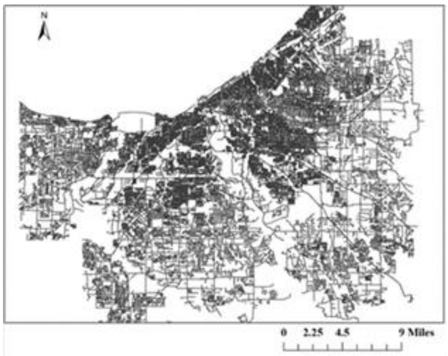
# AI-Platform for Accurate Water Supply Network Leakage Detection and Localization




## 07/29/2024

## Table of Contents

## The Team

**Xiong (Bill) Yu –** Opal J. and Richard A. Vanderhoof Professor and Chair**,** Department of Civil and Environmental Engineering, Case Western Reserve University

**Expertise:**

Intelligent Infrastructure, Smart Community, Infrastructure Health Monitoring, Sensor and Sensing System, AI for Infrastructure Decision Support

**Roles and Responsibility:**

Team lead, Problem Statement and Scope Definition, Results Interpretation, Technical Report

**Ayomide (Zul) Kazeem –** Ph.D. student**,** Department of Civil and Environmental Engineering, Case Western Reserve University

**Expertise:**

Infrastructure Graph Network, ML/AI, Data Analytics, Large Language Model

**Roles and Responsibility:**

Algorithm Development, Programming, Report Drafting

**Xudong Fan -** Postdoctoral researcher, Department of Civil and Environmental Engineering, Princeton University.

**Expertise:**

Hydraulic Modeling, Artificial Intelligence, and Deep Learning

**Roles and Responsibility**

Algorithm Development, Programming, Report Drafting

**Allen Yu –** High school students, Beachwood High School

**Expertise:**

Programming, GUI Design, Data Analyses

**Roles and Responsibility:**

Webpage Design, Programming.

**Alex Margevicius –** Commissioner**,** Cleveland Water Department

**Expertise:**

Water System Management, Hydraulic Modeling, System Engineering

**Roles and Responsibility:**

Utility Partner, Problem Statement and Scope, Results Interpretation.

## 1. EXECUTIVE SUMMARY

The water supply network (WSN) plays an essential community service but its susceptibility to leaks. Nationwide, around 15% treated water are non-revenue water. The loss is more severe for legacy water systems such as the Cleveland Water System, where up to 30% of treated water are non-revenue water. On average, five water main breaks occur in the Cleveland Water System on a typical day, underscoring the urgent need for an effective leak detection solution. Despite of tremendous efforts and progresses, significant challenges remain for early identification of leaks in the WSN.

To address this pressing issue, we have developed a cutting-edge AI-empowered application platform for accurate WSN leakage detection and localization. State of the art AI algorithms are developed in light of unique data features of WSNs, i.e., lack of balanced or labeled data under leaking conditions. The application features a user-friendly application interface that amalgamates WSN hydraulic simulations, water supply network graph analyses & optimization of sensor placements, water pressure query, water pressure monitoring, and advanced AI-powered leak detection. Behind the scenes, state-of-the-art AI algorithms empower the application's capabilities. These include an extended EPANET hydraulic simulator with model updating capabilities forming the backbone to enable digital twin of the WSN, a unique clustering algorithm to guide the optimal deployment of monitoring sensors, an AutoEncoder (AE) Neural Network enabling efficient leak detection, and a Random Forest Algorithm enabling accurate leak localization. Supplementary functions such as pressure query and sensor data management are integrated to the application interface.

Our results indicated that the AI-based leak detection algorithm achieves between 80% to 90% success rate in leakage detection and up to 90% accuracy in leakage localization. Besides, as with AI models, the performance of the system improves over time when more labeled training data of leakage ground truth is available from the engineering team. Techonomic analyses indicate that with the current leakage detection accuracy offered by this AI platform, over $7M revenue associated with non-revenue water can be recouped for the CWD due to early action on leakage reduction. The leak detection app is designed with generality and extensibility in mind and can be real time deployed to utilities with different sizes of WSN and will help these utilities to achieve sustainable operation.

## 2. PROBLEM STATEMENT AND TECHNICAL BACKGROUND

Water supply system provides one of the most essential services for our communities. The United States is served with 171,693 water supply systems, which provide essential water service to 286 million citizens. The water supply systems in Ohio provide services to 4,400 communities. Cleveland Water Division, a major water supplier in Northeast Ohio, provides water service to 88 communities.

The Nation's investment in water infrastructure has created 1.2 million miles of water supply mains, which is 26 times of the total mileage of the interstate highway system. As the rule of Nature, water infrastructure inevitably deteriorates as they age. Our communities experience between 21 to 27 breaks per 100 miles of pipeline per year, up by 27 percent over the past few years. Nationwide, around 15% treated water is leaked into the ground, referred as the non-revenue water [1]. Nearly 1 trillion gallons of drinking water are wasted annually due to water pipe leakages, causing around $2.6 billion lost revenue per year. The loss is more severe for legacy water systems serving a high proportion of underserved communities, i.e., the percentage of non-revenue water are 60% for Jackson, MS, 70% for Detroit, MI, 40% for Chicago Suburbs, IL; 80% for Georgia systems [2].

For legacy water system such as the City of Cleveland Water System, where up to 30% of treated water are non-revenue water. The direct cost to Cleveland Water Division due to water leaks is estimated to be over $10 million in 2021. Leakage from drinking water pipes is also a global problem, with on average of 50% non-revenue water in the water supply systems. The leakages lead to significant amount of lost revenue that otherwise could have been utilized for public services, which are crucially needed for underserved communities.

Many factors can cause leaks, such as pipe corrosion, aging, defects and inappropriate installation [2]. A detailed discussion is presented about the cause of water main failures by [3]. 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 [4], 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 possible damage to property. Technologies to detect leakage and forecast water pipe failure will enable agencies to evolve into preventative strategies with significant potential socio-economic and public health benefits.

