Mechanical Systems and Signal Processing 108 (2018) 276–291

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

An accelerometer-based leak detection system

Samer El-Zahab *, Eslam Mohammed Abdelkader, Tarek Zayed

Department of Building, Civil and Environmental Engineering, Concordia University, Montréal, QC H3G 1M8, Canada

ARTICLE INFO

Article history:
Received 25 June 2017
Received in revised form 13 November 2017
Accepted 14 February 2018

Keywords:
Leak detection
Water main monitoring
Accelerometers
Vibration signal analysis
Asset management
Classification

ABSTRACT

Aging infrastructures, specifically pipelines, that were installed decades ago and currently operating under poor conditions are highly susceptible to the threat of leaks, which pose economic, health, and environmental risks. For example, in the year 2009, the state of Ontario lost 25% of its water supply solely due to leaks. Therefore, a need arises to develop an approach that allows condition monitoring and early intervention. This article proposes a model for a real-time monitoring system capable of identifying the existence of single event leaks in pressurized water pipelines. The model proposes that wireless accelerometers be placed within the network on the exterior of the valves connecting the pipelines. To test the viability of the proposal, experiments were performed on one-inch cast iron pipelines, one-inch and two-inch PVC pipelines using single event leaks and the results were displayed. The vibration signal derived from each accelerometer was assessed and analyzed to identify the Monitoring Index (MI) at each sensor. The data collected from experimentation were analyzed using support vector machines (SVM), Decision Tree (DT), and Naïve Bayes (NB). A leak threshold was determined such that if the signal increased above the threshold, a leak status is identified. The developed models showed promising results with 98.25% accuracy in distinguishing between leak states and non-leak states. The proposed model is aimed at presenting novel approaches to providing municipalities with an affordable real-time monitoring system capable of aiding them in early detection and facilitating the repair process of leaks.

1. Introduction

Water covers 70% of Earth. However, only 3% of the Earth's water is freshwater. The lack of fresh water results in 1.1 billion of people on Earth lack access to fresh water. Moreover, 2.7 billion suffer from a scarcity of water for at least one month a year and two million people die because of diarrheal diseases [1]. The International Water Management Institute (IWMI) stated that 45 countries (33% of the world population) will suffer from water scarcity by 2025 [2]. Most water distribution networks suffer from water leaks which may occur because of corrosion, aging pipelines, or excessive pressure resulting from operational error or opening and closing the valves rapidly. These leaks vary in size and significance. Nevertheless, when they combine with each other, they have a major effect on water loss in water distribution networks. Renzetti and Dupont [3] reported that water supply systems in North America suffer average leaks between 0.5 and 1.4 measurable leaks per mile of pipe and it can reach up to 3 leaks per mile of pipe in some cases. The average water loss in distribution networks has increased from 12.8% in 2006 to 13.3% in 2009 before reaching customers in Canada. The average water loss in

* Corresponding author.
E-mail addresses: samer.elzahab@concordia.ca, samer.alzahab@hotmail.com (S. El-Zahab), eslammohammed.abdelkader@mail.concordia.ca (E. Mohammed Abdelkader), tarek.zayed@concordia.ca (T. Zayed).

https://doi.org/10.1016/j.ymssp.2018.02.030
0888-3270/

larger cities in Canada (14.9%) is almost twice the average in smaller cities (7.6%). The average water loss is 7.5% in Newfoundland and Labrador while it is 22.11% in Quebec City [3]. In the Asian region, the average water loss is 80% and 62% in Singapore and Bangladesh respectively [4]. The average leakage and unaccounted water losses are approximately 23% in developing countries [5].

The amount of lost water is about 20%-30% in transmission networks mainly because of leaks and it can reach to 50% of the production in some older networks [6,7]. Leaks contribute significantly to water loss within transmission networks where they represent 70% and this percentage is expected to increase under low maintenance intensity [8]. Al-Aghbar [9] stated that 40% of the loss in potable water in Canada is because of leaks and deterioration. The overall grade of drinking water infrastructure in the United States is "D", which implies that the infrastructure is in a poor to fair condition with a high risk of failure [10]. Najjaran et al. [11] illustrated that 700 water main breaks are reported daily in North America with a total repair cost around 1 billion $/year. Allouche and Freure [12] stated that municipalities need approximately 1 billion USD repair damage because of water leaks.

Water main leaks can have a severe impact on existing nearby infrastructure systems such as stormwater systems, gas pipes, pavement, sewer systems which consequently may lead to cascading failures for infrastructure systems. Water loss is not the only consequence of leaks, leaks can have social and environmental consequences. For example, the United Kingdom is estimated to open about 4 million holes in the road network for the installation of the pipes and the repair of water leaks. The overall cost of the damages created by leaks in the United Kingdom is estimated to be £7 billion a year (around $10 billion) and it is divided into £1.5 billion (around $2 billion) of direct damage costs, and £5.5 billion (around $8 billion) in social impact costs [13]. Excessive leakage is likely to increase and consequently, it can lead to backflow, which is the intrusion of pathogens and contaminants from the surrounding environment under negative pressure conditions. The intrusion of contaminants may lead to a significant decrease in the water quality as well as harmful effects on human life [14].

Based on the statistics above, accurate leak detection plays a vital role in the overall integrated management of the water pipeline system as it reduces the water losses and diminishes their social and environmental impacts. The damages and negative effects encouraged researchers to develop a real-time monitoring system for water supply networks that is capable of early detection of leaks and consequently optimal time for repair activities. Several models were introduced to address the issue. Nevertheless, several limitations were encountered on the levels of accuracy, device availability, applicability, false alarms, and impacts of external conditions [15–18]. Accordingly, the objectives of this research are: (1) to develop and present a real-time monitoring system for pressurized water networks that is capable of detecting, and localizing leaks by utilizing the data gathered from accelerometers, (2) to develop a model that is capable of identifying the size of a leak, (3) to compare multiple model development techniques to identify the aspects of leak detection using accelerometers, and (4) to propose a set of thresholds to differentiate between leak and non-leak signals, in addition to classifying leaks between big leaks and small leaks.

2. Background

2.1. Accelerometers in leak detection

Accelerometers have drawn the attention of researchers recently where they can be utilized as a full leak detection system to detect vibration signals that are emitted by leaks. Pal et al. presented a methodology to locate leaks in water distribution medium density polyethylene pipes using non-invasive accelerometers and invasive hydrophone sensors with a reasonable accuracy [19]. The detection of the location of water leaks was based on the correlation between the speed of sound in water pipes and the time delay between the leaks signals at two different in the water pipe. They concluded that most of the leak signals lied in the frequency range between 20HZ and 25HZ. Ismail et al. [20] utilized an accelerometer sensor "MPU6050" to investigate the accuracy of the leak detection system of acrylonitrile butadiene styrene (ABS) pipe. The water pipe was a high-pressure pipe of length 10 m. The pipeline system was investigated based on three states namely: (1) no leakage, (2) 1-mm size of leak hole, and (3) 3 mm size of leak hole. Four different pressure levels were investigated which were: 58.84, 78.5, 98.1, and 117.68 kPa. They concluded that the size of pipe leak could be detected when the pressure varied from 58.84 to 98.1 kPa. On the other hand, it was difficult to identify the size of the leak when the pressure was 117.68 kPa.

Shinozuka et al. [21] introduced a methodology to detect the variations in water pressure that occur because of ruptures in the network as well as the damage locations. They used a sensor board that was equipped with Micro-Electro-Mechanical Sensors (MEMS) accelerometer to measure the vibrations on the surface of water pipes without the need for invasive monitoring techniques. They were able to define the damage locations of the water pipes using the generated contour maps, frequency domain analysis, and correlated acceleration data analysis.

Almeida et al. [22] studied the impact of the resonances in the pipeline on the time delay estimate. They stated that the existence of phase changes in the cross-spectral density (CSD) which occurred because of the resonances in the pipeline, caused some errors in the calculation of the time delay. They utilized two correlators which were: (1) basic cross-correlation (BCC), and (2) phase transformation (PHAT). Assume d_1 and d_2 are the two respective distances between the leak and each acoustic sensor. Therefore, d, which represents the total distance between the sensors, can be represented by Eq. (1) as the sum of d_1 and d_2. Then, the correlation formula can be shown in Eq. (2) in terms of d and d_2. After determining d_2, it can be utilized in Eq. (1) to determine the value of d_1.

$$d = d_1 + d_2 \tag{1}$$

$$d2 = (d - cT_0)/2 \tag{2}$$

where d represents the distance between the two sensors, c indicates the propagation speed of leak noise, and T_0 represents the time delay between the two sensors.

