Two-Stage Model for Machine Learning/Transient-based Leak Detection in Pressurized Pipelines

Amir Houshang Ayati (✉ saha.science@gmail.com)

Shahid Chamran University of Ahvaz <https://orcid.org/0000-0002-1524-7677>

Ali Haghighi

Shahid Chamran University of Ahvaz

Hamid Reza Ghafouri

Shahid Chamran University of Ahvaz

Research Article

Keywords: Leak detection, Support vector machine, Ensemble learning, Classification, Regression, Uncertainty analysis

Posted Date: January 24th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1113403/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Two-Stage Model for Machine Learning/Transient-based Leak Detection in Pressurized Pipelines

Amir Houshang Ayati ^{*1}, Ali Haghighi ², Hamid Reza Ghafouri ³

¹ Ph.D. Candidate, Faculty of Civil Engineering and Architecture, Shahid Chamran University of Ahvaz, Ahvaz, Iran (The corresponding author), ORCID ID: 0000-0002-1524-7677, saha.science@gmail.com

² Professor, Faculty of Civil Engineering and Architecture, Shahid Chamran University of Ahvaz, Ahvaz, Iran, ORCID ID: 0000-0002-2765-6929, a.haghighi@scu.ac.ir

³ Professor, Faculty of Civil Engineering and Architecture, Shahid Chamran University of Ahvaz, Ahvaz, Iran, ORCID ID: 0000-0002-3619-9168, ghafouri_h@scu.ac.ir

Abstract

This study introduces a two-stage approach for high-resolution leak localization in large-scale pipelines by coupling machine learning to transient hydraulics. The method includes two stages of leak zone identification and in-zone Leak detection. A transient simulation model using the Method of Characteristics (MOC) is developed to generate the learning data for the pipeline under consideration. Afterward, the problem search space is reduced, and the maximum leak detection error is restricted by determining the most likely leaky zone using Support Vector Regression (SVR). Then, the zone dataset is provided by introducing leak candidates to the identified zone. After that, an ensemble classifier consisting of a set of linear discriminant components is trained to reliably detect the exact location of the leak using the majority voting technique. The models are applied to a theoretical pipeline and an experimental Reservoir-Pipe-Valve (RPV) system. The performance of the applied machine learning algorithms is compared to well-known algorithms considering a variety of kernels and hyperparameters. The impacts of different levels of uncertainty in pipe roughness and initial flow on the models' accuracy are also investigated. The results manifest that the proposed model has high accuracy and is stable, and robust against the hydraulic simulation uncertainties.

Keywords: Leak detection, Support vector machine, Ensemble learning, Classification, Regression, Uncertainty analysis

1. Introduction

At first glance, leaks in pipe systems cause a noticeable loss of clean water ranging from 5% to 50% of the total water supplied (Plath, et al. 2014). Leaks in pipes occur due to various reasons such as poor quality of pipe materials and finish, errors in operation and maintenance, corrosion, internal and external high pressures. Leaks also create severe operational difficulties and impose high costs on the operation management (Shamloo and Haghighi 2009).

Large leaks or bursts may be early visible and usually reported by people. However, tiny leaks that are not visible on the surface can remain undiscovered for a long time while gradually advancing and affecting network performance. The early leak detection would save water and prevent small leaks from turning into bursts and is essential for water companies because of economic, environmental, and reputation reasons (Sophocleous, et al. 2019) and public satisfaction. Fixing leaks at the early stages would also prevent water health and security, especially when inverse leakage is likely. Contamination from underground water, surface flow, and nearby wastewater may be suctioned into the network during an inverse leakage. Hence, the early leak detection facilities are an indispensable part of any pipe system to decrease losses and threats of leaks.

During the last decades, various noninvasive methods have been introduced for leak detection in pipe systems. A group of them are hardware-based methods like acoustic techniques based on the measurement of the sound of water leakage in the pipeline (Tang, et al. 2009; Ozevin and Harding 2012). Other hardware-based techniques are non-acoustic methods that utilize various hardware-based techniques such as infrared thermography (Chunli, et al. 2005),

Ground Penetration Radar (Abouhamad, et al. 2016). Despite the advantages of hardware-based techniques, they are mostly expensive, labor-intensive, and their application is limited to small-scale systems.

Another category of noninvasive methods is model-based techniques that take advantage of hydraulic simulation to model the behavior of pipe systems in the face of leaks. A group of model-based methods is based on steady and extended period analysis, also known as nontransient methods (Wu Zheng and Sage 2006; Wu Zheng, et al. 2010; Farley, et al. 2011; Goulet, et al. 2013). In recent two decades, model-based methodologies based on transient analysis of pipe systems have gained currency among researchers. The transient response signals have higher information content and are more sensitive to changes in leak parameters; therefore, they are preferred to be applied in model-based leak detection models (Xu and Karney 2017; Ayati, et al. 2019). Also, in practice, performing transient tests are not time-consuming; therefore, the results are not affected by daily changes in parameters like demands. In transient-based leak detection, an artificially generated transient wave travels throughout the system and captures the information concerning the leaks in the pipe system as a probe. Then leak detection is performed by interpreting the pressure signal components at measurement sites (Ferrante and Brunone 2003). Xu and Karney (2017) categorized transient methods into four groups: 1) Transient Damping Methods (TDM) (Wang, et al. 2002; Nixon, et al. 2006) which compare the induced damping pattern with the fault-free state of the system; 2) Transient Reflection Methods (TRM) (Jönsson and Larson 1992; Brunone 1999; Lee, et al. 2007; Ferrante, et al. 2009) that focus on the discrepancies between the measured and fault-free hydraulic responses of the system to detect additional reflections in transient (Puust, et al. 2010); 3) System Response Methods (SRM) (Liou 1998; Lee, et al. 2005a; Sattar and Chaudhry 2008; Gong, et al. 2014; Duan and Lee 2016) which use both damping and reflection features of the signal to detect system faults; and 4) Inverse Transient Analysis (ITA) (Liggett and Chen

