Environmental Processes (2024) 11:14
https://doi.org/10.1007/s40710-024-00682-x

RESEARCH

Data-Driven Approaches for Vibroacoustic Localization of Leaks in Water Distribution Networks

Rongsheng Liu[1] · Salman Tariq[1] · Ibrahim A. Tijani[1] · Ali Fares[1] · Beenish Bakhtawar[1] · Harris Fan[2] · Rui Zhang[1] · Tarek Zayed[1]

Received: 23 July 2023 / Accepted: 12 January 2024

Abstract

This study aims to propose Micro-electromechanical System (MEMS) accelerometers for leak localization in the water distribution network and assess the performance of machine learning models in accurately estimating leak locations. Intensive field experimentation was conducted to collect data for model development. Machine learning algorithms were employed to develop leak localization models, specifically artificial neural network (ANN) and support vector machine (SVM). Seventeen time-domain and frequency-domain features were extracted, and feature selection was performed using the backward elimination method. The results indicate that the ANN and SVM models are suitable classifiers for localizing leak distance. Both models achieved leak location predictions with over 80% accuracy, and the mean absolute errors were measured at 0.858 and 0.95 for the ANN and SVM models, respectively. The validation results demonstrated that the models maintained accuracies close to 80% when the distance between sensors and the leak was less than 15 m. However, the performance of the model deteriorates when leaks occur at distances greater than 15 m. This study demonstrates the applicability of MEMS accelerometers for leak localization in water distribution networks. The findings highlight the promising potential of employing MEMS accelerometers-based ANN and SVM models for accurate leak localization in urban networks, even under real-world, uncontrolled conditions. However, the current model exhibits limited performance in long-distance leak localization, requiring further research to address and resolve this issue.

Keywords Data-Driven · Leak Localization · Acoustic Measurements · Water Distribution Networks · Micro-Electromechanical System (MEMS) Accelerometers · Artificial Neural Network · Support Vector Machine

✉ Ali Fares
ali-i.fares@connect.polyu.hk

[1] Department of Building and Real Estate, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China

[2] Development Engineer, Applied Technology Integration, Ltd, Tsuen Wan, New Territories, Hong Kong, China

Published online: 26 February 2024

1 Introduction

According to recent statistics, water lost through leaks and bursts constitutes one-third of the total water supplied through distribution systems (Vrachimis et al. 2021). This non-revenue water (NRW) causes enormous financial losses (Puust et al. 2010). In Hong Kong alone, leaks and bursts cost US$173 million worth of damages annually (Gupta 2017). False alarms and localization errors cost time, money, and considerable effort. Unnecessary interference with the soil-pipeline environment can also cause damage to infrastructure. Therefore, accurate and timely leak localization is one of the pressing challenges in the world of water infrastructure expertise. Early localization of leaks can save repair costs by helping practitioners actively plan pipe repair/replacements (Hu et al. 2021).

Traditionally, leak localization methods can be divided into observation-based methods and data-driven methods (Covas et al. 2005). The observation-based methods require skilled workers to manually survey the pipes using listening sticks and ground microphones for condition assessment and to find the leak location. However, manual observation methods are time-consuming and labor-intensive. Besides, the accuracy largely depends on the experience of workers (Muggleton and Brennan 2004; Fahimipirehgalin et al. 2021). Data-driven methods, on the other hand, are mainly based on data modeling. Such data include water flow, pressure changes, in-pipe acoustics, or pipe wall vibrations. They are collected via sensing technologies either on-site or wirelessly. For example, Candelieri et al. (2014a) proposed a spectral clustering-based approach to analyze the water flow and localize leak points. The dataset was derived from a hydraulic simulation model and transformed into a similarity graph applying spectral clustering. Similarly, Wang et al. (2019b) adopted the spectral-based method to localize pipeline leaks. Alternatively, acoustics sensors have increasingly been used for leak localization (Gao et al. 2022; Cui et al. 2023).

However, the research mentioned above is mainly based on laboratory experiments and has low practicality. The proposed method can find leaks in a single leak scenario and provide a potential alternative for localizing multiple leaks. Lin et al. (2008) recently proposed a feature engineering-based localization method using multiple sensors. Leaks were identified and localized by modeling extracted features. Hu et al. (2021) conducted a critical literature survey to compare the performance of localization methods. They concluded that acoustic methods provide a balanced localization ability, simultaneously considering sensitivity, accuracy, error rate, time, investment cost, and service time.

Acoustic methods localize leaks using signals collected from pipelines non-invasively through acoustic sensors like accelerometers, hydrophones, or noise loggers. Microelectromechanical system (MEMS) accelerometers are low-cost (Guru Manikandan et al. 2021), require less initial deployment cost compared to hydrophones and noise loggers, and are more effective in plastic pipes (Tariq et al. 2021b, a). The application of MEMS-based accelerometers for leak detection was recently investigated by Tariq et al. (Tariq et al. 2021a). Their use for leak detection in the Hong Kong pipeline systems was successfully reported (Tariq et al. 2021a). As a follow-up, this study further investigates using MEMS accelerometers to locate leaks in real urban water systems such as Hong Kong.

The length of the water distribution network (WDN) of Hong Kong is more than 8605 km. Hong Kong Water Supplies Department (WSD) has made great efforts in the last two decades and reduced the leakage rate from exceeding 25% in 2000 to about 15% in 2019 (Water Supplies Department 2020). These efforts include replacing deteriorated pipelines and establishing new monitoring systems (Yue and Tang 2011). Data-driven models based on machine learning assist in increasing efficiency and reducing losses. The current study, therefore, adopts a machine learning-driven methodology for facilitating water leak localization, increasing model accuracy and efficiency. The objectives of this study are three-fold: 1) to investigate the capability of MEMS-based accelerometers for localizing water pipe leaks in real WDNs; 2) to evaluate the performance of support vector machine (SVM) and artificial neural network (ANN) for localizing leaks in WDNs; and 3) to develop a model to localize the pipe leakages in real networks.

This study presents a novel approach utilizing accelerometers for leak localization in water distribution networks. The model result demonstrates the applicability of MEMS accelerometers and machine learning models for leak localization and points out the current limitation. By introducing these innovative methods, this study provides potential solutions for efficient water resource management in other megacities.

2 Literature Review

The vibration signal-based localization methodology can be categorized into traditional signal processing, beamforming, and statistical methods (Hu et al. 2021). Traditional signal processing mainly involves denoising the leak signal and extracting informative signal components for localizing leaks. For example, Mahmutoglu and Turk (2018) proposed an innovative passive acoustic system for leak localization. Based on background noise, detection method, receiver number, signal strength, and measurement number, leaks can be found with low average position errors from several kilometers away. On the other hand, Mahmutoglu and Turk (2019) used signal strength differences to locate pipeline leaks without any information on leak signal strength in absolute terms. Ting et al. (2021) recently introduced a dual-tree complex wavelet transform for water pipe signal processing. The proposed denoising algorithm highlights the peak of the cross-correlation function and improves the localization accuracy manifold. These signal-processing methods have greater robustness and ease of practical implementation. However, the results are sensitive to environmental uncertainties, such as different pipe materials, ambient noise, and soil characteristics.

Beamforming is also an alternative signal processing technique for obtaining directional signals from the sensor array. For example, (Wang and Ghidaoui 2018, 2019; Wang et al. 2019a) adopted an iterative beamform method for locating multiple leaks. Moreover, beamforming has been widely applied to detect the acoustic signal under a noisy background. For example, Maxit et al. (2022) also adopted the beamform technique to improve the signal-to-noise ratio (SNR) of collected vibration data. The array gains are calculated using both the traditional and beamforming techniques. The result shows that the beamforming has improved performance, resulting in a much larger array gain. Overall, the beamforming method benefits from higher accuracy and reliability, requires more sensors and experiment points, and suffers from the high economic cost. Zhi et al. (2023) utilized cross-correlation for leak localization. Agrawal et al. (Agrawal et al. 2023) and Kousiopoulos et al. (2022) utilized the technique of multiple time difference arrival, relying on cross-correlation calculations using signals acquired from pairs of sensors. However, the validation majorily took place through laboratory testing on bare pipes (not buried).

Lastly, statistical methods use extensive data to pinpoint the locations of pipe leaks. By analyzing the collected data, the characteristics of the signal can be summarized and proposed, thereby aiding in the localization of leaks. Advanced statistical techniques like machine learning-based regression and classification algorithms are considered effective for leak localization. Applied algorithms include ANN, SVM, Bayesian, and convolutional neural networks (Poulakis et al. 2003; Jin et al. 2014; El-Abbasy et al. 2016; Tijani and Zayed 2022). For instance, El-Abbasy et al. (2016) developed regression and ANN models based on the acoustic data from noise loggers. The developed models have been validated and proven to have around 90% localization accuracy in the established laboratory testbed. Zhou et al. (2019) innovatively proposed a deep-learning framework to localize pipe bursts further. The validation results show that the model has better robustness and applicability than the linear neural network. Recently, Quiñones-Grueiro et al. (2021) attempted to solve the leak location problem from an inverse perspective and developed a deep learning model considering the topology of the WDN, modifying the location space and combining time series. The result reveals that the proposed model performs well on the extensive pipe network of 268 nodes and 9 sensors. El-Zahab et al. (2022) developed a wireless leak detection and localization system for a building. In the study, vibration sensors were used to test the effectiveness of a wireless system to find leaks in PVC and iron pipes. More details about data-driven approaches and the use of machine learning can be found in recent literature review articles (Yussif et al. 2023; Nimri et al. 2023). Overall, data-driven methods have better performance as compared to other methods.