A significant number 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 [5]. The first two technologies require to use specialized mobile inspection equipment with optical, acoustic, or electromagnetic sensors [6-8], 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 [9]. Technology such as ground penetrating radar can detect leak around pipe but requires heavy human involvement in signal analyses [10-12]. 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 [13]. 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 [14-16] 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 [17], image processing [18], and event forecasting, etc. [19]. Data-driven leak detection, which is based on learning from historical data with statistical or pattern recognition algorithms, is emerging [20]. 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 [21-23]. Other data such as monitored acoustic sensor data was found significantly affected by the environmental noise and limited transmission distance [24, 25]. 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 [26-28]. 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 [29, 30]. Implementation of these technologies, however, indicated there are rooms for further improvements to the agency expectations.

Table 1 summarizes the competitive technologies and the technology area our team aim to make a transformative impact.

**Table 1** summary of competitive technologies for water supply network leakage detection

| Technology | Labor demand (equipment/analyses) | Service Interruption | Cost | Robustness | Accuracy | Overall Rating |
|---|---|---|---|---|---|---|
| **Acoustic Sound detection** | Low/High | No | High | Impaired by environmental noise by traffic | Moderate | +++ |
| **Vibration Monitoring** | Low/High | No | Moderate | Impaired by ambient vibration by vehicles | Moderate | ++ |
| **Ground Penetration Radar** | High/High | No | High | Affected by hydrogeology (ground water) | Moderate | + |
| **In-pipe Robots*** | High/Moderate | Yes | High | Depends on pipe topology | Depends upon sensor modality | ++ |
| **AI-water use analytics** | Low/Low | No | Low | Immune from environmental noise | Leak detection(moderate) Localization (low) | +++ |
| **Proposed technology** | Low/Low | No | Low | Immune from environmental noise | High | ++++ |

## 3. AI-PLATFORM FOR WSN LEAK DETECTION AND LOCALIZATION

**Overview**: Our solution aims to fully leverage the potential of real-time data acquisition systems combined with advanced machine learning (ML) to achieve high sensitivity, accuracy, speed, and reliability in detecting the leaks in the water systems [41, 42] (Figure 1). A user-friendly water system monitoring application seamlessly integrates a variety of functional modules such as 1) hydraulic simulations via EPA standard, 2) water system graph analyses and optimization of sensor placements, 3) pressure monitoring or querying, and 4) AI-algorithms for leak detection and localization.

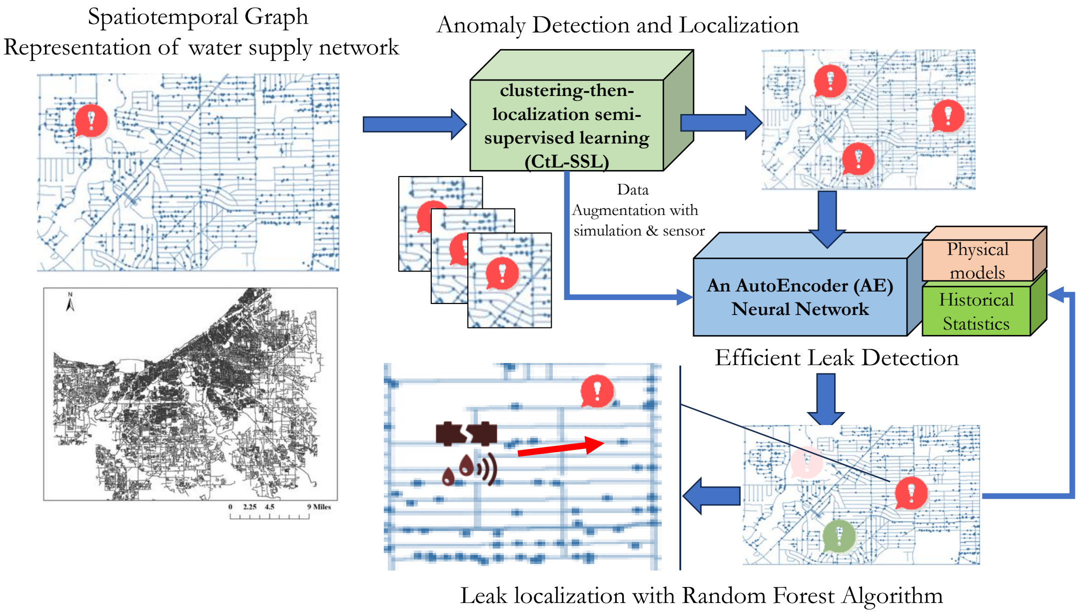



**Figure 1.** Schematic of the workflow of the AI-algorithm for leakage detection and localization in the water supply

Figure 2 illustrates the flowchart depicting the key functions of the AI-based leak detection system. As illustrated in Figure 2, the system builds on *digital twins of WSN* that acquires and manages the hydraulic information from both hydraulic simulation and real-time sensor monitoring data by the SCADA. Enabling to generate and update leak data, optimize sensor placements, and effectively train the AI models. The system provides a novel clustering algorithm to recommend the optimal locations for sensor placements, which provides a cost effective way to monitor the whole water system with minimum number of sensors (i.e., the deployment of sensor to cover the water system as denoted by the red dots). Data for AI model training leverages both the simulation data and real time sensor data. The hydraulic model is based on EPANET which is used to generate different scenarios, including different leak scenarios and normal service scenario (details of the hydraulic simulator and leakage simulation is described in Section 4). A novel cluster algorithm considering WSN topology and hydraulic similarities is used to determine *optimization of sensor placements* (details elaborated in Section 3.1). Our state of the art *AI models for leak detection and localization* and their training processes are described in Section 3.2 and 3.3.