They concluded that BCC correlation function was the most suitable for the leak detection. Nevertheless, the PHAT correlation function was not efficient because it was sensitive to the phase changes that occurred due to the resonances in the pipeline. Martini et al. [23] studied the potential of using vibration signals for detecting leaks automatically. The authors conducted experimental tests with artificially induced leaks. The preliminary experiments provided insight into the potential of using vibration leaks to detect real leaks. Based on the insight provided by the previous work, the authors moved forward towards studying the efficiency of vibration models in real life detection settings. Martini et al. [24] conducted experiments on real leaks that helped to introduce an algorithm that was based on standard deviation to distinguish between leaking and non-leaking states automatically. Moreover, it can be used as a tool to detect the increment of vibration signals that occurred because of water leaks. Further experiments were conducted to improve the proposed algorithm through some filtering techniques, and consequently, the model could deal with some crucial issues such as transient environmental perturbations.

El-Zahab [25] introduced a model that can be utilized for pinpointing the location of single vent leaks in pressurized water pipes based on accelerometers that were placed on the exterior of water pipes. The vibration signals of each accelerometer were analyzed to calculate the Monitoring Index (MI) for each sensor. The experiments were performed on one-inch cast iron pipes as well as two-inch PVC pipes. The model, which was based on regression analysis, provided promising results regarding leak location within a range of 25 cm.

Accelerometers are often coupled with another leak detection technology and that is Acoustic Emission (AE). AE measurement relies on detecting incidental events such as a break or crack growing. The events are then assessed by means of three parameters: (1) Hit, which is when an acoustic signal exceeds the baseline, (2) Counts, which the number of times the threshold was surpassed, and (3) Signal amplitude, which the highest signal amplitude in a detected event. AE sensors can utilize the differences in detection time to propose an origin of the AE event, i.e., a leak location. AE emission is a powerful technique when changes in the network occur, yet when the phenomenon is stable for a sufficient period, accelerometers can prove to be much more beneficial as they aim to monitor any persistent change in the network. The two techniques are promising in the leak detection field when coupled together as they complement one another in terms of functionality, but each technique is capable of filling the role of a real time detection system as well [25–27].

2.2. Classification techniques overview

Classification is regarded as one of the most common machine learning and data mining techniques that can be implemented in intelligent decision making. The proposed model utilizes three techniques to classify leak and non-leak states and to distinguish big leaks from small leaks. Classification is mainly divided into two main stages namely: (1) generating a classifier from a set of historical records of known class values, and (2) applying the classifier to predict the category by knowing the values or the features of the attributes [28]. The training data set, or as it is sometimes referred to as "learning data set," is used to construct the classifier [29]. Each point in the training data set has several attributes. One of the characteristics is the class label or the goal attribute which indicates the class to which the data points belong. After the classifier is constructed and certified, the testing data set is utilized to predict the class label for the unclassified data.

Classification can be either prepared by statistical techniques such as discriminant analysis or artificial intelligence techniques such as K-nearest neighbors, support vector machines, decision trees, and artificial neural networks. The three methods that were utilized for the classification of the collected data are (1) Linear Support Vector Machines, (2) Decision Tree and (3) Naïve-Bayes. The selection of the three supervised learning classifiers for leak detection is due to the difference in their respective natures. This difference can help provide a solid and comprehensive platform for comparison. For instance, SVM is based on defining the optimum hyperplane that maximizes the margin width between positive and negative classes [30]. DT is based on defining classification rules and designing a graphical flowchart-like structure (composed of nodes and connection between nodes) whereas, each node represents a test on the attribute [28]. NB is a probabilistic classifier that depends on the Bayesian theorem whereas an element belongs to the class that has the maximum posterior probability [31]. The three classification techniques will be further discussed in the following sections.

2.2.1. Support vector machines

Support vector machines (SVM) is a supervised learning technique that can be utilized in either classification or regression applications. Support vector machines were initially proposed by Cortes and Vapnik in 1995 for classification purposes [32]. SVM are capable of learning and modeling both linear and complex (non-linear) mapping functions. The most straightforward form for SVM is linear SVM where it can be utilized for the data that are linearly separable in the current space or the original space [33].

The support vector machines technique is based on defining the optimum hyperplane by maximizing the margin between positive and negative classes [34]. The data points that are located on the hyperplane are called "support vectors" [35]. Assume a set of training samples of size N which can be represented as (x_i, y_i). x_i represents the input vectors where y_i

represents the output vectors (label vectors). The support vector machines technique is based on solving an optimization problem which tends to maximize the margin width and minimize the classification error as shown in Eqs. (3) and (4).

$$Minimize \frac{1}{2}||w^2|| + C\sum_{i=1}^{N}\zeta_i \tag{3}$$

$$Subject\ to\ y_i(w^T \cdot x_i + b) \geqslant 1 - \zeta_i,\ \ \zeta_i \geqslant 0 \tag{4}$$

where w represents weighting vector perpendicular to the hyperplane. C denotes a penalty parameter. ζ represents a positive slack variable. C and ζ are introduced to permit misclassification. b denotes a bias term. $||w||$ represents the Euclidean norm for the term w. The optimal hyperplane is the hyperplane that creates the maximum margin ($||w^{-2}||$), i.e., maximum separating distance between the hyperplane and the nearest data points. The objective of the SVM is to define the hyperplane that creates the maximum margin ($||w^{-2}||$), i.e., maximum separating distance between the hyperplane and the nearest data points subject to the two constraints stated in Eq. (4).

In some classification problems, the data points may not be linearly classified and consequently kernel functions have to be utilized. Non-linear classification can be implemented using non-linear functions called "kernel functions." Non-linear classification is performed by mapping the data to a high-dimensional space without knowing this mapping space where the linear classification is feasible. The kernel function is implemented to calculate the inner product of two vectors in a high-dimensional space. The kernel functions need to follows Mercer's condition as shown in Eq. (5). The three most common kernel functions are polynomial, radial basis, and sigmoid functions as shown in Eqs. (6), (7), and (8), respectively [33,34].

$$\Phi(x) \cdot \Phi(x_i) = K(x_i, x_j) \tag{5}$$

$$K(x_i, x_j) = \exp(||x_i - x_j||/2\sigma^2) \tag{6}$$

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^p \tag{7}$$

$$K(x_i, x_j) = \tan(\alpha(x_i \cdot x_j) + \beta) \tag{8}$$

where Φ indicates the mapping function. σ, p, α, and β are adjustable kernel parameters.

2.2.2. Decision trees

A decision tree is a technique that is used for data classification based on a graphical representation of procedures [36]. It consists of a set of nodes that can be either connections or terminal nodes as well as the connections between nodes to provide an unbiased variable selection. The construction of decision trees passes through three main stages: (1) establishing a fitting model for each node, (2) selecting a split point or a variable for each node by using some algorithms to provide unbiased estimates such as Classification rule with an unbiased interaction selection and estimation (CRUISE), and (3) The algorithms tend to prune the prediction error or misclassification error for each subtree [37].

There are two main types of decision trees where the selection of one type over the other depends on the nature of the model under development. Those two categories of decision trees are: (1) regression trees and (2) classification trees. A regression model is constructed at each node in the case of regression trees. As for classification trees, a classification model is implemented to minimize the cost function, i.e., minimize the misclassification error.

2.2.3. Naïve-Bayes

A Naïve-Bayes classifier is a probabilistic classifier that is based on Bayes' theory [31]. Naïve-Bayes has the following advantages: (1) ease of implementation, and (2) the need for only one scan of the training dataset to generate probabilities. The Naïve-Bayes classifier assumes class conditional independence which means that the impact of a specific attribute in a specific class is independent on other characteristics. The Naïve-Bayes classifier can deal with missing attribute values by excluding the probabilities of attributes when calculating the probability of membership for each class [38]. The Naïve-Bayes classifier is conducted by defining the class with maximum posterior probability [39].

Assume a training dataset D which consists of set of records $\{x_1, x_2, x_3, x_4 \ldots x_n\}$ and set of classes $\{C_1, C_2, C_3, C_4 \ldots C_m\}$ where the Naïve-Bayesian classifier predicts that x belongs to the class that has the highest posterior probability.

$$P(C_i|x) = \frac{[P(x|C_i) \times P(C_i)]}{P(x)} \tag{9}$$

where The class C_i to which $P(x|C_i)$ is maximized is known as "Maximum Posteriori Hypothesis". $P(x)$ is constant for all classes and the term $P(x|C_i) \times P(C_i)$ needs to be maximized. In case the class prior probabilities are un-known, it is assumed that all classes are likely equal to occur where $P(C_1) = P(C_2) \ldots = P(C_m)$, and consequently maximizing the term $P(x|C_i)$ is the target of the classification problem.