Previous research has made significant contributions to understanding the pipe leak localization problem. However, most studies are based on lab/testbed experiments (Martini et al. 2015; El-Zahab et al. 2018; Li et al. 2021). Due to the complicated influencing factors on-site, the leak models developed by lab experiments cannot be directly applied to a real WDN (El-Abbasy et al. 2016). First, the actual water distribution has high background noise, which does not conform to assumptions made while conducting experiments in the lab/testbed. Additionally, tests are performed on simplified pipe networks (e.g., a straight pipe) (Mostafapour and Davoudi 2013; El-Abbasy et al. 2016), which cannot reveal the complexity of the in-service pipe network (Tariq et al. 2021a). Thus, there is a need to test methods on real sites with reported leaks. In line with the argumentation and data availability, machine learning has been adopted in the current study as the optimum choice for leak localization in real water networks. In particular, SVM and ANN were adopted and compared to identify their robustness to address the leak localization problem.

3 Research Method

Figure 1 depicts the outline procedures to develop a localization model for water pipe leaks. The adopted research methodology can be divided into data collection, analysis, and machine learning modeling. First, field experiments were conducted on the WDN of Hong Kong in collaboration with the water supplies department (WSD). As the current study has been conducted with the assistance of the WSD, rich field data has been collected. MEMS accelerometers were deployed to collect signals from leak points. The collected information was further analyzed and processed in the data analysis step. In this step, we filtered the collected data to reduce the bias imposed by outliers. In the modeling phase, the feature optimization algorithm was applied to extract the acoustic characteristics of the dataset. Then, the parameters of the machine learning models were optimized. The proposed model can output the leak distances, which are defined as the distance between the deploy site and the leak point. Finally, models with the optimized parameters were validated using samples from other WSD sites.

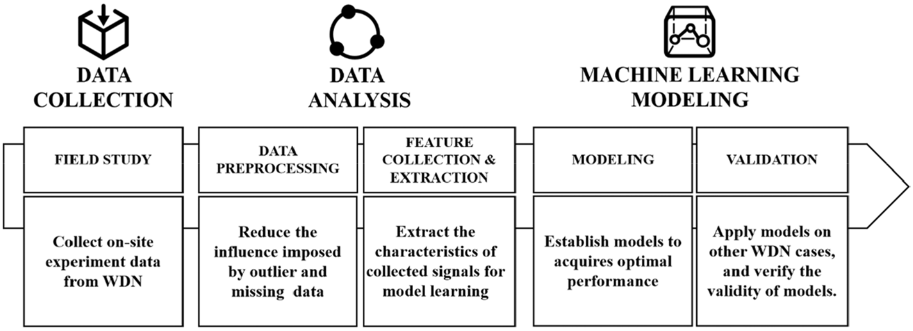

DATA COLLECTION
DATA ANALYSIS
MACHINE LEARNING MODELING
FIELD STUDY
DATA PREPROCESSING
FEATURE COLLECTION & EXTRACTION
MODELING
VALIDATION
Collect on-site experiment data from WDN
Reduce the influence imposed by outlier and missing data
Extract the characteristics of collected signals for model learning
Establish models to acquires optimal performance
Apply models on other WDN cases, and verify the validity of models.

Fig. 1 Research framework for developing leak localization models

3.1 Data Collection

The experiment sites included the municipal water distribution network and the underground pipe leak detection center established by WSD. The experimental duration ranged from Oct 1st, 2020 to Oct 19th, 2021. First, WSD reported leak points and provided the water pipe network map and corresponding pipe information (including pipe diameter, pipe material, and connection), as shown in Fig. 2. The research team subsequently designed the experimental plan and deployed sensors on proposed sites. For each site, 1 to 3 accelerometers were deployed depending on the availability of surrounding chambers. Data collection at each site lasted for one to three days, collecting data at different time points for one minute duration. Leak distance was defined as the distance between the leak point and the location where the sensor was deployed. The field experiments were always conducted at midnight, avoiding the impact of noise caused by surrounding traffic vehicles and passengers. Furthermore, the pipe flow is relatively stable at night.

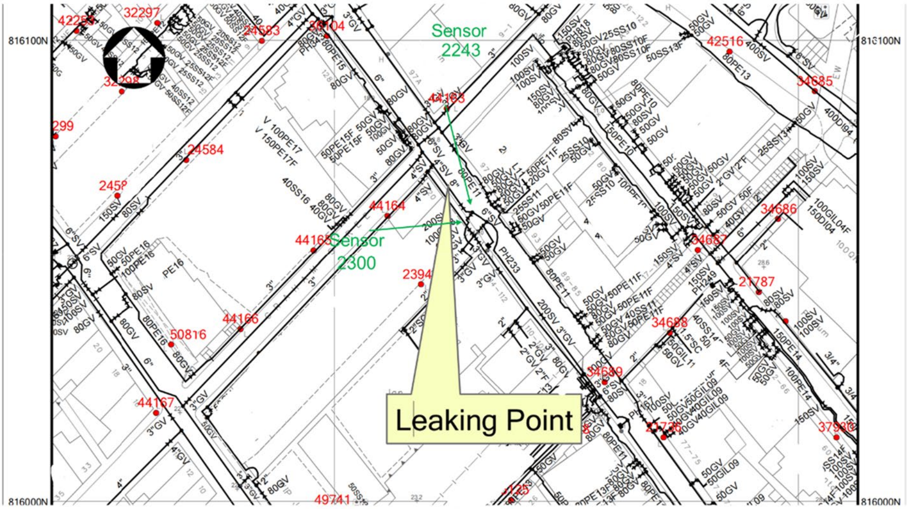


Fig. 2 A typical location in real water supply network where the MEMS accelerometers were deployed

The experiment equipment was MEMS accelerometer brand Beanair, enabling wireless data transmission and collecting time-synchronized data. Figure 3 depicts how the accelerometers were used to collect signals on-site. The sampling frequency was 3000 Hz and the sample duration was 60 s. Therefore, each sample had 180000 data points. The data was collected from the Z-axis (Tariq et al. 2021b). The accelerometer sensors were deployed on the valve within the chamber. The gateway (the central controller) was connected to the laptop, sending commands to sensors through Beanscape software. Overall, 1347 samples were collected and saved in the format of text (g values at every 1/3000 s) for further analysis.

3.2 Data Preprocessing

The quality of training data heavily impacts the performance of machine learning models (Budach et al. 2022). In this study, the collected acoustic dataset suffered from missing data (caused by inconsistent signal transmission) and background noises (emitted by vehicles and pedestrians). Therefore, it is necessary to adopt data preprocess, laying the foundation for the modeling phase.

We adoptrd the outlier removal process to eliminate the errors brought by noise signals. The interquartile rule (IQR) method was applied to find outliers. Specifically, quartiles and maximum and minimum values of signal data were computed. IQR and upper and lower bound values were calculated using Eq. (1), where Q1 represents the 25th percentile of the data; Q3 represents the 75th percentile of the data. Any data values above the upper and below the lower bounds were removed.

$$\begin{gathered} IQR = Q_3 - Q_1 \\ \text{Upper bound} = Q_1 - 1.5IQR \\ \text{Lower bound} = Q_3 + 1.5IQR \end{gathered} \tag{1}$$

Ultimately, a total of 1347 samples remained. Table 1 depicts the details of the data distribution. Most samples (84%) were collected within a 15 m of leak distance from metal

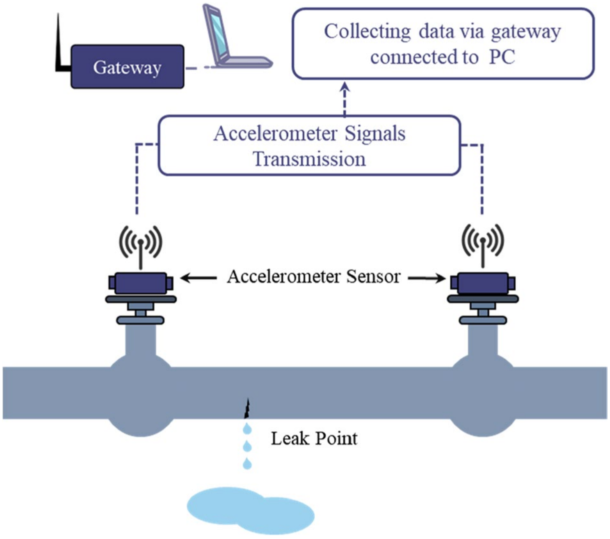

Gateway
Collecting data via gateway connected to PC
Accelerometer Signals Transmission
Accelerometer Sensor
Leak Point

Fig. 3 Data collection system

pipelines. Distance is the label of the dataset, and its unit is meter (m). Considering that this study did not use a deep learning model and the complexity of the model was relatively low, a more significant proportion of data was assigned to the test set and validation set to evaluate the model. A hold-out method was adopted, and 1347 samples were randomly split into the training dataset (808 samples), testing dataset (269 samples), and validation dataset (270 samples), taking up 60%, 20%, and 20%, respectively.