1994; Nash and Karney 1999; Vitkovsky , et al. 2003; Jung and Karney 2008; Covas and Ramos 2010; Haghighi and Shamloo 2011; Haghighi, et al. 2012; Huang, et al. 2015; Lee, et al. 2015; Ranginkaman, et al. 2016; Kim 2018; Sarkamaryan, et al. 2020) in which leaks are detected through solving a least-square problem to minimize the discrepancies between the field measurements and model predictions. Despite the benefits of transient methods, they still have serious drawbacks to be used in practice. In the first three methods, real transient responses of the fault-free system are required, while such data is hardly ever available. Also, studies on the ITA have shown that the raised optimization problem is highly multimodal, ill-posed, affected by uncertainties, and computationally expensive to solve, especially for large-scale and old problems (Vitkovsky John, et al. 2007; Sophocleous, et al. 2019; Sarkamaryan, et al. 2020).

Machine learning (ML) leak detection methods are also noninvasive. They are free from some of the aforementioned computational challenges with the model-based techniques. ML is a branch of artificial intelligence (AI) concerns the computer algorithms that improve automatically through experience (Mitchell 1997). These algorithms build a model based on sample data, known as training data, to make predictions or decisions without explicit programming. The sample data can be obtained from real events from the field or generated by a simulation model. Samples may be labeled or unlabeled. Labeled data are presented in the form of input-output pairs and are suitable for supervised learning tasks like regression and classification. Each input describes a system state, and output corresponds to its expected outcome called a label. Unlabeled data contain pure information about the system and do not include any prior knowledge or judgment from a label assigned by a supervisor. This type of data is used in unsupervised learning tasks like clustering. ML-based leak detection methods exploit ML algorithms to learn the patterns of different leak states from a sample dataset and predict unseen leak events in reality. Compared to the model-based approaches, the ML-based

computational effort is concentrated in the training stage. Once this stage is complete, they can process new inputs and find leaks almost immediately (Ayati, et al. Forthcoming). Additionally, ML algorithms build data-driven models that all they need is a dataset that includes a representative range of leak states. When these data are obtained from real events in the field, the pipe system can be treated as a black box. No additional information about the system hydraulics and configuration is required for fast and reliable leak detection (Bohorquez, et al. 2020).

A group of ML-based leak detection methods is based on experimental learning theory (ELT) such as Bayes identification (Poulakis, et al. 2003; Leu and Bui 2016; Soldevila, et al. 2017) and Artificial Neural Networks (ANN) (Caputo and Pelagagge 2003; Feng and Zhang 2006; Tao, et al. 2014; Bohorquez, et al. 2020). In these methods, acceptable results can be obtained when the size of the training dataset is adequately large (Zhang, et al. 2016). To solve the problems with small datasets, Vapnik and Kotz (1982) established the statistical learning theory (SLT). Boser, et al. (1992) introduced the idea of support vector machines (SVM) based on SLT. The concept of SVM was then extended by Cortes and Vapnik (1995) within the area of SLT and structural risk. Experimental studies (Mukherjee, et al. 1997; Kwok 1998; Zhang, et al. 2009) have depicted that SVMs can achieve superior performance compared to ANNs in some applications. Moreover, SVMs have yielded excellent generalization performance in bioinformatics, text categorization, fault diagnosis, image segmentation, power systems, and financial analysis (Kwok 1998; Ozevin, et al. 2009). Some successful applications of SVM in leak detection are (Ozevin, et al. 2009; Mounce, et al. 2011; Mashford, et al. 2012; Zhang, et al. 2016; Carre~no-Alvarado, et al. 2017). Walt, et al. (2019) conducted a comparison between the Bayesian probabilistic analysis, SVM, and ANN for leak detection in pipe systems. They showed that when the number of training data is limited, Bayesian methods can hardly detect

unique results. Also, SVM and ANN need both pressure and flow measurements for acceptable leak detection accuracy.

The performance of ML algorithms can significantly be affected by the quality and quantity of datasets. Due to the lack of proper data acquisition systems, expensive cost of flow measurement, and low frequency of leak events, an insufficient amount of real data is often available. On this basis, the previous machine learning-based studies have relied on hydraulic simulation modeling for data set generation. Reviewing the literature of ML-based techniques shows that previous studies have mostly relied on datasets generated by steady-state/extended period hydraulic simulation models (Ayati, et al. Forthcoming). However, transient hydraulic responses significantly contain more information about the system and result in more sensitive and reliable leak detection models (Xu and Karney 2017; Ayati, et al. 2019). As a result, they are preferred to be applied in leak detection models. Bohorquez, et al. (2020) showed the high potential of combining transient pressure waves and ANN to detect leaks on a theoretical pipeline. They reported that the training process of ANN is time-consuming, and requires large datasets to predict the leak location. After that, Ayati, et al. (Forthcoming) proposed an ML-assisted model for leak detection in looped WDNs using the frequency domain responses. They showed that in addition to high performance in leak detection, the model is robust against high levels of uncertainty in pipes' friction factors and nodal demands. In that study, the network junctions are considered as leak candidates, thus the applied ML algorithm was aimed to handle a limited number of classes. In practice, the resolution of the leak detection depends on the closeness of leak candidates. Higher resolutions need more candidates of closer distances. Taking such an approach increases the complexity of the ML problem and consequently reduce the leak detection accuracy. To the authors' best knowledge, these are the only studies in joint application of ML and transient analysis of pipe systems. Thus, this research area is still off the

beaten track and needs further development and evaluation to uncover the potentials and address the limitations.