The Naïve-Bayesian classifier assumes conditional independence in order to minimize the computational effort. Accordingly, the attribute values are independent of each other given a certain class. The term $P(x|C_i)$ can be calculated using the following equation.

$$P(x|C_i) = \prod_{k=1}^{N} P(x_k|C_i) \tag{10}$$

where x_k refers to the attribute value.

2.3. Accelerometers signal analysis

During the experimentation process of this paper, multiple challenges were faced regarding deciphering the signal derived from accelerometers. After the development of several models, an alternative model was selected for having more accurate representations of the signal. The model was developed by Martini et al., who proposed a model to utilize accelerometers for leak detection. Additionally, Martini suggested an approach for analyzing the signal received from the sensors [24]. Their mathematical analysis approach can be summed in the following steps:

1. Determine acceleration reading per second in (g).
2. Each t = 100 s, the readings are collected, and their standard deviation is determined.
3. After monitoring for several hours, the lowest ten standard deviations are averaged to determine the lowest monitoring index using Eq. (11). The lowest value, or the baseline value, that is created from the lowest 10 observations is called MI0.

$$MI_j = mean(\sigma_j, 10) \tag{11}$$

4. A value named Monitoring Index Efficiency (MIE) is determined by dividing the current monitoring index of any instant, MI_j, with the lowest monitoring index in no leak state, MI_0, as illustrated in Eq. (12). This equation allows the establishment of sensor-specific values. For example, MI_0 can be unique for each sensor as well as the readings, therefore taking into consideration any pre-existing conditions and external factors. Martini originally utilized the maximum monitoring index of a duration, whereas in this research MIE is determined each t = 100 s.

$$MIE_x = MI_j/MI_0 \tag{12}$$

where
- MIE is the Monitoring Index Efficiency.
- x can be either L or R representing left or right sensor respectively.
- MI_0 is the lowest recorded monitoring index at no leak state.
- MI_j is the monitoring index of a given signal at time j.

3. Methodology and model development

3.1. Overall study methodology

Fig. 1 presents the general methodology that was utilized for model development. The first phase of research was to use the current advancements in leak detection and classification techniques to devise the required experiments. The second phase was to set up a series of related experiments to aid in studying the interaction between leaks and accelerometers. Using the collected data from the trials, the models would be developed using the techniques: (1) Linear SVM, (2) Decision Tree, and (3) Naïve Bayes. After the development of the models, the models would be cross-validated using the available data sets and their accuracy determined. In case the models were inaccurate, the development process would be reassessed as well as the experimentation process. Finally, the accuracy and validity of the developed models would be determined, and the best models would be selected.

The developed models will be assessed using three performance metrics, which are (1) Accuracy, (2) Class recall, and (3) Class precision. Accuracy can be defined as the percentage of correct detections over the total number of detections that the model has performed. Class recall can be defined as "the number of correctly classified positive cases divided by the total number of actual positive cases in the dataset." The class recall is used to calculate the percentage of positive classes that are classified accurately. On the other hand, class precision can be defined as "the number of correctly classified positive cases divided by the number of cases that are classified positive by the model." Class precision is used to calculate the percentage of positive classes which are accurately predicted from the total predicted classes in the positive class [40,41].

3.2. Model development methodology

The production of the models in this paper started with defining the index for assessment (MIE) and defining the required target which is an MIE threshold for identifying various leak states. As illustrated in Fig. 2, the next step would be to perform experimentation on PVC and ductile iron pipelines of sizes one inch and two inches each. The analyzed data of the

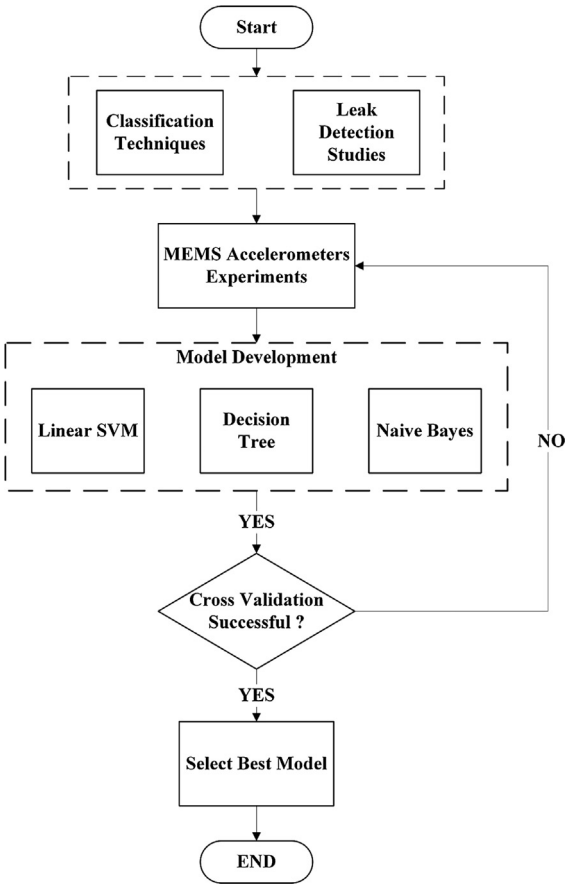


Fig. 1. Overall research methodology.

experiments were collected and prepared to be fed into three model development tools using the Rapid Miner 7.4 platform. Each technique would develop a model using the acquired data; then the respective models would be validated using cross-validation. The models that pass cross-validation with the highest accuracy are selected, and their results would be compared to define the required thresholds.

In this paper, two types of models are targeted. Leak identification models and leak size classification models. Leak identification models aim to find a particular plane that separates the leak state from the no leak state (existent or non-existent). The input for these models is a data set that contains the value of MIE versus the state of the leak. Leak size classification models go a step further in terms of trying to identify the size of the leak. Those models attempt to detect small leaks ranging from 10% of the flow rate up to 25% and big leaks ranging from 25% to 50% of the total flow rate within the pipeline. The assumed ranges resemble by the available markings on the valves utilized in the experiment.

4. Implementation and data collection

4.1. Experiments performed

For the following experiments, the accelerometers used were of the brand Beanair and the model AX3D. The sensors had two main sensitivity modes ±2 g and ±10 g. In the trails, the sensitivity of ±2 g was utilized. Additionally, the sensors operated on a frequency range between 0 Hz and 800 Hz. The sensors operated using low duty cycle data acquisition (LDCDA) mode with a minimum sampling rate of one sample per second (SPS) and it can reach up to 3000 SPS if one axis was used. In the following experiments, the least sampling rate was using and that is one sample per second [42]. The devices where adhered on the exterior of the pipelines, specifically the valves, using strong duct tape to ensure minimum external disruption to the signal. Additionally, a regular house pump was utilized to pump the water into the pipelines. The pump provided a steady flow rate of 3 m^3/h (30 L/min).

To understand the reaction of accelerometers in the presence of leaks, multiple experiments were required on various levels. The first level was carried out to understand the readings noted by the MEMS devices. Thus, at this stage, the first step was to setup any pipeline and place multiple sensors over the pipe. Fig. 3(a) displays the general configuration of all

