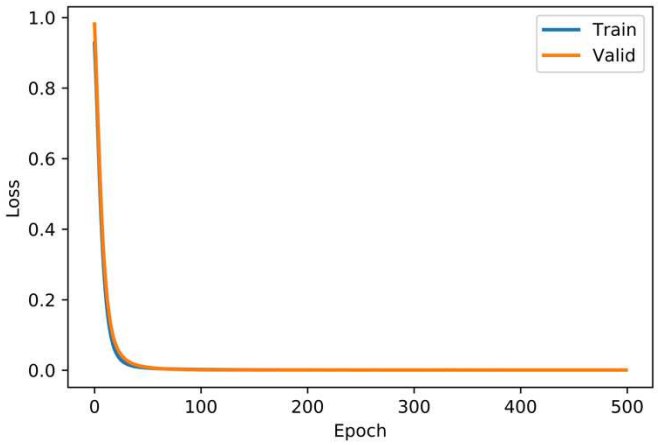


Fig. 9 Loss values (i.e., the reconstruction error) by the AE model during the training process

Figure 9 shows the loss values, defined as the reconstruction error for AE model, during the training process. Small loss values of close to zero are achieved for both the training and testing data, meaning the reconstructed output from the AE model is close to its input of dataset under normal non-leaking conditions. This also implies that the AE model is well trained with the normal non-leaking dataset. When leaking dataset is input to the trained AE model, the model will generate a large reconstruction error, which can be used to detect leaking conditions.

Independent datasets are used to evaluate the performance of the trained AE model to detect leaking conditions. The dataset includes three different scenarios, i.e. 550 datasets from the non-leaking conditions, i.e. there is no leaks anywhere in the water distribution system; 550 cases of dataset where a leak happens at a random pipe inside the monitoring area (pipe '163'); 550 cases of datasets where the leak happens at a random pipe outside the monitoring area ('pipe '198'). The location of example pipes can be found in Figure 3. All of the data samples are normalized based on the mean and variance values of non-leaking dataset (by minimizing the mean value and divided by the variance value). Figure 10 shows the statistics histogram of the reconstruction errors by the AE model for data under the three different scenarios. The rectangles with different color lines indicate the 97.5% range of reconstruction error for normal and two leaking situations. As can be seen in Fig 10, the reconstruction error of data under normal non-leaking situation is small, with 97.5% of reconstruction error less than 0.00015, which is much smaller than those under the other two situations. The reconstruction error of data when leak occurs inside

the monitoring area features largest reconstruction errors. While the reconstruction errors of data for leaking outside the monitoring area lies in between. Overall, dataset corresponding to pipe leaking within the monitoring area leads to large reconstruction error by the trained AE neural network model. The differences in the reconstruction error are clearly differentiable from those by the normal non-leaking cases. This can be used to define the threshold for leak detection. Leaks occurring outside the monitoring areas, however, still has a low probability to be identified as leaking situation. It should be noted that since a certain area is monitored by these sensors, the ideal result is only the leaking inside or very closed to this area can be detected. Leaking faraway from this monitoring area should not be able to trigger the detection. Method to mitigate detection error due to the influence of leaking outside the monitoring areas will be discussed in next sections.

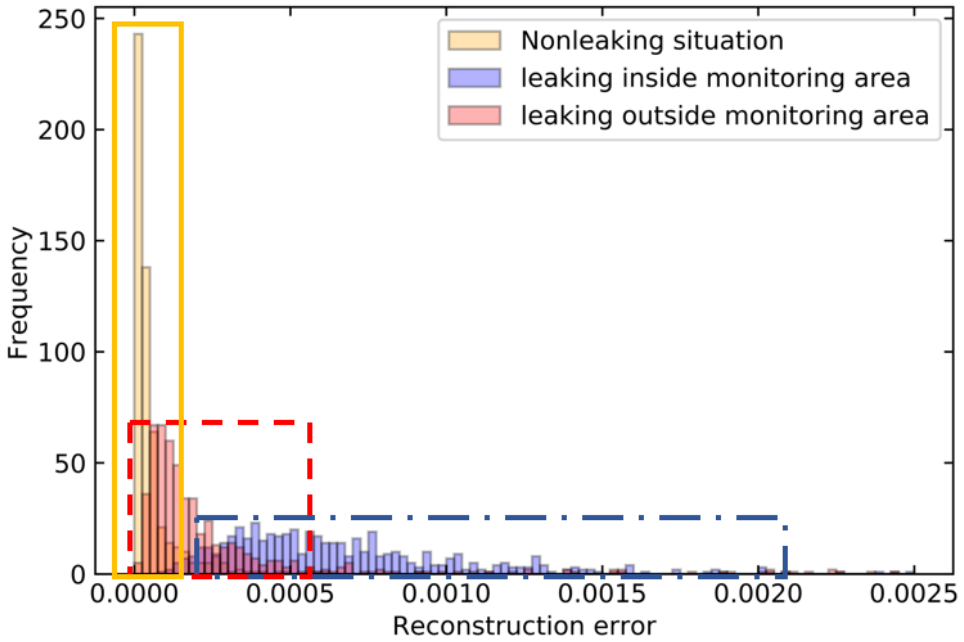
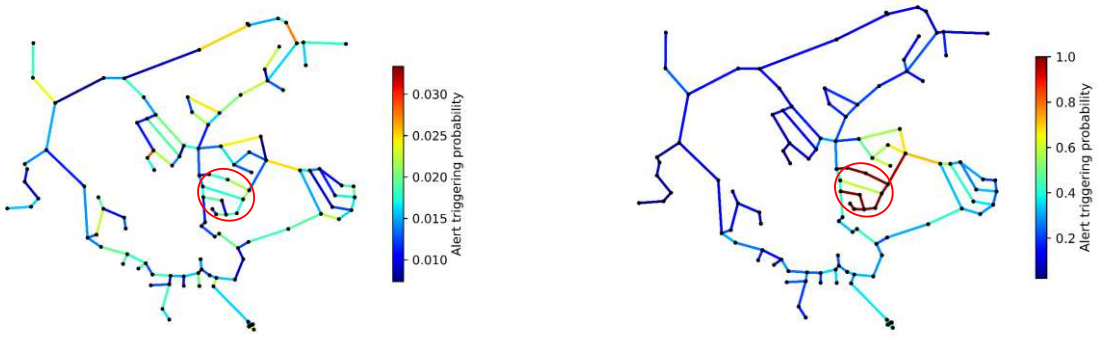


Fig 10. Reconstruction error by AE model for data under normal non-leaking condition versus leaking situation (including leaks inside and outside the monitoring area)

Based on the observation, a threshold of reconstruction error can be defined for the trained AE neural network can be used to differentiate the leak versus non-leak situations. This threshold can be set based on the training process and be further tuned when data under leaking conditions is available. A threshold of 0.000402 is set for this case based on the reconstruction error at the end of AE neural network training. With this trained AE model,

the monitored water pressure data can be fed into the trained model to obtain the reconstruction error. If the reconstruction error is larger than the set threshold value, a leaking alert would be triggered to promote actions such as inspection and replacement. The performance of AE is evaluated at the situations where leaking happens at each pipe. For each single pipe, independent data of 2200 non-leaking cases and 2200 leaking cases are generated. The water pressure data inside the monitoring area is then fed into the trained AE model. Fig 11 summarizes the probability of leaking alert is triggered, i.e. percentage of cases with a reconstruction error larger than the threshold, under each pipe leaking conditions of the WSN. As can be seen from Figure 11 a), the alert triggering probability (i.e., false alert) under non-leaking situation is very low, or only about 3% maximum. Figure 11 b) shows the probability leaking alert is triggered when leak occurs at each pipe. For the leaking happens inside the monitoring area, the alert has 68% to 100% probability to be triggered. For leak happens outside the monitoring area, the chance of triggered the alert is compromised (less than 40% for most parts). These observations imply that AE model can detect leaks from the monitored water pressure data. For the globally monitoring purpose, the monitoring sensors need to be strategically deployed in the WSN to achieve high reliability in leak detection.



a) b)

Fig. 11 Probability of AE model triggers leak by individual water pipe in the WSN under a) non-leaking condition, b) leak condition (encircled are the monitoring area where water pressure data is collected)

Sensitivity Study on Factors Affecting the Accuracy of the AE Model in Leak Detection

Sensitivity study is conducted to evaluate the effects of contributing factors on the performance of AE model for leak detection. These include three independent factors, including the compression ratios of AE, the sizes of leak, and fluctuation/uncertainty of water demand.

The compression ratio is the number of uncompressed data divided by compressed data as calculated in Eq. 12. It is an important hyperparameter of the AE neural network. A large compression ratio can not only save the physical data storage space but also force the AE model to learn the internal pattern of input data. However, too much compression may lead to excess information loss and decrease the detection accuracy. The range of compression ratio is selected between 1 to 6 for the sensitivity study.

$$\text{Compression Ratio} = \frac{\text{Uncompressed Size}}{\text{Compressed Size}} \tag{12}$$

The uncertainty of water usage is the description of water demand fluctuation during a day. A higher fluctuation of water demand increase the difficulty for leaking detection since water demand and leak both affects water pressure in the WSN. To describe its sensitivity, the uncertainties of water usage are assumed to follow a normal distribution and are described with different water usage uncertainty levels, i.e., N(0, 0.001) L/s, N(0.005)L/s, N(0, 0.01)L/s, N(0, 0.05)L/s.

The leaking size is another important factor that influences the detection system performance. Conceptually, detection of small leak is much difficult than large leak, since smaller leak has less influence on the status of WSN and can be inundated with water demand fluctuations. For the sensitivity study, the leaking size is varied from 0.01m to 0.12m.

For each combination of these three factors, the performance is evaluated by a dataset generated by assuming leak occurs in a pipe (pipe '163') inside the monitoring area and a data set with non-leaking. The data is randomly split for independent validation. The final accuracy is calculated as the average accuracy from a 3 rounds of cross-validation processes. Figure 12 shows the leak detect accuracy of the AE models affected by the compression ratios, water usage uncertainty, and leak sizes.

As shown in Fig. 12, the AE model achieved close to 100% accuracy when uncertainty with water usage is small. At a given leaking size, the accuracy of leak detection by the AE model decreases with the increasing water usage uncertainties. As the water usage

uncertainty level increases from 0.001 L/s to 0.015 L/s, the accuracy of the model decreases from 100% to 89.93%. However, even with high variance in water use uncertainty (compared to the baseline water usage at 0.012 L/s), the AE model achieved decent accuracy in leak detection.

The performance of AE model is significantly influenced by the leak size. Small leaks tends not to be detected and WSN is classified under normal non-leaking situations (i.e., 0% correct detection). While normal non-leaking cases are all classified correctly (i.e., 100% correct detection). This gives an accuracy of around 50% for a balanced dataset with equal number of data under both leaking and non-leaking conditions. With increasing leaking sizes, the AE model achieved higher leak detection accuracy. This is reasonable since the larger the leak size, the more disturbance it will have on the pressure distribution in the WSN to allow its detection. A similar conclusion was shown in Zhou et.al (Zhou, Tang et al. 2019).

The compression ratio has a negative influence on the overall detection accuracy. For example, as shown in Fig12 a), at the leaking size of 0.06 m, the accuracy decreased from 85.24% to 67.02% when compression ratio increases from 1 to 6. For leaking size of 0.11 m, the accuracy decreases from 100% to 80.75% when compression ratio increases from 1 to 6. This is reasonable since the higher compression ratio will loss more information of original dataset. However, it is also noticed that the influence of compression ratio is small for compression ratio less than 2. A compression ratio of around 1.5 appeared to achieve the best results. It also should be noted that compared to the other two factors (leaking size and water demand uncertainty), compression ratio has a relatively smaller impact on the detection accuracy. However, for a given leak size and water demand uncertainty, fine tuning the compression ratio of AE model helps to achieve a higher detection accuracy.

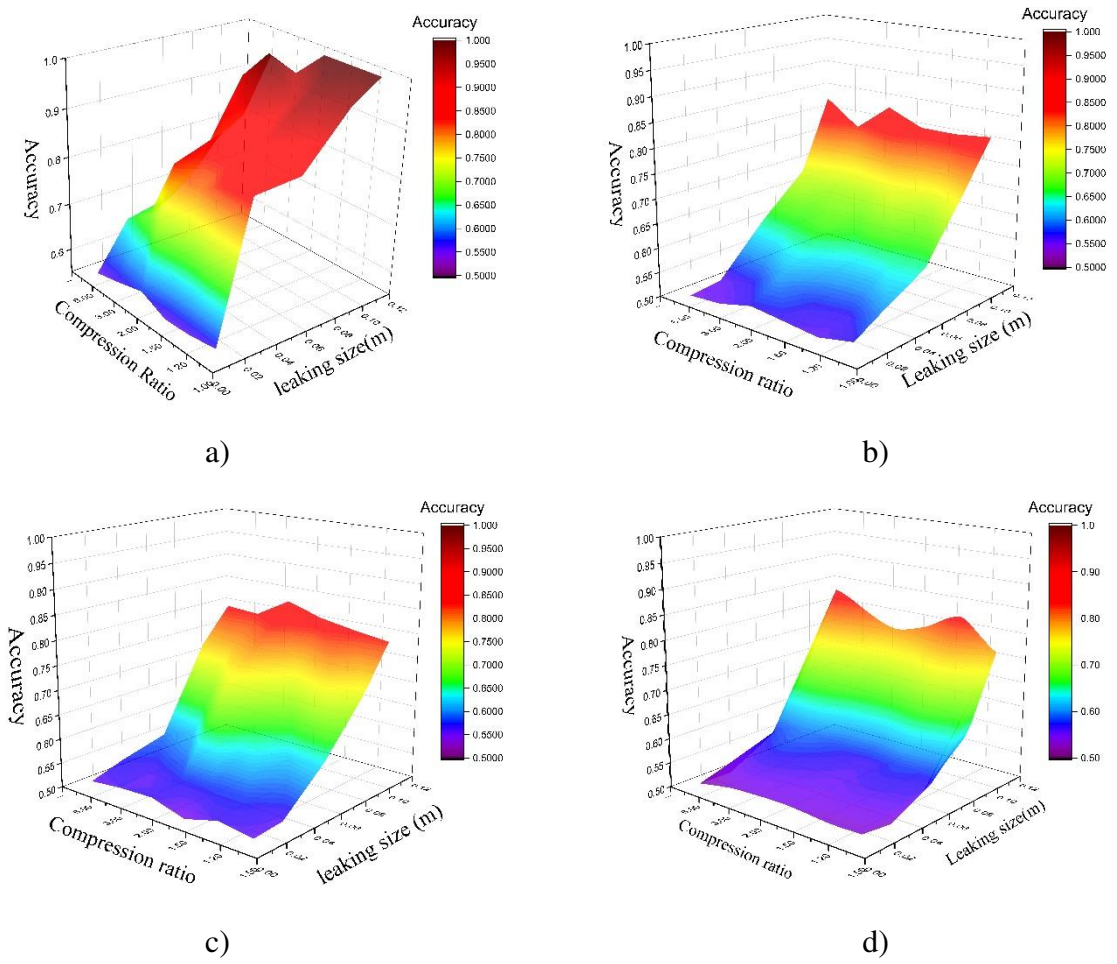


Fig 12. The sensitivity of leak detection accuracy by AE model on the compression ratio, leak size, and water usage uncertainty: (a) water usage uncertainty of N(0,0.001) L/s; water usage uncertainty of N(0,0.005) L/s, (c) water usage uncertainty of N(0,0.01) L/s, (d) water usage uncertainty of N(0,0.015) L/s

Improving Classification Accuracy by Incorporating Multiple Independent Detection

The previous results show that by setting the proper detection threshold, the AE model achieved good leak detection accuracy using unbalanced data. This is a major advantage to the conventional ANN model, which requires balanced data under both non-leaking and leaking conditions. An observation is that the AE model did not achieve as high accuracy as the ANN model. It is desirable to further improve the accuracy of AE model that will help to reduce the amount of false detection (i.e., false leaking alarm or missing detection of leak event) detection. A method is proposed to further increase the leak detection

accuracy by utilizing the probability theory for multiple independent trials. Intuitively, since leaking in the physical world will last for a while before it is repaired, the chance to detect the leak is higher if the effort is attempted multiple times. The leak status is unveiled by a voting strategy. In other words, for *n* attempts in leak detection, the detection outcome is defined as the outcome by the majority (more than 50%) of these attempts. Since each detection attempt is via independent data set, each represents an independent trial. Mathematically, if the probability of correctly detecting a leak under a single attempt is p, then the probability of more than half attempts correctly detect the leak will be

$$\mathrm{p} = \sum_{i=int(\frac{n}{2})}^{n} C_n^i p^i (1-p)^{n-i} \quad (13)$$

where n is the number of the total attempt for identifying a leakage. C_n^i is the set of i combination of set n. n is the total number of monitoring cases. p is the correct detection probability of each case.

