Water Practice & Technology

IWA PUBLISHING

Water Practice & Technology Vol 20 No 9, 1900 doi: 10.2166/wpt.2025.109

Leak detection in water distribution networks using micro-electromechanical systems-based accelerometers: a machine learning approach

Rongsheng Liu[a], Salman Tariq[a], Ibrahim A. Tijani[a], Harris Fan[b], Sherif Abdelmageed[c], Ali Fares [a,*] and Tarek Zayed[a]

[a] Department of Building and Real Estate, The Hong Kong Polytechnic University, Kowloon, Hong Kong
[b] Development Engineer, Applied Technology Integration Ltd, Kowloon, Hong Kong
[c] Faculty of Engineering, Ain Shams University, Cairo, Egypt
*Corresponding author. E-mail: ali-i.fares@connect.polyu.hk

AF, 0000-0002-4286-2755

ABSTRACT

Water distribution networks (WDNs) worldwide are plagued with water losses. These losses cause massive financial damage of US$39 billion per year and further aggravate the existing water scarcity situation faced by half a billion people throughout the year. To curb the damage, our research provides a machine learning-based approach for early detection of leaks in WDNs, constituting a substantial portion of water losses in most WDNs. Experiments were conducted for more than 10 months on real networks using micro-electromechanical system (MEMS) accelerometers. Leak and no-leak signals were collected from metal and non-metal pipes of different sizes. Seventeen time-domain and frequency-domain-based features were extracted using signal processing methods. The most appropriate features were ranked and selected. The selected features were then used to develop an artificial neural network (ANN) and gene expression programming (GEP) based on intelligent models for metal and non-metal pipes. Both ANN and GEP models showed an accuracy of over 99% for leak detection in metal pipes. In contrast, the accuracy in non-metal pipes reached around 89%. Our study represents one of the very few attempts made on leak detection in real WDNs, which will open new avenues of research in this domain.

Key words: artificial neural networks, gene expression programming, leak detection, machine learning

HIGHLIGHTS

- Real-world WDN ML: MEMS accelerometers detect leaks.
- Separate models for metal and non-metal pipe leak detection.
- High field accuracy: >99% metal, ~89% non-metal.
- Systematic acoustic feature selection for robust models.
- Field data advances accelerometer-based leak detection.

1. INTRODUCTION

Water scarcity is a threat to the sustainable development of societies worldwide. Realizing the severity of this threat, the World Economic Forum (2020) report on global risks listed the *water crisis* as the top-most challenging societal risk owing to its potential impact in the coming decade. Mekonnen & Hoekstra (2016) explicitly stated that 4 billion people suffer from water scarcity at least one month a year, out of which half a billion people face this issue throughout the year. Within the context of water scarcity, one of the significant difficulties faced by the urban water lifecycle is the reduction of water loss from the water distribution networks (WDNs) (Molinos-Senante *et al.* 2016). Therefore, prompt detection of leakages is necessary to minimize water loss in WDNs, reducing the severity of the water scarcity problem (Tariq *et al.* 2021).

Besides leak surveys using listening rods, which rely on the operator's experience, vibro-acoustic technologies, such as noise loggers and accelerometers, are most popularly deployed to detect leakages in WDNs due to their ease of use (Tariq *et al.* 2022). Several of these technologies are usually placed non-invasively on utility valves, hydrants, or exposed pipes using magnetic coupling or duct tape for permanent and semi-permanent monitoring of pipelines. Acceleration data collected by these technologies are then transferred to the base station/computer,

where model-based methods, such as coherence (Muntakim *et al.* 2017), cross-correlation (Hu *et al.* 2021), and singular spectrum analysis (Cody *et al.* 2018), are applied to detect leakages in real time. However, model-based methods require accurate information about (1) the system's and pipe's geometry, (2) leak mechanics, and (3) environmental and operational conditions (Cody *et al.* 2020). Obtaining such data for a complex network on a large scale is difficult. Therefore, model-based methods often cause inaccuracies in detection (Cody & Narasimhan 2020) and lead to a high rate of false alarms in real WDNs (Tariq *et al.* 2022).

Data-driven methods have recently gained popularity among researchers as such models require less extensive knowledge of the physical system (Cody *et al.* 2020) (see Table 1). For example, Li *et al.* (2018) developed an artificial neural network (ANN) model for leakage detection in ductile iron pipes based on acoustic emissions generated from laboratory-scale experiments. Similarly, Quy *et al.* (2019) established the *k*-nearest neighbour (KNN) model for stainless steel pipes. Guo *et al.* (2020) built random forest models for leak detection of cast iron and steel pipes. Xu *et al.* (2024) proposed a comprehensive research that evaluated the performance of five machine learning models for acoustic leak detection in Dalian. Furthermore, a deep learning model with complex pattern capture capability was also introduced to acoustic leak detection. Guo *et al.* (2021) adopted convolutional neural networks (CNNs) to analyse the signal components from different frequency ranges and validate their effectiveness on two cities in China. To improve the time-temporal capture capability of the model, Liu *et al.* (2024) adopted a time-transformer, which utilizes the attention regions, reaching higher accuracies than those of other CNN models. Although existing data-driven models have contributed to the advancement in the leakage detection domain, these models have significant shortcomings that question their general applicability to real WDNs, as follows:

(1) Most previous models were developed through laboratory-scale experiments except for a few field-based models, such as El-Zahab *et al.* (2019), Guo *et al.* (2020), Kang *et al.* (2017), and Tariq *et al.* (2022). In their experiment design phase, laboratory-based models ignore the complexity of real WDN, such as the system's topography, valve conditions, and pipes' complexity. Secondly, background noise is introduced artificially and is easier to filter out. In real WDNs, background noise is challenging to eliminate/remove from the acceleration signals collected. Thirdly, the location of the leaks is known to the researchers in advance, which is not the case in real WDNs.
(2) Every existing machine learning model was developed using different features, such as level, spread, and frequency centroid, which were extracted using frequency- and time-domain analyses. There is little to no consensus on which features to use in real WDNs for successful detection. Meanwhile, traditional machine learning models failed to demonstrate interpretability as they lack an equation-based format.
(3) Although deep learning models may capture more complex patterns than traditional models, they suffer from opaque decision-making mechanisms and require greater computational resources. Therefore, traditional models are still a competitive method for leak detection.

To overcome the limitations of the existing models, this study has conducted large-scale experiments in a real WDN to develop more realistic data-driven models. The objectives of the study include (1) exploration and careful selection of features for model development; (2) generalization and easy reproducibility of the methodology; (3) development of real network-based leak detection models, which are highly scarce; (4) comparative analysis and validation of the models; and (5) formulation of simple mathematical equations for easy interpretability.

2. RESEARCH METHODOLOGY

2.1. Overall framework

The research methodology is divided into five phases: (1) research conceptualization and literature review; (2) on-field data collection; (3) signal processing; (4) development of machine learning models; and (5) implementation of machine learning models.

In the first phase, a thorough literature review was conducted to understand (1) the leak detection problem in real networks; (2) the physical meanings of features used in existing leak detection models; (3) machine learning models implemented in this domain; (4) the finalization of research methodology; and (5) selection of acoustic equipment/sensor. In the second phase, an extensive on-site acceleration data collection was carried out over 10.5 months. In the third phase, signal processing of the acceleration data was conducted to generate valuable features for developing machine learning models. The second and the third phases were performed in parallel. The fourth phase developed machine learning models for metal and non-metal pipes separately and was

Table 1 | Data-driven leak detection models

S.N.	Ref.	AI method	Acoustic features	Equipment	Model accuracy	Source of data	Type of pipe
1	Tijani & Zayed (2022)	Gene expression programming	Level Frequency centroid Skewness Kurtosis Under 250 Hz frequency spread Above 1,000 Hz frequency spread	Noise loggers	95%	Field	Metal Non-metal
2	Tijani *et al.* (2022)	*k*-nearest neighbour	Frequency centroid Peak frequency Auto-correlated maximum Lyapunov exponent (MLE) Frequency domain average amplitude Peak amplitude Maximum amplitude Frequency centroid Skewness	Noise loggers	98%	Field	Metal Non-metal
3	Tariq *et al.* (2022)	Random forest	Monitoring index Standard deviations	Accelerometers	100%	Field	Metal Non-metal
4	Guo *et al.* (2021)	Convolutional neural network	RMS Mean of a signal Mean decibel power Spectral density	Piezoelectric accelerometer	98%	Field	CI and Steel DN50 to DN300
5	Guo *et al.* (2020)	Random forest	RMS Mean of a signal Mean decibel power spectral density RMS and Shannon entropy of intrinsic mode functions 1–3 Zero-crossing rate Sub-band spectral entropy Energy operator Energy entropy ratio	Piezoelectric accelerometer	99.45%	Field	CI and Steel DN100 to DN300
6	Shukla & Piratla (2020)	Convolutional neural network	Acceleration signals	Accelerometer	98%	Laboratory	PVC DN76 DN102 DN
7	Cody *et al.* (2020)	Convolutional neural network	N/A	Hydrophones	97%	Laboratory	PVC DN152.4
8	Mysorewala *et al.* (2020)	Support vector machine	Acceleration signals	Accelerometer	93%	Laboratory	PVC DN25.4
9	El-Zahab *et al.* (2019)	K-means clustering	Level Spread	Noise loggers	94.1%	Field	Not given

(*Continued.*)

Table 1 | Continued

S.N.	Ref.	AI method	Acoustic features	Equipment	Model accuracy	Source of data	Type of pipe
10	Liu *et al.* (2019)	Support vector machine	Mean value of the intrinsic mode functions' power spectrum	Accelerometer	98%	Laboratory	Aluminium-plastic composite
11	Quy *et al.* (2019)	*k*-nearest neighbour	Root mean square Short-time energy Average amplitude	Acoustic sensor	99.3%	Laboratory	Stainless steel 304
12	Chuang *et al.* (2019)	Convolutional neural network	Mel-frequency cepstral coefficients	Acoustic sensor	98%	Laboratory	Not given
13	Kang *et al.* (2017)	Convolutional neural network – support vector machine	Acceleration signals	Piezoelectric accelerometer	99.3%	Field	CI DN80 to 300
14	El-Zahab *et al.* (2018)	Decision tree	Monitoring index efficiency	Accelerometer	99.29%	Laboratory	PVC & DI
15	Cody *et al.* (2018)	Support vector machine	Entropy Effective value Spectral peak	Hydrophones	92%	Laboratory	PVC
16	Pan *et al.* (2018)	Support vector machine	Energy Amplitude Average frequency RMS	Acoustic sensor	95%	Laboratory	Steel
17	Li *et al.* (2018)	Artificial neural network	Peak Mean Peak frequency Kurtosis	Acoustic sensor	97.2%	Laboratory	Ductile iron
18	Yang *et al.* (2013)	Artificial neural network	Approximate entropy	Piezoelectric accelerometer	93.8%	Field	Metal
19	Jin *et al.* (2010)	Artificial neural network	Self-similarity features	Piezoelectric accelerometer	92.5%	Field	Metal

accomplished in four consecutive steps. Firstly, features were finalized for metal and non-metal pipes separately; secondly, data balancing techniques were applied to the selected features; thirdly, the machine learning models were selected for research; and fourthly, the performance of the models was computed using different metrics. In the fifth phase, the machine learning models were implemented, and their accuracy was checked. This five-phase methodology is illustrated in Figure 1.