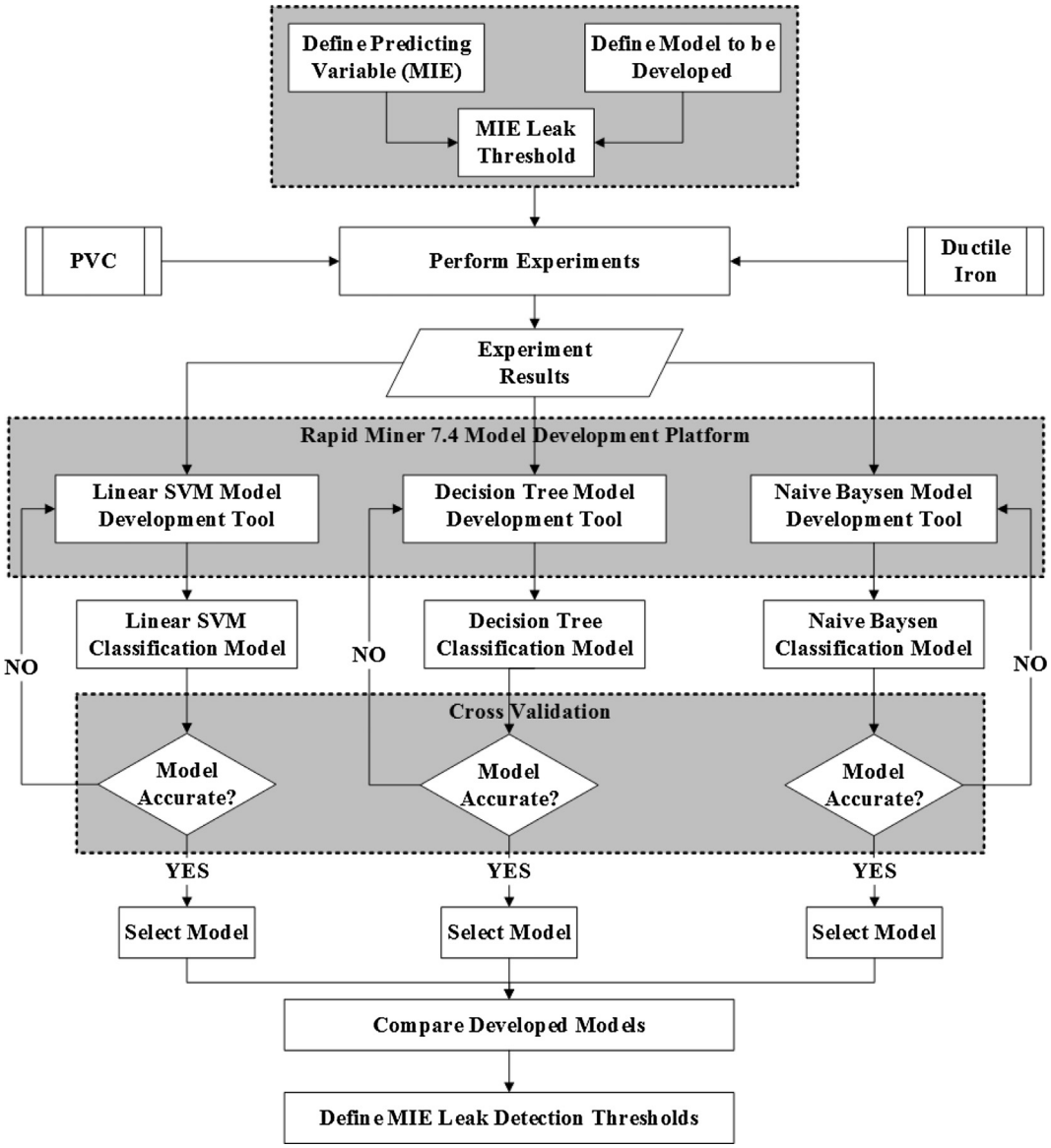
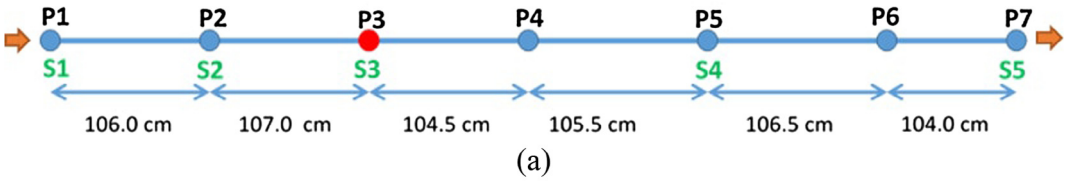


Fig. 2. Model development approach.

experimental pipes along with the relative distances and sensor placements. Each point noted as P(n), is a valve that can be opened at multiple values to simulate the leak. Pressurized water will be inserted through point P1 and will exit through point P7. The image in Fig. 3(b) displays a two-inch ductile iron pipeline supported by two concrete blocks before the installation of the accelerometers and testing. The valves will act as leak simulators as they are slowly opened and closed.

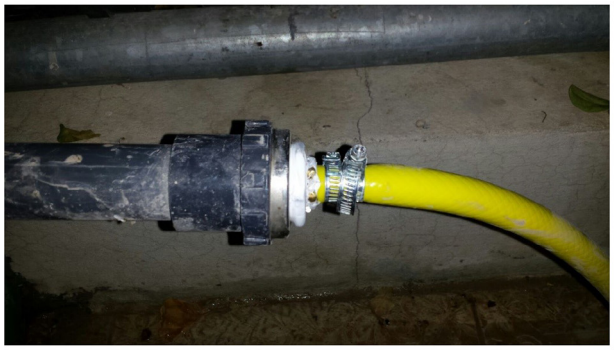
Throughout experimentation, it was noticed that the existence of an immediate open valve at the end could create strong signals that may propagate throughout the whole body of the pipeline and disrupt the data collection process. Thus, the solution was to create a damper that would allow the water to flow outside of the pipe without creating a violent vibration that would disrupt experimentation. Fig. 3(c) shows the solution regarding a hose extension connected to the exit.

After the general installation of the pipeline, Fig. 3(d) displays how the sensors would be placed on each valve before the initiation of the water flow. Once water begins to flow within the pipeline, a certain amount of time is required for it to reach a pressurized state and maintain a uniform flow. Once the inlets and outlets have cohesive uniform flow, the experiments can be commenced. The figure also displays on the right side one of the performed experiments on a one-inch PVC pipeline, where a small leak is simulated via having a small opening in the valve. After the experimentation process was concluded, eight hours of second by second vibration data signals were collected and moved forward to the processing phase. All experiments had only one leak open at any given, but the data was derived from all five sensors for each simulated leak.



(a)

(b)



(c)

(d)

Fig. 3. (a) General experiment setup diagram, (b) two-inch ductile iron pipeline, (c) pipeline exit with release extension and (d) one-inch ductile iron pipeline with one-inch and two-inch PVC pipelines.

Furthermore, the leak was moved from one valve to the next and the next in order to acquire a broader spectrum of the phenomenon. The same experiment was repeated using all four available pipelines which are PVC and ductile iron both of diameters one-inch and two-inch.

4.2. Data analysis

The analysis of the signal data of accelerometers was established as a predominant factor of this research work. The accelerometers would provide the second by second variation of the vibration signal within the pressurized pipeline system as shown in Fig. 4(a). In Fig. 4(a), the signal was collected for an on–off experiment by which a leak was off for 5 min then turned on again for 5 min and so on until a full hour has elapsed. The readings show a level of variation within the system in

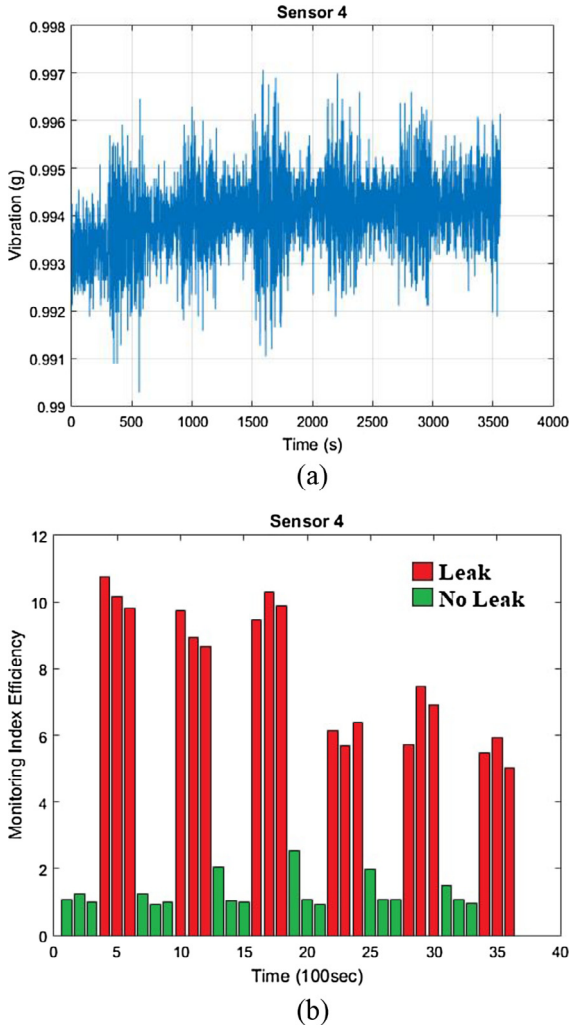


Fig. 4. (a) Accelerometer reading during experimentation and (b) processed accelerometer readings.

reaction to the leak cycle. A MATLAB code was developed to convert the vibration reading into a more comprehensible index using the model mentioned in Section 2.3. As a result, a signal such as the one provided in Fig. 4(a) would be converted to a histogram representing the variation of the Monitoring Index Efficiency (MIE) through time. Fig. 4(b) displays no-leak states in light gray, whereas leak states are shown in dark gray. Fig. 4(b) represents accurately the simulation carried out within this sample experiment. Using the data collected from Section 4.1 and the model described in Section 2.3, 282 values of MIE were collected from all the experiments conducted on the four different pipelines, thus the values of MIE in the dataset are collected from PVC and ductile iron pipelines alike. The values of the collected MIE were classified under three states: (1) No Leak, (2) Small Leak, and (3) Big Leak.