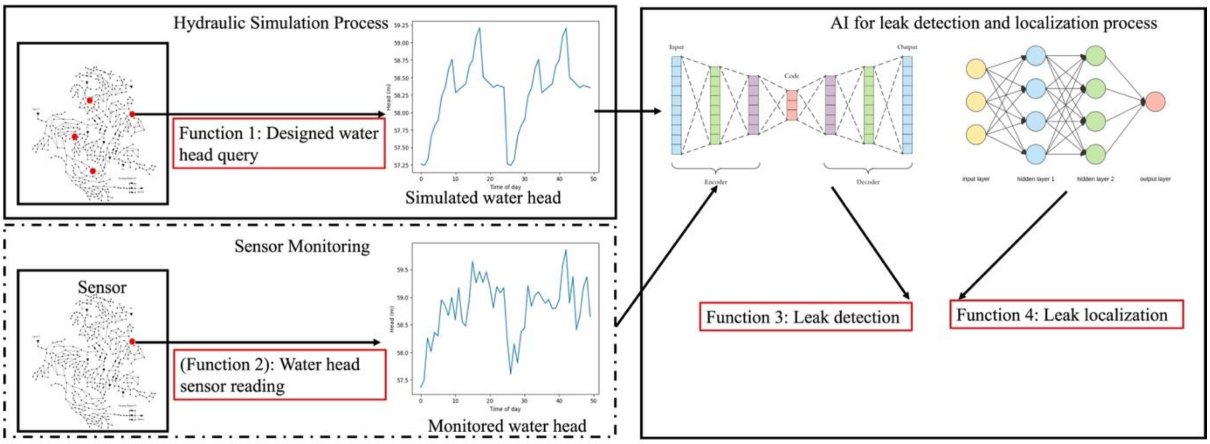



**Figure 2**. Overview of designed intelligent system application (integration of digital twins with functions such as 1) hydraulic simulations via EPA standard, 2) water system graph analyses and optimization of sensor placements, 3) pressure monitoring or querying, and 4) state of the art AI-algorithms for leak detection and localization)

**Generality and Extensibility**: Our application design incorporates ***generality*** and ***extensibility*** in mind, that allows to incorporate further development of AI leak detection technologies and to allow it to be easily adapted to different water systems.

As an example, lack of real-world sensing data put a constraint on the AI model training, however, our digital twin application, which incorporate both simulated data and real-time monitoring data, are fully prepared for inclusion of more labeled leak data in the future. The robustness and accuracy of the AI-application is ready to embrace more real-world data collected by physical sensor deployments. This will further enhance the accuracy and efficacy of this AI-based leak detection methodology.

The strong generality of our app allows it to be deployed to various water utilities. Our team aspires to equip water utilities with a state-of-the-art tool that empowers them to proactively detect and timely address leaks related asset management issues. By integrating the potential of real-time data and cutting-edge AI, our solution aims to enhance the resilience and efficiency of water supply networks, ensuring a sustainable and reliable water distribution system for our communities.

**App Interface**: Figure 3 presents an inviting glimpse of the application's welcome page. This interface is designed to provide users with a seamless experience. Different water utilities can easily upload their hydraulic models in the widely recognized standard 'inp' file format, compatible with the EPANET simulation package.

The welcome page also includes three important functions that cater to the needs of water system management, including hydraulic simulation, sensor placement recommendations, pressure monitoring, and leak detection, making it a comprehensive solution for hydraulic system analysis.

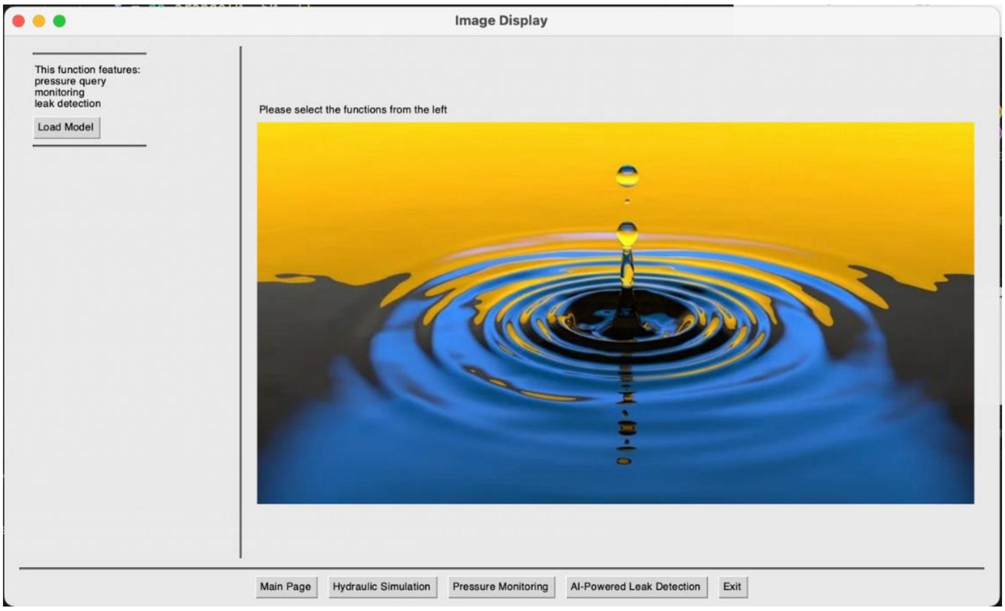



**Figure 3.** The welcome page of the intelligent leak detection app

The remaining documents are organized according to the major functions of the AI platform app. Section 3.1 introduces the hydraulic simulation function for WSN. Section 3.2 briefly describes our novel clustering algorithm for optimal sensor deployment and real-time water pressure monitoring. Section 3.3 introduces the details of our state of the art AI water leak detection and localization models. Section 4 discusses the results of preliminary technomics analyses with WSN under the jurisdiction of CWD. Section 5 provides the final summary and discussion.

### 3.1. Hydraulic Simulation and Water Pressure Query

**Overview of this functional module:** *Digital twins based on holistic WSN hydraulics* provides an effective way to diagnose WSN conditions from the behaviors hydraulic flow. Therefore, this function is an integral part of our application platform. The platform design leverages the industry standard WSN hydraulic simulation models and provides user friendliness.

#### 3.1.1 Interface of hydraulic simulation

Figure 4 shows the example interface of the *hydraulic simulation and query* function. It allows to load any WSN based on EPANET standard input format. The hydraulic simulation is conducted every time when a new hydraulic model is uploaded or modified. The simulation time step is dependent on the settings defined in the hydraulic model. The user interface also includes a water pressure inquiry function, where the user can select to display the water head variations at any nodes in the WSN. An example is shown in Figure 4.