2.2. Description of the intelligent ML-based leak detection models for WDNs

Abdelmageed *et al.* (2022) summarized the characteristics of artificial intelligence applications in water leak management. They found that KNN has better interpretability and can avoid overfitting but it cannot handle noise and outliers, which are generally caused by running vehicles and nearby work. Support vector machine (SVM) avoids model overfitting and imposes less burden on computation, but does not handle large datasets well, and the interpretation of results is difficult.

An ANN is an ML algorithm inspired by the configuration of the human brain. In this configuration, a multilayer feedforward neural network is the most broadly adopted and most straightforward type of ANN and can handle non-linear problems (Fine 1999; Lawal *et al.* 2020; Lawal & Kwon 2021). ANN is good at analysing a

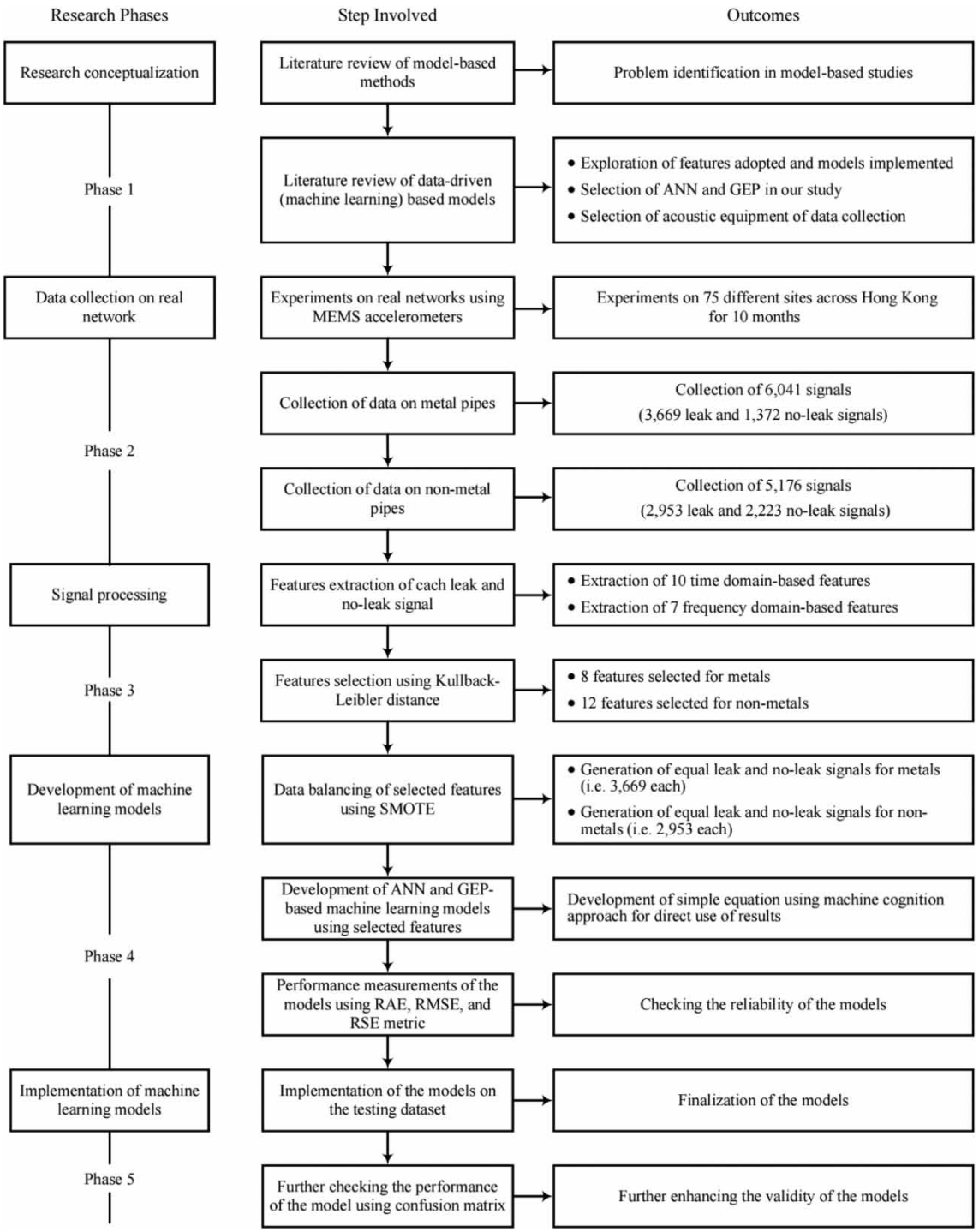


Figure 1 | Research methodology for the proposed ML framework.

large amount of data, missing data, and structured and unstructured data (Abdelmageed *et al.* 2022). Besides, ANN has higher robustness and can handle noise and outliers.

Meanwhile, gene expression programming (GEP) is a supervised ML technique that expresses relationships concealed between independent and dependent variables according to the principles of Darwinian evolution (Ferreira 2001). It harnesses the merit of genetic algorithms and genetic programming and advances on their shortcomings (Lawal *et al.* 2021). In the literature, a remarkable performance of the GEP for leak detection using noise loggers was reported in another study published by our research team (Tijani & Zayed 2022). Thus, ANN and GEP are potentially powerful computing techniques employed herein for modelling, testing, and validation.

3. DATA COLLECTION AND SIGNAL PROCESSING

3.1. Selection of the acoustic sensor

Selection of the appropriate acoustic sensor for a successful leak detection campaign is an *absolute must*. Issues, such as on-site deployment (permanent/semi-permanent, placement in/out of the pipe, etc.), type of monitoring (real-time/non-real-time), site accessibility, pipe characteristics, accuracy, sensitivity, cost, etc., impact the selection of a suitable sensor. In this regard, a literature review of real-time sensors, including noise loggers, accelerometers, hydrophones, MEMS sensors, fibre optics sensors, and wireless sensor networks, is conducted.

Their advantages and limitations are noted. For example, (1) noise loggers are suitable for permanent monitoring but result in high false alarms. Besides, their applicability in plastic pipes can only be rated as not-so-effective; (2) accelerometers can detect weak leaks, but are unsuitable for the low-frequency range. In addition, accelerometer systems are often wired, which poses another challenge for their deployment in real networks; (3) hydrophones are suitable for the low-frequency range and high attenuation pipes. However, they are not effective for weak leaks, and their on-site deployment is an issue as they have to be placed inside the fire hydrants; (4) MEMS sensors are inexpensive and consume less power but are not widely available commercially; (5) fibre-optic sensors are small, light, and highly sensitive even to a very small-scale leak but their cost of deployment for existing pipes is very high; and (6) wireless sensor networks are non-invasive and more flexible. It is cheaper than wired systems but consumes more energy.

Considering the advantages of MEMS technology, accelerometers, and wireless connectivity, MEMS accelerometers were selected for this study. Their cost-effectiveness (relative to other acoustic sensors) and ease of deployment made them ideal for research. Additionally, these accelerometers offer high time synchronization. As demonstrated by Tariq *et al.* (2022), they are highly effective for leak detection in real-world WDNs.

3.2. Leak detection experimental campaign

Pilot studies were conducted in September 2020 to understand the system's working principle and design the experimental protocols. The working principle of the system is displayed in Figure 2, which comprises four consecutive steps: (1) placing the accelerometers on the valves for data collection, (2) collecting acceleration data, (3) transmitting the data to the gateway wirelessly, and (4) storing the data through corresponding software. Pilot studies revealed that sensing only along the radial direction successfully detected leaks. Thus, the *z*-axis of the accelerometers was selected for data acquisition, with the sampling frequency at 3,000 Hz in streaming mode.