3.3 Feature Extraction and Selection

Adopted features constitute the most critical building block for machine learning modeling, assisting to summarize the characteristics of signals and promote model learning (Li et al. 2017). The quality of modeling training data and feature extraction determines the performance optimality of machine learning models. Parameter optimization and experiment testing facilitate achieving more optimum performance. Feature extraction consists of two steps: feature collection and feature selection.

A literature review was conducted to select water pipe leak localization features in the feature collection step. Various studies adopted different combinations of acoustic-based features. These features consist of time, frequency, and power dimensions. A total of 17 features were initially introduced from the literature.

Table 2 depicts the corresponding expression of these features, where x_i and x_f refers to the signal in the time specturm and frequency spectrum. f represents the frequency band. df is the resolution of the spectrum. N denotes the number of data samples. F is the maximum frequency in a spectrum. T is the time length of the signal. j is the index when $x_i - x_j$ is the smallest over the signal. R_{xx} is the autocorrelation function of x. μ_{xx} and σ_{xx} are the mean and standard deviation of R_{xx}. n is the number of data where x_f is larger than 33 percent of max x_f. μ_f and σ_f are mean and the standard deviation of x_f.

Table 3 depicts the value distribution of extracted features. The distribution is categorized by leak distances, including lower than 3.5 m, between 3.5 m to 15 m, and larger than 15 m. The comparison of the values of each feature was visualized through box plots appended in Supplementary Material (SM) Figures SM.1 to SM.17. The result shows that there is generally a distinction in values for different leak distances. Though values of features under different leak distances overlap, machine learning models can utilize various features and fit the multiple dimensions dataset. In the next stage, the initially collected 17 features would be analyzed and finalized.

Though features help to develop the relationship between the input dataset and target output (Mitra et al. 2002; Naghibi et al. 2015), inappropriate feature selection would decrease model accuracy and performance (Li et al. 2017). Thus, feature selection and optimization is urgently needed for establishing the water pipe leak localization models. Current feature optimization methodologies can be divided into three

Table 1 Details of the finalized dataset

Distance (m)	Total Number	Proportion
<3.5 m	850	62%
3.5 m to 15 m	282	21%
>15 m	215	17%
Total	1347	100%

Table 2 Result and expression of feature collection

Features	Expression	Features	Expression
Root mean square (*RMS*)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$	Average Amplitude (*FDAvgAmp*)	$\frac{1}{N}\sum_{f=0}^{F} x_f$
Level	$20log_{10}\frac{RMS}{20\times10^{-6}}$	Peak Frequency (*PF*)	f when $x_f = max\left(x_f\right)$
Spread (*Sp*)	$max\left(x_i\right) - min\left(x_i\right)$	Maximum Amplitude (*MA*)	$max\left(x_f\right)$
Average amplitude (*TDAvgAmp*)	$\frac{1}{N}\sum_{i=1}^{N} \lvert x_i\rvert$	Frequency Centroid (*FC*)	$\frac{\sum_{f=0}^{F} x_f f}{\sum_{f=0}^{F} x_f}$
Peak amplitude (*PA*)	$max\left(\lvert x_i\rvert\right)$	Skewness (*Sk*)	$\frac{1}{N}\sum_{f=0}^{F}\left(\frac{x_f-\mu_f}{\sigma_f}\right)^3$
The crest factor (*CF*)	$\frac{max(\lvert x_i\rvert)}{RMS}$	Kurtosis (*Ku*)	$\frac{1}{N}\sum_{f=0}^{F}\left(\frac{x_f-\mu_f}{\sigma_f}\right)^4$
Energy (*En*)	$\frac{1}{N}\sum_{i=1}^{N} x_i^2$	Frequency Spread (*FS*)	$df\sum_{i=0}^{N} x_f > 0.33max\lvert x_f\rvert$
Maximum Lyapunov Exponent (*MLE*)	$\frac{4}{T}\sum_{i=\frac{N}{4}}^{\frac{N}{2}}\frac{1}{5}\sum_{k=1}^{5}\frac{1}{k}ln\left\lvert\frac{x_{i+k}-x_{j+k}}{x_i-x_j}\right\rvert$		
Kurtosis of Autocorrelation Function (*AKu*)	$\frac{1}{N}\sum_{f=0}^{F}\left(\frac{R_{xx}-\mu_{xx}}{\sigma_{xx}}\right)^4$		
Maximum Lyapunov Exponent of Autocorrelation Function (*AMLE*)	$\frac{4}{T}\sum_{i=\frac{N}{4}}^{\frac{N}{2}}\frac{1}{5}\sum_{k=1}^{5}\frac{1}{k}ln\left\lvert\frac{R_{xx,i+k}-R_{xx,j+k}}{R_{xx,i}-R_{xx,j}}\right\rvert$		

Table 3 Range of the values of the extracted features

Feature	< 3.5 m		3.5–15 m		> 15 m	
	MIN	MAX	MIN	MAX	MIN	MAX
Level	34.40	54.16	34.35	66.73	36.11	38.92
Spread	0.24	12.24	0.00	22.83	0.24	5.23
RMS	1.06E-03	1.06E-02	1.05E-03	4.29E-02	1.31E-03	1.75E-03
TD Avg. Amp	8.53E-04	8.39E-03	8.26E-04	3.28E-02	1.04E-03	1.39E-03
Peak Amp	2.81E-03	5.33E-02	2.93E-03	4.21E-01	3.37E-03	7.05E-03
Crest Factor	2.42	5.08	2.15	11.98	2.41	4.36
Energy	2.43E-07	1.12E-03	2.11E-07	1.34E-02	3.59E-07	1.62E-06
MLE	-5.68	431.97	-3.49	2759.21	0.00	172.33
Autocorr. Kurt	72.71	502.61	3.22	309.82	40.29	269.95
Autocorr. MLE	0.00	4149.71	0.00	4229.28	0.00	293.26
FD Avg. Amp	5.63E-05	1.51E-04	7.23E-05	5.52E-03	7.26E-05	1.57E-04
Peak Freq	3.04	489.58	59.08	332.75	2.57	410.83
Max. Amp	1.83E-04	2.09E-03	3.17E-04	3.50E-02	2.57E-04	7.99E-04
Freq. Centroid	251.55	343.04	172.40	366.52	271.12	314.84
Skewness	0.41	3.72	0.65	6.40	0.67	2.15
Kurtosis	2.26	22.39	3.32	75.94	2.63	12.04
Freq. Spread	2.00E-02	2.25E-01	1.33E-03	1.73E-01	1.07E-02	1.14E-01

types, Filter (Peng et al. 2005; Wang et al. 2008), Wrapper (Yang and Ong 2011; Mafarja and Mirjalili 2018), Embedded method (Guyon et al. 2002).

Considering that the acoustic features are generally co-correlated and data volume is limited, this study has adopted the Wrapper method to optimize the features. Regarding Wrapper, this method would packaged all features as a feature set. Each sub-feature set would be used to establish a model to learn or fit the target dataset. The performances of these models were compared to find the optimum feature set (Liu and Wang 2021). Though Wrapper considers the combination of features, it requires a significant additional time when there are numerous features. In addition, Wrapper might suffer from overfitting when there are limited data samples (Maldonado and Weber 2009).

Specifically, the backward selection, as a category of Wrapper method, was selected. First all features were used for modeling. Then, it temporarily discards one feature to check whether the model performance has decreased. If such a decrease is observed, the feature is retained; otherwise, it is dropped. This process is repeatedly conducted for each feature. Because the feature optimization results are varied for different models, the backward selections are respectively conducted on ANN and SVM. The 17 features that are presented in Table 2 were used as the initial feature set for the backward selection algorithm.

Feature selection for the SVM yielded 15 critical features, including *Level*, *Spread*, *RMS*, *TD Avg. Amp*, *Peak Amp.*, *Energy*, *MLE*, *Autocorr. MLE*, *FD Avg. Amp.*, *Peak Freq.*, *Max. Amp.*, *Freq. Centroid*, *Skewness*, *Kurtosis*, *Autocorr. MLE*, *Freq. Spread*. Only four features were selected for the ANN model, which included *Skewness*, *RMS*, *Level*, and *Freq. Centroid*. The optimized features were adopted to train the models.