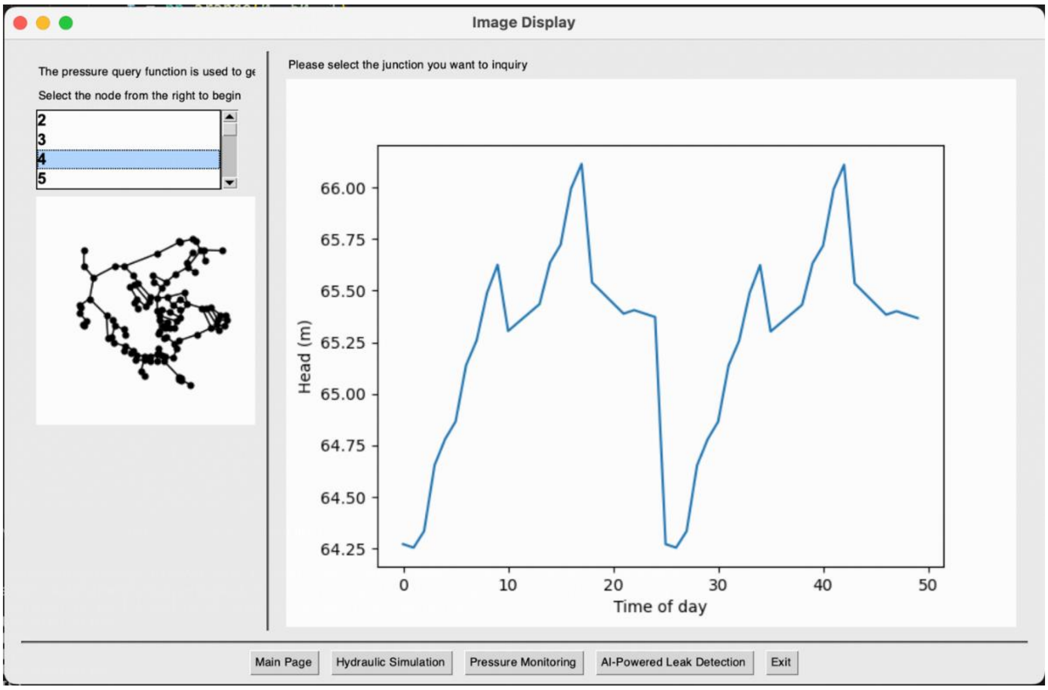



**Figure 4.** Illustrate of automatic hydraulic simulation and query based on automation of EPANET simulation flow with python

### 3.1.2 Hydraulic model

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 [31], 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, which augment real monitoring data from the WSN, 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.

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

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 [32], 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 [33], 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 [34]. A *pressure-driven water demand model* is used in our framework to consider the effects of losing pressure due to change of water demand or leaks.

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

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 leakages are 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) [35].

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

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 $\P$ is the discharge coefficient, which is 0.5 for steel pipe, and $r$ is water density.

### 3.2. Optimization of Monitoring Sensor Placement and Management Monitoring Data

**Overview of this functional module:** in the deployment of monitoring sensors, water agencies, most face financial and manpower constraints, typically raise questions such as how many sensors need to be installed to cover the WSN? which locations to place the sensors to get the maximum values? and how to effectives utilizing the sensor data for WSN decisions? This functional module is designed to optimize the sensor placements and harvest the maximum value of sensor monitoring data for WSN management (i.e., leakage detection and localization, as well as WSN system condition assessment).

#### 3.2.1 AI-algorithm for optimal sensor placement

Our app includes a novel AI clustering algorithm to recommend the optimal sensor placements to achieve the maximum values of monitoring data. Figure 4 illustrates the user interface for determination of the optimal sensor placement. This user-friendly interface allows users to customize the number of sensors they wish to utilize and recommend the optimal locations to deploy sensors. The optimized sensor locations are visually represented by larger nodes, each distinguished by different colors. For instance, in Figure 4, we demonstrate a system equipped with 4 sensors, and their strategically placed locations to achieve the maximum value of sensor data collection. In addition to identifying the best sensor positions, the sensor optimization algorithm automatically groups the system into various clusters, each denoted by a unique color. These groups play a crucial role in leak localization algorithms, as they aid in narrowing down the potential leak areas. Detailed algorithms about the sensor optimization process are expanded in the subsequent paragraphs.

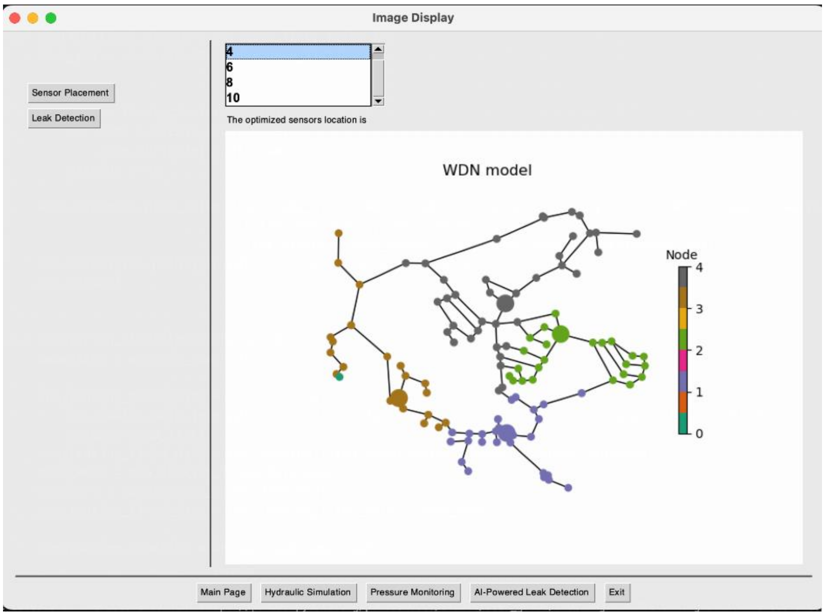



**Figure 5.** Interface that automatically run the AI clustering algorithm to determine the optimal sensor placements based on the number of sensors to be installed (in light of financial resource available etc.)