5. Implementation and results

5.1. Leak state detection model

On the level of leak state identification, the primary criterion for assessment was the capability of the devices and the models of finding a separating plane between the leak state and the no-leak state. After cross-validating each model, the results were collected and displayed in this section, and detection thresholds for each model were identified.

5.1.1. Linear SVM model

On the level of SVM, the output of the training of the model was determined using the RapidMiner platform. The platform presented an equation for the threshold plane that is MIE = 1.018 in Eq. (13). The model was cross-validated against the

Table 1
Leak identification model results using three techniques.

Model	Results			
Support Vector Machines	Accuracy = 96.44% ± 5.76%	True No Leak	True Leak	Class Precision (%)
	Predicted as No Leak	144	9	94.12
	Predicted as Leak	1	127	99.22
	Class Recall (%)	93.31	93.38	
Decision Tree	Accuracy = 99.29% ± 1.43%	True No Leak	True Leak	Class Precision (%)
	Predicted as No Leak	143	0	100
	Predicted as Leak	2	136	98.55
	Class Recall (%)	98.62	100	
Naïve Bayes	Accuracy = 98.57% ± 2.37%	True No Leak	True Leak	Class Precision (%)
	Predicted as No Leak	144	3	97.96
	Predicted as Leak	1	133	99.25
	Class Recall (%)	99.31	97.79	

original 282 MIE data points, and the results are summarized in the first row of Table 1. Table 1 shows that the accuracy of the Linear SVM model is averaged at 96.44% with an average deviation of 5.76% increasing or decreasing. Table 1 also shows that the Linear SVM model has missed nine leaks and categorized them under no-leak states. Furthermore, the SVM model is capable of identifying and retrieving No Leak and Leak data accurately. Hence the class recall values are 93.31% and 93.38% for No Leak and Leak states respectively. Additionally, the quality of the retrieved data per class is deemed to be high as the class precisions are 94.12% for the No Leak state and 99.22% for the Leak state.

$$\textbf{\textit{State}} = \begin{cases} \textbf{\textit{No Leak}}, & \textit{if } MIE \leqslant 1.018 \\ \textbf{\textit{Leak}}, & \textit{if } MIE > 1.018 \end{cases} \tag{13}$$

5.1.2. Decision tree model

The output of the RapidMiner decision tree model presented the value of MIE = 1.052 be the threshold as shown in Fig. 5 (a) as well as Eq. (14). The figure also illustrates that any amount equal or less than 1.052 is considered a no-leak state with 100% confidence, whereas any value higher than 1.052 is supposed to be a leak. The cross-validation results of this model are summarized in the second row of Table 1. The accuracy of the model was calculated to be at 99.29% with an average deviation of 1.43%. The model did not classify any leaks as no leaks, yet on the other hand, the model ranked two no-leak states as leaks. The model had the highest percentage of leak data retrieval amidst the developed models with a 100% class recall and 98.55% of the collected data were precisely identified as leaks. As for the No Leak state, the model was capable of accurately obtaining No Leak data with a 98.62% class recall and classify the received data into the No Leak state with a 100% class precision.

$$\textbf{\textit{State}} = \begin{cases} \textbf{\textit{No Leak}}, & \textit{if } MIE \leqslant 1.052 \\ \textbf{\textit{Leak}}, & \textit{if } MIE > 1.052 \end{cases} \tag{14}$$

5.1.3. Naïve Bayes model

As for the Naïve Bayes model, Fig. 6(a) and Eq. (15) show that the separating point between the leak and no-leak states is estimated to be at the intersection point between the two curve where MIE = 1.07. The capabilities of the model were validated using cross-validation, and the results were summarized in the third row of Table 1. Table 1 shows that the average accuracy of the NB model is estimated to be 98.57% with an average deviation of 2.37%. The NB model had three missed leaks and one false alarm during testing and validation. Furthermore, the Naïve Bayes model has shown high levels of proper data retrieval with 99.31% and 97.79% class recall for No Leak and Leak states respectively. Additionally, the collected data were classified with high precision. The class precision for the No Leak state was 97.96% whereas the class precision for the Leak state was 99.25%.

$$\textbf{\textit{State}} = \begin{cases} \textbf{\textit{No Leak}}, & \textbf{\textit{if MIE}} \leqslant \mathbf{1.07} \\ \textbf{\textit{Leak}}, & \textit{if } MIE > 1.07 \end{cases} \tag{15}$$

5.2. Leak size identification model

Regarding leak size identification, the models are expected to identify planes that separate between three different states (1) No-Leak, (2) Small Leak, and (3) Big Leak. Additionally, to reassess the consistency of the leak detection models, the data of no-leak states were used in the development of leak size classification models. A small leak is assumed to have

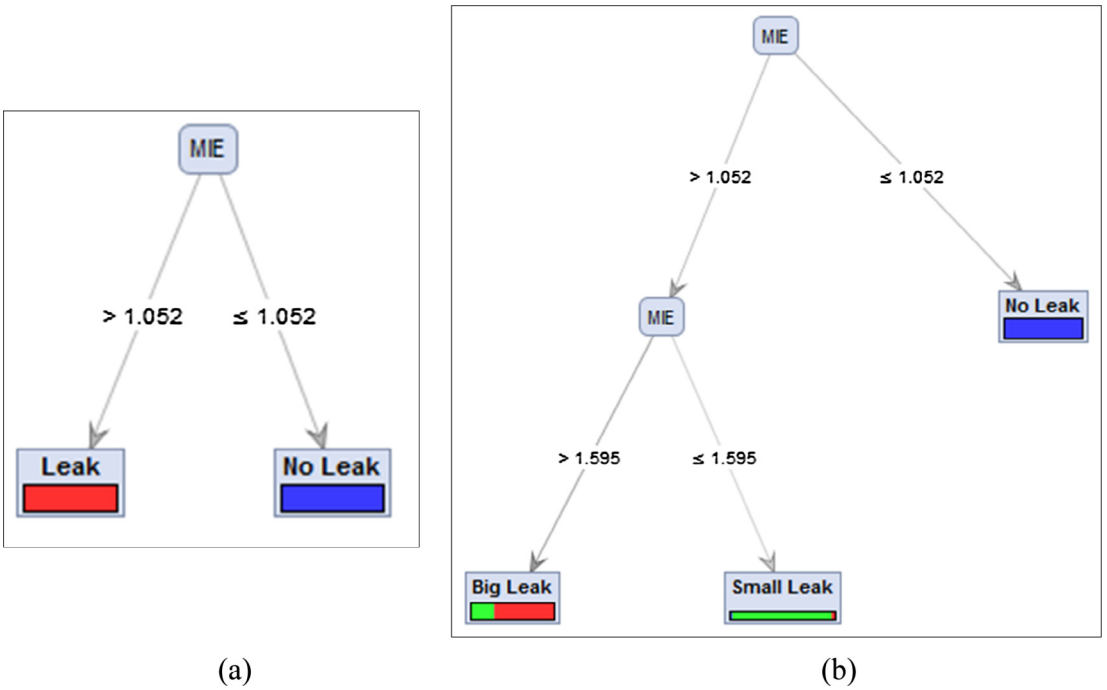


Fig. 5. (a) Decision tree leak identification model and (b) decision tree leak size classification model.

a discharging leak between 10% and 25% of the overall flow rate, whereas a big leak has a discharge rate of 26% and up to 50% of the total flow rate.