According to Equation (13), the probability of correct detection approaches to 1 when n approach infinite $\mathrm{n} \to \infty$, under the condition p is larger than 0.5.

According to the principle described by Equation 13, multiple attempts were made for leak detection, i.e., multiple datasets under a given leaking or non-leaking condition are fed into the AE leak detection model. The final designation of leaking or non-leaking condition is based on if more than half of the detection attempts give that result.

To evaluate the performance of multiple attempts, three scenarios are considered, i.e. non-leaking situation, leaking in pipe 163 located inside the monitoring area, and leaking in pipe 214 located outside the monitoring. 2200 set of pressure data are generated under each scenario. For each scenario, the accuracy with n times of attempts is calculated by the following procedures (also illustrated in Figure 13):

1) n sets of data are randomly selected from the 2200 cases.
2) Each of the dataset is fed into the AE model to generate an output of Leak or Non-leak condition based on the set threshold.
3) The final designation of Leaking versus Non-Leaking condition based on more than half of the n attempts give that condition.
4) Determine if the defection is correct or wrong by comparing the detected condition by 3) with the actual condition of the pipe.

The procedure from 1) to 4) are repeated 1000 times. From this, the overall accuracy in correctly detecting the pipe condition is calculated.

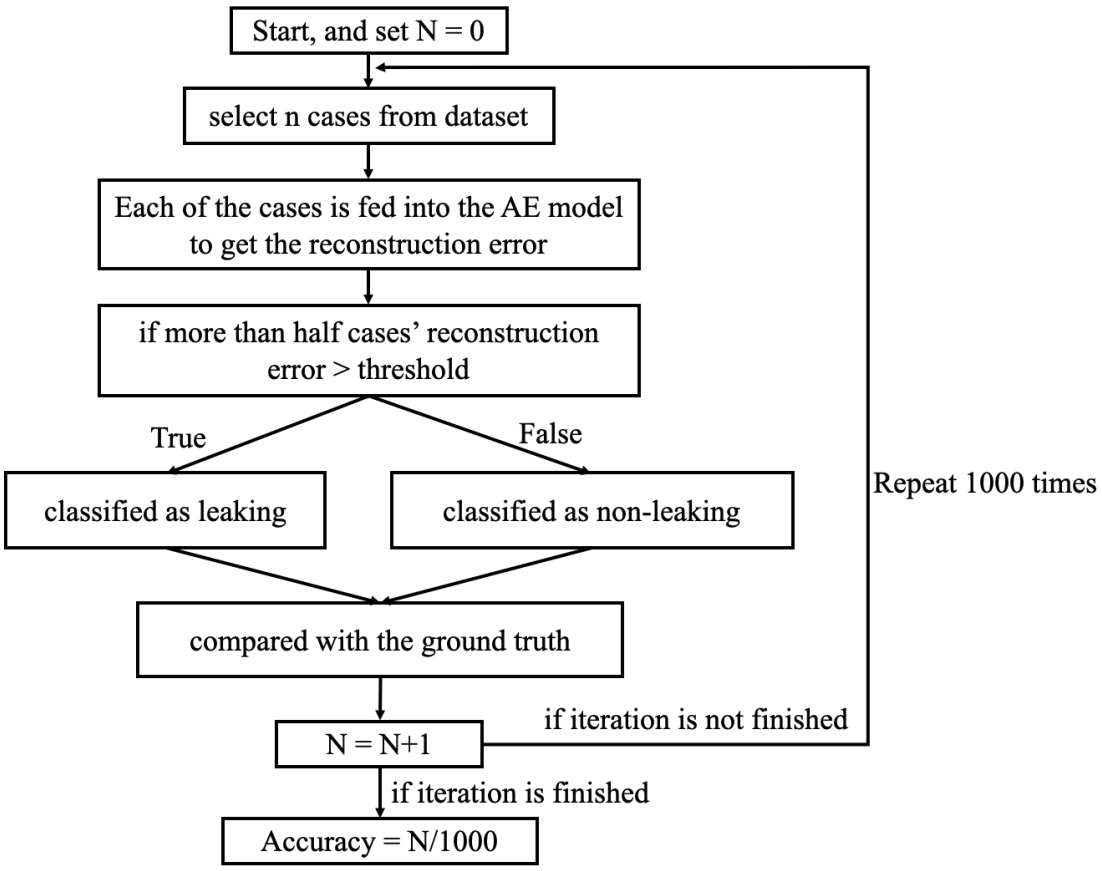


Fig 13 Flow chart of evaluation process with *n* attempts

1) Effects of multiple attempts

The results of accuracy under multiple attempts of detections using the pre-trained AE model are shown in Fig 14. The detection threshold is set as 0.000402. The vertical axis is the detection accuracy of the pipe condition. The horizontal axis indicate the number of attempts in detection. As seen from this figure, the detection accuracy improved with multiple attempts and achieved close to 100% detection accuracy, regardless where leak occurs. This is consistent with what is predicted by Eq. 13.

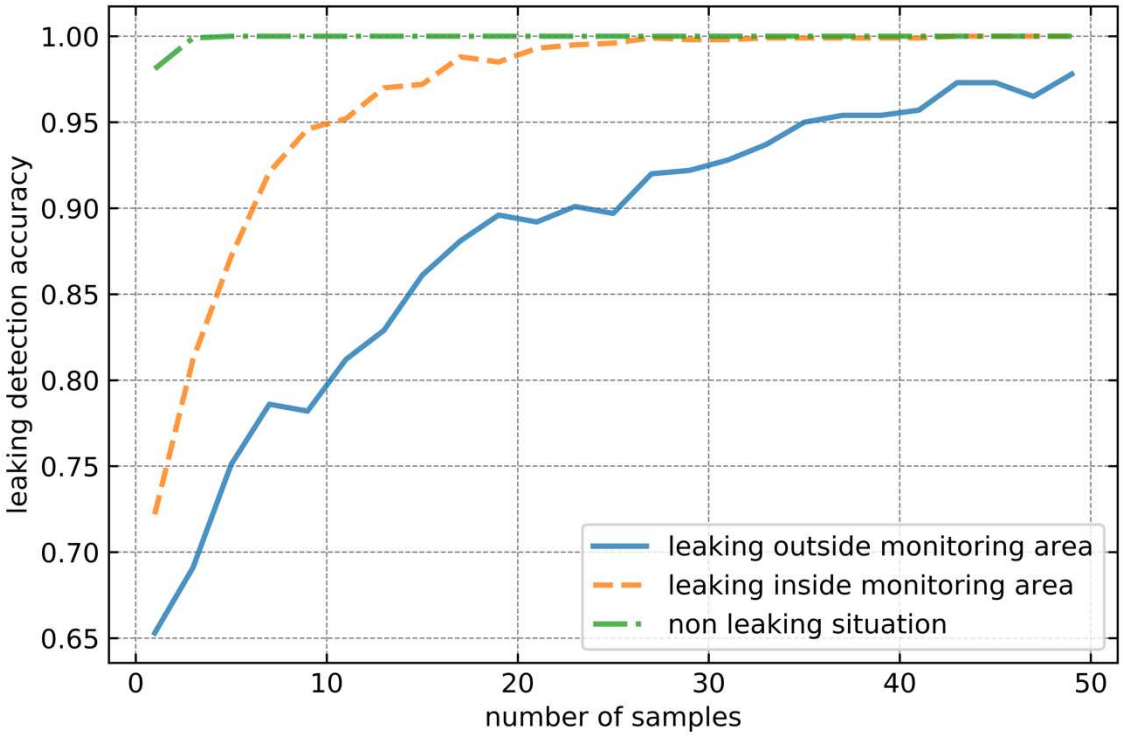


Fig 14 Accuracy of correct pipe condition detection with multiple attempts with pre-trained AE model with threshold set as 0.000402

2) Effects of detection threshold by the AE model

Threshold is a critical parameter for the AE leak detection model. Sensitivity analyses are conducted on the influence of threshold on the model detection accuracy. The results for the three different scenarios defined as in the previous context are shown in Fig. 15. 2200 cases of dataset were generated for each scenery and are fed into the AE model to determine the reconstruction errors. From this, the percentage reconstruction error by AE model larger than the reconstruction error is determined. The horizontal axis are the thresholds and vertical axis is the percentage of cases with reconstruction error larger than the threshold (i.e., the case is identified as leaking by the AE model). The 50% line is also indicated in the figure. As can be seen from the figure, for all the three pipe condition sceneries, a smaller threshold corresponds to larger chance for the condition to be identified as leaking condition. For no leaking condition, this presents as false alarm. A larger threshold reduces the false alarm but may miss leaking cases. According to the equation 13, leaks can be properly identified with multiple attempts as long as a detection accuracy

is larger than 0.5. Based on this criteria, leak within the monitoring area can be accurately detected with any threshold between 0.00029 to 0.00057, by use of the multiple attempts strategy.

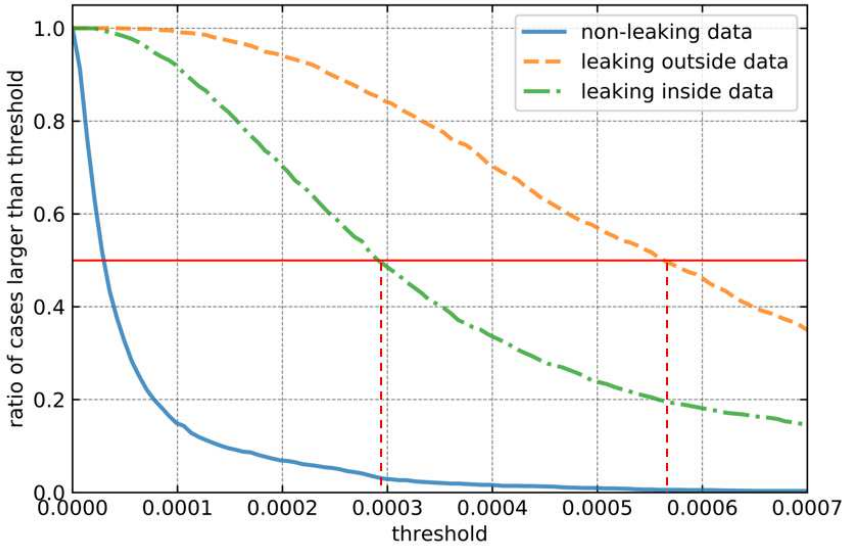


Fig. 15 Percent of leak warning at different detection thresholds of AE model under different pipe conditions

By using the proposed post-processing method with 50 time steps data. The final detection result is shown in Fig 16. The result clearly demonstrates that when leaking happens inside or nearby the monitoring area, the AE model is able to detect such leaking happens correctly. For leaking which is far away from the monitoring area, the model can differential it from inside leaking situation to mitigate the false alert. There are two pipe leaking situations inside the monitoring area not detected. The main reason is the unappropriated threshold selection since all the pipes are using the same threshold. However, when more and more data available during the operation stage, this threshold can be tuned for each pipe, which will increase the detection ability eventually.

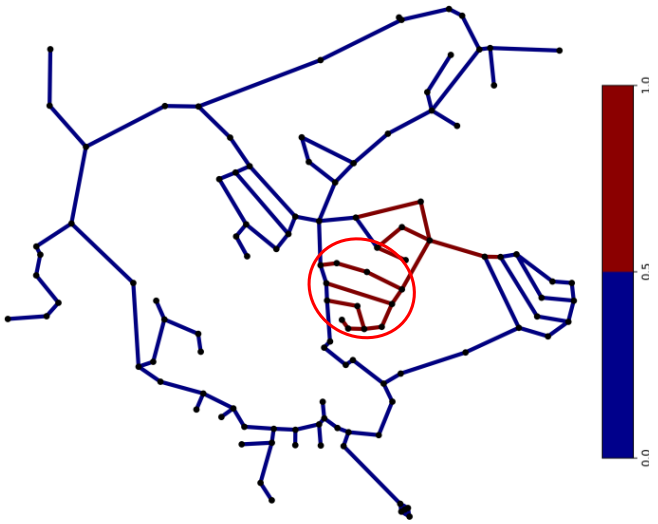


Fig. 16 Final leak detection result for each pipe leaking situation

4.2 Case Study – II: C-Town water distribution network

C-Town water distribution network is a virtual network that was used for calibration competition in Battle of the Water Calibration Networks (BWCN) (Ostfeld, Salomons et al. 2012). The topology and mode of operations of the network are described in details and the true network data are made public after the competition. This well-characterized water distribution network allows to test the proposed leaking detection method under a more complex scenario. From this, the performance of the proposed Autoencoder model based leak detection model is evaluated.

The topology of C-Town water distribution network is shown in Fig 17. There are 1 reservoir and 7 water supply tanks. This network including 388 user nodes, 432 pipes, 11pumps, and 4 valves, which are divided into 5 district meter areas (DMA). The water demand at each node is provided. In this study, 4 predefined monitoring areas are chosen as shown in Figure 17.

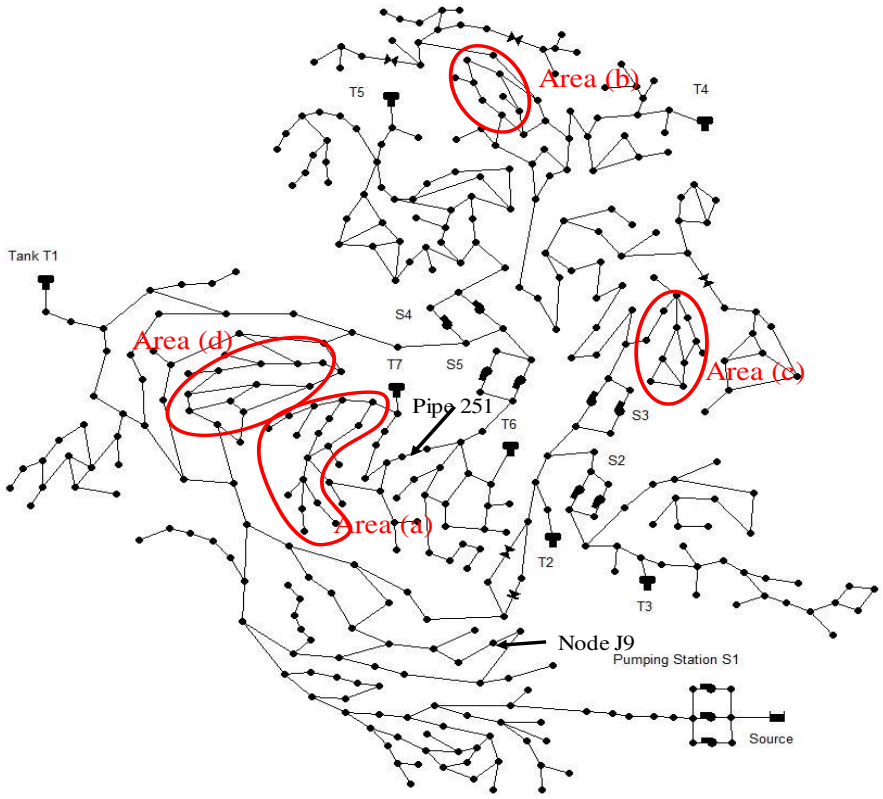


Fig. 17 Topology of the C-Town WSN with DMA areas noted.