### 3.2.2 Modified *k-means* algorithm for sensor's optimal placement.

To achieve the optimal sensor placement, we have developed a unique modified *k*-means clustering algorithm for WDN partition. Compared to the standard *k*-means WDN clustering which only considers the topology (i.e., graph distance) or leakage characteristics of junctions (i.e., disturbance of water pressure pattern associated with leakage), our new cluster algorithm considers both the topology of the WSN and leakage characteristics of junctions. The pseudocode of the new *k-means* clustering algorithm is shown in the following Table 2.

**Table 2**. New k-means cluster algorithm for WDN partition and optimal sensor placement

| **Algorithm**: Modified *k-means* algorithm for WDN sensor placement |
|---|
| **step 1**: Initialize parameters: set the number of cluster *k*, tolerance, maximum iteration number<br>**step 2**: Randomly select *k* junctions from the WDN as the first group of centroids.<br>**step 3**: Data preparation<br>**step 3.1:** Prepare the leakage characteristics matrix.<br>• I.a) For the conventional leakage characteristics matrix, use Table 1.<br>• I.b) For PCA- or AE- based leakage characteristics matrix, follow Figs 2 and 3 respectively.<br>• II. Normalize the leakage characteristics matrix by dividing its maximum value.<br>**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.<br>**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 $(\nu \in J)$ to its nearest clusters based on the total Euclid distance defined as

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

$$\nu \in cluster_i, \text{ if } L_{\nu,c_i} \le L_{\nu,c_l}, \forall l \in \{1, 2, 3, \dots, k\}$$

*where $J$ is the set of all junctions, $L_{\nu,c_i}$ 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 $(v_k, c_2, c_3, \dots, c_k)$

**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.,

$$c_i^{new} = \nu_k, \text{ if } \textstyle\sum_{m=1}^{M} L_{\nu_k,\nu_m} \text{ is minimum.}$$

*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.

$$abs\left(sum(L_{\nu,c^{new}}) - sum(lL_{\nu,c})\right) < tolerance$$

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

*where $l_{\nu,c}$ 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.*

It is noted that in Step 3.1, the leakage characteristics matrix can be obtained by using different definitions, i.e., conventional leakage characteristic matrix based on pressure change or feature extraction with ML algorithms. Although Principal Component Analyses (PCA) and Autoencoder (AE) models are used in the clustering algorithm, other ML models can also be integrated into this framework, such as the Mahalanobis classification system (MCS) [36]. In Step 3.2, the physical distance between pairs of junctions is obtained by using Dijkstra's shortest pathfinding algorithm [37]. Other shortest path algorithms could also be considered when dealing with different types of graphs, such as Floyd-Warshall algorithm [38]. 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 represent distance between junctions is defined as the unweighted average of physical distance and leakage characteristics distance. The algorithm allows to assign different weights of the leakage characteristics distance and the physical distance. In Step 6, the process of centroid redistribution of each cluster requires re-acquiring 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 optimal junctions (that minimize distance within the cluster) is set as the new centroids so that centroids remain on the junctions in the WDN.

With clusters of nodes determined, the recommended sensor placements are at the centroid of each cluster to maximize the value of sensor data acquisition. The clustering of nodes based on their hydraulic similarities and network distance also facilitates improved accuracy of leak detection and localization.

### 3.2.3 Sensor data visualization

An interface has been created to manage the sensor monitoring data. The data flow stream are compatible to the data protocol of the existing SCADA system of the WSN. This allows to store, retrieve, and display sensor data according to the user demand. Automatic data quality check is implemented, which gives warning on erroneous data such as abnormal value of the data. An example display of data collected by pressure monitoring sensors is illustrated in Figure 6.

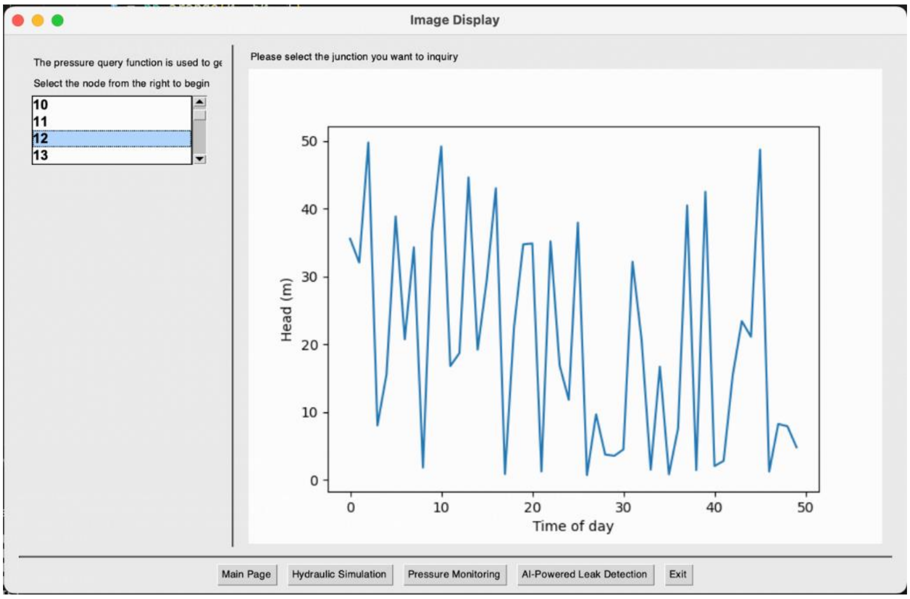



**Figure 6**. Interface to manage and display of data by water pressure monitoring sensor

### 3.3. Novel AI Models for WSN Leak Detection and Localization

**Overview of this functional module:** Data-driven AI models features unique advantages in detecting abnormal phenomena in the WSN when physical models fail due to complex contributing factors. However, in developing AI model for WSN leak detection purpose, agencies typically face paradox situations with unbalanced dataset (due to that WSNs mostly work in normal conditions and leakage events are relatively rare events) or uncertainty with leakage locations (or data are unlabeled). This means common successful AI algorithms based on balanced labeled data won't work reliability. Our team advance the state of the art by developing unique AI algorithms for leakage detection and localizations.