5.2.1. Linear SVM model

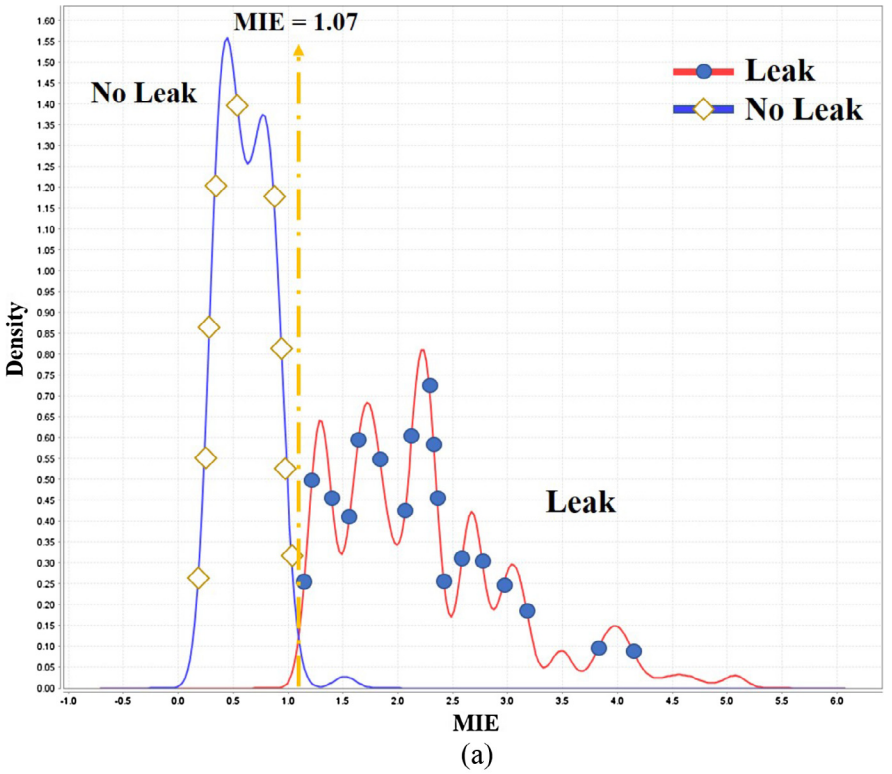
On the level of leak size classification, the RapidMiner Linear SVM model provided two main equations as thresholds. The first threshold value presented is MIE = 1.018, as presented in Eq. (16), and it is utilized to separate the small leaks from no leak condition. The supplied value is consistent with the previously developed model. Additionally, the model provided the value of MIE = 2.24 to separate between small leaks and big leaks. The average accuracy of the model was 80.06% with an average deviation in accuracy equal to 7.55% as displayed in the first row of Table 2. The model has retained its capacity to accurately retrieve and identify No Leak data with a class recall of 100% and class precision of 84.8%. On the level of big leaks, the model has an equal class recall and precision amounting to 80%. On the other hand, the model faced issues in classifying small leaks often misinterpreting them as no leaks or big leaks and thus having a low class-recall for small leaks equal to 32.79% and a low class-precision of 57.14%.

$$\textbf{\textit{State}} = \begin{cases} \textbf{\textit{No Leak}}, & \textit{if } MIE \leq 1.018 \\ \textbf{\textit{Small Leak}}, & \textit{if } MIE \in]1.018, 2.24] \\ \textbf{\textit{Big Leak}}, & \textit{if } MIE > 2.24 \end{cases} \tag{16}$$

5.2.2. Decision tree model

The decision tree model remained consistent regarding the separation of leak states from no-leak states by retaining the threshold of MIE = 1.052 as shown in Fig. 5(b) and Eq. (17). Additionally, the decision tree model specified the value of 1.595 for MIE to represent the threshold separating small leaks from big leaks. If the value of MIE is less than or equal to 1.595 yet above 1.052, then the leak has a very high degree of confidence to be a small leak. Above an MIE of 1.595, the leak would most probably be a big leak yet there would be a possibility of having small leaks at this range as well.

The second row of Table 2 summarizes the results of the cross-validation of the model. The model has an accuracy of 85.39% with an average deviation of 3.02%. The model has had two false alarms and two small leaks that went undetected. Furthermore, the decision tree model behaved better than the SVM model with the class-precision and the class-recall both equal to 98.62% for the No Leak state plus a class recall rate of 49.18% for small leaks and a class precision of 75%. On the level of big leaks, the class recall showed an improvement in decision trees compared to the SVM model by 9.33% whereas the class precision was 10% less than the SVM model.



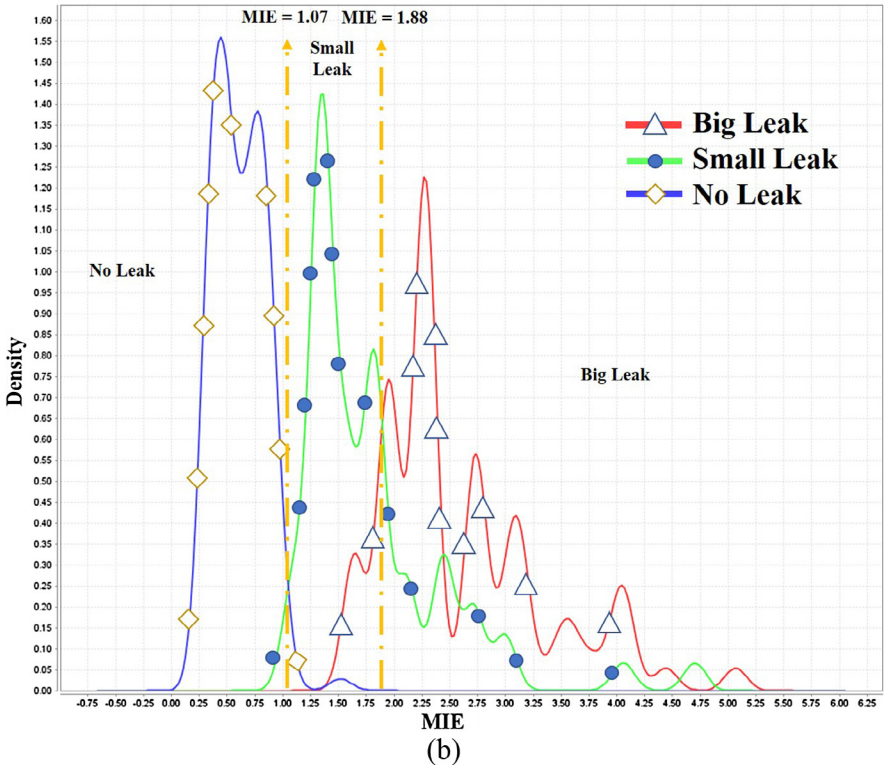


Fig. 6. (a) Naïve Bayes leak identification model and (b) Naïve Bayes leak size classification model.

Table 2
Leak size classification model results using three techniques.

Model	Results				
Support Vector Machines	Accuracy: 80.06% ± 7.55%	True No Leak	True Small Leak	True Big Leak	Class Precision (%)
	Predicted as No Leak	145	26	0	84.8
	Predicted as Small Leak	0	20	15	57.14
	Predicted as Big Leak	0	15	60	80.00
	Class Recall (%)	100	32.79	80.00	
Decision Tree	Accuracy: 85.39% ± 3.02%	True No Leak	True Small Leak	True Big Leak	Class Precision (%)
	Predicted as No Leak	143	2	0	98.62
	Predicted as Small Leak	2	30	8	75.00
	Predicted as Big Leak	0	29	67	69.79
	Class Recall (%)	98.62	49.18	89.33	
Naïve Bayes	Accuracy: 86.8% ± 6.01%	True No Leak	True Small Leak	True Big Leak	Class Precision (%)
	Predicted as No Leak	144	2	0	98.63
	Predicted as Small Leak	1	39	14	72.22
	Predicted as Big Leak	0	20	61	75.31
	Class Recall (%)	99.31	63.93	81.33	

$$\textbf{\textit{State}} = \begin{cases} \textbf{\textit{No Leak}}, & \textit{if } MIE \leq 1.052 \\ \textbf{\textit{Small Leak}}, & \textit{if } MIE \in]1.052, 1.595] \\ \textbf{\textit{Big Leak}}, & \textit{if } MIE > 1.595 \end{cases} \tag{17}$$