This study introduces a two-stage Machine Learning/Transient-Based (ML/TB) model for fast and reliable leak detection in pipelines and addressing the aforementioned issues. Here, the leak detection problem is defined as determining the location of the leak. The method exploits transient pressure heads as the system state vector and includes two general stages: 1) leak zone identification; 2) in-zone leak detection. In the first stage, the number of leak candidate locations is reduced, and the maximum leak detection error is restricted. For this purpose, the problem search space is reduced by determining the most likely leaky zone utilizing Support Vector Regression (SVR). Afterward, in the second step, a fast ensemble classifier based on Linear Discriminant Analysis (LDA) is applied to the identified zone for reliable and precise leak localization within the zone. The method's performance is evaluated using Mean Absolute Error (MAE), Maximum Leak Detection Error (MLDE), and the Accuracy measure. Accuracy is calculated by dividing the number of correct detections by the total number of predictions.

In the following sections, first, the applied hydraulic simulation model and dataset generation are explained. After that, the leak zone identification stage is clarified based on the concept of SVR. Next, the application of the ensemble learning approach to in-zone leak detection is explained. The model is applied to two case study pipelines. The first one is a numerical pipeline to evaluate the method in large-scale problems with a large number of leak candidates. The second is an experimental reservoir-pipe-valve (RPV) system to study the method in an actual situation. Finally, the results are discussed, and the study is concluded. Figure (1) presents the general framework of the proposed method.

2. Dataset generation based on transient hydraulic simulation

The performance of the data-driven models can be highly affected by the quality of the dataset. In leak detection, the training set should contain a sufficient number of samples, including a variety of leak scenarios, to comprehensively capture the behavior of pipelines in the face of leaks (Zhang, et al. 2016). In the context of ML, the input vector of each sample data is denoted as the feature vector. It should include the most effective elements that represent the problem's behavior, thus leading to higher performance of the applied learning algorithm. In this study, a vector containing the transient pressure heads of the pipe system is used as the input to describe a certain leak state, and the location of the corresponding leak state is considered as the label. The length of applied pressure trace is $3L/a$, where L and a are pipe length and wave speed, respectively. The reason for this choice is that this part of the response signal contains information on the complete pipeline without significant energy dissipation (Bohorquez, et al. 2020). Accordingly, to set up the required datasets, the following steps are taken:

1. To compute the transient pressure heads, a hydraulic simulation model is developed based on the transient governing equations of continuity and momentum (Chaudhry 2016):

$$\frac{\partial H}{\partial t} + \frac{a^2}{gA} \frac{\partial Q}{\partial x} = 0 \quad (1)$$

$$\frac{\partial Q}{\partial t} + gA \frac{\partial H}{\partial x} + \frac{fQ|Q|}{2DA} = 0 \quad (2)$$

where x is the distance along the pipe, t is time, a is wave speed, g is gravitational acceleration, A is cross-sectional pipe area, D is pipe diameter, Q is instantaneous discharge, H is the instantaneous piezometric head, and f is friction factor. The governing equations are solved using the method of characteristics (MOC). In the MOC, the equations are linearly combined and then integrated, resulting in two positive

and negative characteristic equations. The boundary conditions like reservoirs, valves, network junctions, and leaks are also added to the models based on Chaudhry (2016) 's formulations.

The friction factor in Equation (2) is the sum of two terms of quasi-steady and unsteady frictions as the following.

$$f = f_q + f_u \quad (3)$$

in which f_q and f_u are quasi-steady and unsteady friction factors, respectively. f_q is obtained implicitly through the Colebrook-White Equation:

$$\frac{1}{\sqrt{f_q}} = -2\log\left(\frac{\varepsilon/D}{3.7} + \frac{2.51}{Re\sqrt{f_q}}\right) \quad (4)$$

where ε is pipe roughness, D is pipe diameter, $Re = VD/\nu$ is Reynold's number, V is pipe flow velocity, and ν is Kinematics viscosity of the fluid. In this study, the unsteady friction factor is calculated by the following model initially introduced by Brunone, et al. (1995) and revised by Vitkovsky (2000).

$$f_u = \frac{kD}{V|V|} \left(\frac{\partial V}{\partial t} + a \times \text{sign}(V) \left| \frac{\partial V}{\partial x} \right| \right) \quad (5)$$

In which

$$\text{sign}(V) = \begin{cases} +1 & V \geq 0 \\ -1 & V < 0 \end{cases} \quad (6)$$

k is Brunone's decay coefficient and is calculated using the Vardy and Brown's shear decay coefficient, C^* as follows.

$$k = \frac{\sqrt{C^*}}{2} \quad (7)$$

where C^* is

$$C^* = \begin{cases} 0.00476 & Re \leq 2000 \\ \frac{7.41}{Re^{\log(14.3/Re^{0.05})}} & Re > 2000 \end{cases} \quad (8)$$

In Equation (5), terms $\frac{\partial V}{\partial t}$ and $a \times \text{sign}(V) \left| \frac{\partial V}{\partial x} \right|$ are known as the local and convective accelerations, respectively.