Hydraulic model of this C-Town WSN is built with WNTR. Since the water demand at each node is already defined, the actual water demands are used for hydraulic model rather than using the water demand ratio and uncertainty. The WDN under non-leaking situation and pipe failure (leaking size of 0.05 meter) are simulated. The corresponding water pressure data at the monitoring nodes in the C-Town WSN are produced following the designed data generation framework described in the earlier part of this paper.

Fig 18 shows the water pressure at node 'J9' with and without leaking. The leaking situation corresponding to leaking at pipe 'P251'. The exact position of node and pipe can also be found in Figure 17.

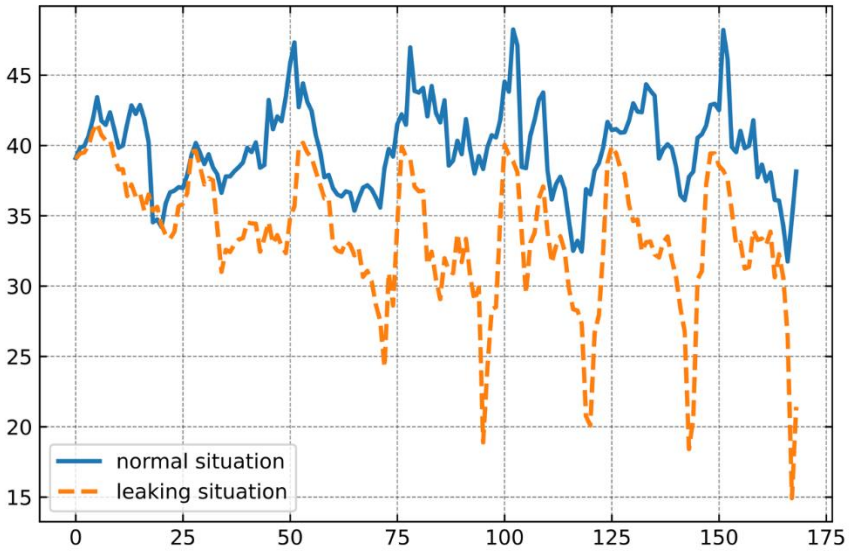
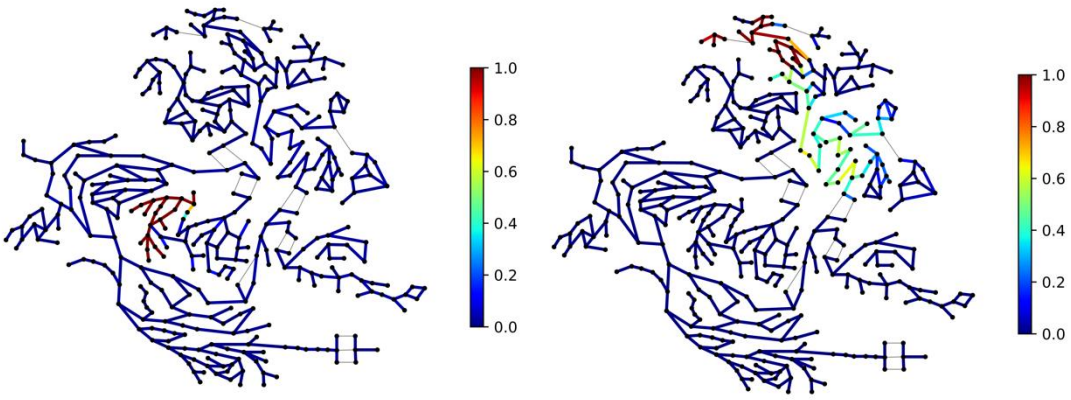


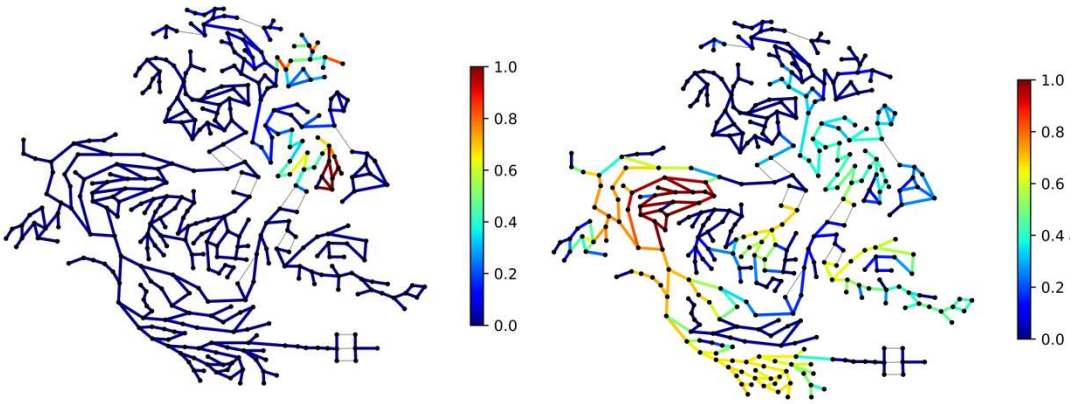
Fig. 18 Water pressure (water head) at 'J9' with and without leaking

Performance of AE model in leak detection

The performance of the AE leaking detection model is evaluated by calculating the probability of AE model triggering an alarm when leaking happens at each pipe, by sensors installed at different DMAs as shown in Fig. 17. For each pipe in the network, 169 sets of leaking and non-leaking water pressure data at nodes within the monitored area are generated. By feeding the water pressure data into the AE model, the probability successfully detection of each pipe leaking is shown in Fig 19.



a) b)



c) d)

Fig. 19 The probability of leak alert by AE detection model when leak happens on pipe in the WSN: a) monitoring sensors located in area a of C-town, c) monitoring sensors located in area b of C-town, c) monitoring sensors located in area c of C-town, d) monitoring sensors located in area d of C-town

A few observations can be made from Fig. 19: 1) The probability of a successfully leak detection of a pipe is affected by the location of the pipe and the distribution of monitoring sensors. 2) the AE leak detection model achieved a high detection accuracy for leaks of pipe inside the monitoring area. 3) The probability of detecting leak in pipes outside the monitoring area is typically smaller and is affected by the topological structure of the WSN and setting of the AE model.

The proposed multiple detection attempts strategy is utilized to further improve the leaking detection accuracy and mitigate the false alarm. 100 independent attempts are used. Leak alert will be triggered if more than 50 attempts indicated leaking (or reconstruction error by the AE model larger than the threshold). Figure 20 shows the updated result of probability of detecting leaks in pipes in the WSN using this strategy. The results indicated leaks in pipes located in the monitoring area are all detected with 100% accuracy. Compared with the results shown in Figure 18, the false alarm is significantly mitigated. In the meanwhile, leaks outside the monitoring area is not detected, except for Figure 19 b). This is due to a conditional valve nearby and therefore leaks in these pipes have a larger disturbance to the WSN. The implication is that sensors need to be deployed in a strategic way to ensure the full coverage of the complete WSN.

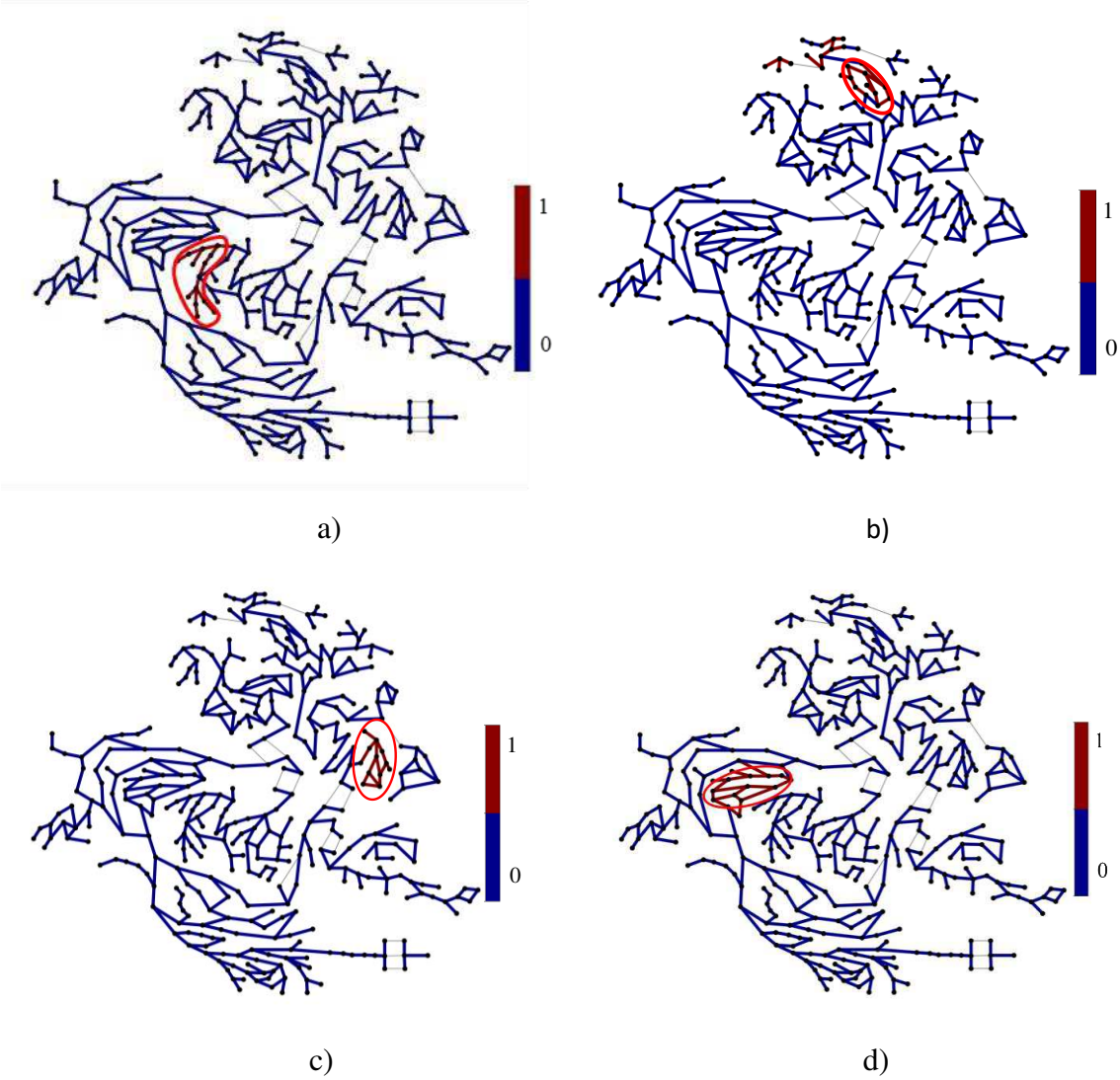


Fig. 20 The probability of leak alert by AE detection model incorporating multiple attempts strategies for leak happens on pipe in the WSN: a) monitoring sensors located in DMA 4, c) monitoring sensors located located in DMA 2, c) monitoring sensors located in DMA 3, d) monitoring sensors located in DMA 1

6. Conclusions and Discussions

Real time detection of leaks in the water supply network (WSN) bears important socio-economic benefits and is, however, challenging. Innovative data-driven machine learning (ML) models for leak detection are developed in this study. The spatial relationship of water pressure at multiple nodes in a water distribution network was learned and used for leak detection. Model-based data generation strategy is developed where data was

generated by an industry certified hydraulic model for testbed WSNs. Factors such as the fluctuation of water demand are considered under both non-leaking and leaking situations. Artificial Neural Network (ANN) is found to achieve high accuracy in leak detection. It, however, requires balanced dataset including similar amount of data collected under leaking or non-leaking conditions, which is difficult to implement in the real WSNs where data under leaking condition is relative rare than that under normal conditions. The Autoencoder Neural Network (AE) model, an unsupervised ANN model, is developed to learn from the unbalanced data to classify the leaking versus non-leaking conditions. The results indicate AE achieved decent accuracy in leak detection. Factors affecting AE leak detection model are analyzed. An innovative strategy is proposed to further improve the reliability in leak detection by use of multiple independent attempts. The results show that this significantly reduces the false alarm. The AE based leak detection strategy is applied to a mature WSN to further evaluate its performance. The results indicated the AE ML model-based strategy achieves high accuracy in detecting leaks in pipes located within the monitoring area. The accuracy in detecting leaks in a pipe is dependent upon the location of the water pipe and the distribution of monitoring sensor. The use of multiple attempts strategy significantly reduced the false alarm.

Discussions:

Detecting leaks in the WSN is challenging due to the complex topology, fluctuations in user demand, and lack of monitoring data. Detecting leaks with inspection tools is expensive and labor-intensive, and cannot achieve real-time detection. Traditional model updating approach for structural health monitoring is difficult to implement for WSN due to complex topology and uncertainty in the hydraulic conditions. Detecting leaks based on the transient responses of WSN requires to capture the transient signals over a very short period when leak occurs, which requires high sampling rate. Data-driven approach using ML models is promising to achieve quick and reliable leak detection. The rationale is that the spatial pattern of water pressure and its variations under leak are affected by the network structure of water distribution offers information about the conditions of the WSN. The ML models developed in this study allow to detect leak from unbalanced data, i.e., with only data under normal operational conditions. Compared to other leak detection algorithms, the methods have the following advantages:

1) Provide quick leak detection with high accuracy.
2) Unlike the conventional transient-based leak detection, which requires sensor with high sampling rates to capture the transient process. The AE leak detection model learn from the spatial pattern contained in the data and only needs sensor with low sampling rate (and therefore inexpensive).
3) By using data from multiple nodes, the detection is more robust than data-driven models that only use data at single node.
4) The data driven approach does not require strong domain expertise to implement.

While data used for model training and validation in this study are from generated data by high fidelity model for WSN. The framework is readily applied to real world data. With the development of the Internet of Things (IoT), more and more sensors, i.e., smart meters, will be deployed into the water distribution network to monitor its health conditions. This will allow to obtain data to be used by the developed model.

It is noted that the AE leak detection model identifies leak within a certain monitoring area. Precisely locate the leaking position is important for timely intervention, which requires further investigation.

Declaration

Ethics approval and consent to participate: No human subject or animals are involved in the study.

Consent for publication: the authors consent the publications of this paper.

Availability of data and material: the data and code are available upon request

Competing interests: N/A

Funding: This research is partially supported by the US National Science Foundation.

Authors' contributions: Xiong (Bill) Yu: envision the research, guide research activities; Xudong Fan: conduct analyses; Xijin Zhang: provide assistance

Acknowledgements: The author acknowledge the help from the staff members of the Cleveland Water Department led by Mr. Alex Margevicius during the study.