3.4 Modeling and Optimization

Leak distance estimation in water distribution networks presents a complex challenge due to the various distinct characteristics of this task (Tyagi et al. 2023). The multi-dimensional nature of sensor data, including parameters like pressure, flow rate, and acoustic signals, increases the complexity (Cody and Narasimhan 2020). Additionally, the relationships between these sensor parameters and leak distances are often non-linear, making traditional linear methods less effective. Real-world data introduces noise and variability into the equation, further complicating accurate predictions.

In this context, SVM and ANN offer unique strengths compared to other machine learning models. SVM excels at recognizing complex patterns within multi-dimensional data, making it suitable for capturing intricate relationships between sensor readings and leak distances. ANN, especially deep neural networks, is highly adept at modeling non-linear relationships, which is essential for accurate leak distance prediction (Fan et al. 2021).

In summary, SVM and ANN are well-suited for leak distance estimation due to their capacity to address the multi-dimensional, non-linear, and noisy aspects of this task. Thus, this study adopted SVM and ANN for subsequent modeling.

3.4.1 Basic Theory

(i) *Artificial Neural Network (ANN)*: The artificial neural network is inspired by the structure of bio-neuron (Abiodun et al. 2018). It is designated to solve multi-dimensional and complicated problems (Nagajothi and Elavenil 2020). An ANN model mainly consists of three layers: input, hidden, and output, as shown in Fig. 4. The input layer represents dataset features, and the hidden layer, which can vary in complexity, is the core of the ANN, and the output layer produces model results. In literature, ANN has been widely applied to solve the leak problems in gas (Wang

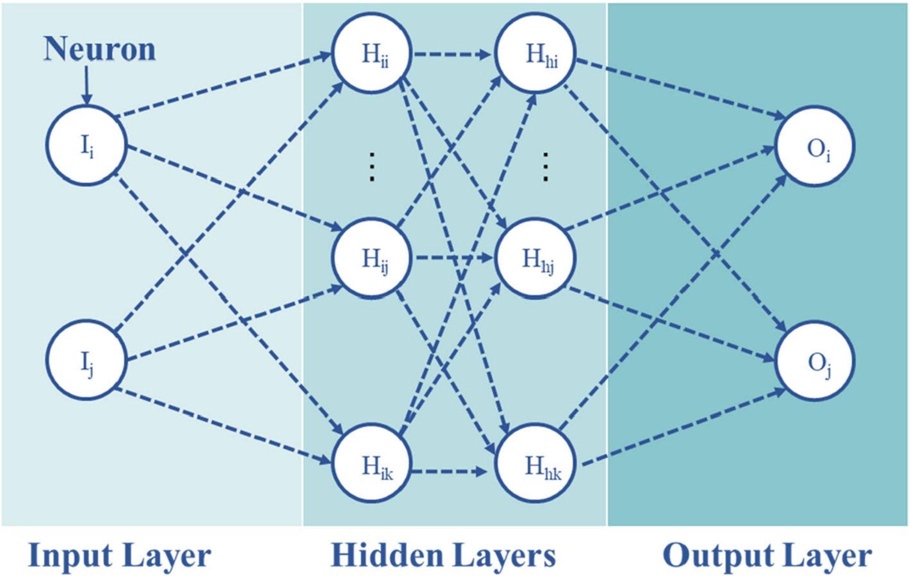


Fig. 4 Basic Structure for artificial neural network (ANN) model

et al. 2021) and water (Sattar et al. 2019; Almheiri et al. 2020) pipelines. Thus, it is introduced as an effective method for pinpointing water leaks in the actual water distribution network case of Hong Kong.

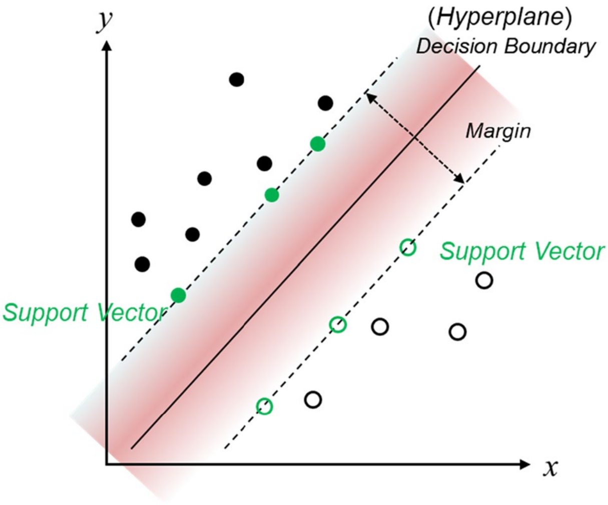
(ii) *Support vector machine (SVM)*: Support vector machine is a supervised learning approach based on the linear classifier (Cortes and Vapnik 1995). As shown in Fig. 5, the SVM model tries to establish a hyperplane to classify different samples. The samples closest to the hyperplane are defined as the support vector. Margin is the distance between the support vector and hyperplane. SVM model is trained by maximizing the margin. Meanwhile, introducing the kernel helps the model to establish a more complicated hypersurface (decision boundary), solving the linear inseparable problem (Jain et al. 2018). In the literature, SVM reaches promising performance in solving water pipe leak problems (Mounce et al. 2011; Mashford et al. 2012).

(iii) *Performance Evaluation of Machine Learning Models*: The performance of the classification models can be directly evaluated using explicit indicators such as prediction accuracy, recall rate, and precision. However, the regression model output is continuous and requires different types of performance indicators, including the mean square error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). These indicators have been widely used in significant regression models (Chicco et al. 2021) and are presented in Eqs. (2) to (4).

The values of MAE, MSE and RMSE denote the model error. Thus, the model with the lower MAE, MSE, and RMSE has a better model prediction. In Eqs. (2) to (4), $\hat{y}_t$ denotes the value deduced by the model. y_t denotes the real value of samples. and T denotes the total number of samples.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}\left(\hat{y}_t - y_t\right)^2}{T}} \tag{2}$$

Fig. 5 The basic concept of SVM



$$MSE = \frac{\sum_{t=1}^{T} \left(\hat{y}_t - y_t\right)^2}{T} \tag{3}$$

$$MAE = \frac{\sum_{i=1}^{T} \left|\hat{y}_t - y_t\right|}{T} \tag{4}$$

MAE directly reveals the real bias or error between real and fake values. Its robustness to outliers and interpretability make it suitable for practical understanding. RMSE emphasizes large errors and retains the same unit of measurement compared to MSE (Chai and Draxler 2014; Hodson 2022).

This study also introduces the model accuracy as the indicator. If the leakage distance output by the model deviates from the actual leakage distance by less than 1 m, it will be considered as correctly identified. Therefore, the model accuracy reflects the proportion of cases where the leakage prediction error is less than 1 m. Thus, the current study adopts RMSE, MAE, and Accuracy as the model performance indicators.

3.4.2 Model Optimization

(i) *ANN model optimization:* In the current study, ANN was developed using RapidMiner. The proposed model adopts Sigmoid as the activation function between the input and hidden layers. In the structure of ANN, the number of neurons in the input layer is 4, corresponding to the features mentioned in the last section. However, there is no specific methodology to determine the structure of ANN (Jin et al. 2021). In addition, the learning rate and the training recycling also significantly affect the model performance. Thus, experiments and tests were conducted to regulate the model structure to reach the optimum result.

Considering ANN only uses four features, this study adopted one hidden layer to simplify and avoid overfitting. Figure 6 depicts the performance of ANN models with different neurons. The architectures featuring 6 and 9 hidden neurons demonstrate superior performance, reaching closely comparable RMSE and MAE. Considering that complex structures demand higher computational resources and may potentially give rise to overfitting issues (Sun et al. 2017), this study adopts hidden layer comprising 6 neurons for the subsequent modeling phase. The overall structure of ANN is shown in Fig. 7. Based on parametric experiments and empirical rules, the learning rate is 0.03, and the training epochs are 300.

The established ANN models were transformed into a mathematical expression to describe the relationship among input attributes, variables, and the final output. Hence, the weights and biases of the selected optimum architectures are presented in Table 4, which were used accordingly to develop the prediction expression.