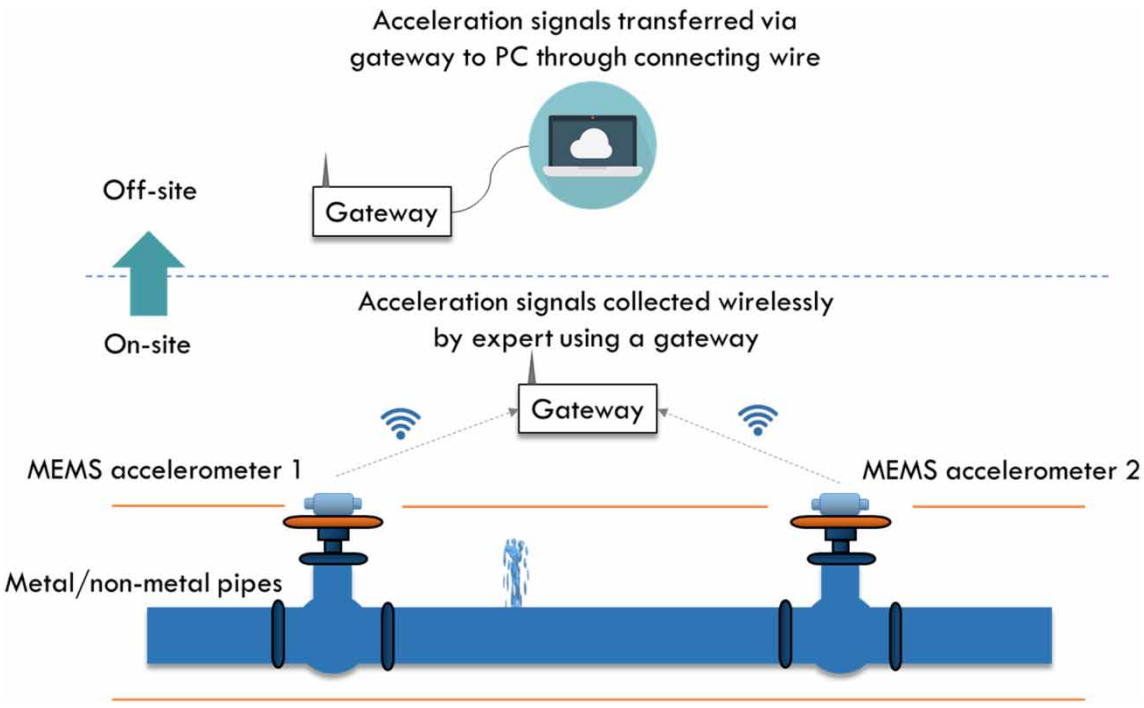


Figure 2 | Working principle of the equipment.

Meanwhile, to ensure uninterrupted signal transmission, the chamber cover must be opened during data acquisition. Otherwise, signal transmission may be disrupted, compromising data quality. The accelerometers we used are lift-and-shift type sensors. These sensor types are gaining popularity in practice because of the flexibility they offer in surveying larger areas within a reasonable time frame. Accelerometers are placed in one area, data are collected, and then they are moved to another location. Since data are gathered over a short duration, it is practical to keep valve or hydrant chambers open for data transmission.

The experimental campaign on real WDN in Hong Kong began on the 1st of October, 2020, and lasted till the 15th of August, 2021. This rigorous campaign was conducted in collaboration with the Water Supply Department (WSD) of Hong Kong and a local contractor involved in leak detection for over 20 years. All the experiments were conducted at midnight to minimize the effects of water flow, traffic noise, and other perturbations on the collected data. Besides, most leak detection surveys in Hong Kong were traditionally conducted at night. Experiments were conducted across pipes of different materials, including polyethylene, asbestos cement, plasticized polyvinyl chloride, galvanized iron, ductile iron, and stainless steel. The diameter of pipes varied between 50 and 300 mm. A typical deployment of accelerometers on-site is shown in Figure 3.

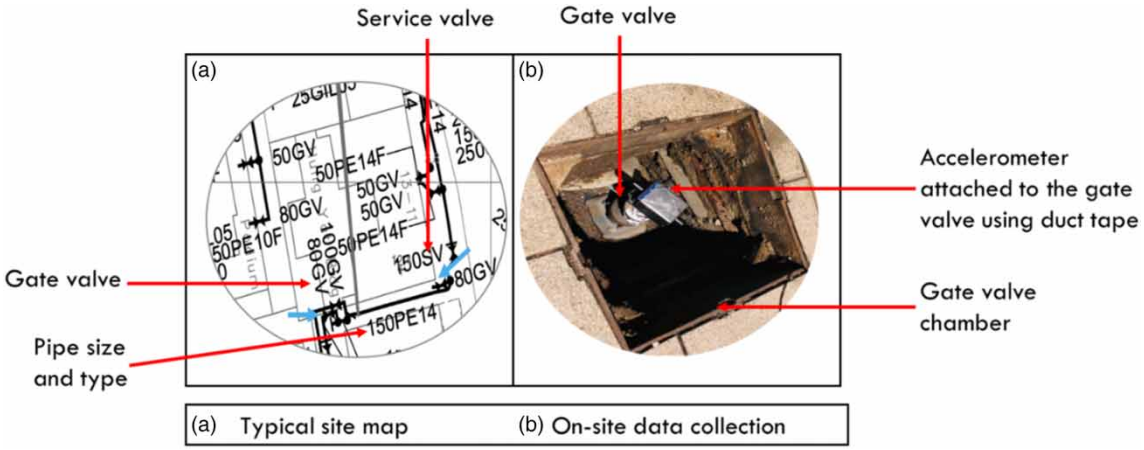


Figure 3 | Data collection process.

The WSD would notify the local contractor in case of any potential leak. Afterwards, this research team and the local contractor visited the reported site at midnight the same day, along with the data acquisition equipment and collected data. If the WSD later confirms the leak, the collected data would be labelled as leak data. Otherwise, the collected data would be labelled as no-leak data.

Following the protocols mentioned above, 75 sites were visited over more than 10 months, and 993 cases (leak and no-leak) were collected. The data collection for each case varied from 3 min (minimum) to 1.5 h (maximum), depending on the site condition and stability of the wireless connection between accelerometers and the gateway.

The distance between the leak and accelerometers was between 1 and 16 m, depending upon the availability/usability of valves and the number of accelerometers deployed per site. Two types of leakages were primarily considered: single leakage (85%) and joint leakage (15%). This approach was deemed appropriate as it was confirmed by the WSD that joint leakages typically do not exceed 15%. The collection of many signals at different locations fulfilled the most boundary conditions of the network, thus enhancing the representativeness of the whole WDN. To remove the inconsistency and the impact of externalities in the data and obtain sufficient data points for model development, the data collected from each case were divided into smaller intervals of continuous signals. Inconsistent intervals were removed, and every remaining continuous signal was considered a separate signal. Fourteen thousand and twenty-one signals were collected. Seven thousand, five hundred and fifty-one signals (4,586 leak signals and 2,965 no-leak signals) were collected from metal pipes, and 6,470 (3,691 leak signals and 2,779 no-leak signals) were collected from non-metal pipes. These signals were further subjected to signal processing techniques and data preprocessing approaches to develop machine learning models.

4. PROCESS OF INTELLIGENT ML-BASED LEAK DETECTION MODELS

4.1. Features extraction

Signal processing was carried out to extract useful features from the collected signals. A thorough literature review was carried out of the acoustics-based data-driven and machine learning models to select the features. Throughout the literature, it was revealed that different models used different features in time and frequency domains, and there is not much consensus on which features should be used for real networks. Level and spread are the most used features for leak detection. However, pilot studies showed that two features alone were insufficient for leak detection. Therefore, 15 features were also selected from the literature apart from level and spread. A total of 10 features were selected from the time domain, and seven features were selected from the frequency domain (Lim 2015; Li *et al.* 2018; Liu *et al.* 2018; Martini *et al.* 2018; El-Zahab *et al.* 2019; Quy *et al.* 2019; Tijani & Zayed 2022). The features from the time domain include level, spread, root mean square, average amplitude, peak amplitude, Crest factor, energy, maximum Lyapunov exponent (MLE), Kurtosis of the autocorrelation function, and MLE of the autocorrelation function. The features from the frequency domain include average amplitude, peak amplitude, maximum amplitude, frequency centroid, skewness, Kurtosis, and frequency spread. The formulas of features are given in Table 2.

4.2. Dataset balancing

The inputs of the detection models were the extracted features from the recorded signals from metal and non-metal pipes. Most ML algorithms work on the balanced class principle, and an imbalanced dataset should be pre-processed for reliable classification (Choi *et al.* 2020; Tariq *et al.* 2022). Imbalanced datasets are typically challenging for most ML techniques because the techniques are inclined to neglect minority cases. Hence, imbalanced data should be resized to produce a balanced dataset. Typically, for binary classification problems, a balanced dataset can be realized through under-sampling and oversampling (Chawla *et al.* 2002; Wang *et al.* 2021). Under-sampling sympathizes with the dataset by deleting most observations randomly. Hence, the main demerit of this method is that under-sampling tends to eliminate helpful information in the data, thus deteriorating the majority class distribution. In contrast, oversampling does not have such a demerit, as this method tends to increase minority observations. The synthetic minority oversampling technique (SMOTE) is the most typical oversampling method that generates artificial minority cases by interpolating existing minority information (Chawla *et al.* 2002). It also has the advantage of overcoming overfitting (Cheng *et al.* 2019). In the literature, Choi *et al.* (2020) conducted the SMOTE analysis to enrich project information, predicting the possibility of project accidents. Similarly, Wang *et al.* (2021) used SMOTE to enrich the dataset, predicting the successful implementation of infrastructure projects. Besides, Tariq *et al.* (2022) also used SMOTE to enrich the accelerometer signals for leakage detection.

The current study applied SMOTE to the modelling dataset to address the class imbalance problem. The leak cases were segregated from the modelling dataset and duplicated to attain the same number of leak and no-leak cases.