Also, in hydraulic simulation models, a leak is simulated by the orifice equation as follows,

$$Q_L = \xi A_e \sqrt{2g|H_L - Z_L|} \quad (9)$$

where Q_L is leak discharge, $A_e = C_d A_L$ is effective leak area, C_d is coefficient of discharge, A_L is apparent leak area, Z_L and H_L are elevation and instantaneous piezometric head at the leak location, respectively. Also, $\xi = 1$ if $H_L > Z_L$ and is equal to 0 otherwise. The calibrated simulation model functions leak parameters and returns the transient pressure heads at the measurement sites. This hydraulic model is exploited to provide the transient pressure heads of the system at the measurement site for different leak scenarios.

2. To provide a uniform sampling over the problem space, various artificial leaks are introduced to the model. It is done by changing the position of leaks with different sizes and locations. Then, their corresponding hydraulic responses at the measurement site are computed by the hydraulic simulation.
3. Taking the aforementioned attitude, different training and test datasets are generated for each case study.

3. Leak zone identification using Support Vector Regression (SVR)

In pipelines leak detection, the number of leak candidates is dependent on the desired resolution and accuracy. The higher resolutions need more leak candidates along the pipeline. There are

a few leak candidates for short pipelines or low-resolution problems, and conventional ML algorithms can handle the problem complexity. The leak detection problem can be defined as a classification problem or a regression problem in this condition. In the classification approach, the leak's location is considered a discrete value, and some locations along the pipeline are considered leak candidate locations. Each location is a label, and the classifier is supposed to find the correct label for new inputs. In the regression approach, the leak location is a continuous real number. Thus, the location of possible leaks is not limited to certain places. The regressor would return a real number as the leak location. In long pipelines, the number of leak candidate locations is increased to preserve the leak localization resolution. This will increase the ML complexity by increasing the number of classes (leak candidate locations), and consequently, standard ML algorithms depict poor performance in terms of accuracy and reliability.

In this study, to address the issue mentioned above, a regression-classification scheme is introduced. In this approach, leak candidates are noticeably reduced through the Leak zone identification utilizing Support Vector Regression (SVR). Then, a fast classifier is exploited to the detected zone for more accurate leak localization. To this aim, a regressor based on SVR is trained and tested using datasets including various leak scenarios distributed over the pipeline with different leakage sizes. This regressor is utilized as the leak zone identification tool, and its performance is evaluated using Mean Absolute Error (MAE) and Maximum Leak Detection Error (MLDE). The leaky zone is defined as a symmetric neighborhood of the zone center. The radius of this neighborhood is supposed to be fixed and is considered as the MLDE of the SVR algorithm on the train and test set. The regression model is applied to new data or measurements to determine the zone center. The closest leak candidate to the predicted location is considered the center of the leaky zone.

SVM is primarily developed for discriminating samples of two categories called binary classification. SVR is a version of SVM introduced by Drucker, et al. (1997) for regression problems. The central idea of SVR is to estimate a function that maps an input to a real number based on training data (Figure 2). Let $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$ be a set of training samples, where $\mathbf{x}$ is a d -dimensional vector of inputs, $y \in \mathbb{R}$ denotes the output corresponds to each $\mathbf{x}$ and N is the number of samples. The SVR function is

$$y = f(\mathbf{x}) = (\mathbf{w}^T \cdot \mathbf{x}) + b \quad (10)$$

where $\mathbf{w}$ is weight vector; $\mathbf{w}^T$ is the transposition of $\mathbf{w}$; and b is the bias. Function $f(\mathbf{x})$ approximates all pairs $(\mathbf{x}_i, y_i)$ while maintaining the differences between estimated values and real values under ε precision. The nonlinear estimation function, $f(\mathbf{x})$, should minimize the following objective function:

$$\text{minimize: } L(\mathbf{w}, \xi) = \frac{1}{2}(\mathbf{w}^T \cdot \mathbf{w}) + C \sum_i (\xi_i^2, \hat{\xi}_i^2), \quad C > 0 \quad (11)$$

$$\text{subject to: } y_i - \mathbf{w} \cdot \mathbf{x}_i - b \leq \varepsilon + \xi_i, \quad \forall (\mathbf{x}_i, y_i) \in S \quad (12)$$

$$\mathbf{w} \cdot \mathbf{x}_i + b - y_i \leq \varepsilon + \hat{\xi}_i, \quad \forall (\mathbf{x}_i, y_i) \in S \quad (13)$$

$$\xi_i, \hat{\xi}_i \geq 0 \quad (14)$$

where C is the trade-off parameter between the margin size ($1/(\mathbf{w}^T \cdot \mathbf{w})$) and errors. Variables ξ and $\hat{\xi}$ deal with infeasible constraints of the optimization problem by imposing the penalty to the excess deviations which are larger than ε . The dual form of the objective function (Equation 9) can be written in terms of the training data as

$$\text{Maximize: } L(\alpha) = \sum_i y_i (\alpha_i - \hat{\alpha}_i) - \varepsilon \sum_i (\alpha_i + \hat{\alpha}_i) - \frac{1}{2} \sum_i \sum_j (\alpha_i - \hat{\alpha}_i)(\alpha_i - \hat{\alpha}_i) K(\mathbf{x}_i, \mathbf{x}_j)$$

$$(15)$$

$$\text{Subject to: } \sum_i (\alpha_i - \hat{\alpha}_i) = 0 \quad (16)$$

$$\hat{\alpha}_i \geq 0, \hat{\alpha}_i \geq 0, \alpha \geq 0, \hat{\alpha} \leq C \quad (17)$$