#### 3.3.1 Autoencoder ML Model for Accurate Leak Detection with Unbalanced Data

Common AI applications requires sufficient amount of labeled data. This presents a major barrier for WSN leak detection due to relative rare amount of labeled data corresponding to leak conditions. To overcome this technical barriers, we have designed an unsupervised ML model with Autoencoder (AE) neural network. The AE neural network model 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 required to learn the hidden patterns between the input data.

A typical architecture of the autoencoder neural (AE) network is shown in Fig 6. 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.

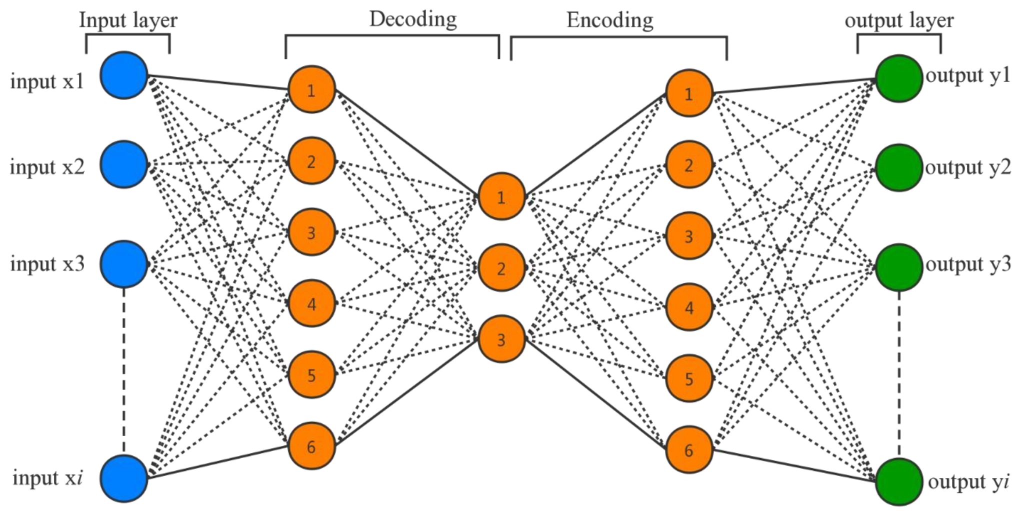



**Figure 7**. Schematic of architecture of an autoencoder neural network (the numbers of neurons in the decoding layers and encoding layers are conceptual)

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:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{6}$$

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.

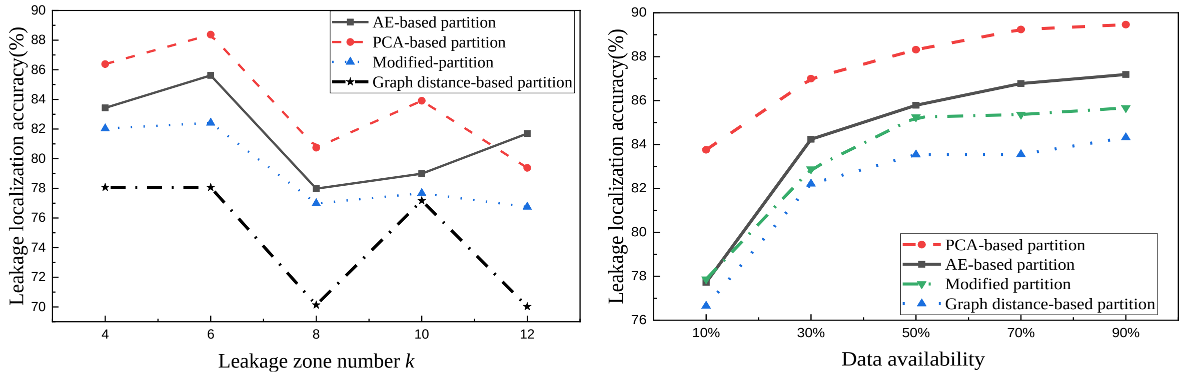
For a model trained by dataset using normal (non-leak) condition, a large reconstruction error occurs it is inputted with data under leaking condition because the relationship described by the trained AE neural network is not valid under such condition. By setting a threshold in the construction error,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

**3.3.2 Novel Random Forest ML Model for Leak Localization based on Limited Labeled Data**

For leakage localization purpose, a challenge for ML models is the amount of labeled data where accurate leakage location information is rare. This requires ML model that is primarily based on unlabeled data. For this purpose, we defined the leakage localization a classification problem, i.e., the leakage conditions are classified into different WDN partition zones (or clusters). There are various types of ML models for classification problems, such as the Artificial Neural Network, Support Vector Machine, Decision Tree, Random Forest (RF), etc. We employed Random Forest (RF) model 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.

We also validated the performance of leak localization considering the impact of number of sensors. The effects of the number of sensors on the leakage zone localization accuracy are summarized in Figure 8 (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. It is noted that the accuracy of leak localization is worst when only considers the physical distance of junctions with no consideration of the leakage characteristics, which is the case with conventional ML model. 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.

As with ML models, the more data used for model training, the better the ML performance. For the case of leak detection, the higher the percentage of nodes with historical failures, the more accurate the detection results. Figure 8 (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 that with leaking data available at more junctions, the leakage localization accuracy improves. The results also showed that the best performance in leak localization is achieved with RF-based leakage zone localization over PCA-based partition.




(a) (b)

**Figure 8.** Leakage localization accuracy with RF model under a) different numbers of partition zones, and b) percentage of junctions with leakage data available under 10 partitions

### 3.3.3 Sensitivity Analysis of the Leak Detection Algorithm

Sensitivity study has been conducted to evaluate the effects of leak size on the performance of the AI-based leak detection algorithm. The leaking size is an 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 noises such as the water demand fluctuations. For the sensitivity study, the leaking size is varied from 0.01m to 0.12m.

Small leaks tend to be 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.

Additionally, we also examined the influence the compression ratio of the AE algorithm. The compression ratio is the number of uncompressed data divided by compressed data when constructing the AE neural network. 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.