4.3. Features selection

Generally, some of the independent variables – extracted features – are usually distorted and hardly dependent on the leakage phenomenon. Hence, there is a need to examine the sensitivity of the proposed intelligent machine learning models to all the independent variables. Sensitivity analysis provides a perception of the influence of input parameters on the target variable (Tunkiel *et al.* 2020; Tijani & Zayed 2022). Feature selection plays a substantial role in detecting the extracted features that are most appropriate for pattern recognition before developing the models (Li *et al.* 2018). To determine the sensitivity of the independent variables to the leak state, the Kullback–Leibler (KL) distance – was used to rank the independent variables from best to worst. KL distance is defined as:

$$\mathrm{KL} = \sum p(x|a_1) \log \frac{p(x|a_1)}{p(x|a_2)} + \sum p(x|a_2) \log \frac{p(x|a_2)}{p(x|a_1)} \tag{1}$$

where a_1 and a_2 are the classes illustrative of the cases with leak and no-leak; $x = (x_1, x_2, x_3, \ldots, x_w)^T$ are the features for the w samples. KL measures the distance between the likelihood functions $p\,(x|a_1)$ and $p\,(x|a_2)$. The values of the KL distance increase as the two distributions move further apart. Meanwhile, when the two

Table 2 | Expressions of features extracted

Features in the time domain	Expression	Features in the frequency domain	Expression
Level (Lv)[a,b]	$20\log_{10}\frac{\text{RMS}}{20\times10^{-6}}$, where RMS is the root mean square	Average amplitude ($FDAvgAmp$)[b,c]	$\frac{1}{N}\sum_{f=0}^{F}x_f$, where F is the maximum frequency in a spectrum
Spread (Sp)[a]	$\max(x_i)-\min(x_i)$, where x_i is the i-th data in the signal	Peak frequency (PF)[d]	f when $x_f=\max(x_f)$
Root mean square (RMS)[c,d]	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2}$, where N is the total amount of data in the signal	Maximum amplitude (MA)[d]	$\max(x_f)$
Average amplitude ($TDAvgAmp$)[c,d]	$\frac{1}{N}\sum_{i=1}^{N}\lvert x_i\rvert$	Frequency centroid (FC)[b]	$\frac{\sum_{f=0}^{F}x_f f}{\sum_{f=0}^{F}x_f}$
Peak amplitude (PA)[d]	$\max(\lvert x_i\rvert)$	Skewness (Sk)[b,d]	$\frac{1}{N}\sum_{f=0}^{F}\left(\frac{x_f-\mu_f}{\sigma_f}\right)^3$, where μ_f and σ_f are the mean and the standard deviation of x_f
Crest factor (CF)[b,d,e]	$\frac{\max(\lvert x_i\rvert)}{\text{RMS}}$	Kurtosis (Ku)[b,d]	$\frac{1}{N}\sum_{f=0}^{F}\left(\frac{x_f-\mu_f}{\sigma_f}\right)^4$
Energy (En)[b,d]	$\frac{1}{N}\sum_{i=1}^{N}x_i^2$	Frequency spread (FS)[d]	$df\times n$, where n is the amount of data where $x_f\geq0.33\times\max(x_f)$
Maximum Lyapunov exponent (MLE)[f,g]	$\frac{4}{T}\sum_{i=\frac{N}{4}}^{\frac{N}{2}}\frac{1}{5}\sum_{k=1}^{5}\frac{1}{k}\ln\left\lvert\frac{x_{i+k}-x_{j+k}}{x_i-x_j}\right\rvert$, where T is the time length of the signal, and j is the index when x_i-x_j is the smallest over the entry signal	–	–
Kurtosis of autocorrelation function (AKu)[h]	$\frac{1}{N}\sum_{f=0}^{F}\left(\frac{R_{xx}-\mu_{xx}}{\sigma_{xx}}\right)^4$, where R_{xx} is the autocorrelation function of x, μ_{xx} and σ_{xx} are the mean and the standard deviation of R_{xx}	–	–
Maximum Lyapunov exponent of autocorrelation function ($AMLE$)[f]	$\frac{4}{T}\sum_{i=\frac{N}{4}}^{\frac{N}{2}}\frac{1}{5}\sum_{k=1}^{5}\frac{1}{k}\ln\left\lvert\frac{R_{xx,i+k}-R_{xx,j+k}}{R_{xx,i}-R_{xx,j}}\right\rvert$	–	–

[a]El-Zahab *et al.* (2019).
[b]Tijani & Zayed (2022).
[c]Quy *et al.* (2019).
[d]Li *et al.* (2018).
[e]Lim (2015).
[f]Liu *et al.* (2018).
[g]Tijani *et al.* (2022).
[h]Tijani & Zayed (2022).

probability functions completely overlapped, the KL distance equals zero. Generally, the features with a higher value of the KL distance are appropriate for the model development.

Besides, correlation analysis is a statistical method to study the closeness of the relationship between two or more variables in the same position. The equation for correlation analysis is shown in Equation (2), where R represents the correlation degree. Besides, X and Y represent the values of two measured features. The higher

correlation values between two features mean that they can replace each other. Thus, followed by KL distance analysis, the correlation analysis was adopted to remove the overlap between features:

$$R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 (Y_i - \bar{Y})^2}} \tag{2}$$

4.4. Normalization and performance indicators

Before developing the detection models for leak detection, the datasets were normalized between 0 and 1 using max normalization to avoid overfitting and remove dimensional variation between the input parameters (Tijani & Zayed 2022).

For a leak detection problem (a typical case of a binary classification problem), the performance of detection results can be evaluated using a confusion matrix, as illustrated in Figure 4. This matrix provides information about the detection class (by the detection models) and the target class measured from the field. In this study, the leak is regarded as positive, while a no-leak is regarded as negative. True positive (TP) represents the true leak point and is also detected as a leak. It can be understood as the corrected alarm. False positive (FP) is the true no-leak point but is mistakenly detected as the leak; it can be understood as a false alarm. In contrast, a true negative (TN) refers to the no-leak point and is correctly detected as the no-leak. It can be understood as the correct no alarms. The false negative (FN) refers to the leak point but is mistakenly detected as the no-leak point. It can be understood as a missed leak.

		Actual	
		Positive	Negative
Model Result	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Figure 4 | Confusion matrix.

After that, four indicators –accuracy, error rate, sensitivity, and specificity – are used to assess the validation of the detected results. The accuracy presented how often the models correctly detected true positive and true negative as true phenomena. The error rate presented how often the models wrongly classified the true phenomenon as a false scenario. Meanwhile, the sensitivity presented how often the model classified a leak as a true condition. At the same time, specificity indicated the rate at which the model classified no-leak correctly. The equations for these indicators are given in Equations (3)–(6):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{3}$$

$$\text{Error rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{4}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{6}$$

5. ML-BASED MODELS FOR LEAK DETECTION

5.1. Features extraction

This study mainly divides the dataset into a training and testing dataset (80%) and a validation dataset (20%). The total dataset for metal pipe contains 6,041 samples for training and testing, and 1,510 samples for validation. Regarding the non-metal dataset, 5,906 samples remain for training and testing, and 1,294 samples for validation. Initially, 17 features were extracted from the recorded acoustic signals. Generally, it is impossible to present the

values of these features because of the sample size of the data points. The range values of the extracted features for leak and no-leak for metal and non-metal pipe WDNs are presented in Table 3.

Table 3 | The range of features extracted

	Metals		Non-metals	
Features	**Leak**	**No-leak**	**Leak**	**No-leak**
Lv	32.02–40.32	34.48–39.60	34.04–45.31	35.28–39.29
Sp	0.00–6.25	1.32–7.81	0.18–5.79	0.24–8.02
RMS	0.00080–0.0022	0.0011–0.0020	0.0011–0.0037	0.0012–0.0020
TDAvgAmp	0.00064–0.0017	0.00085–0.0016	0.00080–0.0030	0.00092–0.0015
PA	0.0021–0.0079	0.0035–0.0088	0.0027–0.012	0.0034–0.0080
CF	2.07–4.26	2.83–4.97	2.41–4.27	2.29–5.18
En	$0.00–3.21 \times 10^{-06}$	4.52×10^{-07}–3.21×10^{-05}	2.11×10^{-07}–7.99×10^{-06}	3.03×10^{-07}–1.1×10^{-05}
MLE	−8.25 to 601.74	−1.56 to 50.11	−5.68 to 460.96	0.00–124.41
AKu	8.00–192.12	21.00–3,061.15	8.20–349.70	21.91–1,436.27
AMLE	−19.43 to 950.14	11.16–3,220.00	−1.27 to 1,899.60	0.00–956.32
FDAvgAmp	0.000050–0.00033	1.47×10^{-05}–8.2×10^{-05}	3.47×10^{-05}–0.00032	2.78×10^{-05}–0.000167
PF	2.78–500	3.33–237.65	2.23–489.58	1.33–113.73
MA	0.00026–0.0012	0.00013–0.00065	0.00020–0.0018	0.00015–0.0089
FC	267.09–331.21	264.00–331.11	244.82–361.05	262.69–319.77
Sk	−0.0081 to 3.55	0.80–5.26	0.41–3.42	-0.69–2.82
Ku	1.62–23.45	2.99–52.70	2.26–18.86	2.70–16.24
FS	0.0033–0.072	0.0013–0.16	0.0060–0.12	0.0073–0.33

Table 3 shows a clear difference in the values of some features for leak and no-leak (e.g., *AKu* for metal pipes). At the same time, there is a substantial overlap in the values of some other features for both leak and no-leak (such as *Ku* and *CF* for non-metal pipes). The current study highlighted that the overlap in the values of the extracted features for the leak and no-leak could be attributed to the effect of water usage or other irregular disturbances within the WDNs (Tijani & Zayed 2022). Notwithstanding this overlap, ML techniques can understand the multifaceted phenomena secreted in the dataset and reach an appropriate description that best demonstrates the circumstance.

After that, SMOTE was applied to balance the size of the datasets. Hence, 3,669 leak samples and 2,372 no-leak points for metal pipes became 3,669 leak and 3,669 no-leak data points, respectively. Meanwhile, 2,953 leak and 2,223 no-leak data points became 2,953 leak and 2,953 leak data points for non-metal pipes. It should be noted that the minimum and maximum values of the extracted features, as given in Table 3, remain the same after applying the SMOTE for balancing the size of the data points. Balanced datasets are used in the following sections for feature selection analysis.

5.2. Features selection

Undoubtedly, database cleansing is a common practice of minimizing the effects of outliers in the datasets and reducing the number of input variables. Meanwhile, it upholds the proposed machine learning models' high detection capability and high accuracy. The ranking of the extracted features is calculated using Equation (1), and the ranking of the features for metal and non-metal pipes is listed in Table 4. It can be observed that seven and 12 features have an essential value of above 50% for metal and non-metal, respectively. The proposed ML models were developed using features with an essential value of at least 50%.