Subsequently, the resulting mathematical expression based on the weights and biases of the model for WDNs is presented in Eq. (5). *LD* denotes the leak location formulation. x_i can be obtained based on the weights and bias in Table 4. $f(x)$ denotes the activation function:

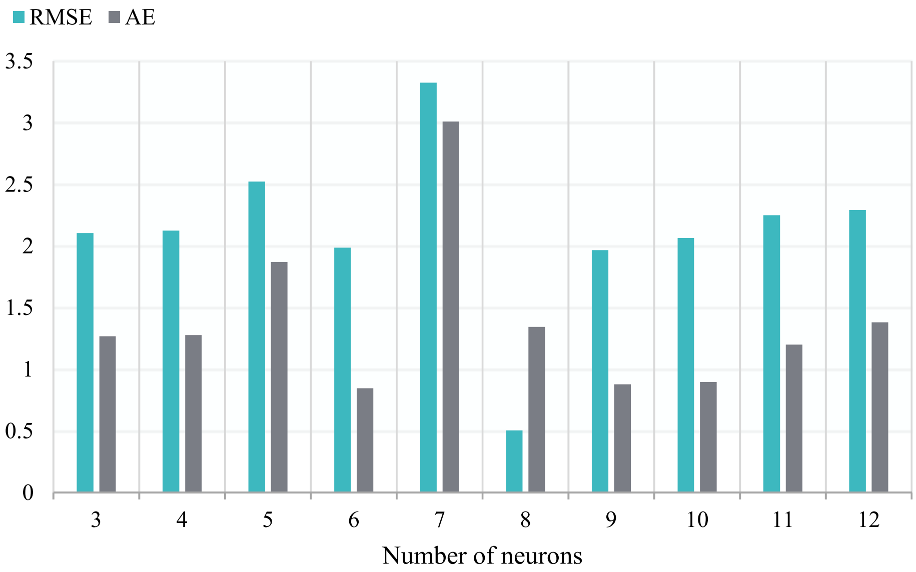


Fig. 6 Impacts of the number of neurons on ANN performance

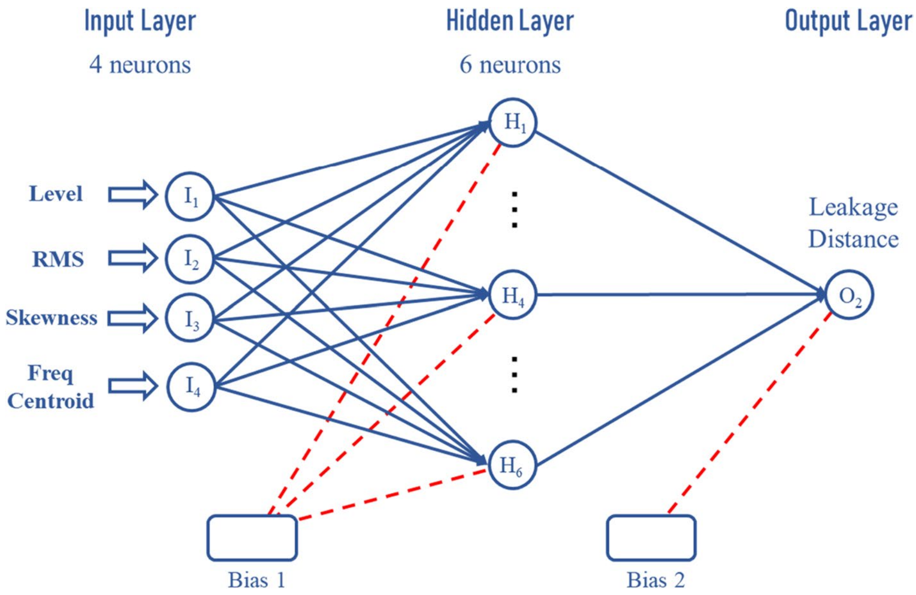


Fig. 7 Selected optimum ANN structure

Table 4 Weights and biases of the selected optimum ANN structure

No. of the node in the hidden layer	Weights					Bias	
	W$_1$				W$_2$		
	Level	RMS	Freq. Centroid	Skewness		B$_1$	B$_2$
1	6.643	0.236	-1.638	4.248	-2.128	1.514	0.872
2	1.267	2.248	9.14	1.251	0.887	-3.09	
3	-24.439	13.47	16.253	-8.222	-1.976	-18.919	
4	0.996	-1.075	0.433	1.46	0.438	-3.186	
5	1.334	-0.7	0.64	0.889	0.311	-3.174	
6	1.017	-0.795	-2.721	-0.005	1.007	-2.136	

$$\begin{aligned}
x_1 &= w_{1-1} \times f\begin{bmatrix} 6.643\ \mathit{level} + 0.23\ \mathit{RMS} - 1.638\ \mathit{Freq.Centroid} \\ +4.248\ \mathit{Skewness} + 1.514 \end{bmatrix} \\
x_2 &= w_{1-2} \times f\begin{bmatrix} 1.267\ \mathit{level} + 2.248\ \mathit{RMS} + 9.14\ \mathit{Freq.Centroid} \\ +1.251\ \mathit{Skewness} - 3.09 \end{bmatrix} \\
x_3 &= w_{1-3} \times f\begin{bmatrix} -24.439\ \mathit{level} + 13.47\ \mathit{RMS} + 16.253\ \mathit{Freq.Centroid} \\ -8.222\ \mathit{Skewness} - 18.919 \end{bmatrix} \\
x_4 &= w_{1-4} \times f\begin{bmatrix} 0.996\ \mathit{level} - 1.075\ \mathit{RMS} + 0.433\ \mathit{Freq.Centroid} \\ +1.46\mathit{Skewness} - 3.186 \end{bmatrix} \\
x_5 &= w_{1-5} \times f\begin{bmatrix} 1.334\ \mathit{level} - 0.7\ \mathit{RMS} + 0.64\ \mathit{Freq.Centroid} \\ +0.889\mathit{Skewness} - 3.174 \end{bmatrix} \\
x_6 &= w_{1-6} \times f\begin{bmatrix} 1.017\ \mathit{level} - 0.795\ \mathit{RMS} - 2.721\ \mathit{Freq.Centroid} \\ -0.005\ \mathit{Skewness} - 2.136 \end{bmatrix} \\
LD &= \textstyle\sum_{i=1}^{6} x_i - B_2
\end{aligned} \tag{5}$$

(ii) *SVM model optimization:* The performance of SVM is greatly dependent on the type of kernel function k and the corresponding parameters (Liu et al. 2016). This study adopted the radial kernel, the Gaussian kernel function (Liu et al. 2011). The main parameter of the radial kernel is the penalty coefficient C and kernel gamma, γ.

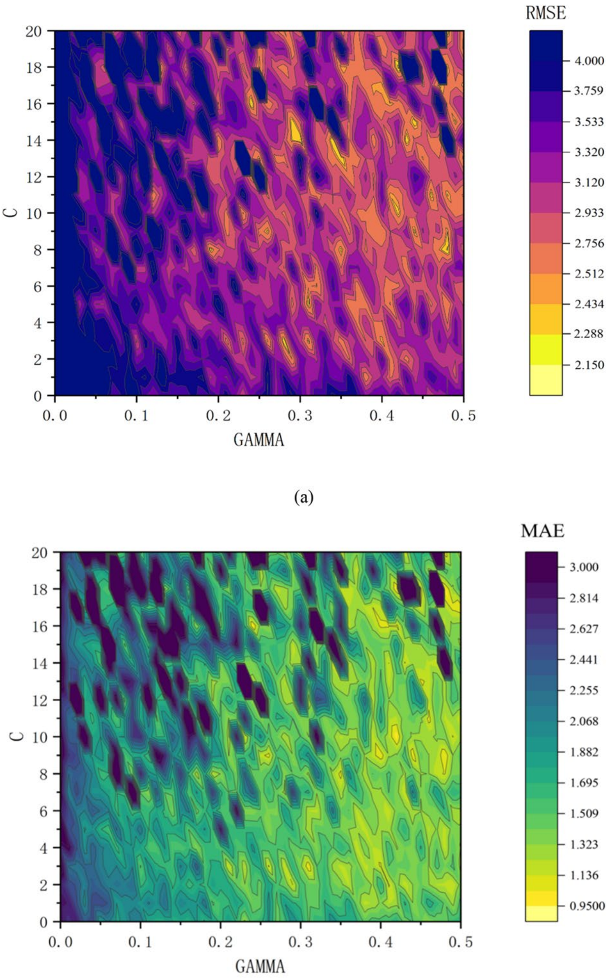
The above two parameters mainly balance the relationship between the model complexity and the error rate. When C is large, the loss function will be more significant, and the model gives up distant outliers. SVM establishes a more complex hyperplane to fit the distant sample and reach higher accuracy. However, it is also easier to cause the overfitting problem.

On the other hand, γ is the other parameter of the kernel function. It mainly defines the influence of a single sample on the whole classification hyperplane. When γ is small, a single sample imposes a greater impact on the hyperplane, and the sample is easier to be selected as a support vector.

Overall, SVM is more complicated and have more support vector when C and γ are large. Conversely, the model would be simpler when C and γ are minor. Thus, the setting of C and γ should be optimized by comparing the RMSE and MAE of models.

Figure 8 depicts the optimization result of SVM. The analysis reveals that RMSE falls within the range of 2 to 4 for most SVM models. As observed in Fig. 8, models exhibited lower RMSE when γ and C values are high. However, it is crucial to exercise caution with

Fig. 8 Optimization results of the SVM model for different parameters: **a** RMSE; **b** MAE



(a)

(b)

elevated γ and C values, as they may induce overfitting. To balance the model performance and structure complexity, SVM parameters were configured as follows: γ is 0.42, and C is 10. Table 5 shows the weights of 15 adopted features and biases for better describing the established signal. The total number of support vectors is 809.