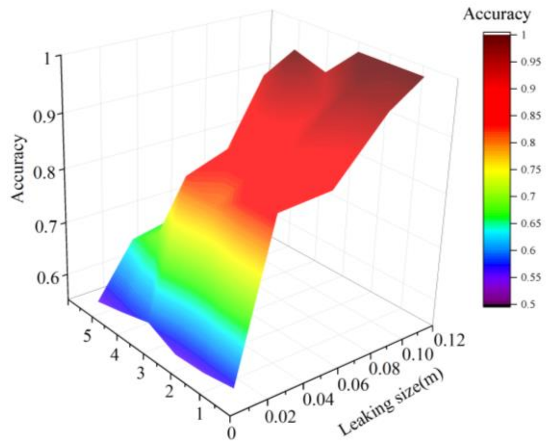



**Figure 9.** The sensitivity of leak detection accuracy over the leak size and the compression ratio of AE neural network model

The compression ratio is found to have a negative influence on the overall detection accuracy. As shown in Figure 9, 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 leaking size, compression ratio has a relatively smaller impact on the detection accuracy.

## 4. TECHNOMICS ANALYSES

Water leaks have led to significant direct and indirect costs. According to the Cleveland Water Department, the average cost for treated water is $487.53 per million gallons. In the year 2021 alone, the total annual water production reached 73,559 million gallons [39]. Considering the aging water system have an average leaking rate around 30%, the direct cost attributed to water leaks amounted to a staggering $10,758,665 in 2021. Additionally, the entire United States experiences an average daily water leak of 6 billion gallons [40]. As a result, the total direct cost is approximately $2,925,180 each day.

In order to estimate the economic savings of the proposed algorithm, it is assumed the leak size distribution follows a Weibull distribution as shown in Eq. 7, with the shape parameter (a) set to 1.5 and the scale parameter (b) to 20. Figure 10 displays the plot of the leak size distribution.

$$f(x,a,b) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} e^{-\left(\frac{x}{b}\right)^{a}} \tag{7}$$

where *a* is the shape parameter, and *b* is the scale parameter.

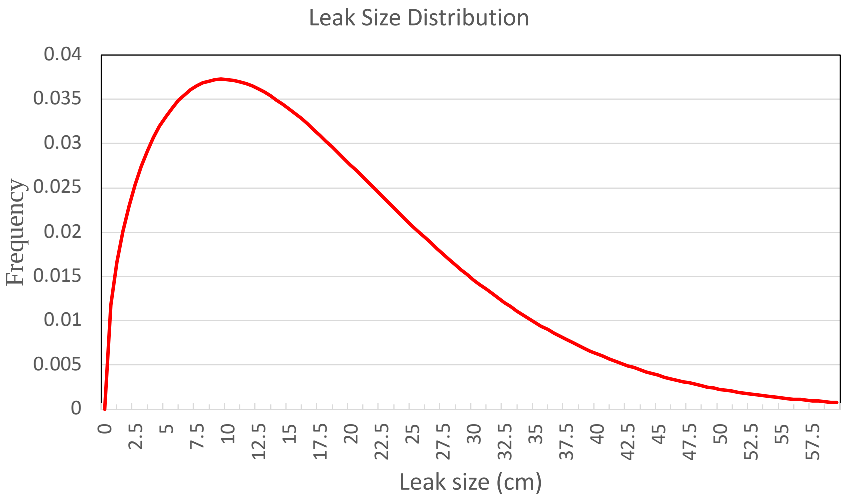



**Figure 10**. Leak size distribution of a water system

Upon applying the accuracy of developed leak detection algorithm to different sized leaks (Figure 9), the probability of water leak prevention can be estimated. Given the assumption of the leak size distribution (Figure 10) and the leak detection accuracy (Figure 9), the proposed method can prevent 68.4% water leak if the detected leaks can be timely fixed. For water system with water supply such as the Cleveland water system, the potential economic savings is estimated to be $ 7,358,926.86 per year.

Besides the financial benefits to recoup the lost revenue, prevention of water leaks also help to mitigate the health risks to the public and improve public perception. Contamination from external sources can enter the water system in the event of leaks and pipe failures, particularly when the pipe's internal pressure is lower than the pressure by the groundwater. Efficient leak detection and mitigation measures can significantly reduce the chance of backflow, thereby minimizing the risks to public health.

## 5. SUMMARY

Our team presents a solution to the intelligent water supply challenge (IWS) by developing an innovative AI platform to accurately detect and localize leakages in the WSN. The platform starts with a novel AI clustering algorithm that guide the optimal sensor placements. This allows to use a limited number of sensors to completely cover the conditions of the whole WSN, therefore, maximizing the values of the sensor data. This AI module meets water agencies needs due to financial constraints.

Our innovation is further driven by the state-of-the-art AI algorithms for accurate leak detection and localization. These algorithms are designed in light of the unique challenges of data-driven AI approach for WSN, i.e., lack of balanced data or labeled data. Our novel ML model allows leaks to be detected without a vast amount of labeled dataset under leak scenarios. The unsupervised AI algorithm learns the patterns of non-leaking scenarios, making it a highly efficient and reliable approach adapted to address the unique challenge in WSN leak detection. Besides, our ML model is based on learning the changes in the water pressure patterns among multiple sensors (rather than pressure differences from a single sensor as with many existing ML models). Therefore, it achieves much more robust and reliable results.

Besides, our platform includes semi-supervised ML approach for leak localization. The significant of this approach is that it only requires data at a small portion of nodes with documented leakage history to achieve accurate locate leakages across the whole WSN. This makes it an effective solution under practical constraints with WSN operations.

Our AI platform application features user-friendly interface. It is designed with excellent extensibility and generality in mind for WSN of different sizes. The system can be integrated with data from the existing water SCADA system, provides automatic hydraulic simulation via EPANET, sensor data management and query functions, as well as the ML models to accurately detect and localize the leakages.

The implications of this innovation are vast, as it will empower public utilities to promptly identify leakages or other factors leading to pipe failures. By supporting effective and timely maintenance measures, this system will recoup the economic values associated with the vast amount of non-revenue water and significantly reduces the risk of leak-related health issues. This presents a cutting-edge solution to support sustainable intelligent water system management.