Then, correlation analysis was conducted on metal and non-metal pipe features, and the results are shown in Figure 5. It can be observed that some extracted features have high correlation values and are redundant. Thus, 0.8 is set as the statistical threshold to eliminate the overlapping features. Eventually, six features, Peak amp, Crest factor, Energy, FD Avg. amp, Freq. centroid, and Freq. spread remained for non-metal pipe leakage detection,

Table 4 | Ranking of extracted features

Metals		Non-metals	
Features	**Ranking**	**Features**	**Ranking**
En	79.02	*MLE*	77.46
MLE	75.32	*AKu*	76.25
AMLE	72.35	*FS*	72.65
AKu	68.23	*AMLE*	69.68
FS	67.32	*En*	68.35
MA	67.02	*FDAvgAmp*	63.25
FDAvgAmp	58.26	*FC*	62.98
Sp	48.25	*TDAvgAmp*	61.53
Lv	46.83	*CF*	60.15
PA	46.32	*PA*	58.78
CF	45.96	*RMS*	55.36
Sk	45.62	*Lv*	53.65
TDAvgAmp	45.36	*MA*	47.02
Ku	44.24	*Ku*	46.86
FC	43.26	*Sp*	46.32
RMS	39.82	*PF*	45.23
PF	36.24	*Sk*	44.45

while four features, Energy, Auto. Corr. Kurtosis, FD Avg. amp, and Freq. spread, remained for metal pipe leakage detection.

5.3. Development of the intelligent ML-based leak detection models

The data points of the selected features were randomly assembled such that no specific data point was considered a reference point.

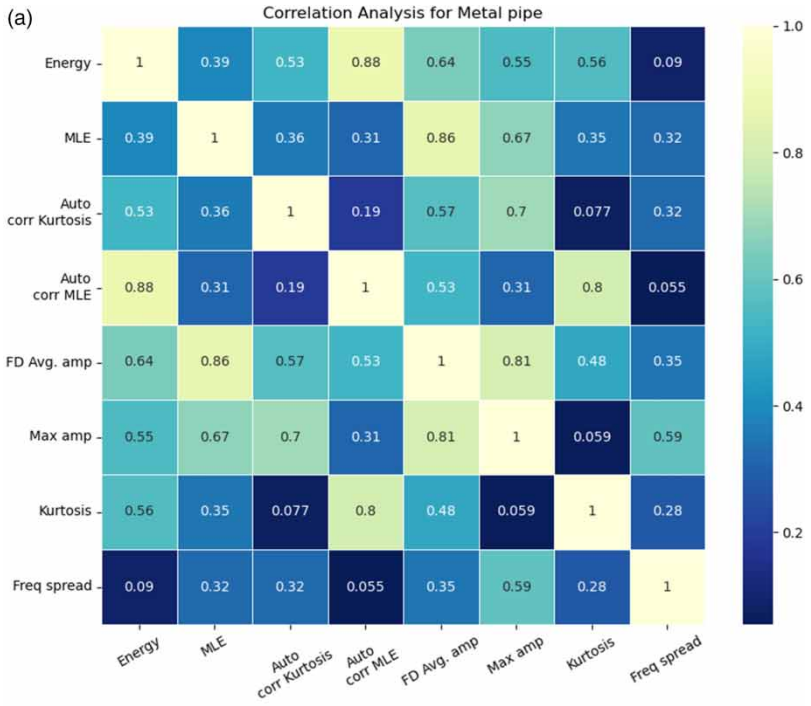
5.3.1. Artificial neural network

In the current study, the ANN-based model was developed using RapidMiner. The training and testing dataset consists of 80% of the training set and 20% of the test set. No fixed method exists to determine the number of neurons in the hidden layers (Alpaydin 2020; Lawal 2020; Lawal & Kwon 2021). Therefore, different ANN architectures were tried using an exponential rectifier linear unit function as the activation function between the input and hidden layers. This function was selected because of its superiority over other functions (Clevert *et al.* 2016). Meanwhile, the softmax activation function was used between the hidden and output layers. Therefore, 4-input, one hidden layer with a different number of nodes, and 2-output ANN architectures were tried for metal pipes, while 6-input with a different number of nodes in one hidden layer were tried to produce 2-output for non-metal pipes. The performance of the tried different ANN architectures is presented in (a) and (b) for metal and non-metal pipes, respectively. The training cycles are set to 200 times. The learning rate is 0.01. Momentum is 0.9.

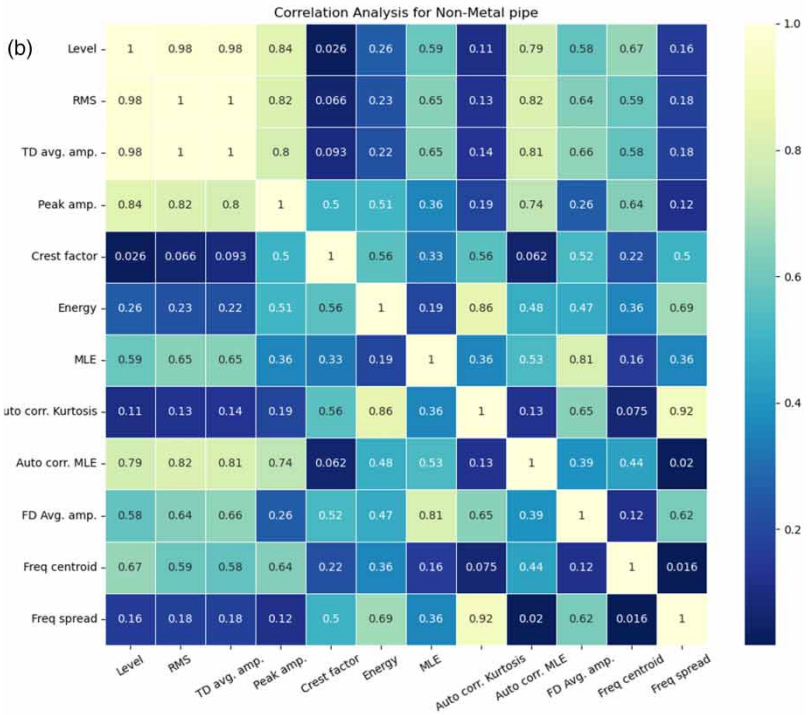
Based on the performance of the architectures (as illustrated in Figure 6(a) and 6(b)), the selected optimum ANN architectures are 4-12-2 and 6-7-2 architectures for metal and non-metal pipes, respectively.

Generally, one of the drawbacks of ANN models is that the relationship between the input parameters and the dependent variable is usually not understood. This problem can be solved by transforming the proposed ANN models into mathematical expressions that can be used for easy prediction of the targeted variable – leak state – without the need to reconstruct a new ANN simulation (Adesanya *et al.* 2021).

Subsequently, the resulting mathematical expression based on the weights and biases of the model for metal WDNs is presented in Equation (7), where $\text{net}_i^{(k)}$ represents the weighted sum of the input layer vector and offset vector of the i-th neural node in the i-th layer. Y is the output value of the neural node, and W is the weight of the corresponding node. b is the bias. m represents the number of nodes in a specific layer.



Correlation analysis on the features in the metal pipe



Correlation analysis of the features in the non-metal pipe

Figure 5 | Correlation analysis for developing leakage detection in the (a) metal pipe and (b) non-metal pipe.

$f^{(k)}(\text{net}_i^{(k)})$ represents the activation function:

$$\text{net}_i^{(k)} = \sum_{i=1}^{m_{k-1}} W_{i,j}^{(k)} Y_j^{(k-1)} + b_i^{(k)}$$
$$Y^{(k)} = f^{(k)}(\text{net}_i^{(k)}) \quad (7)$$

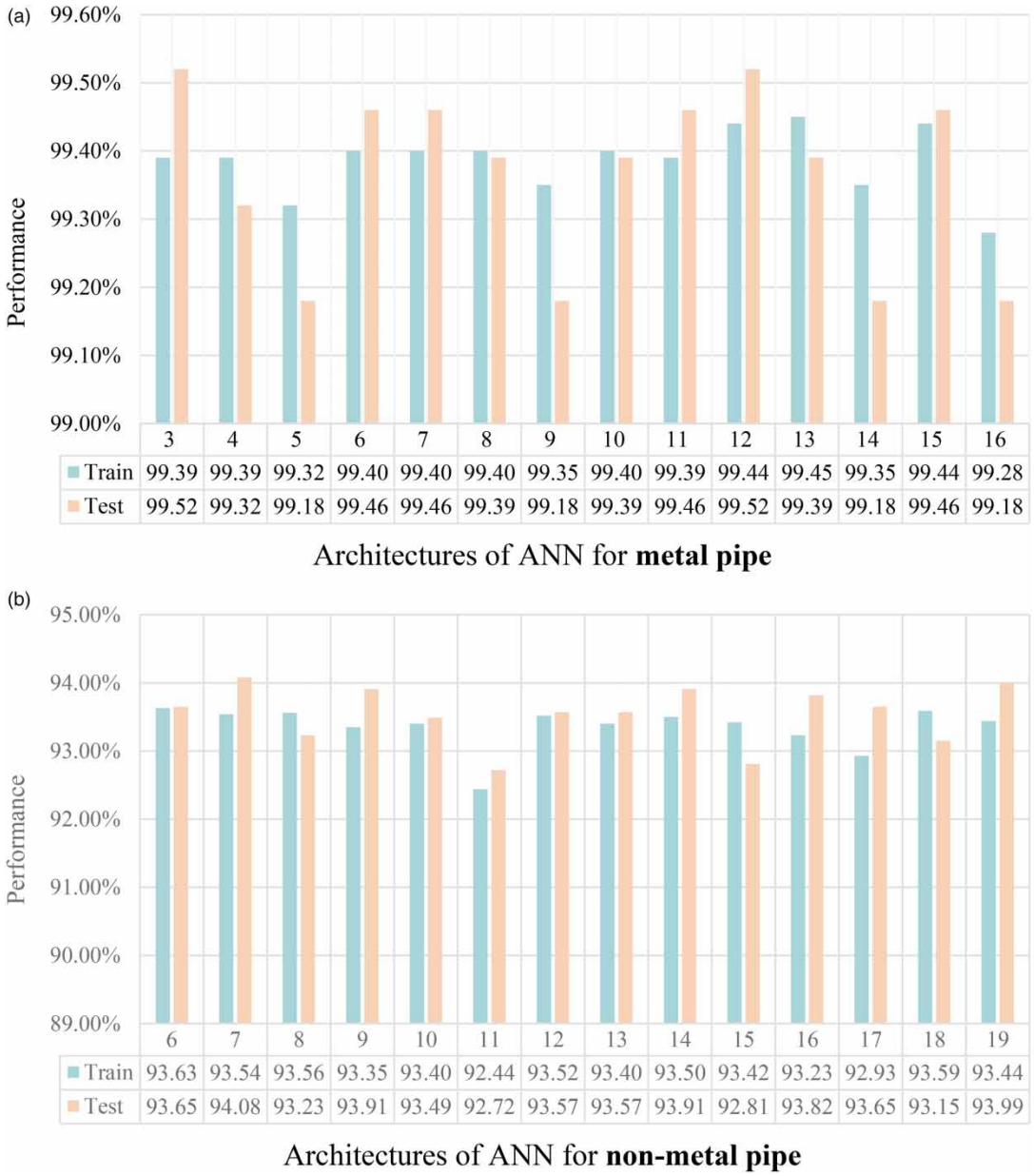


Figure 6 | Performance of different ANN architectures: (a) metal pipe and (b) non-metal pipe.