The corresponding nonlinear SVR function is

$$y = f(\mathbf{x}) = \sum_i (\hat{\alpha}_i - \alpha_i) K(\mathbf{x}_i, \mathbf{x}_j) + b \quad (18)$$

where y is the target value; α , $\hat{\alpha}_i$, $\hat{\alpha}_i$ are Lagrangian multipliers; C is penalty factor corresponds to the trade-off between the margin size and errors; b is the bias; ε is the precision and $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function. The kernel function is a similarity function between two vectors that the function output is maximized when the two vectors become equivalent (Yu and Kim 2012). The kernel function is applied to data to map the original nonlinear observations to a higher-dimensional space to enhance their separability. This process is performed implicitly by calculating the inner products between the images of all pairs of data in the feature space. This approach is called the kernel trick and is often computationally cheaper than the explicit computation of the coordinates (Shawe-Taylor and Cristianini 2004). The performance support vector machines depend on kernel type and its parameters (Chamasemani and Singh 2011; Mashford, et al. 2012). In the present study, the performance of the following well-known kernel functions is investigated:

$$1. \text{ Linear: } K(\mathbf{x}_i, \mathbf{x}) = \mathbf{x}^T \cdot \mathbf{x}_i$$

$$2. \text{ polynomial: } K(\mathbf{x}_i, \mathbf{x}) = (1 + \mathbf{x}^T \cdot \mathbf{x}_i)^p$$

$$3. \text{ Radial basis function (RBF): } \exp(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2)$$

in which p is the order of the polynomial, γ is the scaling parameter and $\|\mathbf{x} - \mathbf{x}_i\|$ returns the norm of the vector $\mathbf{x} - \mathbf{x}_i$. In the case of polynomial, the quadratic type ($p = 2$) and the cubic type ($p = 3$) are used. Also, three RBFs with scaling parameters of $\sqrt{q}/4$, $\sqrt{q}$, and $\sqrt{q} \times 4$ are

defined and applied as fine, medium, and coarse RBFs, respectively, where q is the length of the applied feature vector in each case.

In this study, the K-folds Cross-validation method (Refaeilzadeh, et al. 2009) is applied to tuning the model parameters and handling the issues related to overfitting and noise in samples.

In this technique, the training set is randomly split into K subsets of the same size. Each of these disjoint subsets is called a fold. Then for each fold, the classifier is trained using out-of-fold samples, and after that, the model performance is assessed using in-fold data as a validation set. In the end, the average accuracy is computed over all folds. It is worth mentioning that the optimum hyperparameters of employed algorithms are determined through the grid search, in the present work.

4. In-zone leak localization using Discriminant-based Ensemble Classification

After determining the leaky zone by SVR, the zone dataset is provided. This dataset is a modified version of the initial training set containing only samples of leak candidates in the identified zone. The in-zone leak detection problem is defined as a classification problem. The zone dataset is used to train a fast ensemble classification algorithm to perform precise and reliable leak detection within the zone. Ensemble learning is a learning paradigm in ML that imitates the second nature of humans to seek several opinions and combine them to make more reliable decisions in crucial situations (Polikar 2006). For ensemble learning, several predictive models are combined to create a more accurate and reliable one (Mohri, et al. 2018).

In ensemble classification, each component model is a classifier (also called ensemble classifier), and the final decision is made based on the outputs of the ensemble classifiers. Accordingly, a properly trained system of ensemble classifiers provides most likely correct predictions with higher confidence when a vast majority of the classifiers support a decision

(Muhlbaier, et al. 2005). In a successful ensemble system, the errors made by some component models are corrected by others through the combination of ensemble classifiers outputs. For this purpose, applied classifiers should make different errors on different instances (Kuncheva and Whitaker 2003; Banfield, et al. 2005; Brown, et al. 2005). Such an attribute is denoted as the diversity of component classifiers. To obtain diverse classifiers, several approaches have been introduced so far. In this study, a common technique called the random subspace method (Ho 1998) is applied because of its efficiency and simplicity (Duda, et al. 2001). Different subsets of existing features (pressure heads in each sample data) are used to train individual classifiers in this technique. Finally, the outputs of different ensembles are integrated by the majority voting. In majority voting, the prediction made by each component model is treated as a vote for a candidate class, and the class that receives the largest total vote is considered as the final prediction.

Ayati, et al. (Forthcoming) compared the performance of Linear Discriminant Analysis (LDA), SVM, and Nearest Neighbor (K-NN) classifiers in transient-based leak detection of pipe systems. They showed that both SVM and LDA have high performance on the leak detection problem. However, in terms of computational speed, LDA is much faster than SVM. Therefore, in the present study, LDA is selected as the component classifier in the ensemble scheme. The core idea of LDA is to maximize the distance between classes and minimize the distance within classes. Let $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_i, y_i), \dots, (\mathbf{x}_N, y_N)\}$ be a set of training samples, where $\mathbf{x}_i$ is a d -dimensional feature vector, $y_i \in \{1, \dots, C_M\}$ is its corresponding class label, M is the number of classes, and N is the number of samples. Assuming conditional distribution $(P(\mathbf{x}|y = k))$ for data from class k , LDA predicts the class label of each training sample using Bayes's rule.