5.2.3. Naïve Bayes model

As for Naïve Bayes, the leak versus no leak separation threshold remained consistently at MIE = 1.07 as displayed in Fig. 6 (b) and Eq. (18). On the other hand, the threshold between a small and a big leak is estimated to be at the meeting point between the downward curve of the small leak and the upward going cure of the big leak state. Thus the threshold would be the straight-line MIE = 1.88. Regarding accuracy, the third row of Table 2 shows that the NB model has an average accuracy of 86.8% with an average deviation of 6.01%. The model had one false alarm and two small leaks going undetected. The table shows as well that the Naïve Bayes model provided a more stable model than the other techniques with a class recall of 99.31% for the No Leak state and a class precision of 98.63%. The NB model provided the highest classification for small leaks with a 63.93% class recall and 72.22% class precision. The classification of big leaks using the NB model is acceptable with an 81.33% class recall and a 75.31% class precision.

$$\textbf{\textit{State}} = \begin{cases} \textbf{\textit{No leak}}, & \textit{if } MIE \leq 1.07 \\ \textbf{\textit{Small Leak}}, & \textit{if } MIE \in]1.07, 1.88] \\ \textbf{\textit{Big Leak}}, & \textit{if } MIE > 1.88 \end{cases} \tag{18}$$

5.3. Results discussion

The results shown in this section, display the capabilities of accelerometers in detecting the vibrations caused by leaks as well as accurately identify them. In terms of leak identification, the three techniques utilized presented high results as summarized in Fig. 7(a). The box and whisker plot shows the median as a thick line that cuts the box into two partitions, each partition representing a quartile of the distribution of the accuracies presented by the model. Additionally, the distance from the periphery of the box to each outer line represents a quartile of the data as well. For example, the decision tree data depicted in Fig. 7(a) show a median of 99.29% and a maximum possible accuracy of 99.8% and minimum accuracy of 97.8%. The illustration also indicates that 75% of the time the decision tree model would have an accuracy ranging from 98.2% to 99.8%. In terms of leak identification accuracy, the decision tree algorithm provided the model with the highest accuracy with the minimum deviation at 99.29% accuracy and a deviation of 1.43%. The decision tree model was followed by the Naïve Bayes model and afterward the linear SVM model.

On the level of leak size identification, the Naïve Bayes model provided the highest average accuracy at 86.8% but with a deviation of 6.01% followed by the decision tree model with 85.39% and finally the Linear SVM model. Using the box-and-whisker plot in Fig. 7(b), it can be deduced that the decision tree model had a lesser average accuracy than that of the Naïve Bayes model, but the deviation of the DT model is ±3.02% which represents a more consistent model development than that of the two other techniques compared to ±6.01% that was presented by the NB model.

The determined thresholds are summarized in Table 3. The lowest threshold to separate the leak states from the no-leak state was provided the linear SVM with a value of MIE = 1.018. Whereas the highest threshold was provided by the NB model with an MIE of 1.07. Since the three models showed high levels of accuracy, it is possible to view the three thresholds as

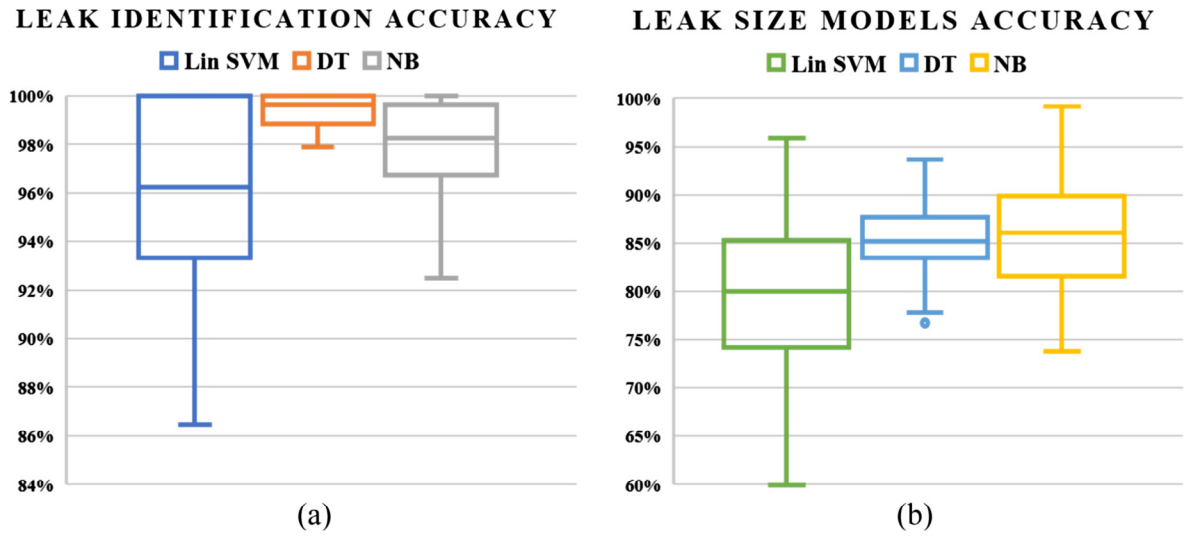


Fig. 7. Box-and-Whisker plot summary of the results for (a) leak identification models and (b) leak size classification models.

Table 3
Derived leak and leak size thresholds for each technique.

		Algorithm results (MIE = x)		
		Linear SVM	Decision Tree	Naive Bayes
State	No-Leak	1.018	1.052	1.07
	Small Leak	2.24	1.595	1.88
	Big Leak			

three levels of conservativeness with 1.018 being the highly conservative threshold and 1.07 being the least conservative threshold separating the leak states from the no-leak state.

Accordingly, the same view can be employed to assess the leak size identification thresholds with MIE equal to 1.595 as the most conservative threshold and MIE of 2.24 as the least conservative threshold for separating small leaks from big leaks. Evidentially, the moderately conservative selections for thresholds would be an MIE of 1.052 to identify the existence of a leak and an MIE of 1.88 to signify the presence of a big leak. The midpoint solution is considered to be a valid starting point for network assessment. On the other hand, if one of the techniques was to be selected, the suggested criterion for selection is consistency and it is characterized by a low deviation. Therefore, the technique with the least deviation from its median is DT, as displayed in Fig. 7(a) and 7(b). The bars concerned with decision tree spread the least compared to the other two techniques. The second technique would be NB and finally SVM having the biggest distance between its extremities in both cases. Additionally, some of the models mentioned above, presented the possibility of making false detection which lowered their respective accuracy. The decision of model selection against their susceptibility for false alarms is in the hands of the users, namely water service organizations. Some organizations may have sufficient resources that allow them to endure the process of checking for false alarms. Comparatively, some organizations may have a narrow resource pool that make on-site checks an added expense and would prefer missing the detection of some smaller leaks over mobilizing for false alarm detection.

6. Conclusions

A leak detection model was developed using accelerometers. Experiments were conducted using one-inch and two-inch pipelines of types PVC and ductile iron. Using the collected data and vibration signal analysis, a qualitative index (MIE) was developed. Through the established index and 8 h of experimental signal data two models were developed using three classification techniques: (1) Linear SVM, (2) Decision Trees, and (3) Naïve Bayes. The first model defined a threshold to identify the existence of leaks within a pipeline system. The leak identification model had an average accuracy of 98% amongst the three classification techniques. The second model proposed a threshold to separate small leaks from big leaks. The average accuracy of the second model amidst the three classification techniques was 83.71%. Amidst the three techniques, DT performed the best as it provided a high accuracy in the two models with minimum deviation, followed by Naïve Bayes, and SVM was the least performing among the three techniques with a lower average accuracy and a wide range of deviation.

On the level of contributions, this research paper proved the utility of Martini's vibration signal analysis model for PVC and ductile iron pipelines. Additionally, this article presented a leak detection model with high accuracy and a potential leak size identification model. Furthermore, three level of thresholds were identified based on conservativeness and the degree of accuracy required by the user. Regarding further development, signal filtering can be used to improve the data collection process. Also, additional experimentation is necessary to identify the causes of anomalies and the depletion of vibration signals throughout pipelines.