References

Abokifa, A. A., K. Haddad, et al. (2018). "Real-Time Identification of Cyber-Physical Attacks on Water Distribution Systems via Machine Learning–Based Anomaly Detection Techniques." Journal of Water Resources Planning and Management **145**(1): 04018089.

Adedeji, K. B., Y. Hamam, et al. (2017). "Towards achieving a reliable leakage detection and localization algorithm for application in water piping networks: An overview." IEEE Access **5**: 20272-20285.

Amran, T. S. T., M. P. Ismail, et al. (2017). Detection of underground water distribution piping system and leakages using ground penetrating radar (GPR). AIP Conference Proceedings, AIP Publishing LLC.

Bakker, M., J. Vreeburg, et al. (2013). "A fully adaptive forecasting model for short-term drinking water demand." Environmental Modelling & Software **48**: 141-151.

Bimpas, M., A. Amditis, et al. (2010). "Detection of water leaks in supply pipes using continuous wave sensor operating at 2.45 GHz." Journal of Applied Geophysics **70**(3): 226-236.

Braun, M., O. Piller, et al. (2017). "Limitations of demand-and pressure-driven modeling for large deficient networks." Drinking Water Engineering and Science **10**(2): 93-98.

Buch, N., S. A. Velastin, et al. (2011). "A review of computer vision techniques for the analysis of urban traffic." IEEE Transactions on Intelligent Transportation Systems **12**(3): 920-939.

Butler, D. (2000). Leakage Detection and Management: A Comprehensive Guide to Technology and Practice in the Water Supply Industry, Palmer Environmental.

Caputo, A. C. and P. M. Pelagagge (2003). "Using neural networks to monitor piping systems." Process Safety Progress **22**(2): 119-127.

Chan, T. K., C. S. Chin, et al. (2018). "Review of Current Technologies and Proposed Intelligent Methodologies for Water Distributed Network Leakage Detection." IEEE Access **6**: 78846-78867.

Chim, T. W., S.-M. Yiu, et al. (2011). "SPECS: Secure and privacy enhancing communications schemes for VANETs." Ad Hoc Networks **9**(2): 189-203.

Colombo, A. F., P. Lee, et al. (2009). "A selective literature review of transient-based leak detection methods." Journal of hydro-environment research **2**(4): 212-227.

Crowl, D. A. and J. F. Louvar (2001). Chemical process safety: fundamentals with applications, Pearson Education.

De Coster, A., J. P. Medina, et al. (2019). "Towards an improvement of GPR-based detection of pipes and leaks in water distribution networks." Journal of Applied Geophysics **162**: 138-151.

Funk, A. and W. B. DeOreo (2011). "Embedded energy in water studies study 3: End-use water demand profiles." Prepared by Aquacraft, Inc. for the California Public Utilities Commission Energy Division, Managed by California Institute for Energy and Environment, CALMAC Study ID CPU0052.

Gao, J., B. Shi, et al. (2006). "Monitoring the stress of the post-tensioning cable using fiber optic distributed strain sensor." Measurement **39**(5): 420-428.

Hernadez, E., S. Hoagland, et al. (2016). Water distribution database for research applications. World Environmental and Water Resources Congress 2016.

Housh, M. and Z. Ohar (2018). "Model-based approach for cyber-physical attack detection in water distribution systems." Water research **139**: 132-143.

Jain, A. K., J. Mao, et al. (1996). "Artificial neural networks: A tutorial." Computer(3): 31-44.

Kang, J., Y.-J. Park, et al. (2017). "Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems." IEEE Transactions on Industrial Electronics **65**(5): 4279-4289.

Kiliç, R. (2016). "Effective Management of Leakage in Drinking Water Network." Acta Physica Polonica, A. **130**(1).

Kim, J.-H., G. Sharma, et al. (2010). SPAMMS: A sensor-based pipeline autonomous monitoring and maintenance system. 2010 Second International Conference on COMmunication Systems and NETworks (COMSNETS 2010), IEEE.

Klise, K. A., R. Murray, et al. (2018). AN OVERVIEW OF THE WATER NETWORK TOOL FOR RESILIENCE (WNTR), Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Liao, S.-H., P.-H. Chu, et al. (2012). "Data mining techniques and applications–A decade review from 2000 to 2011." Expert systems with applications **39**(12): 11303-11311.

Lin, W.-Y., Y.-H. Hu, et al. (2011). "Machine learning in financial crisis prediction: a survey." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **42**(4): 421-436.

Liou, C. P. (1998). "Limitations and proper use of the Hazen-Williams equation." Journal of Hydraulic Engineering **124**(9): 951-954.

Loth, J. L., G. J. Morris, et al. (2004). Acoustic detecting and locating gas pipe line infringement, West Virginia University (US).

Mashford, J., D. De Silva, et al. (2012). "Leak detection in simulated water pipe networks using SVM." Applied Artificial Intelligence **26**(5): 429-444.

Mazzolani, G., L. Berardi, et al. (2017). "Estimating leakages in water distribution networks based only on inlet flow data." Journal of Water Resources Planning and Management **143**(6): 04017014.

Mounce, S., J. Boxall, et al. (2010). "Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows." Journal of Water Resources Planning and Management **136**(3): 309-318.

Mounce, S. R., A. J. Day, et al. (2002). "A neural network approach to burst detection." Water science and technology **45**(4-5): 237-246.

Mounce, S. R., R. B. Mounce, et al. (2010). "Novelty detection for time series data analysis in water distribution systems using support vector machines." Journal of hydroinformatics **13**(4): 672-686.

Ostfeld, A., E. Salomons, et al. (2012). "Battle of the Water Calibration Networks." Journal of Water Resources Planning and Management **138**(5): 523-532.

Ozevin, D. and H. Yalcinkaya (2012). Reliable monitoring of leak in gas pipelines using acoustic emission method. Proc. Civil Struct. Health Monit. Workshop (CSHM).

Pal, A. and K. Kant (2019). "Water Flow Driven Sensor Networks for Leakage and Contamination Monitoring in Distribution Pipelines." ACM Transactions on Sensor Networks (TOSN) **15**(4): 1-43.

Prasad, T. D., N.-S. J. J. o. W. R. P. Park, et al. (2004). "Multiobjective genetic algorithms for design of water distribution networks." **130**(1): 73-82.

Romano, M., Z. Kapelan, et al. (2012). "Automated detection of pipe bursts and other events in water distribution systems." Journal of Water Resources Planning and Management **140**(4): 457-467.

Sadeghioon, A., N. Metje, et al. (2014). "SmartPipes: smart wireless sensor networks for leak detection in water pipelines." Journal of sensor and Actuator Networks **3**(1): 64-78.

Sadiq, R., B. Rajani, et al. (2004). "Probabilistic risk analysis of corrosion associated failures in cast iron water mains." Reliability Engineering & System Safety **86**(1): 1-10.

Shanker, M., M. Y. Hu, et al. (1996). "Effect of data standardization on neural network training." Omega **24**(4): 385-397.

Srirangarajan, S., M. Allen, et al. (2013). "Wavelet-based burst event detection and localization in water distribution systems." Journal of Signal Processing Systems **72**(1): 1-16.

Srirangarajan, S., M. Iqbal, et al. (2010). Water main burst event detection and localization. Water Distribution Systems Analysis 2010**:** 1324-1335.

Stoianov, I., L. Nachman, et al. (2007). PIPENETa wireless sensor network for pipeline monitoring. Proceedings of the 6th international conference on Information processing in sensor networks.

Tao, T., H. Huang, et al. (2013). "Burst detection using an artificial immune network in water-distribution systems." Journal of Water Resources Planning and Management **140**(10): 04014027.

Taormina, R., S. Galelli, et al. (2018). "Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks." Journal of Water Resources Planning and Management **144**(8): 04018048.

Twort, A. C., D. D. Ratnayaka, et al. (2000). Water supply, Elsevier.

Walski, T. M. and J. W. Male (2000). Maintenance and rehabilitation/replacement, McGraw-Hill**:** 17.11-17.28.

Wu, Y. and S. Liu (2017). "A review of data-driven approaches for burst detection in water distribution systems." Urban Water Journal **14**(9): 972-983.

Ye, G. and R. A. Fenner (2011). "Kalman filtering of hydraulic measurements for burst detection in water distribution systems." Journal of pipeline systems engineering and practice **2**(1): 14-22.

Zhou, X., Z. Tang, et al. (2019). "Deep learning identifies accurate burst locations in water distribution networks." Water Res **166**: 115058.

Figures

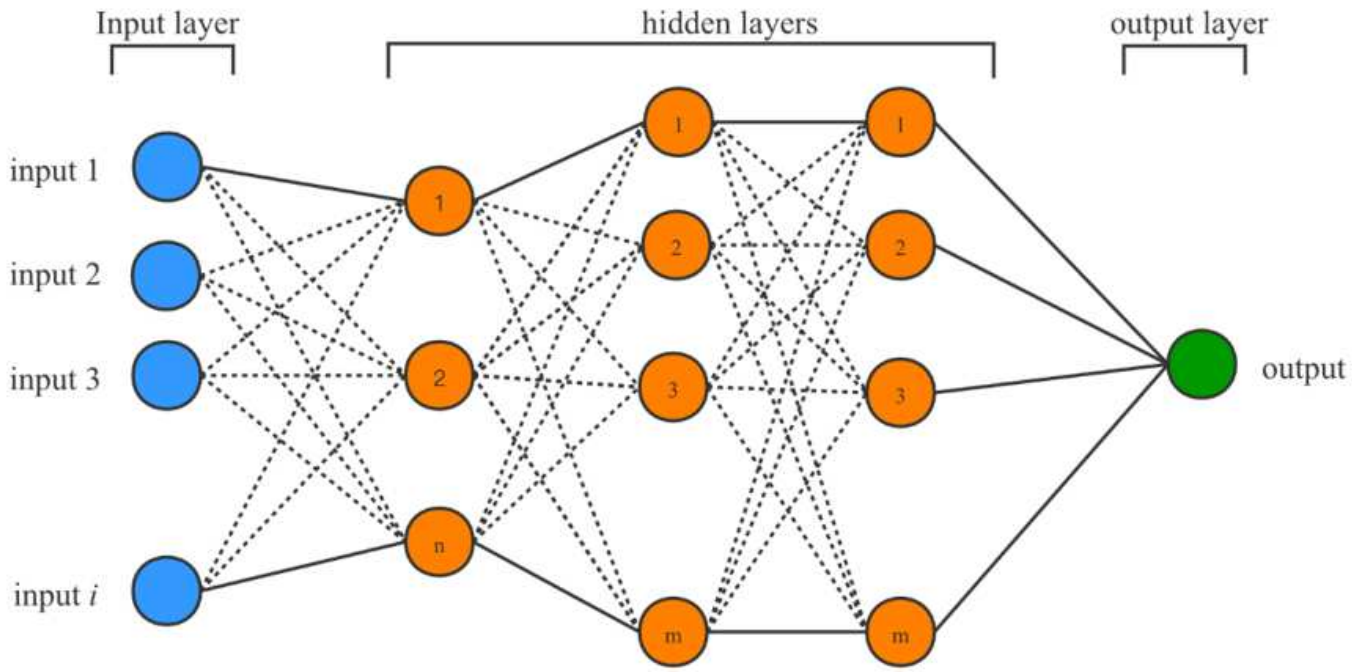


Figure 1

Schematic of ANN architecture

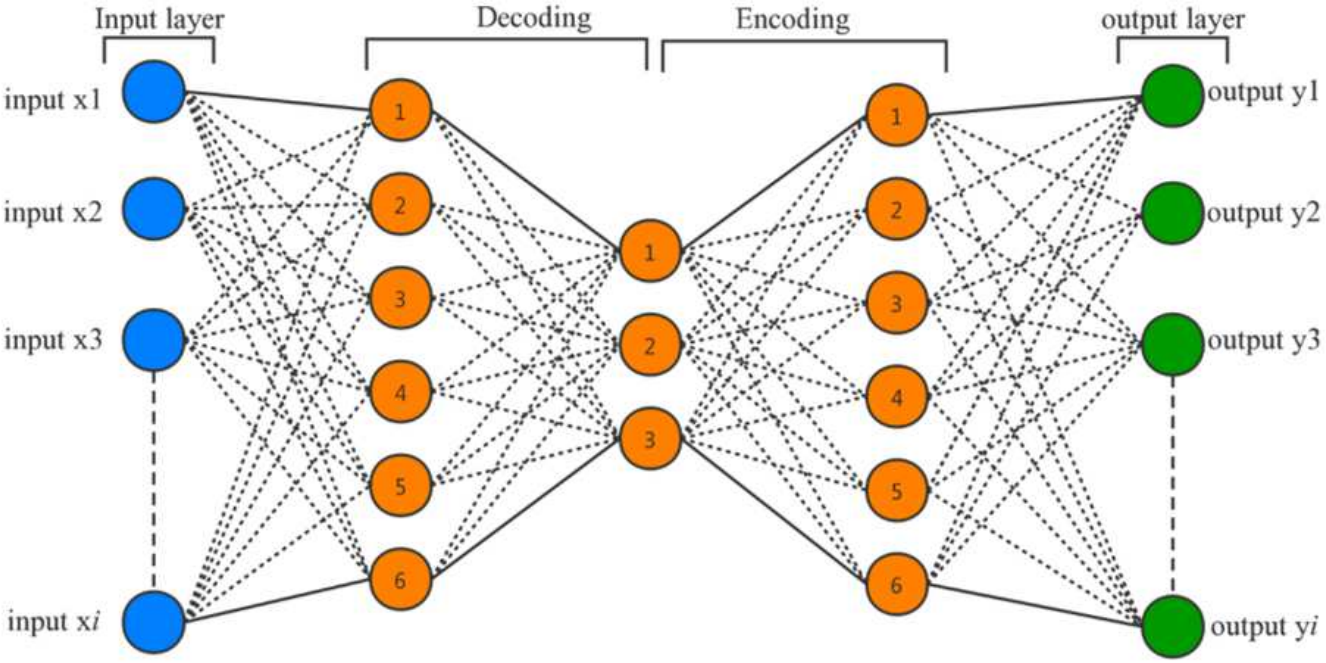


Figure 2

Schematic of architecture of an autoencoder neural network (the numbers of neurons in the decoding layers and encoding layers are conceptual)