$$P(y_i = k|\mathbf{x}_i) = \frac{P(\mathbf{x}_i|y_i = k)P(y_i=k)}{P(\mathbf{x}_i)} \quad (19)$$

where $P(y_i = k|\mathbf{x}_i)$ is called posterior probability. The goal of LDA is to find the class label that maximizes the posterior probability. LDA assumes that all classes share the same covariance matrix (Σ). Considering the multivariate Gaussian distribution function, the probability of k th class samples can be presented as follows.

$$P(\mathbf{x}_i|y_i = k) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{(\mathbf{x}_i - \mu_k)^T \Sigma^{-1} (\mathbf{x}_i - \mu_k)}{2}\right) \quad (20)$$

where T denotes matrix transpose operator, μ_k is the mean of k th class, and $|\Sigma|$ and Σ^{-1} are determinant and the inverse of the covariance matrix of the class, respectively. Taking the natural logarithm of the Equations (20) and (19) and then substituting Equation (20) in Equation (19) yields

$$\ln(P(y_i = k|\mathbf{x}_i)) = -\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \Sigma^{-1} (\mathbf{x}_i - \mu_k) + \ln(P(y_i = k)) + Cst \quad (21)$$

where term $Cst = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|) - \ln(P(\mathbf{x}_i))$ and is the same for all classes; hence, it is considered as a constant. In Equation (18), the term $\ln(P(y_i = k|\mathbf{x}_i))$ is called the log-posterior of LDA. Compared to the original posterior term, it is more convenient to be maximized. Therefore, the applied LDA scheme is based on the maximization of the LDA log-posterior. To prevent overfitting, the K-fold cross-validation technique is applied in the training of individual classifiers. As mentioned before, the model's performance is evaluated using MAE, MLDE, and Accuracy measures.

5. Case study one: large-scale theoretical model

5.1. model configuration

To evaluate the performance of the proposed model on a large-scale problem with a high number of leak candidates, a numerical Reservoir-Pipe-Valve (RPV) system (Figure 3) is

studied here. The pipeline is connected at the upstream end to a reservoir with a fixed 80 m head and the downstream end to a side discharge valve. The pipe's length, inner diameter, and wave speed are 1000 m and 250 mm, and 1000 m/s, respectively. The initial discharge in the pipeline is 99.02 l/s. To generate the transient state flow, the downstream valve is closed in 1 s for a 10% reduction in the initial flow. This closure time is less than $2L/a = 2$ s, and could be considered the instantaneous valve maneuver (Chaudhry 2016). In this theoretical example, only steady-state friction is taken into account with a 0.02 Darcy–Weisbach friction factor.

5.2. Dataset generation

The transient pressure heads immediately at the upstream face of the valve are sampled for the ML models dataset. In each sample, the transient trace of $3L/a = 3$ s is considered as the feature vector. This part of the response signal includes the complete pipe information and is less affected by energy dissipation (Bohorquez, et al. 2020). In the MOC analysis, the distance between the characteristic nodes is 2 m. All characteristic nodes except the reservoir and the valve nodes are candidates of a leak resulting in 499 leakage candidates. Accordingly, a training dataset of 4990 samples is generated supposing 10 leak sizes between zero and 5 cm² (1.02% of the pipe's cross-sectional area). The leak size is changed stepwise with the precision of 0.5 cm². In addition, two test sets of 1500 samples were produced randomly, considering different leak sizes and locations in the ranges mentioned above. The setting of transient simulations is similar to one for the training set. The size and location of artificial leaks are generated using two independent uniform distributions to preserve the randomness. Together with the training set, one of these test sets is used in the training process of algorithms to control the overfitting and determine the leak zone parameters. The second test set is utilized to evaluate the performance of the model on unseen data. In both test sets, the leak size is changed randomly with the precision of 0.001 cm². Considering $\Delta x = 2$ m, wave speed=1000 m/s, and a transient trace of 3 s, the $\Delta t = 0.002$ s each sample in datasets contains 3000 head values.

5.3. Leak zone identification

The first stage in the proposed method is to train the leak zone identification algorithm and determine the extent of the leaky zone according to its MLDE. To do so, the SVR of different kernels is trained using the training set and tested using the first test set. Table (1) presents the results of training the SVR algorithm for leak zone identification using different kernels in terms of MAE, MLDE, and Maximum Error Ratio (MER). MER is calculated by dividing the maximum error of train and test by the whole pipe length. The closeness of the performance on train and test sets depicts that 5-fold cross-validation can successfully prevent overfitting. As explained in previous sections, leak zone identification aims to reduce the problem search space by decreasing the number of leak candidates that should be applied in the next stage, the precise leak localization through the ensemble classification. Therefore, a proper zone detection is one with less MAE and MLDE. Among the explored kernels, the best performance belongs to the Cubic kernel with average MAE and MLDE of 20.8 m and 41.2 m, respectively. The MER of 4.22% shows that in the face of unseen data, the radius of the leaky zone is 4.22% of the whole pipe length, and the selected algorithm can make a 91.56% reduction in the problem's search space. Accordingly, the SVR with a Cubic kernel is found efficient for leak zone identification in this case study.

To this point, the applied datasets contain samples with 3000 head values. Handling such high-dimensional data is time-consuming and can increase the risk of overfitting. As a remedy, the dimension of data samples in all datasets is reduced in a timewise down-sampling process. In this process, the initial $\Delta t = 0.002$ s is increased to $\Delta t = 0.01$ s by keeping the first head value and then every five samples after the first. Accordingly, the dimension of each sample is reduced to 600 head values (80% reduction). To investigate the impact of such down-sampling, the same process of SVR training is repeated using down sampled data sets. The results of this process are presented in Table (2) in terms of MAE, MLDE, and MER. Comparison of the

results from Table (1) and (2) manifests that in the case of SVR with Cubic kernel, the performance does not noticeably change by down-sampling. Therefore, to enhance the computational speed, the study continues using down-sampled datasets in the following.