Hence, the weights and biases of the selected optimum architectures presented in Tables 5 and 6 and were used accordingly to develop the prediction expression. To better understand, the artificial intelligence models for non-metal pipeline are illustrated as Equations (8) and (9):

$$\begin{aligned}
\text{net}_1^1 &= w_{2,1} \times f\begin{bmatrix} -0.849\,PA + 2.339\,CF + 1.097\,En + 3.231\,AKu \\ -3.273\,FC - 1.318\,FS + 0.098 \end{bmatrix} \\
\text{net}_2^1 &= w_{2,2} \times f\begin{bmatrix} -14.272\,PA + 5.758\,CF + 11.826\,En - 11.311\,AKu \\ +5.599\,FC - 1.028\,FS - 5.504 \end{bmatrix} \\
\text{net}_3^1 &= w_{2,3} \times f\begin{bmatrix} -0.506\,PA + 1.969\,CF + 0.767\,En + 2.62\,AKu \\ -2.493\,FC - 1.052\,FS - 0.093 \end{bmatrix} \\
&\vdots \\
\text{net}_7^1 &= w_{2,7} \times f\begin{bmatrix} -14.242\,PA + 6.377\,CF + 2.027\,En + 6.718\,AKu \\ -5.729\,FC - 3.859\,FS - 0.067 \end{bmatrix}
\end{aligned} \tag{8}$$

Table 5 | Weights and biases of the selected optimum ANN architecture for non-metal pipes

Number of nodes in the hidden layer	Weights								Bias		
	w1						w2			B2	
	Peak amp	Crest factor	Energy	Auto corr Kurtosis	Freq. centroid	Freq. spread	Leak	No-leak	B1	Leak	No-leak
1	−0.849	2.339	1.097	3.231	−3.273	−1.318	−2.421	2.361	0.098	1.016	− 1.008
2	−14.272	5.758	11.826	−11.311	5.599	−1.028	7.501	−7.502	−5.504		
3	−0.506	1.969	0.767	2.62	−2.493	−1.052	−1.789	1.752	−0.093		
4	−0.351	1.782	0.656	2.183	−2.061	−0.747	−1.397	1.443	−0.213		
5	−0.214	0.104	0.379	0.86	−0.223	0.807	0.458	−0.449	−0.458		
6	−3.213	3.984	2.027	6.718	−6.703	−3.054	−5.634	5.675	0.846		
7	−14.242	6.377	−2.951	11.233	−5.729	−3.859	−8.213	8.21	−0.067		

Table 6 | Weights and biases of the selected optimum ANN architecture for metal pipes

Number of nodes in the hidden layer	Weights						Bias		
	w1				w2			B2	
	Energy	Auto corr Kurtosis	FD Avg. amp	Freq. spread	Leak	No-leak	B1	Leak	No-leak
1	2.897	1.405	−3.719	1.226	−3.292	3.315	1.434	− 0.24	0.241
2	−2.613	−1.198	3.425	−1.116	2.732	−2.733	−1.436		
3	2.817	1.389	−3.612	1.135	3.241	3.178	1.377		
4	−3.001	−1.382	3.985	−1.331	3.202	−3.161	−1.609		
5	0.169	0.223	0.094	0.042	−0.174	0.211	−0.175		
6	−0.79	−0.291	1.096	−0.326	0.731	−0.752	−0.42		
7	0.799	0.435	−0.638	0.324	−0.83	0.859	0.088		
8	0.649	0.414	−0.514	0.269	−0.72	0.697	−0.009		
9	−4.479	−2.128	5.996	−2.103	4.917	−4.908	−2.424		
10	−1.827	−0.824	2.417	−0.781	1.858	−1.909	−1.039		
11	5.446	2.596	−7.192	2.572	−6.289	6.336	2.945		
12	3.667	1.791	−4.758	1.575	−4.201	4.178	1.899		

$$Y^{\text{leak}} = f\left[\begin{array}{c} -2.421\,\text{net}_1 + 7.501\,\text{net}_2 - 1.789\,\text{net}_3 - 1.397\,\text{net}_4 + 0.458\,\text{net}_5 \\ -5.634\,\text{net}_6 - 8.213\,\text{net}_7 + 1.016 \end{array}\right]$$
$$Y^{\text{no-leak}} = f\left[\begin{array}{c} 2.361\,\text{net}_1 - 7.502\,\text{net}_2 + 1.752\,\text{net}_3 + 1.443\,\text{net}_4 - 0.449\,\text{net}_5 \\ +5.675\,\text{net}_6 + 8.21\,\text{net}_7 - 1.008 \end{array}\right] \quad (9)$$

5.3.2. Gene expression programming

The GEP-based models were implemented in the GeneXproTools 5.0 software (Ferreira 2006) with the following GEP characteristics: Eight head sizes, 30 chromosomes, four genes, and different functions, such as +, –, ×, ÷, sqrt, exp, log, max, min, and avg. The cognition models were developed using a 'logistic' function that returns a binary output – zeros for no leaks and one for leaks. The circumstance under the current study consideration is programmed for the solution. The technique randomly creates an initial population of feasible chromosomes. Subsequently, it converts the chromosomes into expression trees corresponding to mathematical expressions. After that, the fitness criteria for the chromosomes are determined. If the fitness criteria are sufficiently good, a solution is deemed fit. At this stage, the analysis stops, and a typical model is obtained. Otherwise, using roulette wheel sampling, some chromosomes are selected and then transformed to obtain a new generation. Figure 7 illustrates the iterative process, which is continuous until achieving the desired fitness criteria. Then, the chromosomes are deciphered for the best solution to the problem (Teodorescu & Sherwood 2008).

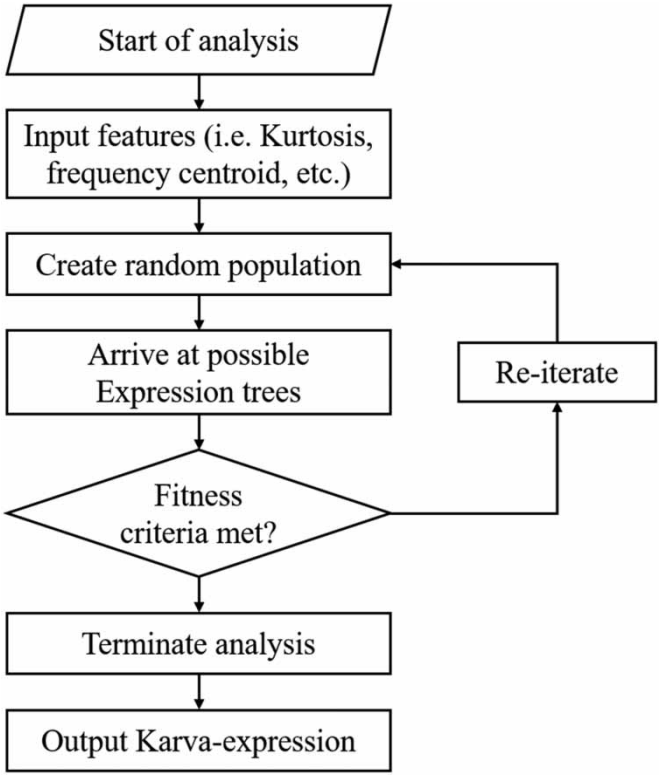


Figure 7 | Typical GEP procedure (Teodorescu & Sherwood 2008; Tijani & Zayed 2022).

Eventually, the optimal models for metal and non-metal pipe were trained 3,053 times and 3,853 times, respectively. Regarding training performance, the metal pipe reaches an overall accuracy of 99.54% in the testing set, with sensitivity equal to 100.00% and specificity equal to 99.16%. Meanwhile, the non-metal pipe reaches an overall accuracy of 89.64%, with sensitivity equal to 90%, and specificity equal to 89.5%.

The obtained expression tree for the candidate solutions for metal and non-metal WDNs composed of four sub-expression trees (sub-ETs) is illustrated in Figures 8 and 9, respectively. Typically, each sub-ET might comprise one or more input parameters, constants, and different mathematical operators (such as +, /, and log). The set of mathematical functions in Equations (10) and (11) is listed in Table 7.