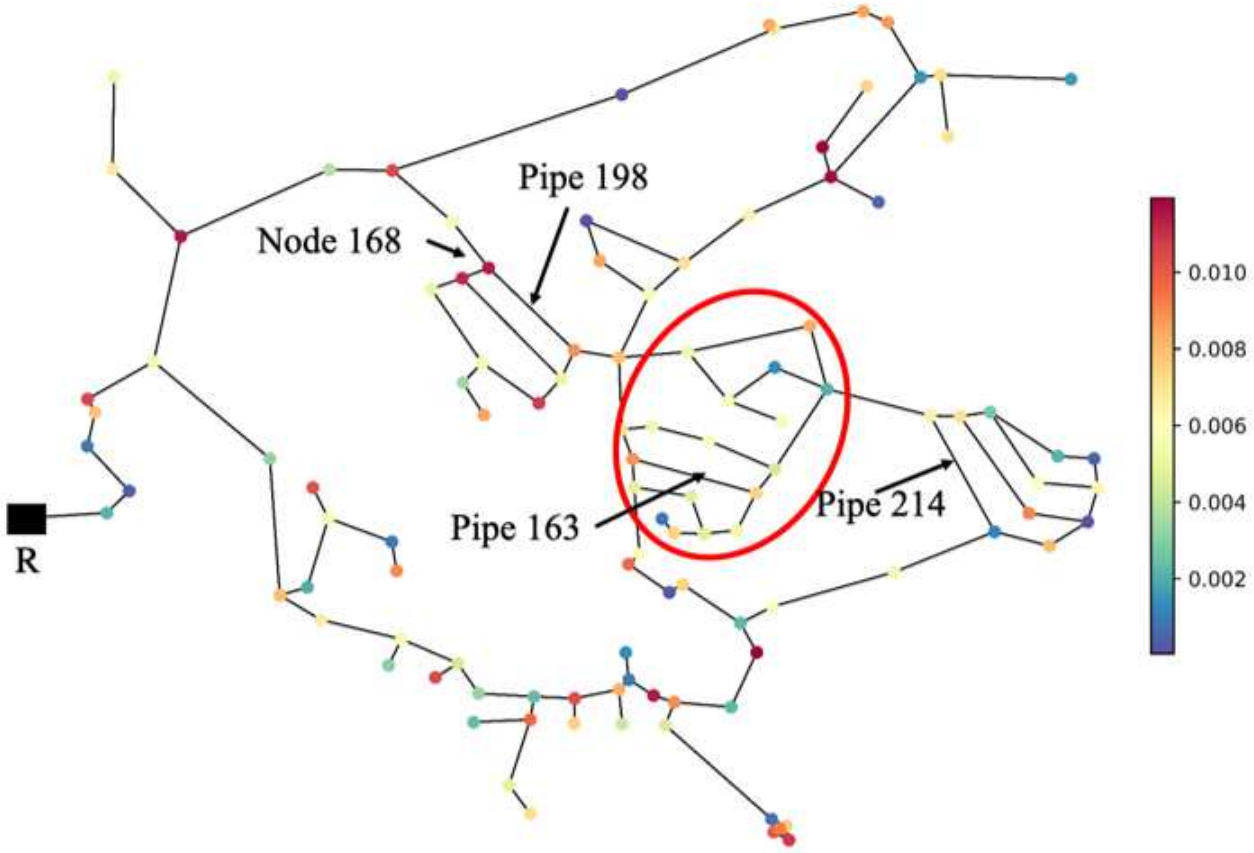


Figure 3

Illustration of the water supply network and water demand at each node (color code corresponds to basic water demand in the unit of L/s)

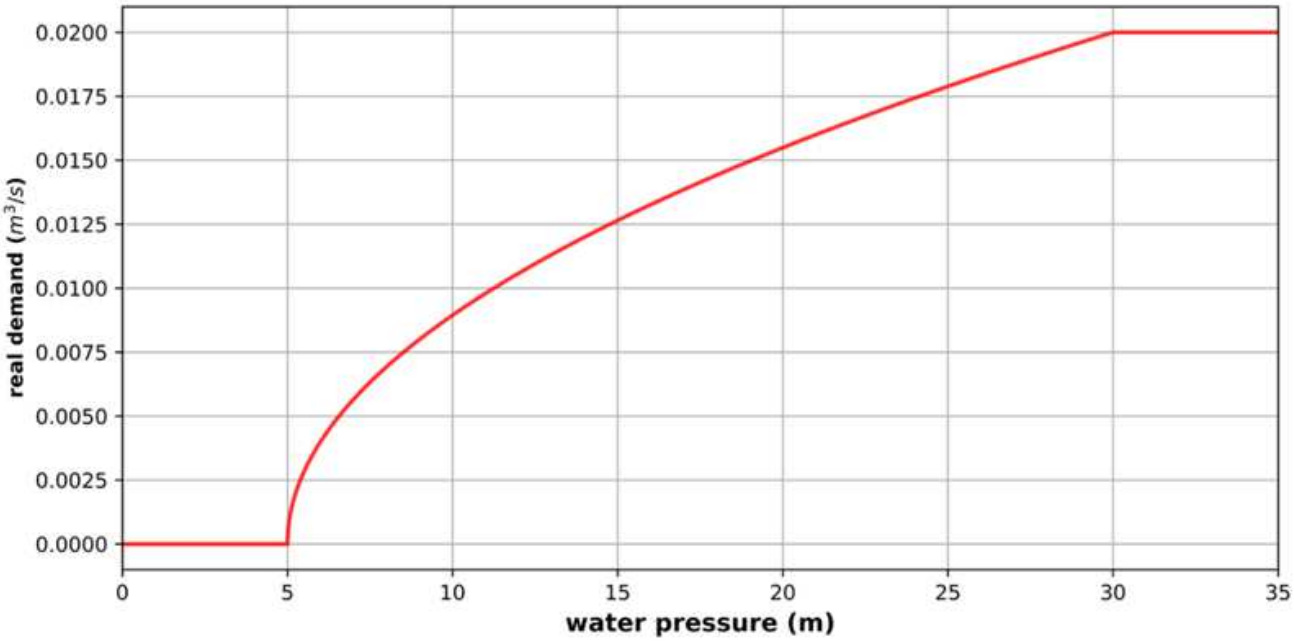


Figure 4

Example of the relationship between water demand and pressure head at a node with base demand of 0.02 m3/s

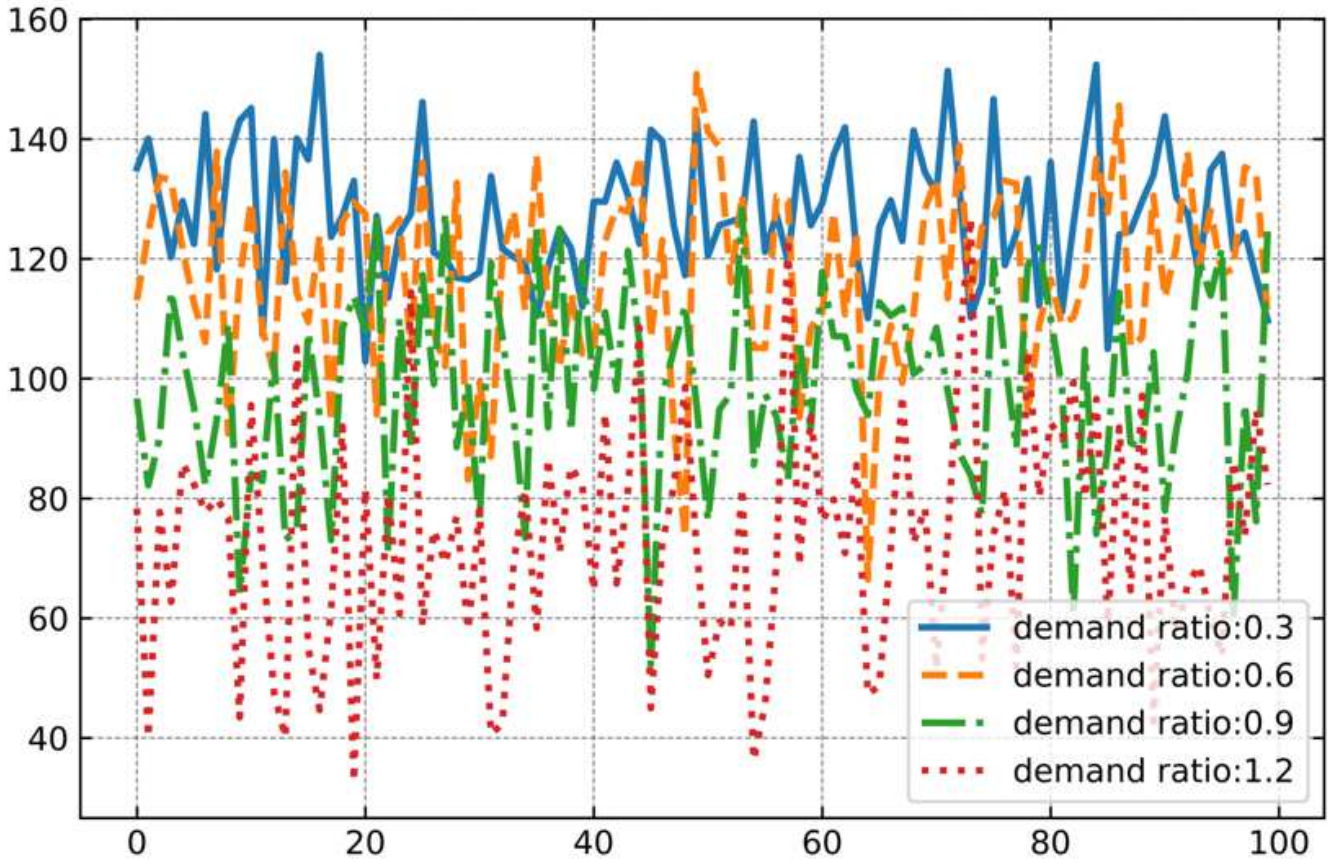


Figure 5

Example of water head at node ‘168’ under different water demands (note: demand ratio is defined as the average water demand to the predefined baseline demand at the node (Eq. 11); the fluctuation is due to fluctuations in the real demand).

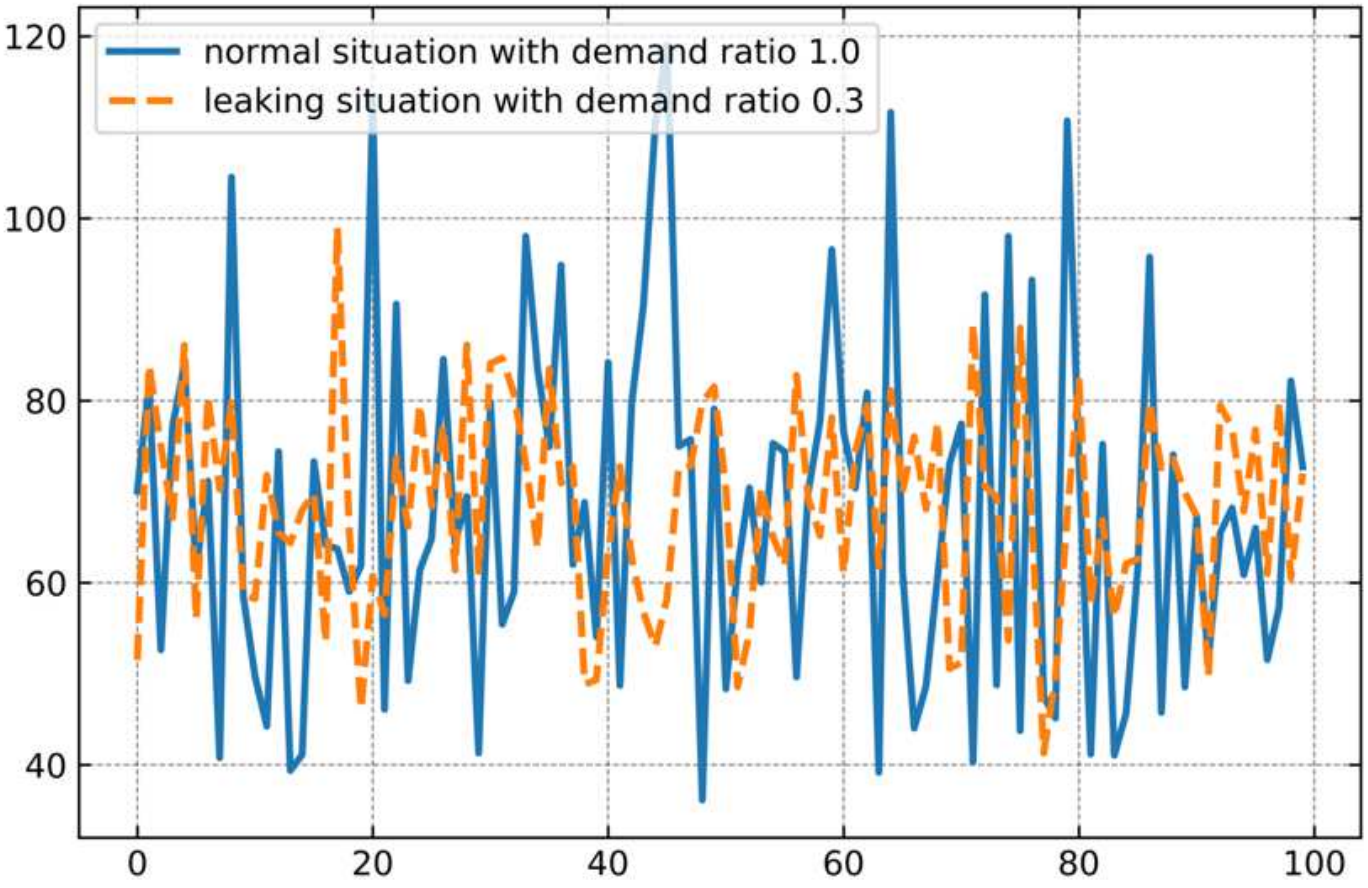


Figure 6

Illustration of water head at node ‘168’: the average water head is similar for non-leaking condition with high water demand versus leaking condition with low water demand