**Reference**


1. Sadeghioon, A., et al., *SmartPipes: smart wireless sensor networks for leak detection in water pipelines.* Journal of sensor and Actuator Networks, 2014. **3**(1): p. 64-78.
2. AP News, https://apnews.com/article/water-loss-infrastructure-broken-pipes-poor-neighborhoods-2d747180d294ba62cdbf0906f9305802
3. Sadiq, R., B. Rajani, and Y. Kleiner, *Probabilistic risk analysis of corrosion associated failures in cast iron water mains.* Reliability Engineering & System Safety, 2004. **86**(1): p. 1-10.
4. Walski, T.M. and J.W. Male, *Maintenance and rehabilitation/replacement.* 2000, McGraw-Hill. p. 17.1-17.28.
5. Chan, T.K., C.S. Chin, and X. Zhong, *Review of Current Technologies and Proposed Intelligent Methodologies for Water Distributed Network Leakage Detection.* IEEE Access, 2018. **6**: p. 78846-78867.
6. Kang, J., et al., *Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems.* IEEE Transactions on Industrial Electronics, 2017. **65**(5): p. 4279-4289.
7. Ozevin, D. and H. Yalcinkaya. *Reliable monitoring of leak in gas pipelines using acoustic emission method.* in *Proc. Civil Struct. Health Monit. Workshop (CSHM).* 2012.
8. Gao, J., et al., *Monitoring the stress of the post-tensioning cable using fiber optic distributed strain sensor.* Measurement, 2006. **39**(5): p. 420-428.

9. Butler, D., *Leakage Detection and Management: A Comprehensive Guide to Technology and Practice in the Water Supply Industry*. 2000: Palmer Environmental.
10. Amran, T.S.T., et al. *Detection of underground water distribution piping system and leakages using ground penetrating radar (GPR)*. in *AIP Conference Proceedings*. 2017. AIP Publishing LLC.
11. Bimpas, M., A. Amditis, and N. Uzunoglu, *Detection of water leaks in supply pipes using continuous wave sensor operating at 2.45 GHz*. Journal of Applied Geophysics, 2010. **70**(3): p. 226-236.
12. De Coster, A., et al., *Towards an improvement of GPR-based detection of pipes and leaks in water distribution networks*. Journal of Applied Geophysics, 2019. **162**: p. 138-151.
13. Srirangarajan, S., et al., *Wavelet-based burst event detection and localization in water distribution systems*. Journal of Signal Processing Systems, 2013. **72**(1): p. 1-16.
14. Mashford, J., et al., *Leak detection in simulated water pipe networks using SVM*. Applied Artificial Intelligence, 2012. **26**(5): p. 429-444.
15. Colombo, A.F., P. Lee, and B.W. Karney, *A selective literature review of transient-based leak detection methods*. Journal of hydro-environment research, 2009. **2**(4): p. 212-227.
16. Adedeji, K.B., et al., *Towards achieving a reliable leakage detection and localization algorithm for application in water piping networks: An overview*. IEEE Access, 2017. **5**: p. 20272-20285.
17. Liao, S.-H., P.-H. Chu, and P.-Y. Hsiao, *Data mining techniques and applications–A decade review from 2000 to 2011*. Expert systems with applications, 2012. **39**(12): p. 11303-11311.
18. Buch, N., S.A. Velastin, and J. Orwell, *A review of computer vision techniques for the analysis of urban traffic*. IEEE Transactions on Intelligent Transportation Systems, 2011. **12**(3): p. 920-939.
19. Lin, W.-Y., Y.-H. Hu, and C.-F. Tsai, *Machine learning in financial crisis prediction: a survey*. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2011. **42**(4): p. 421-436.
20. Romano, M., Z. Kapelan, and D.A. Savić, *Automated detection of pipe bursts and other events in water distribution systems*. Journal of Water Resources Planning and Management, 2012. **140**(4): p. 457-467.
21. Chim, T.W., et al., *SPECS: Secure and privacy enhancing communications schemes for VANETs*. Ad Hoc Networks, 2011. **9**(2): p. 189-203.
22. Kim, J.-H., et al. *SPAMMS: A sensor-based pipeline autonomous monitoring and maintenance system*. in *2010 Second International Conference on COMmunication Systems and NETworks (COMSNETS 2010)*. 2010. IEEE.
23. Stoianov, I., et al. *PIPENETa wireless sensor network for pipeline monitoring*. in *Proceedings of the 6th international conference on Information processing in sensor networks*. 2007.
24. Loth, J.L., et al., *Acoustic detecting and locating gas pipe line infringement*. 2004, West Virginia University (US).

25. Srirangarajan, S., et al., *Water main burst event detection and localization*, in *Water Distribution Systems Analysis 2010*. 2010. p. 1324-1335.

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

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

28. Zhou, X., et al., *Deep learning identifies accurate burst locations in water distribution networks.* Water Res, 2019. **166**: p. 115058.

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

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

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

32. Twort, A.C., D.D. Ratnayaka, and M.J. Brandt, *Water supply.* 2000: Elsevier.

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

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

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

36. Cheng, L., et al., *Quality inspection of complex-shaped metal parts by vibrations and an integrated Mahalanobis classification system.* Structural Health Monitoring, 2020: p. 1475921720979707.

37. Dijkstra, E.W., *A note on two problems in connexion with graphs.* Numerische mathematik, 1959. **1**(1): p. 269-271.

38. Floyd, R.W., *Algorithm 97: shortest path.* Communications of the ACM, 1962. **5**(6): p. 345.

39. *Cleveland Water 2021 ANNUAL REPORT.* 2021.

40. ASCE, *2021 INFRASTRUCTURE REPORT CARD.* www.infrastructurereportcard.org, 2021.

41. Fan, X. and Yu, X., 2022. *An innovative machine learning based framework for water distribution network leakage detection and localization. Structural Health Monitoring*, *21*(4), pp.1626-1644.

42. Fan, X., Zhang, X. and Yu, X.B., 2021. *Machine learning model and strategy for fast and accurate detection of leaks in water supply network*. Journal of Infrastructure Preservation and Resilience, 2(1), pp.1-21.