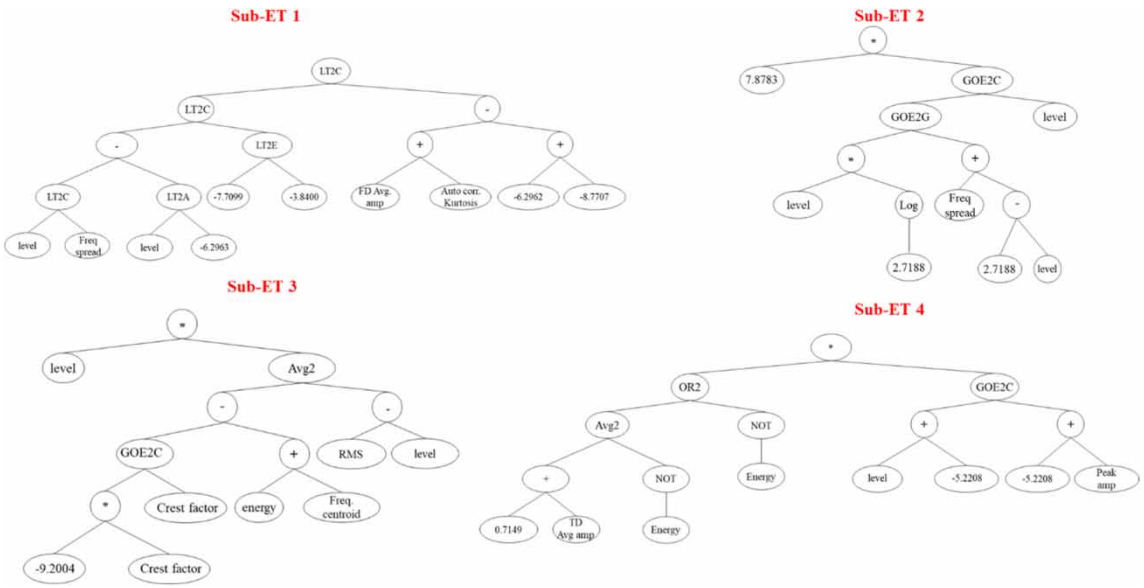


Figure 8 | Sub-ETs of the non-metal model.

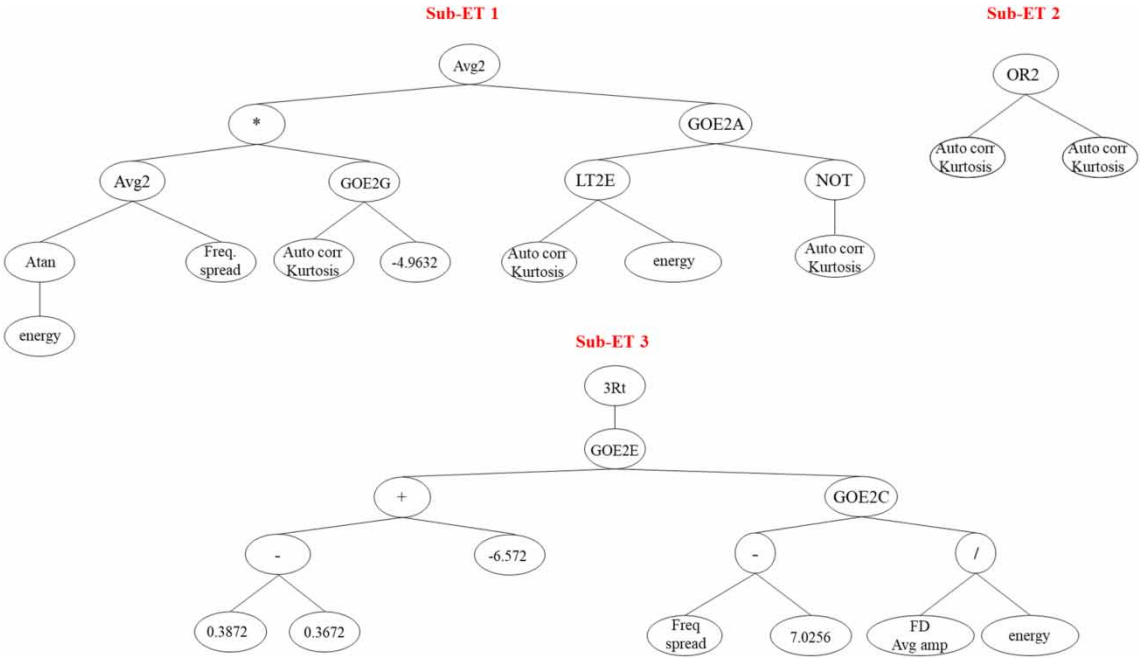


Figure 9 | Sub-ETs of the metal model.

Table 7 | Input mathematical functions

Functions	Definition
LT2C	If $x < y$, then $(x + y)$, else $(x - y)$
LT2E	If $x < y$, then $(x + y)$, else $(x * y)$
LT2A	If $x < y$, then x, else y
GOE2E	If $x \geq y$, then $(x + y)$, else $(x * y)$
GOE2C	If $x \geq y$, then $(x + y)$,else $(x - y)$
GOE2G	If $x \geq y$, then $(x + y)$, else atan$(x * y)$
GOE2A	If $x \geq y$, then x, else y
OR2	If $x > 0$ OR $y > 0$, then 1, else 0
Avg2	Avg(x,y)
NOT	$(1 - x)$
Atan	Arctan(x)
3Rt	x^(1/3)
logi	logistic

Based on the sub-ETs illustrated in Figures 8 and 9, the definitions of abbreviations are described in Table 7. Based on that, the detection models for metal and non-metal WDNs are formulated and presented in Equations (10) and (11), respectively:

$$LS_{nm} = logi\left[\begin{array}{l} LT2C\left(\left[LT2C\left(\begin{array}{c} LT2C([LT2C(Lv,\ FS) - LT2A(Lv,\ -6.2963)]), \\ LT2E(-7.7099 - 3.8400), \\ [FD\ Avg.\ amp + AKu + 6.2962 + 8.7707] \end{array}\right)\right]\right) \\ +7.78783{*}GOE2C\left(\begin{array}{c} GOE2G([Lv{*}\log(2.7188)], [FS + 2.7188 - Lv]), \\ Lv \end{array}\right) \\ +level{*}Avg\left(\left[GOE2C\left(\begin{array}{c} -9.2004{*}CF, \\ CF \end{array}\right)\right] - En - FC \right) \\ , (RMS - level) \\ +OR2\left(\begin{array}{c} Avg\ (0.7149 + TD\ Avg\ amp,\ NOT(En)), \\ NOT(En) \end{array}\right){*}GOE2C\left(\begin{array}{c} Lv - 5.2208, \\ PA - 5.2208 \end{array}\right) \end{array}\right] \quad (10)$$

$$LS_m = logi\left[\begin{array}{c} Avg\left[\begin{array}{c} Avg(\arctan(En), FS) * GOE2G(AKu, \ -4.9632), \\ GOE2A(LT2E(AKu, \ En), NOT(AKu)) \end{array}\right] \\ +OR2(AKu, \ AKu) \\ +3Rt\left(GOE2E\left([0.3872 - 0.3672 - 6.572], GOE2C\left(\begin{array}{c} FS - 7.0256, \\ FD\ avg\ amp/En \end{array}\right)\right)\right) \end{array}\right] \quad (11)$$

6. VALIDATION RESULTS OF ML-BASED MODELS

Subsequently, the confusion matrix of the results for metal and non-metal pipe systems is shown in Table 8.

Table 8 | Confusion matrix for detected results for pipes

		ANN		GEP	
		Actual leak	Actual no-leak	Actual leak	Actual no-leak
Metal pipe	Detected leak	913	5	915	2
	Detected no-leak	4	588	5	588
Non-metal pipe	Detected leak	673	60	674	70
	Detected no-leak	65	496	64	486

The accuracy, error rate, sensitivity, and specificity of the detected results for both metal and non-metal pipes are presented in Table 9.

Table 9 | Performance indicators of the models in the application

	Metal		Non-metal	
	ANN	GEP	ANN	GEP
Accuracy	99.40%	99.54%	90.34%	89.64%
Error rate	0.60%	0.46%	9.66%	10.36%
Sensitivity	99.56%	99.78%	91.19%	87.41%
Specificity	99.16%	99.16%	89.21%	91.33%

Generally, as presented in Tables 8 and 9, the accuracy of the metal pipe system models is about 99%. In comparison, the error rates are less than 1%. Meanwhile, the accuracy of the models is about 90%, while the error rates are around 10% for non-metal pipe systems. The sensitivity and specificity of metal pipe systems based on the two ML-based models are close to 99%. Meanwhile, the sensitivity and specificity for non-metal pipe systems range from 86 to 92%. Hence, these indicators show that the developed machine learning models for leak detection of the pipe systems based on the acoustic signals recorded by MEMS sensors are reliable and accurate for leakage detection, which means that a significant number of leaks and no-leak conditions are classified as true conditions.

7. CONCLUSIONS

Water scarcity has constantly threatened sustainable development. In this regard, this research presented an effort towards minimizing the water losses in urban water supply networks through early detection of leakages.

This research presents intelligent machine-learning models trained on acoustic signals from real WDNs using MEMS sensors. Extensive fieldwork was conducted at midnight at the lowest noise time of the day. The research team recorded a total of 7,551 signals (4,586 leak signals and 2,965 no-leak signals) for metal and 6,470 (3,691 leak signals and 2,779 no-leak signals) for non-metal using MEMS sensors from real WDNs. The signals were collected from 75 different sites in Hong Kong over more than 10 months. The recorded signals were processed, and 17 features were extracted from the signals. After that, the size of the datasets was balanced to ensure

unification in the size of the datasets. Based on the sensitivity analysis, four and six features were most significant for model development of metal and non-metal pipes, respectively.

The main recurring features for both pipes are energy, MLE, Kurtosis of the autocorrelation function, MLE of the autocorrelation function, average amplitude in the frequency domain, and frequency spread. Intelligent machine learning models that can detect the phenomenon of water leakage in real WDNs were established. The performance metrics obtained for the machine learning models are substantially satisfactory and adequate because of lower error values. The models detected the leakage phenomenon with sufficient accuracy, ranging from 99% for metal pipes and 89% for non-metal pipes. The intelligent machine learning models are significantly sensitive to leak and no-leak, as the models substantially detect the phenomenon as true positive and true negative, respectively.