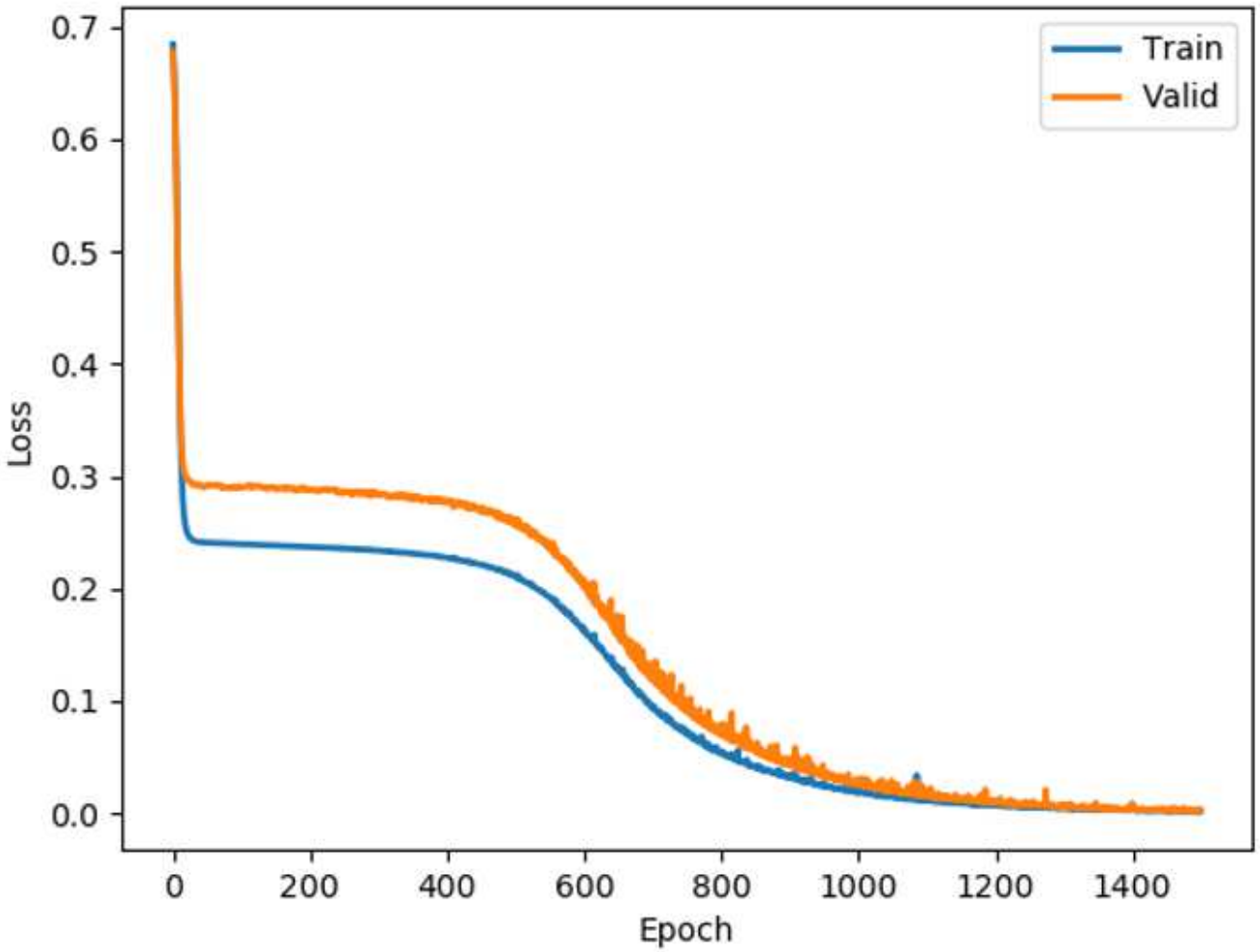


Figure 7

Loss values during the ANN model training process

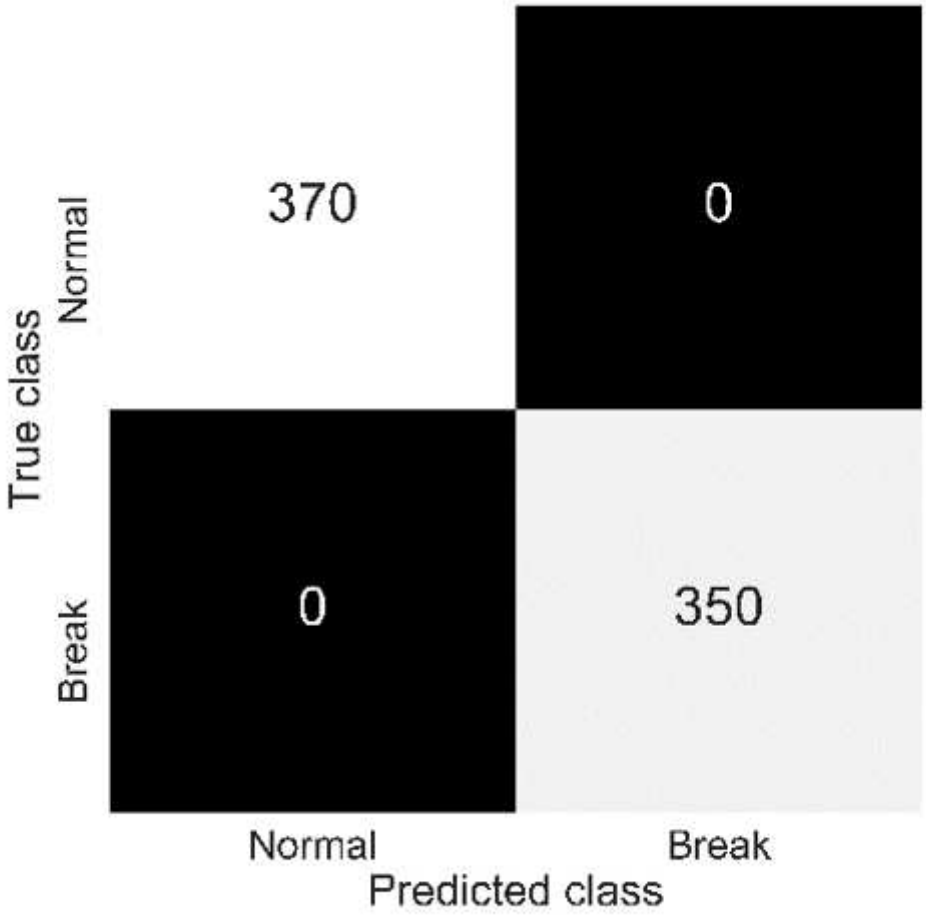


Figure 8

The confusion matrix of the classification result of leaking and non-leaking cases by the trained ANN model (achieved 100% accurate with no misclassification)

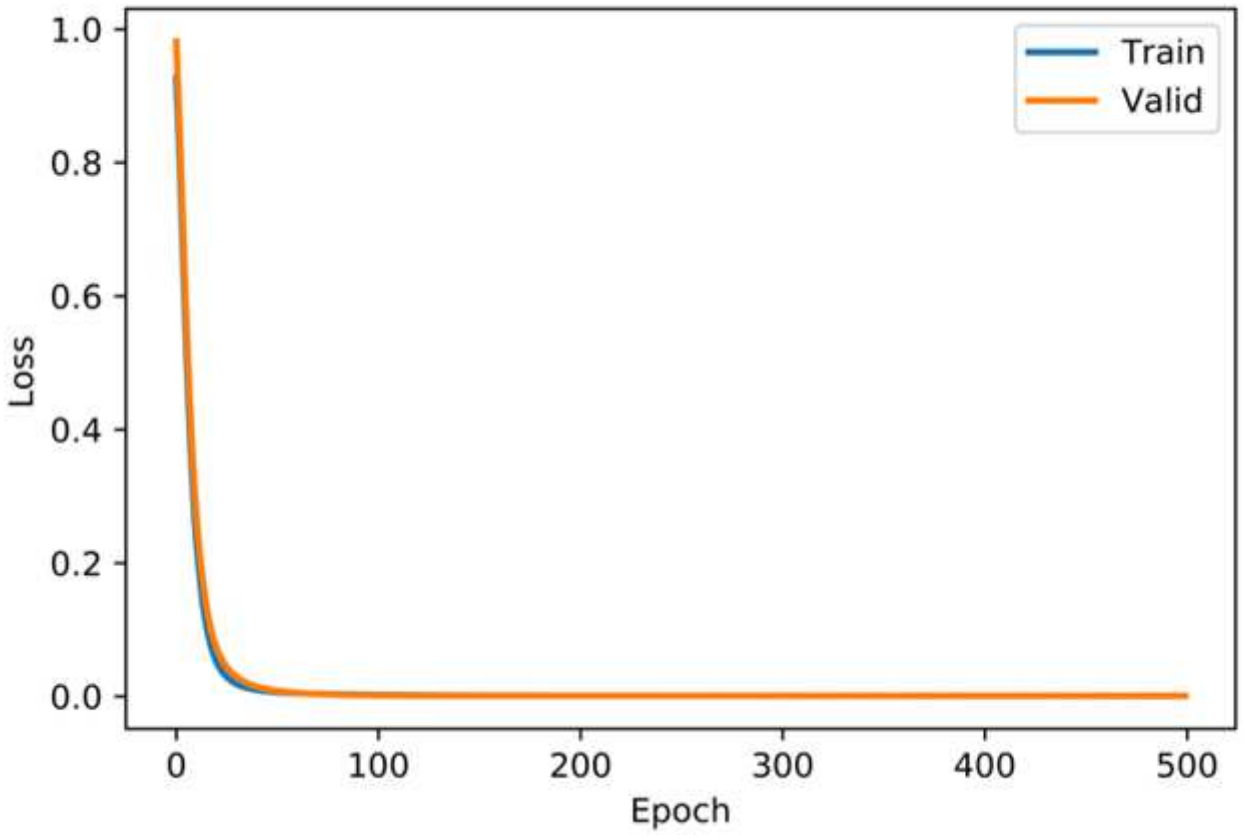


Figure 9

Loss values (i.e., the reconstruction error) by the AE model during the training process

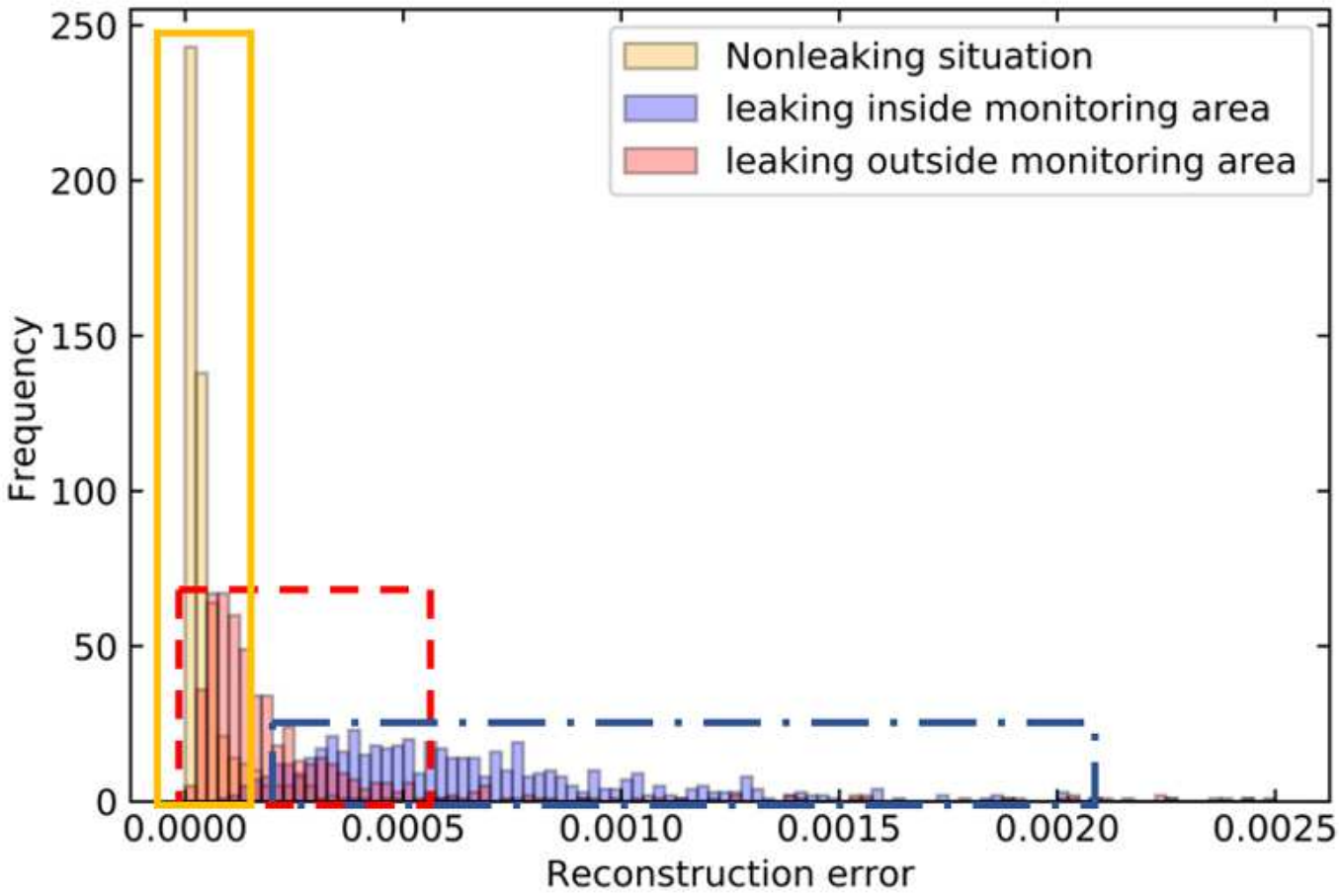


Figure 10

Reconstruction error by AE model for data under normal non-leaking condition versus leaking situation (including leaks inside and outside the monitoring area)

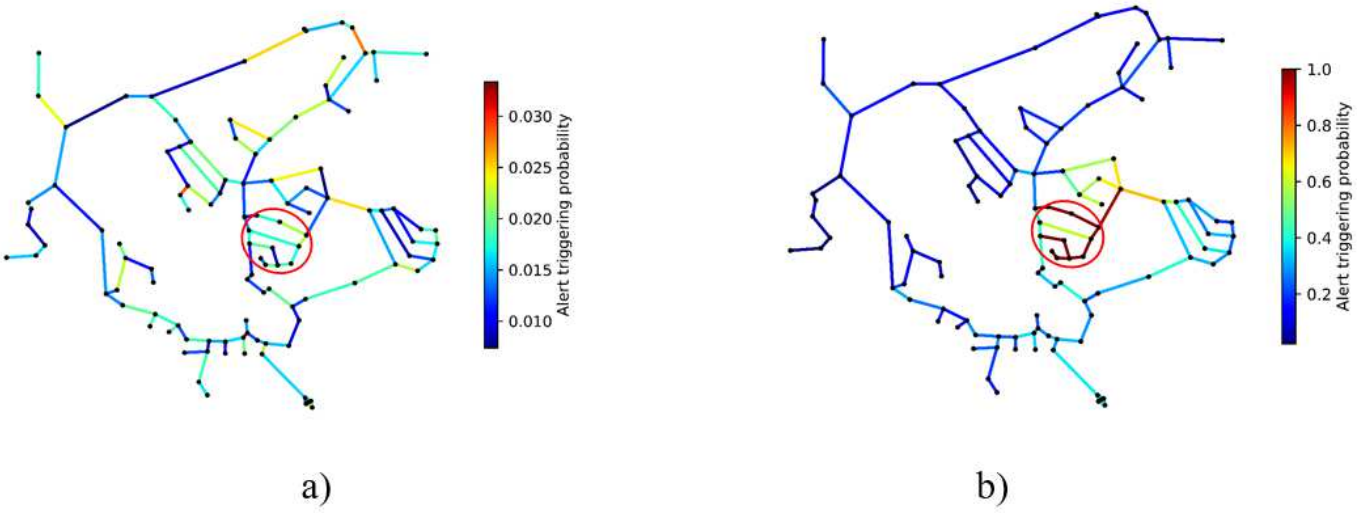


Figure 11

Probability of AE model triggers leak by individual water pipe in the WSN under a) non-leaking condition, b) leak condition (encircled are the monitoring area where water pressure data is collected)