4 Results

4.1 Model Testing Results

Table 6 presents the test results of ANN and SVM. Regarding RMSE, the value of ANN (1.9960) is about 20% less than SVM (2.5410). The mean average error of ANN is ±0.8580m,

Table 5 Weights of SVM model

Feature	Weight
Level	-0.9463
Spread	-19.2997
RMS	-3.1598
TD Avg. Amp	-4.4032
Peak Amp	-0.5368
Energy	-4.6384
MLE	-1.7758
Autocorr. MLE	-26.5902
FD Avg. Amp	12.5322
Peak Freq	-50.8835
Max. Amp	1.3475
Freq. Centroid	-13.2211
Skewness	1.5192
Kurtosis	-4.9189
Freq. Spread	-68.0868
Bias	3.555
Total number of Support Vector	809

Table 6 Testing results of ANN and SVM model

	RMSE	MAE	True Prediction	False Prediction	Accuracy Rate
ANN	1.996	0.858	232	37	86.25%
SVM	2.541	0.95	217	52	80.67%

Table 7 Validation results of ANN and SVM models

	RMSE	MAE	Accuracy Rate	True Prediction	False Prediction
ANN	2.486	1.074	80.37%	217	53
SVM	2.699	1.088	78.89%	213	57

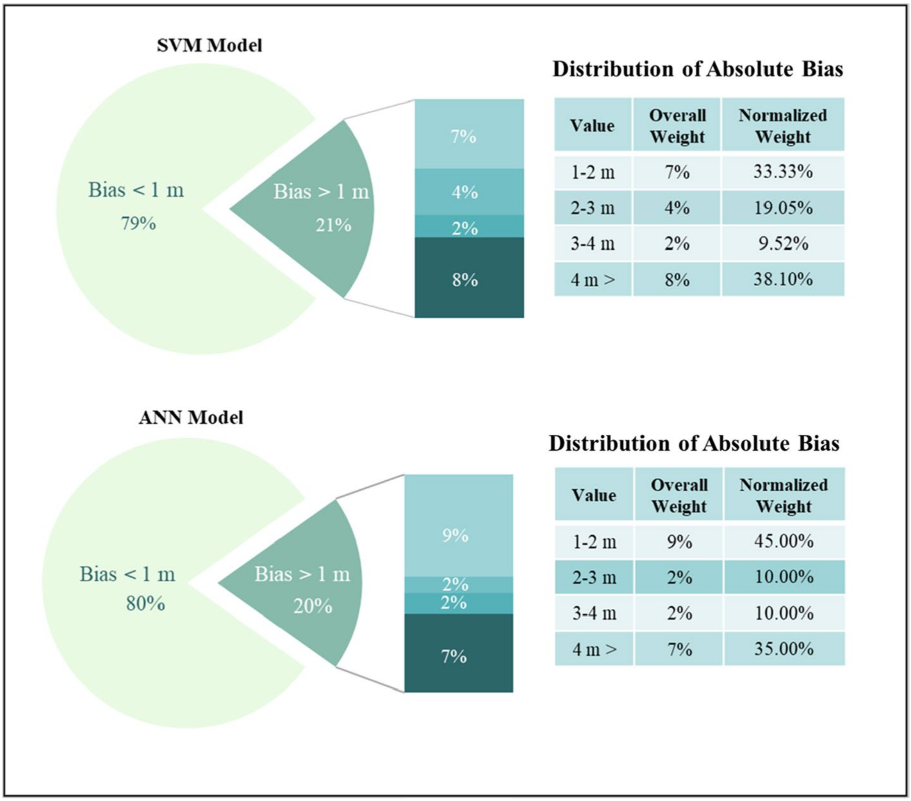
which is more than 0.1 m lower than the respective value of SVM ($\pm 0.9500m$). The accuracy of ANN (86.25%) is more than 6% better than SVM (80.67%). Regarding the training results, ANN might have a more accurate prediction ability than SVM. On the other hand, both the MAE of ANN and SVM are lower than 1 m. The prediction accuracies of the two models are over 80%. The testing result reveals that the proposed ANN and SVM models have been well-trained and achieved promising localization ability. However, it still needs further validation.

4.2 Model Validation Results

For model validation, this study randomly extracted 20% of the dataset before model training, including 270 samples (167 samples were collected within 3.5 m, and 57 samples were collected from leak distances between 3.5 m to 5 m; 46 samples were collected from leak distance larger than 15 m). The trained ANN and SVM models have been respectively applied to the validation set. When models fail to meet the validation requirements, it necessitates revisiting the data preprocessing stage to identify and address issues in order to enhance model performance.

Table 7 shows the validation result of ANN and SVM. Regarding RMSE, MAE, and accuracy, the performance of ANN was slightly better than SVM. The accuracy gap between ANN and SVM is lower than 2%, with 80.37% and 78.89%, respectively. Overall, the MAE of ANN and SVM are close to 1 m, and the accuracy rates are close to 80%. The model validation performance is generally lower than the performance in the testing set. However, the performance of ANN and SVM is still acceptable, with MAE 1.074 and 1.088, denoting that the performance of SVM and ANN initially meets the standard requirement.

Figure 9 depicts the distribution of absolute bias from the SVM and ANN models. The absolute bias denotes the absolute value of the difference between the model distance and the real distance. From the perspective of the SVM model, nearly 21% of the bias was larger than 1 m. Only 7% of bias was between 1 and 2 m. The bias among 2 m to 3 m and 3 m to 4 m only took up 4% and 2% of the total sample, respectively. Nearly half of the samples were from 1 to 2 m. The bias distribution of ANN was similar to SVM, but the



Distribution of Absolute Bias

Value	Overall Weight	Normalized Weight
1-2 m	7%	33.33%
2-3 m	4%	19.05%
3-4 m	2%	9.52%
4 m >	8%	38.10%

Distribution of Absolute Bias

Value	Overall Weight	Normalized Weight
1-2 m	9%	45.00%
2-3 m	2%	10.00%
3-4 m	2%	10.00%
4 m >	7%	35.00%

Fig. 9 The absolute bias distribution of ANN and SVM model

Table 8 The validation result from the perspective of leak distance

Model	Distance (m)	Total Number	Number of correct	Number of Error	Accuracy	MAE
ANN	< 3.5 m	167	20	20	88.02%	0.661315
	3.5–15	58	10	10	82.76%	0.75357
	> 15 m	45	23	23	48.89%	3.031617
SVM	< 3.5 m	167	22	22	86.83%	0.502375
	3.5–15	58	3	3	94.83%	0.179026
	> 15 m	45	32	32	28.89%	4.430931

overall bias of ANN was slightly lower than SVM. In this regard, most biases were below or close to 1 m, and only nearly 7% to 8% of the samples had more than 4 m bias.

In the validation set, the location of the water leak mainly consisted of three ranges, including lower than 3 m, between 3.5 and 15 m, and larger than 15 m. The validation results categorized into different leak distances are shown in Table 8. Regarding ANN, the model reached 88.02% accuracy for leaks within the range below 3.5 m and 82.76% accuracy for leaks within the range of 3.5 m to 15 m. However, the localization accuracy was decreased to 48.89% in the leak distance larger than 15 m. When the leak distance was lower than 15 m, MAE in other points was all under or close to 1 m. Regarding SVM, most leaks between 3.5 and 15 m have been correctly predicted, with 94.83% accuracy. Regarding leaks lower than 3.5 m, SVM achieved over 86.83% accuracy. However, the model was invalid in finding leaks larger than 15 m, with only 28.89% accuracy. When the distances of leaks were lower than 15 m, MAE values were all below 0.51. When the distances of leaks were larger than 15 m, the MAE value was larger than the value in other points, with 4.43.

5 Discussion

The validation results show that ANN and SVM reach nearly 80% accuracy and the MAE close to 1 m. However, it is worth noting that other studies have achieved a higher localization performance. For instance, Candelieri et al. (2014b) have developed SVM and achieved nearly 98% accuracy. Besides, Fan and Yu (2021) also developed a WDN machine learning model with over 83% localization accuracy. Though previous models reached high-level prediction accuracy, most were conducted based on hypothetical conditions or simulated experiments. The modeling and validation data are characterized by clarity and less noise. This controlled setting may explain the higher prediction accuracy they achieved.

In contrast, when the laboratory-based model is applied to real-world scenarios, its performance may significantly deteriorate. Real-world conditions introduce various complexities and uncertainties that are not present in controlled laboratory settings, leading to decreased model performance. The performance of the laboratory-based model was significantly deteriorate when applied to other scenarios (Terao and Mita 2008; Tariq et al. 2021a). It is crucial to consider the limitations and potential performance degradation when comparing laboratory-based models to field-experiment-based models. Thus, the proposed models still demonstrate their applicability in accurately localizing leaks in real-world scenarios.

Regarding leak distance, both models demonstrate limited performance in finding leaks at long distances (larger than 15 m). This is caused by the attenuation of the acoustic signal during propagation. As the acoustic signal travels longer distances, it gradually weakens and loses its strength, making it more challenging for the models to identify and accurately locate leaks (Muggleton and Brennan 2004; Almeida et al. 2015). Similarly, Bui Quy and Kim (2020) also find difficulty locating pipeline leaks influenced by attenuation. Regarding accuracy using the mean absolute error (MAE) and the root mean square error (RMSE), the ANN and SVM models demonstrated comparable levels of accuracy.