Although this study adopted a generalized and reproducible methodology, the models presented are based on MEMS accelerometers; comparison with other technologies, such as noise loggers and hydrophones, is in the data collection process.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support from the Innovation and Technology Fund (Innovation and Technology Support Programme (ITSP)), Hong Kong, and the Water Supplies Department, Hong Kong, under grant number ITS/067/19FP.

DATA AVAILABILITY STATEMENT

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

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abdelmageed, S., Tariq, S., Boadu, V. & Zayed, T. (2022) Criteria-based critical review of artificial intelligence applications in water-leak management, *Environmental Reviews*, **30** (2), 280–297.

Adesanya, E., Aladejare, A., Adediran, A., Lawal, A. & Illikainen, M. (2021) Predicting shrinkage of alkali-activated blast furnace-fly ash mortars using artificial neural network (ANN), *Cement and Concrete Composites*, **124**, 104265.

Alpaydin, E. (2020) *Introduction to Machine Learning*. London: MIT Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002) SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, **16**, 321–357.

Cheng, K., Zhang, C., Yu, H., Yang, X., Zou, H. & Gao, S. (2019) Grouped SMOTE with noise filtering mechanism for classifying imbalanced data, *IEEE Access*, **7**, 170668–170681.

Choi, J., Gu, B., Chin, S. & Lee, J.-S. (2020) Machine learning predictive model based on national data for fatal accidents of construction workers, *Automation in Construction*, **110**, 102974.

Chuang, W.-Y., Tsai, Y.-L. & Wang, L.-H. (2019) Leak detection in water distribution pipes based on CNN with mel frequency cepstral coefficients, *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*, New York: Association for Computed Machinery (ACM), pp. 83–86. DOI: 10.1145/3319921.3319926.

Clevert, D. A., Unterthiner, T. & Hochreiter, S. (2016) Fast and accurate deep network learning by exponential linear units (ELUs), *4th International Conference on Learning Representations, ICLR 2016 – Conference Track Proceedings*, May 2–4, 2016, Caribe Hilton, San Juan, Puerto Rico. (Preprint/open-access version available on arXiv: arXiv:1511.07289).

Cody, R. A. & Narasimhan, S. (2020) A field implementation of linear prediction for leak-monitoring in water distribution networks, *Advanced Engineering Informatics*, **45**, 101103.

Cody, R., Harmouche, J. & Narasimhan, S. (2018) Leak detection in water distribution pipes using singular spectrum analysis, *Urban Water Journal*, **15** (7), 636–644.

Cody, R. A., Tolson, B. A. & Orchard, J. (2020) Detecting leaks in water distribution pipes using a deep autoencoder and hydroacoustic spectrograms, *Journal of Computing in Civil Engineering*, **34**, 04020001.

El-Zahab, S., Mohammed Abdelkader, E. & Zayed, T. (2018) An accelerometer-based leak detection system, *Mechanical Systems and Signal Processing*, **108**, 58–72.

El-Zahab, S., Asaad, A., Abdelkader, E. M. & Zayed, T. (2019) Development of a clustering-based model for enhancing acoustic leak detection, *Canadian Journal of Civil Engineering*, **46** (6), 278–286.

Ferreira, C. (2001) Gene expression programming: a new adaptive algorithm for solving problems, *arXiv preprint* cs/0102027. https://doi.org/10.48550/arXiv.cs/0102027.

Ferreira, C. (2006) *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*. Berlin: Springer.

Fine, T. L. (1999) *Feedforward Neural Network Methodology*. New York: Springer.

Guo, G., Yu, X., Liu, S., Xu, X., Ma, Z., Wang, X., Huang, Y. & Smith, K. (2020) Novel leakage detection and localization method based on line spectrum pair and cubic interpolation search, *Water Resources Management*, **34** (12), 3895–3911.

Guo, G., Yu, X., Liu, S., Ma, Z., Wu, Y., Xu, X., Wang, X., Smith, K. & Wu, X. (2021) Leakage detection in water distribution systems based on time–frequency convolutional neural network, *Journal of Water Resources Planning and Management*, **147** (2), 4020101.

Hu, Z., Tariq, S. & Zayed, T. (2021) A comprehensive review of acoustic based leak localization method in pressurized pipelines, *Mechanical Systems and Signal Processing*, **161**, 107994.

Jin, Y., Yumei, W. & Ping, L. (2010) Approximate entropy-based leak detection using artificial neural network in water distribution pipelines, *2010 11th International Conference on Control Automation Robotics & Vision*. Piscataway, NJ: IEEE, pp. 1029–1034.

Kang, J., Park, Y.-J., Lee, J., Wang, S.-H. & Eom, D.-S. (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.

Lawal, A. I. (2020) An artificial neural network-based mathematical model for the prediction of blast-induced ground vibration in granite quarries in Ibadan, Oyo State, Nigeria, *Scientific African*, **8**, e00413.

Lawal, A. I. & Kwon, S. (2021) Application of artificial intelligence to rock mechanics: an overview, *Journal of Rock Mechanics and Geotechnical Engineering*, **13** (1), 248–266.

Lawal, A. I., Aladejare, A. E., Onifade, M., Bada, S. & Idris, M. A. (2021) Predictions of elemental composition of coal and biomass from their proximate analyses using ANFIS, *ANN and MLR, International Journal of Coal Science & Technology*, **8** (1), 124–140.

Li, S., Song, Y. & Zhou, G. (2018) Leak detection of water distribution pipeline subject to failure of socket joint based on acoustic emission and pattern recognition, *Measurement: Journal of the International Measurement Confederation*, **115**, 39–44.

Lim, J. (2015) *Underground Pipeline Leak Detection Using Acoustic Emission and Crest Factor Technique*. Berlin: Springer.

Liu, J., Wang, J., Liu, S. & Qian, X. (2018) Feature extraction and identification of leak acoustic signal in water supply pipelines using correlation analysis and Lyapunov exponent, *Vibroengineering Procedia*, **19**, 182–187.

Liu, Y., Ma, X., Li, Y., Tie, Y., Zhang, Y. & Gao, J. (2019) Water pipeline leakage detection based on machine learning and wireless sensor networks, *Sensors (Switzerland)*, **19** (23), 5026.

Liu, R., Zayed, T., Xiao, R. & Hu, Q. (2024) Time-transformer for acoustic leak detection in water distribution network, *Journal of Civil Structural Health Monitoring*, **15**, 759–775.

Martini, A., Rivola, A. & Troncossi, M. (2018) Autocorrelation analysis of vibro-acoustic signals measured in a test field for water leak detection, *Applied Sciences (Switzerland)*, **8** (12), 2450.

Mekonnen, M. & Hoekstra, A. (2016) Four billion people facing severe water scarcity, *Science Advances*, **2**, e1500323–e1500323.

Molinos-Senante, M., Mocholí-Arce, M. & Sala-Garrido, R. (2016) Estimating the environmental and resource costs of leakage in water distribution systems: a shadow price approach, *Science of the Total Environment*, **568**, 180–188.

Muntakim, A. H., Dhar, A. S. & Dey, R. (2017) Interpretation of acoustic field data for leak detection in ductile iron and copper water-distribution pipes, *Journal of Pipeline Systems Engineering and Practice*, **8** (3), 05017001.

Mysorewala, M. F., Cheded, L. & Ali, I. M. (2020) Leak detection using flow-induced vibrations in pressurized wall-mounted water pipelines, *IEEE Access*, **8**, 188673–188687.

Pan, S., Xu, Z., Li, D. & Lu, D. (2018) Research on detection and location of fluid-filled pipeline leakage based on acoustic emission technology, *Sensors (Switzerland)*, **18**, 3628.

Quy, T. B., Muhammad, S. & Kim, J. M. (2019) A reliable acoustic emission based technique for the detection of a small leak in a pipeline system, *Energies*, **12**, 1472.

Shukla, H. & Piratla, K. (2020) Leakage detection in water pipelines using supervised classification of acceleration signals, *Automation in Construction*, **117**, 103256.

Tariq, S., Hu, Z. & Zayed, T. (2021) Micro-electromechanical systems-based technologies for leak detection and localization in water supply networks: a bibliometric and systematic review, *Journal of Cleaner Production*, **289**, 125751.

Tariq, S., Bakhtawar, B. & Zayed, T. (2022) Data-driven application of MEMS-based accelerometers for leak detection in water distribution networks, *Science of the Total Environment*, **809**, 151110.

Teodorescu, L. & Sherwood, D. (2008) High energy physics event selection with gene expression programming, *Computer Physics Communications*, **178** (6), 409–419.

Tijani, I. A. & Zayed, T. (2022) Gene expression programming based mathematical modeling for leak detection of water distribution networks, *Measurement: Journal of the International Measurement Confederation*, **188**, 110611.

Tijani, I. A., Abdelmageed, S., Fares, A., Fan, K. H., Hu, Z. Y. & Zayed, T. (2022) Improving the leak detection efficiency in water distribution networks using noise loggers, *Science of the Total Environment*, **821**, 153530.

Tunkiel, A. T., Sui, D. & Wiktorski, T. (2020) Data-driven sensitivity analysis of complex machine learning models: a case study of directional drilling, *Journal of Petroleum Science and Engineering*, **195**, 107630.

Wang, Y., Shao, Z. & Tiong, R. L. K. (2021) Data-driven prediction of contract failure of public-private partnership projects, *Journal of Construction Engineering and Management*, **147** (8), 4021089.

World Economic Forum (2020) Global Risk Report. Available at: www.weforum.org/publications/the-global-risks-report-2020/ (accessed August 24, 2025).

Xu, Z., Liu, H., Fu, G., Zeng, Y. & Li, Y. (2024) Feature selection of acoustic signals for leak detection in water pipelines, *Tunnelling and Underground Space Technology*, **152**, 105945.
Yang, J., Wen, Y., Li, P. & Wang, X. (2013) Study on an improved acoustic leak detection method for water distribution systems, *Urban Water Journal*, **10** (2), 71–84.

First received 11 May 2025; accepted in revised form 21 July 2025. Available online 22 August 2025