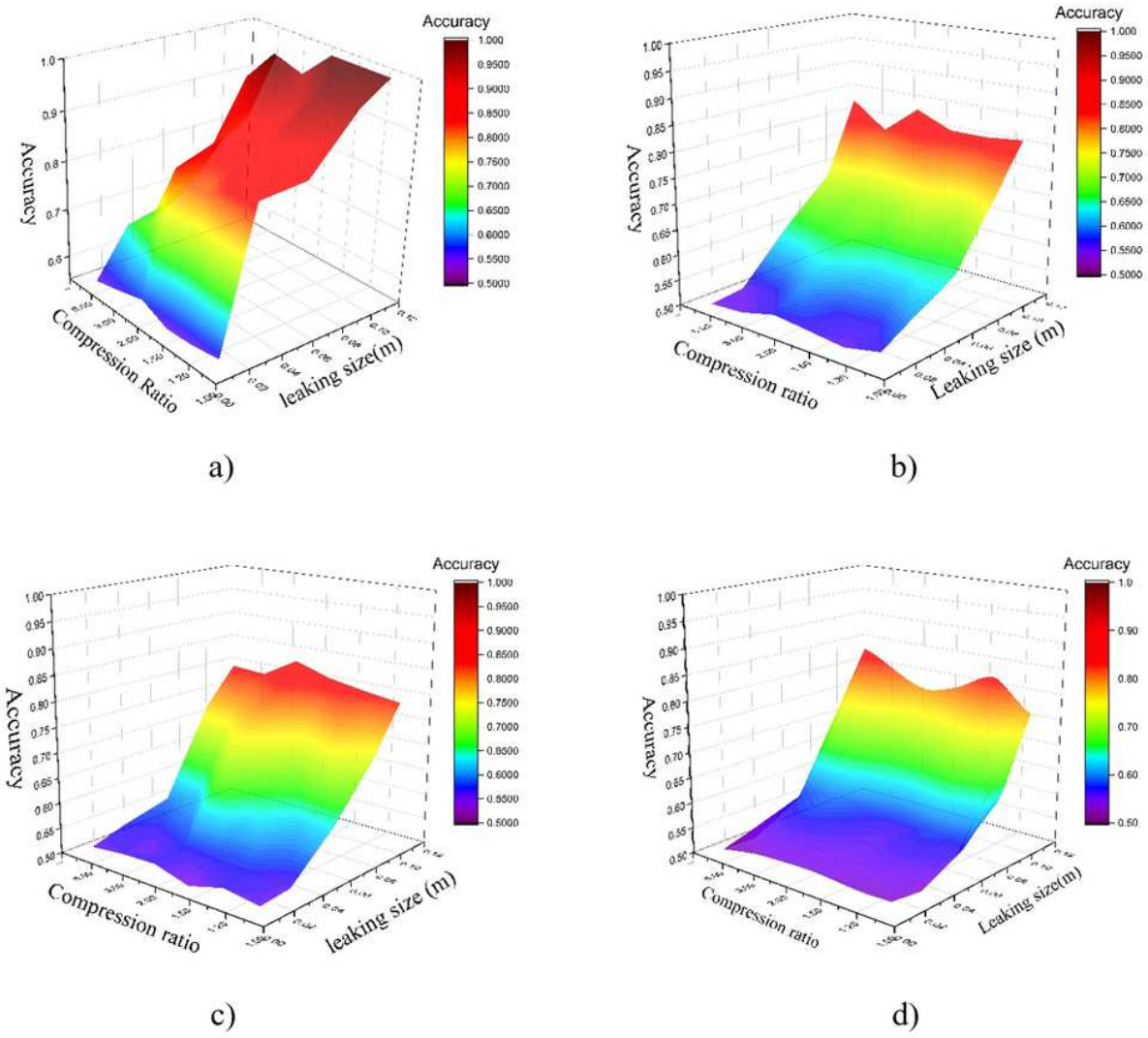


Figure 12

The sensitivity of leak detection accuracy by AE model on the compression ratio, leak size, and water usage uncertainty: (a) water usage uncertainty of N(0,0.001) L/s; water usage uncertainty of N(0,0.005) L/s, (c) water usage uncertainty of N(0,0.01) L/s, (d) water usage uncertainty of N(0,0.015) L/s

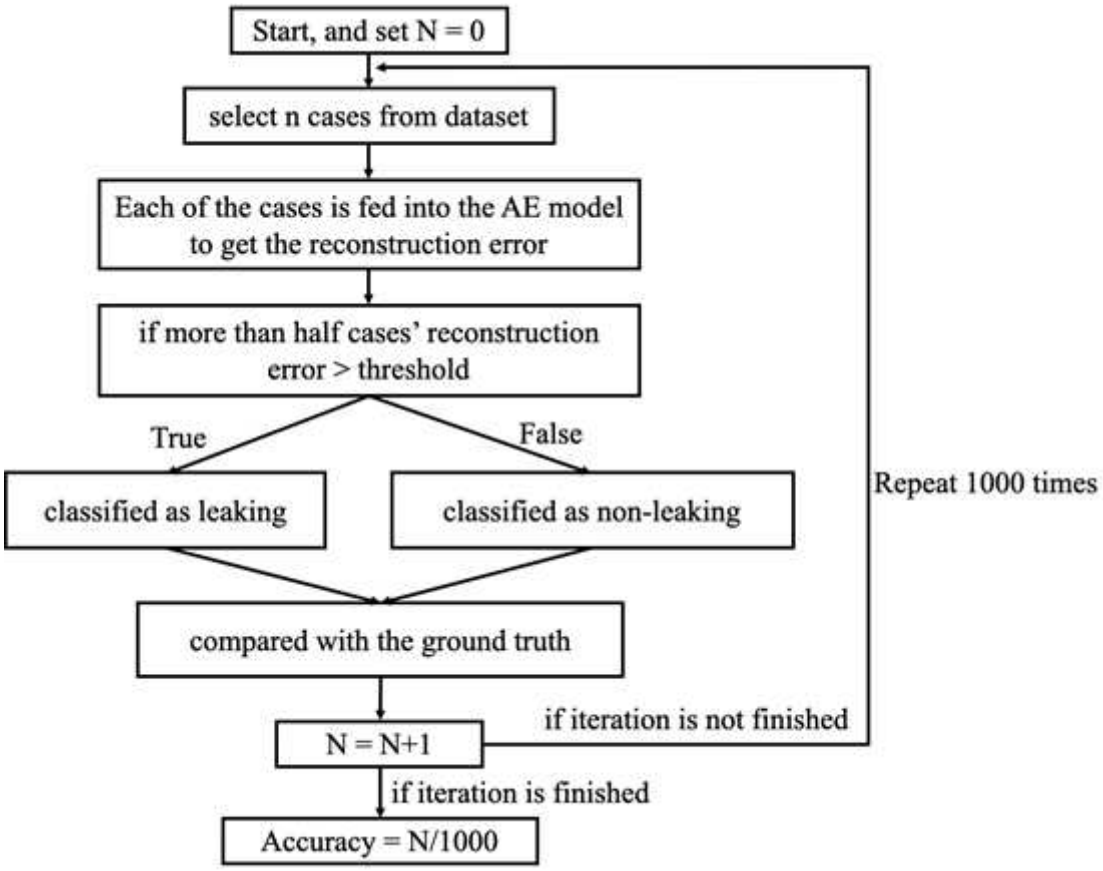


Figure 13

Flow chart of evaluation process with n attempts

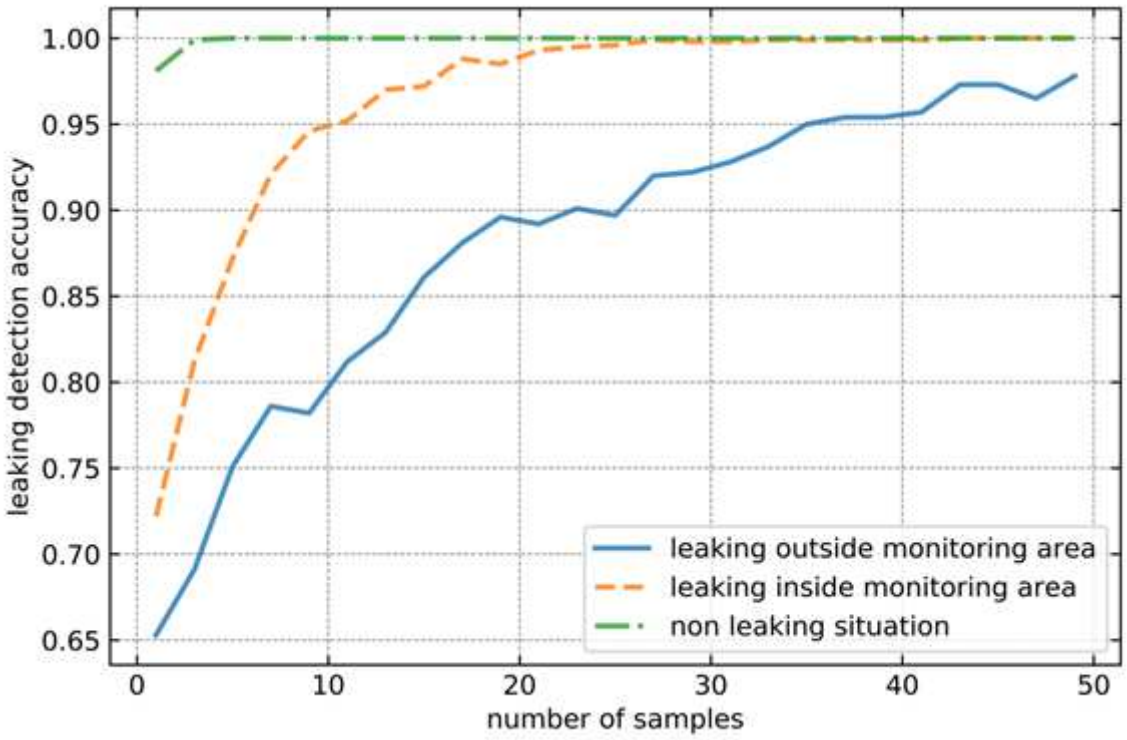


Figure 14

Accuracy of correct pipe condition detection with multiple attempts with pre-trained AE model with threshold set as 0.000402

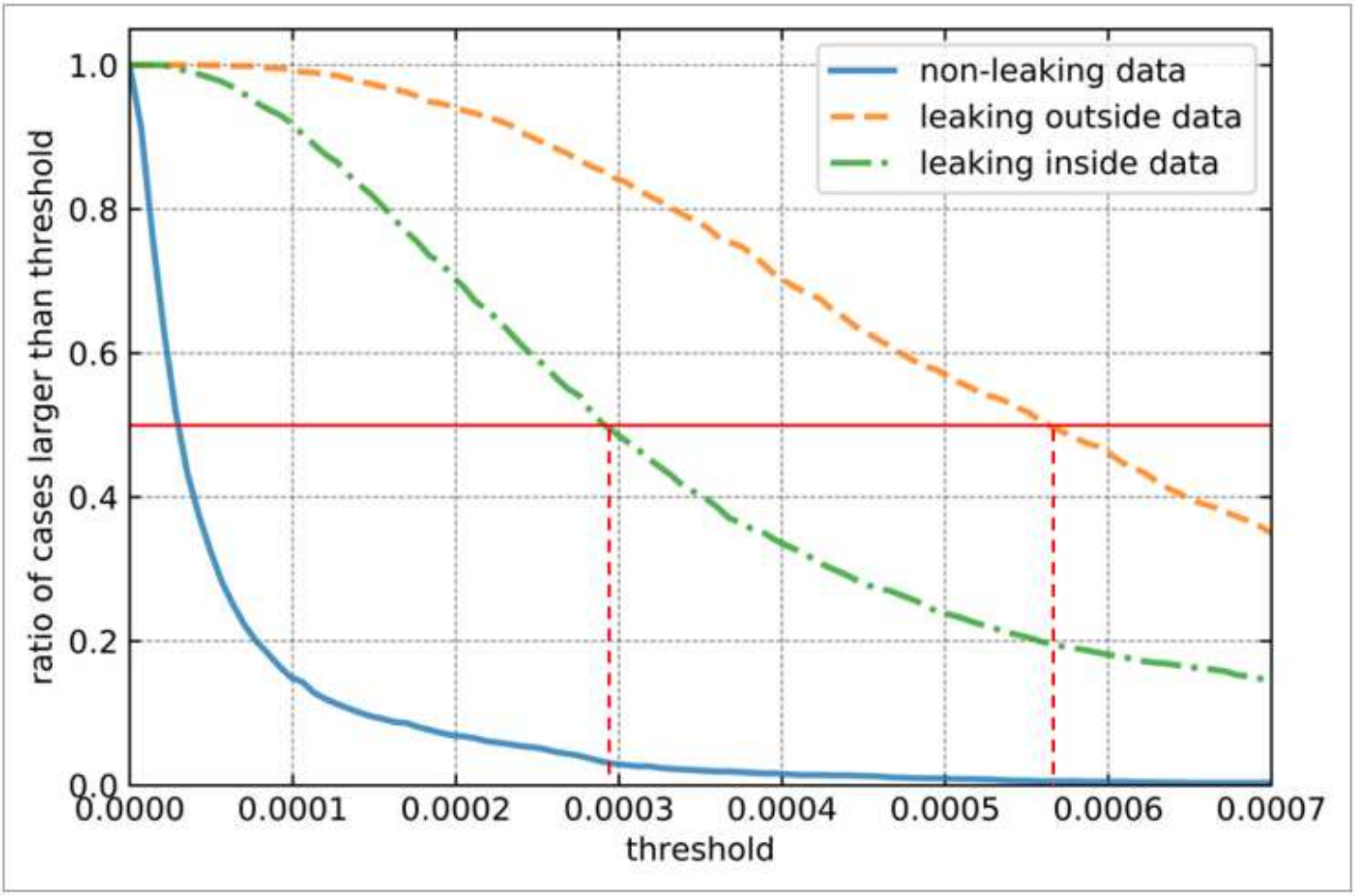


Figure 15

Percent of leak warning at different detection thresholds of AE model under different pipe conditions