6 Conclusions

This study adopted an acoustic methodology, using the Micro-electromechanical System (MEMS) accelerometer to localize pipe leaks in the Hong Kong water distribution network. First, field experiments were conducted to collect accelerometer signals. Subsequently, backward selection was respectively applied to ANN and SVM models for feature selection. As a result, four features were selected for the ANN model, while the SVM model utilized fifteen features. Through model optimizing experiments, parameters of ANN and SVM were finalized, reaching the optimum performance. Finally, models have been tested and validated through various perspectives.

According to modeling results, ANN and SVM reach promising performance on the testing set, with over 80% accuracy (86.25% for ANN; 80.67% for SVM), and MAE lower than 1 m (0.858 for ANN; 0.95 for SVM). However, the model performance decreased in the validation cases, indicating that models have an encouraging result when finding long-distance leaks. Overall, the above results have proved that the applicability of implementing MEMS accelerometer signals in localizing water pipe leaks, which reveals the potential of applying machine learning models, ANN and SVM, to pinpoint water pipe leaks.

However, certain limitations still hinder the validity of current models. First, the experiments were conducted in the city center, and the maximum detection range of the method in real situations has not been thoroughly tested. Second, because of the limitation of the collected dataset, the model cannot correctly handle the long-distance leaks. Third, the localization accuracy can be further improved by utilizing larger databases and enhancing data. Thus, future research should investigate using MEMS accelerometers that can be permanently installed and enable wireless remote distance signal transmission. Transfer learning algorithms can be used in water leak localization, which might help to combine and best use the previous models and water pipe leak datasets.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40710-024-00682-x.

Author contributions Conceptualization and writing - original draft preparation: RongSheng Liu and Salman Tariq; Formal analysis and investigation: RongSheng Liu; Salman Tariq and Ibrahim A. Tijani; Writing - review and editing: Ali Fares, Harris Fan; and Rui Zhang; Funding acquisition and supervision: Tarek Zayed.

Funding Open access funding provided by The Hong Kong Polytechnic University. 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 The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Competing interests The authors declare no competing interests.

Competing interest The authors report there are no competing interests to declare.

References

Abiodun OI, Jantan A, Omolara AE et al (2018) State-of-the-art in artificial neural network applications: A survey. Heliyon 4:e00938. https://doi.org/10.1016/j.heliyon.2018.e00938

Agrawal P, Fong S, Friesen D, Narasimhan S (2023) Maximum Likelihood Estimation to Localize Leaks in Water Distribution Networks. J Pipeline Syst Eng Pract 14:04023038. https://doi.org/10.1061/JPSEA2.PSENG-1494

Almeida FCL, Brennan MJ, Joseph PF et al (2015) Towards an in-situ measurement of wave velocity in buried plastic water distribution pipes for the purposes of leak location. J Sound Vib 359:40–55. https://doi.org/10.1016/j.jsv.2015.06.015

Almheiri Z, Meguid M, Zayed T (2020) Intelligent Approaches for Predicting Failure of Water Mains. J Pipeline Syst Eng Pract 11:04020044. https://doi.org/10.1061/(asce)ps.1949-1204.0000485

Budach L, Feuerpfeil M, Ihde N et al (2022) The effects of data quality on machine learning performance. arXiv preprint arXiv:220714529. https://doi.org/10.48550/arXiv.2207.14529

Bui Quy T, Kim J-M (2020) Leak detection in a gas pipeline using spectral portrait of acoustic emission signals. Measurement 152:107403. https://doi.org/10.1016/j.measurement.2019.107403

Candelieri A, Conti D, Archetti F (2014a) Improving Analytics in Urban Water Management: A Spectral Clustering-based Approach for Leakage Localization. Procedia Soc Behav Sci 108:235–248. https://doi.org/10.1016/j.sbspro.2013.12.834

Candelieri A, Soldi D, Conti D, Archetti F (2014b) Analytical Leakages Localization in Water Distribution Networks through Spectral Clustering and Support Vector Machines. Icewater Approach Procedia Eng 89:1080–1088. https://doi.org/10.1016/j.proeng.2014.11.228

Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci Model Dev 7:1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chicco D, Warrens MJ, Jurman G (2021) The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci 7:1–24. https://doi.org/10.7717/PEERJ-CS.623

Cody RA, Narasimhan S (2020) A field implementation of linear prediction for leak-monitoring in water distribution networks. Adv Eng Inform 45:101103. https://doi.org/10.1016/j.aei.2020.101103

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018

Covas D, Ramos H, de Almeida AB (2005) Standing Wave Difference Method for Leak Detection in Pipeline Systems. J Hydraul Eng 131:1106–1116. https://doi.org/10.1061/(asce)0733-9429(2005)131:12(1106)

Cui X, Gao Y, Ma Y et al (2023) Time delay estimation using cascaded LMS filters fused by correlation coefficient for pipeline leak localization. Mech Syst Signal Process 199:110500. https://doi.org/10.1016/j.ymssp.2023.110500

El-Abbasy MS, Mosleh F, Senouci A et al (2016) Locating leaks in water mains using noise loggers. J Infrastruct Syst 22:04016012. https://doi.org/10.1061/(asce)is.1943-555x.0000305

El-Zahab S, Al-Sakkaf A, Mohammed Abdelkader E, Zayed T (2022) A machine learning-based model for real-time leak pinpointing in buildings using accelerometers. J Vibr Control 107754632110662. https://doi.org/10.1177/10775463211066247

El-Zahab S, Mohammed Abdelkader E, Zayed T (2018) An accelerometer-based leak detection system. Mech Syst Signal Process 108:58–72. https://doi.org/10.1016/j.ymssp.2018.02.030

Fahimipirehgalin M, Trunzer E, Odenweller M, Vogel-Heuser B (2021) Automatic Visual Leakage Detection and Localization from Pipelines in Chemical Process Plants Using Machine Vision Techniques. Engineering 7:758–776. https://doi.org/10.1016/j.eng.2020.08.026

Fan X, Yu X (2021) An innovative machine learning based framework for water distribution network leakage detection and localization. Struct Health Monit. https://doi.org/10.1177/14759217211040269

Fan X, Zhang X, Yu XB (2021) Machine learning model and strategy for fast and accurate detection of leaks in water supply network. J Infrastruct Preserv Resil 2:1–21. https://doi.org/10.1186/s43065-021-00021-6

Gao Y, Piltan F, Kim J-M (2022) A Hybrid Leak Localization Approach Using Acoustic Emission for Industrial Pipelines. Sensors 22:3963. https://doi.org/10.3390/s22103963

Gupta A (2017) Hong Kong is wasting a third of its water. https://chinadialogue.net/en/cities/9803-hong-kong-is-wasting-a-third-of-its-water/. Accessed 17 Jul 2023

Guru Manikandan K, Pannirselvam K, Kenned JJ, Suresh Kumar C (2021) Investigations on suitability of MEMS based accelerometer for vibration measurements. Mater Today: Proc 45:6183–6192. https://doi.org/10.1016/j.matpr.2020.10.506

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422. https://doi.org/10.1023/A:1012487302797

Hodson TO (2022) Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geosci Model Dev 15:5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Hu Z, Tariq S, Zayed T (2021) A comprehensive review of acoustic based leak localization method in pressurized pipelines. Mech Syst Signal Process 161:107994. https://doi.org/10.1016/j.ymssp.2021.107994

Jain DK, Dubey SB, Choubey RK et al (2018) An approach for hyperspectral image classification by optimizing SVM using self organizing map. J Comput Sci 25:252–259. https://doi.org/10.1016/j.jocs.2017.07.016

Jin H, Zhang L, Liang W, Ding Q (2014) Integrated leakage detection and localization model for gas pipelines based on the acoustic wave method. J Loss Prev Process Ind 27:74–88. https://doi.org/10.1016/j.jlp.2013.11.006

Jin Y, Wang H, Sun C (2021) Introduction to machine learning. Data-driven evolutionary optimization: integrating evolutionary computation, machine learning and data science, pp 103–145. https://doi.org/10.1007/978-3-030-74640-7_4

Kousiopoulos G-P, Kampelopoulos D, Karagiorgos N et al (2022) Acoustic Leak Localization Method for Pipelines in High-Noise Environment Using Time-Frequency Signal Segmentation. IEEE Trans Instrum Meas 71:1–11. https://doi.org/10.1109/TIM.2022.3150864

Li J, Cheng K, Wang S et al (2017) Feature Selection: A Data Perspective. ACM Comput Surv 50:1–45. https://doi.org/10.1145/3136625

Li Y, Zhou Y, Fu M et al (2021) Analysis of propagation and distribution characteristics of leakage acoustic waves in water supply pipelines. Sensors 21(16):5450. https://doi.org/10.3390/s21165450