Considering the fixed number of leak intensity (10 leak sizes) at each leak candidate, the size and the resolution of generated datasets are a function of the distance between leak location candidates. The smaller distance between leak candidates leads to datasets of higher resolution and larger size. To investigate the impact of this phenomenon, a sensitivity analysis is performed. For this purpose, several training datasets of different distances between leak candidates are produced based on the initial training dataset. In each case, a Cubic SVR is trained, and its performance is evaluated on training and test set correspondingly. Here, the applied test set is used previously with a leak candidate distance of 2 m. The results of this sensitivity analysis are presented in Table (3). Comparing the MAE and MLDE on train and test sets depicts that as the resolution of the training set grows, the performance on training and test set gets closer, and consequently, the risk of overfitting will be decreased. Moreover, it shortens the corresponding leaky zone by decreasing the MER. The right column of Table (3) shows the number of in-zone leak candidates assuming the target leak detection resolution of 2 m. The number of in-zone candidates can affect the performance of the applied classification algorithm in the second stage of the method. The more candidates in a zone, the more classes will exist to be discriminated with the classifier, and consequently, the classification performance will be reduced. This issue is significant for long pipelines. This sensitivity analysis shows that in the case of longer pipelines, the number of in-zone classes can be reduced by increasing the resolution of the applied training dataset.

To compare the efficiency of SVR with other regression algorithms, a set of well-known regression algorithms; Linear regression, Decision trees, Ensemble Boosted Trees, and Ensemble Bagged Trees, are explored. These algorithms are investigated considering different

hyperparameters, and the results are presented in Table (2). The closeness of the results for training and test sets confirms the success of 5-fold cross-validation in preventing the overfitting issue. According to these results, among the investigated regression algorithms, SVR with Cubic kernel has the highest performance in terms of MAE, MLDE, and MER. Therefore, the SVR with Cubic kernel and average performance of MAE=21.29 m, MLDE=42.36 m, and MER=4.35% is selected to be applied as the leak zone identification module. This module has a zone radius of $0.0435 * 2 \times 1000 = 87 \text{ m}$, and in the next stage of the methodology (the precise leak localization using ensemble LDA-based classification), will decrease the number of candidate leak locations from 499 to at most 45 (90.98% reduction).

Figure (4) shows the leak detection error of Cubic SVR on train and test datasets. It seems that extreme errors are almost uniformly distributed over the whole pipe length. To analyze the distribution of errors more precisely, the percentage exceedance associated with the absolute error in the leak location prediction is also presented in Figure (5). The percentage exceedance helps to find the proportion of cases with a leak detection error that surpasses a certain error size. For example, in the train set, 50% of the sample has an error of 25 m or larger while this proportion is reduced to 40% for the test set.

5.4. In-zone leak detection

In this stage, an ensemble classifier with subspace LDA components is used to find the precise location of the leak in the identified zone. As explained before, two test sets were generated in this case study. The first one was applied in the zone identification stage to determine the radius of the leaky zone. In this section, the second test set is used to evaluate the method's performance on unseen data. Accordingly, in each test, based on the dataset resolution, the closest leak candidate location to the detected location by Cubic SVR is considered the center

of the leaky zone. A new subset of the initial training set that includes only leak candidates within the zone is provided. Hereafter, this subset dataset is called the zone dataset. Next, an LDA-based ensemble classification algorithm is trained using the zone dataset and exploited to the second test set, or field measured data. The output of this classifier will be the most likely location of the corresponding leak.

First of all, to compare the performance of the LDA-based ensemble classifier with some well-known classification algorithms, a hypothetical leak zone is assumed around the center of the pipeline. It is assumed that the Cubic SVR zone identifier reported the middle of the pipeline as the leak zone center. Accordingly, concerning MER=4.35% for the SVR zone identifier, the leak zone includes 45 leak candidate locations, and the problem is a multi-class classification problem of 45 classes. The zone dataset contains 450 data samples. A new test set is generated consisting of 450 samples with random leaks within the zone and random leak sizes between zero and 5 cm^2 . To solve this classification problem, different classifiers are applied and their performance is compared in Table (4). To prevent overfitting, all algorithms are trained through the 5-fold cross-validation.

The results manifest that simple LDA and LDA-based Ensemble classifiers have the highest accuracies. The accuracy measure is an important measure that directly affects the reliability of the leak detection. For instance, in the case of LDA-based ensemble, the test accuracy=100% tells that in 100% percent of cases, the classifier has successfully detected the exact location of the leak concerning predefined resolution. Another important measure that is especially important in field application is MER. It shows that in case that the classifier fails to detect the correct location of the leak, to what extent the predicted location is far from the real leak location. As can be seen, the MER for the LDA-based Ensemble classification algorithm is zero. At the same time, the simple LDA has the MER=100%, which means when the simple LDA fails to detect the leak within the zone, the distance between the predicted location and

the real location of the leak can be equal to the zone length. On this basis, the LDA-based Ensemble classifier is selected for in-zone leak detection in this study. The applied algorithm includes 30 simple LDA classifiers; each is trained using a random subset of 300 head values (50% of the whole head values in each sample). To avoid overfitting, each component classifier is trained using 5-fold cross-validation. The training time of this model is 94 s with is about 30 times the 3s for a simple LDA classifier.