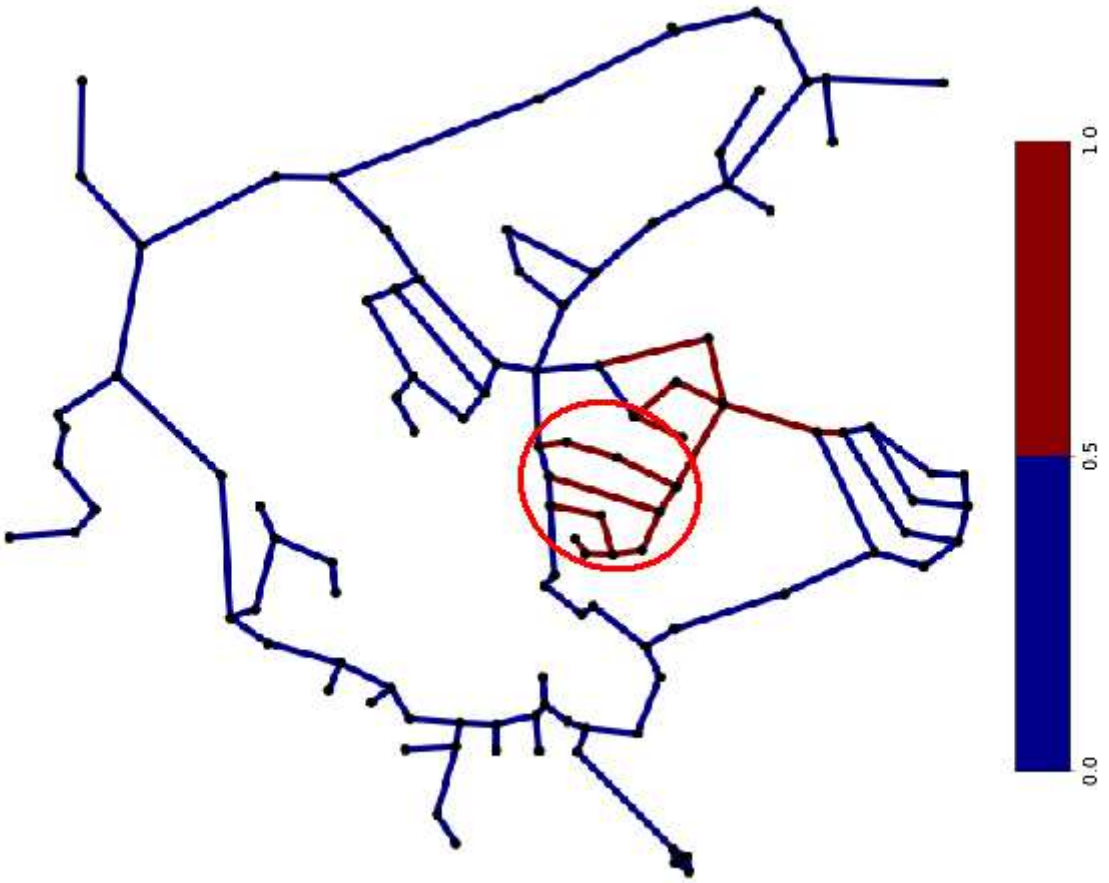


Figure 16

Final leak detection result for each pipe leaking situation

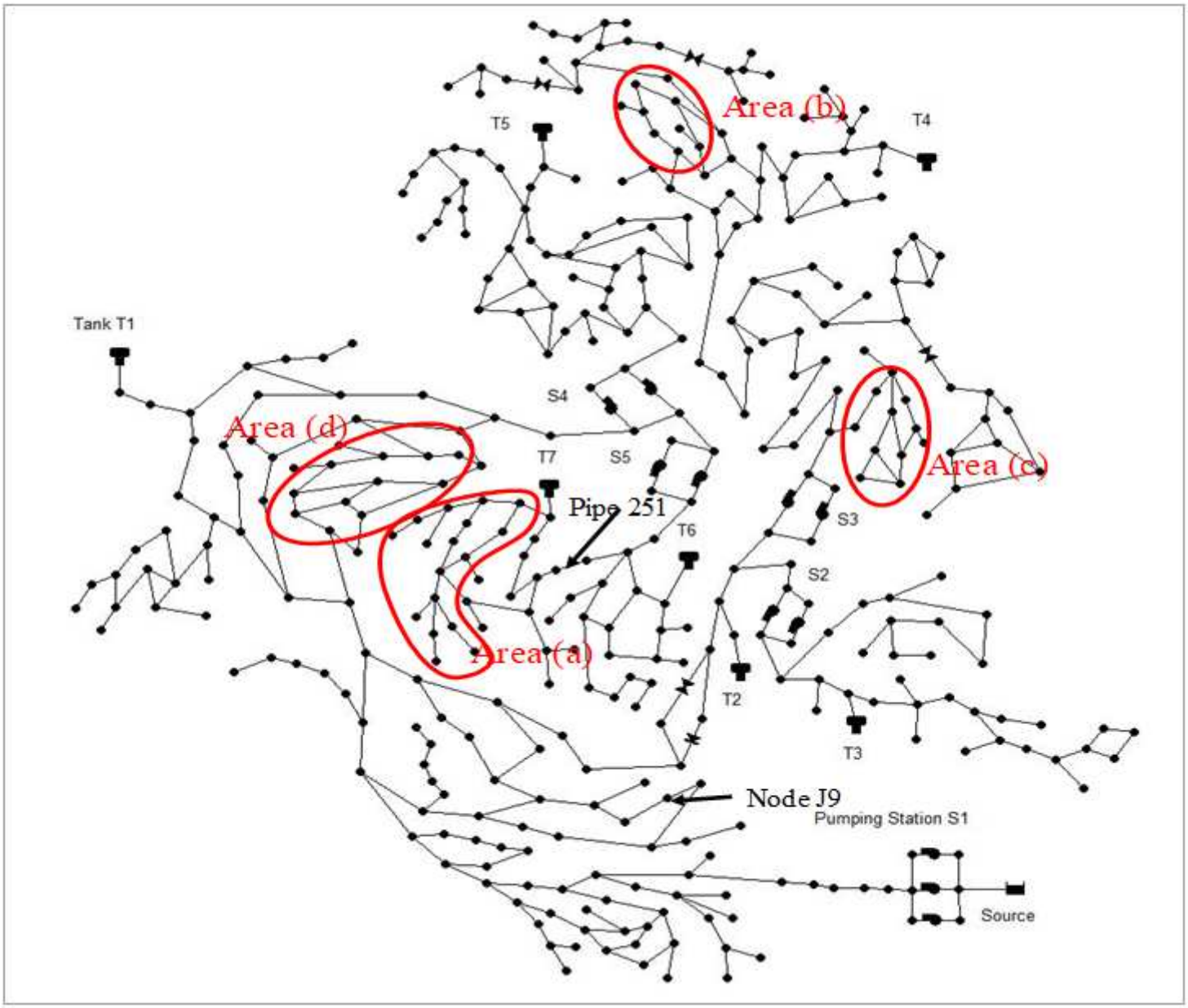


Figure 17

Topology of the C-Town WSN with DMA areas noted.

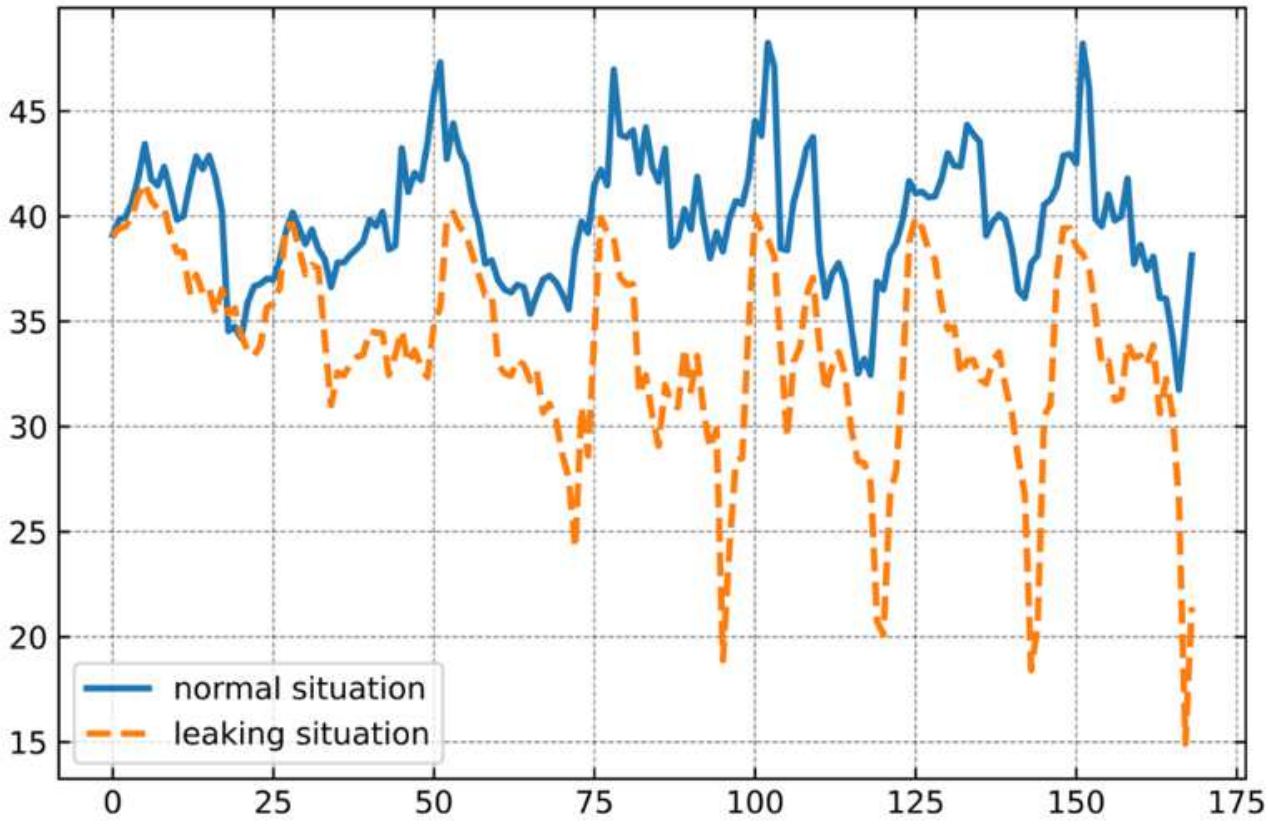


Figure 18

Water pressure (water head) at ‘J9’ with and without leaking

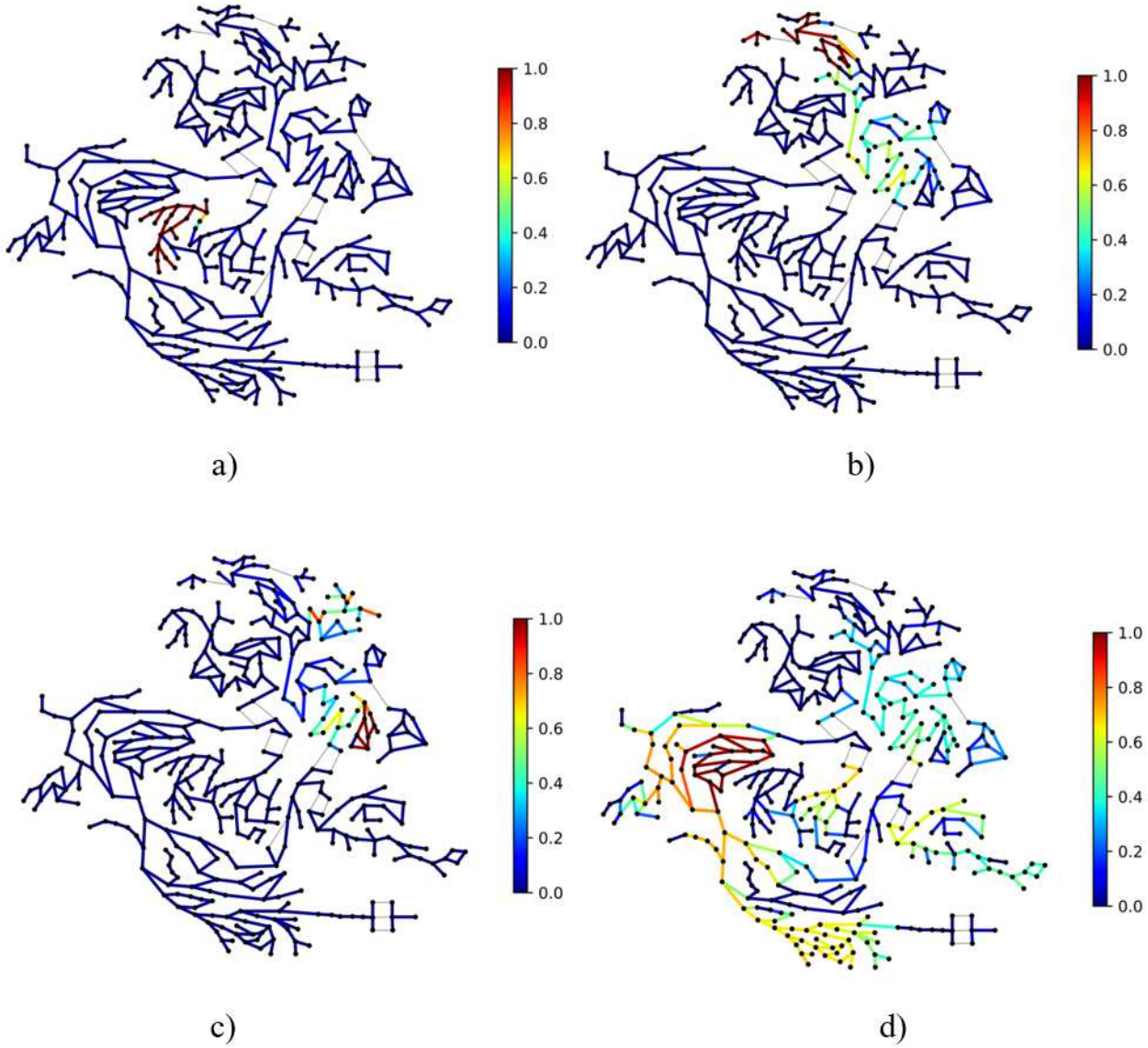


Figure 19

The probability of leak alert by AE detection model when leak happens on pipe in the WSN: a) monitoring sensors located in area a of C-town, c) monitoring sensors located in area b of C-town, c) monitoring sensors located in area c of C-town, d) monitoring sensors located in area d of C-town

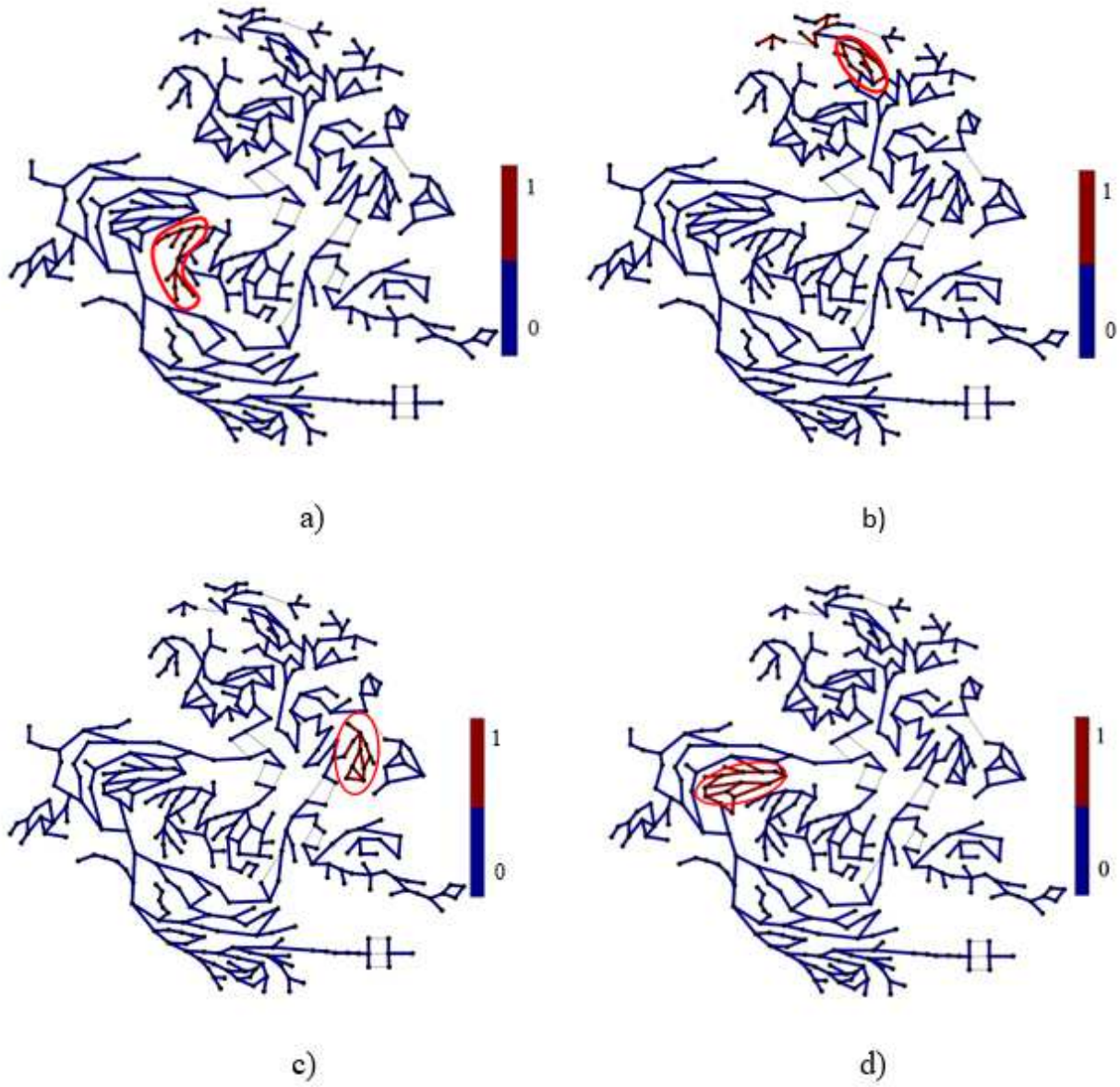


Figure 20

The probability of leak alert by AE detection model incorporating multiple attempts strategies for leak happens on pipe in the WSN: a) monitoring sensors located in DMA 4, c) monitoring sensors located located in DMA 2, c) monitoring sensors located in DMA 3, d) monitoring sensors located in DMA 1