Lin S-W, Ying K-C, Chen S-C, Lee Z-J (2008) Particle swarm optimization for parameter determination and feature selection of support vector machines. Exp Syst Applic 35:1817–1824. https://doi.org/10.1016/j.eswa.2007.08.088

Liu Q, Chen C, Zhang Y, Hu Z (2011) Feature selection for support vector machines with RBF kernel. Artif Intell Rev 36:99–115. https://doi.org/10.1007/s10462-011-9205-2

Liu W, Wang J (2021) Recursive elimination–election algorithms for wrapper feature selection. Appl Soft Comput 113:107956. https://doi.org/10.1016/j.asoc.2021.107956

Liu Y, Pi D, Cheng Q (2016) Ensemble kernel method: SVM classification based on game theory. J Syst Eng Electron 27:251–259. https://doi.org/10.1109/JSEE.2016.00025

Mafarja M, Mirjalili S (2018) Whale optimization approaches for wrapper feature selection. Appl Soft Comput 62:441–453. https://doi.org/10.1016/j.asoc.2017.11.006

Mahmutoglu Y, Turk K (2018) A passive acoustic based system to locate leak hole in underwater natural gas pipelines. Digital Signal Process: Rev J 76:59–65. https://doi.org/10.1016/j.dsp.2018.02.007

Mahmutoglu Y, Turk K (2019) Received signal strength difference based leakage localization for the underwater natural gas pipelines. Appl Acoust 153:14–19. https://doi.org/10.1016/j.apacoust.2019.04.006

Maldonado S, Weber R (2009) A wrapper method for feature selection using Support Vector Machines. Inf Sci 179:2208–2217. https://doi.org/10.1016/j.ins.2009.02.014

Martini A, Troncossi M, Rivola A (2015) Automatic Leak Detection in Buried Plastic Pipes of Water Supply Networks by Means of Vibration Measurements. Shock and Vibration 2015. https://doi.org/10.1155/2015/165304

Mashford J, De Silva D, Burn S, Marney D (2012) Leak detection in simulated water pipe networks using SVM. Appl Artif Intell 26:429–444. https://doi.org/10.1080/08839514.2012.670974

Maxit L, Karimi M, Guasch O, Michel F (2022) Numerical analysis of vibroacoustic beamforming gains for acoustic source detection inside a pipe conveying turbulent flow. Mech Syst Signal Process 171:108888. https://doi.org/10.1016/j.ymssp.2022.108888

Mitra P, Murthy CA, Pal SK (2002) Unsupervised feature selection using feature similarity. IEEE Trans Pattern Anal Mach Intell 24:301–312. https://doi.org/10.1109/34.990133

Mostafapour A, Davoudi S (2013) Analysis of leakage in high pressure pipe using acoustic emission method. Appl Acoust 74:335–342. https://doi.org/10.1016/j.apacoust.2012.07.012

Mounce SR, Mounce RB, Boxall JB (2011) Novelty detection for time series data analysis in water distribution systems using support vector machines. J Hydroinf 13:672–686. https://doi.org/10.2166/hydro.2010.144

Muggleton JM, Brennan MJ (2004) Leak noise propagation and attenuation in submerged plastic water pipes. J Sound Vib 278:527–537. https://doi.org/10.1016/j.jsv.2003.10.052

Nagajothi S, Elavenil S (2020) Influence of Aluminosilicate for the Prediction of Mechanical Properties of Geopolymer Concrete – Artificial Neural Network. SILICON 12:1011–1021. https://doi.org/10.1007/s12633-019-00203-8

Naghibi T, Hoffmann S, Pfister B (2015) A semidefinite programming based search strategy for feature selection with mutual information measure. IEEE Trans Pattern Anal Mach Intell 37:1529–1540. https://doi.org/10.1109/TPAMI.2014.2372791

Nimri W, Wang Y, Zhang Z et al (2023) Data-driven approaches and model-based methods for detecting and locating leaks in water distribution systems: a literature review. Neural Comput Applic 35:11611–11623. https://doi.org/10.1007/s00521-023-08497-x

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Poulakis Z, Valougeorgis D, Papadimitriou C (2003) Leakage detection in water pipe networks using a Bayesian probabilistic framework. Probab Eng Mech 18:315–327. https://doi.org/10.1016/S0266-8920(03)00045-6

Puust R, Kapelan Z, Savic DA, Koppel T (2010) A review of methods for leakage management in pipe networks. Urban Water J 7:25–45. https://doi.org/10.1080/15730621003610878

Quiñones-Grueiro M, Ares Milián M, Sánchez Rivero M et al (2021) Robust leak localization in water distribution networks using computational intelligence. Neurocomputing 438:195–208. https://doi.org/10.1016/j.neucom.2020.04.159

Sattar AMA, Ertuğrul ÖF, Gharabaghi B et al (2019) Extreme learning machine model for water network management. Neural Comput Appl 31:157–169. https://doi.org/10.1007/s00521-017-2987-7

Sun X, Sun W, Ma S et al (2017) Complex structure leads to overfitting: a structure regularization decoding method for natural language processing. arXiv preprint arXiv:1711.10331. https://doi.org/10.48550/arXiv.1711.10331

Tariq S, Bakhtawar B, Zayed T (2021a) Data-driven application of MEMS-based accelerometers for leak detection in water distribution networks. Sci Total Environ 151110. https://doi.org/10.1016/j.scitotenv.2021.151110

Tariq S, Hu Z, Zayed T (2021b) Micro-electromechanical systems-based technologies for leak detection and localization in water supply networks: A bibliometric and systematic review. J Clean Prod 289:125751. https://doi.org/10.1016/j.jclepro.2020.125751

Terao Y, Mita A (2008) Robust water leakage detection approach using the sound signals and pattern recognition. In: Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems (vol 6932, pp 697–705). SPIE. https://doi.org/10.1117/12.775968

Tijani IA, Zayed T (2022) Gene expression programming based mathematical modeling for leak detection of water distribution networks. Measurement 188:110611. https://doi.org/10.1016/j.measurement.2021.110611

Tyagi V, Pandey P, Jain S, Ramachandran P (2023) A Two-Stage Model for Data-Driven Leakage Detection and Localization in Water Distribution Networks. Water 15:2710. https://doi.org/10.3390/w15152710

Vrachimis SG, Timotheou S, Eliades DG, Polycarpou MM (2021) Leakage detection and localization in water distribution systems: A model invalidation approach. Control Eng Pract 110. https://doi.org/10.1016/j.conengprac.2021.104755

Wang S, Wang KY, Zheng L (2008) Feature selection via analysis of relevance and redundancy. J Beijing Institute Technol (English Ed) 17:300–304

Wang W, Mao X, Liang H, et al (2021) Experimental research on in-pipe leaks detection of acoustic signature in gas pipelines based on the artificial neural network. Measurement 183. https://doi.org/10.1016/j.measurement.2021.109875

Wang X, Ghidaoui MS (2018) Identification of multiple leaks in pipeline: Linearized model, maximum likelihood, and super-resolution localization. Mech Syst Signal Process 107:529–548. https://doi.org/10.1016/j.ymssp.2018.01.042

Wang X, Ghidaoui MS (2019) Identification of multiple leaks in pipeline II: Iterative beamforming and leak number estimation. Mech Syst Signal Process 119:346–362. https://doi.org/10.1016/j.ymssp.2018.09.020

Wang X, Ghidaoui MS, Lin J (2019a) Identification of multiple leaks in pipeline III: Experimental results. Mech Syst Signal Process 130:395–408. https://doi.org/10.1016/j.ymssp.2019.05.015

Wang X, Palomar DP, Zhao L et al (2019b) Spectral-Based Methods for Pipeline Leakage Localization. J Hydraul Eng 145:04018089. https://doi.org/10.1061/(asce)hy.1943-7900.0001572

Water Supplies Department (2020) WSD - Water Loss Management. In: The website of Water Supplies Department. https://www.wsd.gov.hk/en/core-businesses/operation-and-maintenance-of-waterworks/reliable-distribution-network/index.html. Accessed 26 Sep 2023

Yang JB, Ong CJ (2011) Feature selection using probabilistic prediction of support vector regression. IEEE Trans Neural Networks 22:954–962. https://doi.org/10.1109/TNN.2011.2128342

Yue DPT, Tang SL (2011) Sustainable strategies on water supply management in Hong Kong. Water Environ J 25:192–199. https://doi.org/10.1111/j.1747-6593.2009.00209.x

Yussif A-M, Sadeghi H, Zayed T (2023) Application of Machine Learning for Leak Localization in Water Supply Networks. Buildings 13:849. https://doi.org/10.3390/buildings13040849

Zhi B, Wu Z, Chen C et al (2023) A High Sensitivity AlN-Based MEMS Hydrophone for Pipeline Leak Monitoring. Micromachines 14:654. https://doi.org/10.3390/mi14030654

Zhou X, Tang Z, Xu W et al (2019) Deep learning identifies accurate burst locations in water distribution networks. Water Res 166:115058. https://doi.org/10.1016/j.watres.2019.115058

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.