References

[1] K. Krchnak, Water Scarcity, Water Scarcity - World Wild Life, 2016. <https://thewaterproject.org/water-scarcity/> (accessed December 18, 2016).
[2] D.W. Seckler et al., World Water Demand and Supply, 1990 to 2025: Scenarios and Issues, vol. 19, IWMI, 1998.
[3] S. Renzetti, D.P. Dupont, D.P. Dupont, Buried Treasure: The Economics of Leak Detection and Water Loss Prevention in Ontario. Rep No ESRC-2013, 2013, p. 1.
[4] X.J. Wang, A.R. Simpson, M.F. Lambert, J.P. Vítkovský, Leak detection in pipeline systems using hydraulic methods: a review, in: Conf. Hydraul. Civ. Eng., Institution of Engineers, Barton, ACT, Australia, 2001.
[5] P. Rao, K. Sridharan, J. Liggett, C. Li-Chung, Discussion and closure: inverse transient analysis in pipe networks, J. Hydraul. Eng. 122 (1996) 287–289.
[6] L.C. Cheong, Unaccounted for water and economics of leak detection, in: 8th Int. Proc. Int. Water Supply Congr. Exhib., vol. 9, International Water Supply Association, Copenhagen, 1991, pp. 11–16.
[7] AWWA, Leaks in water distribution systems: a technical/economic overview. first ed., AWWA, Denver, Colorado, USA, 1987.
[8] J.E. Van Zyl, C.R.I. Clayton, The effect of pressure on leakage in water distribution systems, Proc. Inst. Civ. Eng. Manage. 160 (2007) 109–114.
[9] A. Al-Aghbar, Automated Selection of Trenchless Technology for Rehabilitation of Water Mains, Concordia University, 2005.
[10] American Society of Civil Engineers, ASCE 2013 Report Card for America's Infrastructure, Washington, DC, USA, 2013.
[11] H. Najjaran, R. Sadiq, B. Rajani, Modeling pipe deterioration using soil properties-an application of fuzzy logic expert system, Pipeline Eng. Constr. What's Horizon?, 2004, pp. 1–10.
[12] E.N. Allouche, P. Freure, Management and Maintenance Practices of Storm and Sanitary Sewers in Canadian Municipalities, Institute for Catastrophic Loss Reduction, 2002.
[13] A.C.D. Royal, P.R. Atkins, M.J. Brennan, D.N. Chapman, H. Chen, A.G. Cohn, et al, Site assessment of multiple-sensor approaches for buried utility detection, Int. J. Geophys. 2011 (2011) 1–19, https://doi.org/10.1155/2011/496123.
[14] J.M.A. Alkasseh, M.N. Adlan, I. Abustan, H.A. Aziz, A.B.M. Hanif, Applying minimum night flow to estimate water loss using statistical modeling: a case study in Kinta Valley, Malaysia, Water Resour. Manage. 27 (2013) 1439–1455.
[15] S. Eyuboglu, H. Mahdi, H. Al-Shukri, L. Rock, Detection of water leaks using ground penetrating radar, in: 3rd Int. Conf. Appl. Geophys. 2003, 2003.
[16] H. Schempf, E. Mutschler, V. Goltsberg, G. Skoptsov, A. Gavaert, G. Vradis, Explorer: untethered real-time gas main assessment robot system, in: Proc. Int. Work. Adv. Serv. Robot. ASER, vol. 3, 2003.
[17] M. Fahmy, O. Moselhi, Automated detection and location of leaks in water mains using infrared photography, J. Perform. Constr. Facil. 24 (2010) 242–248, https://doi.org/10.1061/(ASCE)CF.1943-5509.0000094.
[18] M.S. El-Abbasy, F. Mosleh, A. Senouci, T. Zayed, H. Al-Derham, Locating leaks in water mains using noise loggers, J. Infrastruct. Syst. 4016012 (2016).
[19] A. Pal, K.Y.-H. Gin, A.Y.-C. Lin, M. Reinhard, Impacts of emerging organic contaminants on freshwater resources: review of recent occurrences, sources, fate and effects, Sci. Total Environ. 408 (2010) 6062–6069.
[20] M.I.M. Ismail, R.A. Dziyauddin, N.A.S. Ahmad, Performance evaluation of wireless accelerometer sensor for water pipeline leakage, in: 2015 IEEE Int. Symp. Robot. Intell. Sensors, 2015, pp. 120–125.
[21] M. Shinozuka, P.H. Chou, S. Kim, H.R. Kim, E. Yoon, H. Mustafa, et al., Nondestructive monitoring of a pipe network using a MEMS-based wireless network, in: Proc. SPIE - Nondestruct. Charact. Compos. Mater. Aerosp. Eng. Civ. Infrastructure, Homel. Secur., vol. 7649, 2010, p. 76490P. http://doi.org/10.1117/12.848808.
[22] F.C. Almeida, M.J. Brennan, P.F. Joseph, Y. Gao, A.T. Paschoalini, The effects of resonances on time delay estimation for water leak detection in plastic pipelines, J. Sound Vib. (2017).
[23] A. Martini, M. Troncossi, A. Rivola, D. Nascetti, Preliminary investigations on automatic detection of leaks in water distribution networks by means of vibration monitoring, Lect. Notes Mech. Eng. 5 (2014) 535–544, https://doi.org/10.1007/978-3-642-39348-8_46.
[24] A. Martini, M. Troncossi, A. Rivola, Automatic leak detection in buried plastic pipes of water supply networks by means of vibration measurements, Shock Vib. 2015 (2015) 1–13.
[25] S. El-Zahab, F. Mosleh, T. Zayed, An accelerometer-based real-time monitoring and leak detection system for pressurized water pipelines, Pipelines 2016 (2016) 257–268, https://doi.org/10.1061/9780784479957.025.
[26] T.M. Juliano, J.N. Meegoda, D.J. Watts, Acoustic emission leak detection on a metal pipeline buried in sandy soil, J. Pipeline Syst. Eng. Pract. 4 (2012) 149–155.
[27] A. Martini, M. Troncossi, A. Rivola, Leak detection in water-filled small-diameter polyethylene pipes by means of acoustic emission measurements, Appl. Sci. 7 (2016) 2.
[28] I. Jenhani, Amor N. Ben, Z. Elouedi, Decision trees as possibilistic classifiers, Int. J. Approx. Reason. 48 (2008) 784–807.
[29] Q. Ding, Q. Ding, W. Perrizo, Decision tree classification of spatial data streams using peano count trees, in: Proc. 2002 ACM Symp. Appl. Comput., Madrid, Spain, 2002, pp. 413–417.
[30] I. Steinwart, A. Christmann, Support Vector Machines, 2008. http://doi.org/10.1007/978-0-387-77242-4.
[31] W. Jang, J.K. Lee, J. Lee, S.H. Han, Naive Bayesian classifier for selecting good/bad projects during the early stage of international construction bidding decisions, Math. Probl. Eng. (2015) 1–12.
[32] V.R. Kohestani, M. Hassanlourad, Modeling the mechanical behavior of carbonate sands using artificial neural networks and support vector machines, Int. J. Geomech. 16 (2015) 4015038.
[33] S. Park, D.J. Inman, J. Lee, C. Yun, Piezoelectric sensor-based health monitoring of railroad tracks using a two-step support vector machine classifier, J. Infrastruct. Syst. 14 (2008) 80–88.
[34] C. Feng, S. Ju, H. Huang, D. Ph, Using a simple soil spring model and support vector machine to determine bridge scour depth and bridge safety, J. Perform. Constr. Facil (2015) 1–14, https://doi.org/10.1061/(ASCE)CF.1943-5509.0000837.
[35] H. Chen, L. Wei, R. Ning, Z. Cai, H. Shao, Application of factor analysis and SVM technique in expressway condition pattern recognition, CICTP 2015 (2015) 2073–2085.
[36] D. Wu, Supplier selection: a hybrid model using DEA, decision tree and neural network, Expert Syst. Appl. 36 (2009) 9105–9112, https://doi.org/10.1016/j.eswa.2008.12.039.
[37] M. Lee, A.S. Hanna, W. Loh, Decision tree approach to classify and quantify cumulative impact of change orders on productivity, J. Comput. Civ. Eng. 18 (2004) 132–144.
[38] D. Farid, L. Zhang, C. Mofizur, M.A. Hossain, R. Strachan, Expert systems with applications hybrid decision tree and na{ï}ve Bayes classifiers for multi-class classification tasks, Expert Syst. Appl. 41 (2014) 1937–1946, https://doi.org/10.1016/j.eswa.2013.08.089.
[39] M. Qady, Kandil A. Al, M. Asce, automatic classification of project documents on the basis of text content, J. Comput. Civ. Eng. (2014) 1–10, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000338.

[40] M. Hossin, M.N. Sulaiman, A review on evaluation metrics for data classification evaluations, Int. J. Data Min. Knowl. Manage. Process. 5 (2015) 1.
[41] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manage. 45 (2009) 427–437.
[42] BeanAir Inc., BeanDevice AX-3D, BeanDevice AX-3D, 2017. <http://www.beanair.com/wireless-accelerometer-spec.html> (accessed November 4, 2017).