According to the general methodology proposed in this research, the in-zone classifier is trained in each new test sample. Two sensitivity analysis is performed to investigate the impact of the ensemble parameters on the performance of the LDA-based Ensemble classifier. The first one is done on the effect of subspace size (Table 5), and the second is to investigate the impact of the number of components LDA classifiers on its performance (Table 6). According to Table (5), considering the 30-component model, the subspace size larger than 30% of all pressure head values result in 100% accuracy with MER=0%. Also, based on Table (6), the least configuration for 100% accuracy and MER=0 is a subspace size of 20% and more than 40 component models. Considering that these sensitivity analyses are performed on one zone, to enhance the robustness and reliability of the model in various zones along the pipeline, the algorithm with 40 models and subspace of 50% is selected for in-zone leak detection.

Finally, according to the general flow of the method in Figure (1), a two-stage leak detection model is developed by coupling the leak zone identification module (the trained Cubic SVR regressor) with the in-zone leak detection module (the LDA-based ensemble classifier). Then, the performance of the model has evaluated on the train set and two test sets. As explained before, the train set and the first set have participated in the leak zone detection stage; however, the second test set was not used previously and is treated as the unseen data. Figure (6) presents the performance of the model on the train and test sets. According to Figure (6.c), the model

successfully detects the unseen leak cases in the second test set with the accuracy of 99.9%,
MAE=0.002 m, MLDE=2 m, and MER=0.2%.

According to Bohorquez, et al. (2020), considering a pipeline of a similar scale, ANN needs
large datasets of 50000 samples to give the MAE=1.15m and MLDE=98.21m on the test set
while utilizing train and test set of 4990 and 1500 samples (87% less than ANN), the method
proposed in this study depicts MAE=0.087m and MLDE=4m on unseen data. Also, in 95.86%
of cases, the model predicted the exact location of the leak with no error. In addition, using the
proposed method, the training time of the applied model is fewer than five minutes (Table 4)
which is significantly less than the ANN training time of several hours reported by Bohorquez,
et al. (2020).

6. Case study two: Experimental model

6.1. Model configuration

To investigate the performance of the proposed method in the experimental condition, a
Reservoir-Pipe-Valve (RPV) model was constructed in the Hydraulics Lab of Shahid Chamran
University of Ahvaz. The setup of the applied model is presented in Figure (7). The model
includes a copper pipe with a diameter of 19.05 mm and a length of 64 m. The pipe forms a
spiral-like shape of 1.05 m diameter with 22 loops that are well anchored to a set of vertical
frames. At the pipe outlet (downstream end), a Brass ball valve, a Globe valve, and a Pressure
Transmitter, model WIKA A10 with an accuracy of $\pm 0.3\%$ of the calibrated span, are installed
for transient generation, flow adjustment, and pressure measurement, respectively. The
upstream end includes a pressure vessel of 0.8 m³ and a compressor to fix the pressure head
around 50m. The upstream end of the pipe, the pressure vessel junction, is 0.56 m lower than

the downstream end. At the downstream end, a centrifugal pump is used to recirculate the water from the reservoir to the pressure vessel forming a closed-loop system.

Two artificial leak locations were created using a short pipe segment forming a T-junction with the main pipe diverting the leaked water away from the pipe. In each case, a Ball valve was installed along the segment to impose the desired leak discharge. The aforementioned leaks are located at 32 m (Leak 1) and 48 m (Leak 2) from the upstream at 57cm and 81cm levels, correspondingly. At each location, utilizing the corresponding valve, an artificial leak was imposed (leak sizes of 2.3 mm^2 at location 1 and 1.88 mm^2 at location 2). Considering the initial flow of 0.09 l/s (Reynolds number $Re \approx 4800$), the transient flow is generated by the instantaneous full closure of the downstream Brass ball valve in 0.03 s . The measured quantities of the system pressure response were concurrently transmitted via a data logger to a PC with a sampling rate of 1000 Hz . Then the Butterworth Low pass Filter (Butterworth 1930) is applied to denoise the measured signal. Figure (8) presents the measured signal after denoising. As the case study one, the first $3L/a$ of the signal (here $3 \times 64/1205 = 0.1593\text{s}$) was used as the feature vector.

6.2. Dataset generation

A hydraulic simulation model considering unsteady-state friction (Equations (3) to (8)) is developed based on MOC, and the transient response at the upstream face of the downstream valve (the location of Pressure Transmitter) is applied in dataset generation. For calibration of the model, an experiment is done with no leak, and the difference between the measured transient pressure heads at the valve and simulated head values was minimized. In the mentioned optimization problem, a least-squares objective function is defined based on the discrepancy between the measured and model-predicted pressure heads. Then, a Genetic Algorithm is coupled to the simulation model to calibrate the parameters of pipe wall

roughness, wave speed, and two correction coefficients of α and β introduced to the local and convective accelerations terms of the Brunone unsteady friction equation. The optimum values of the above parameters are obtained 0.005 mm, 1205 m/s, 1.012, and 1.015, respectively. Figure (8) shows the numerical and experimental results after calibration.

In the MOC analysis, the distance between the characteristic nodes is 0.5 m, and all characteristic nodes except the first (the reservoir node) and the last (the valve node) are considered leak candidates. Thus, considering 127 leakage candidates, a training dataset of 1270 training samples was generated, supposing 10 leak effective areas between zero and 5 mm² (1.75% of pipe cross-sectional area). The leak size is changed stepwise with the precision of 0.5 mm². Also, two test sets of 635 samples were randomly generated with different leak sizes and locations within the ranges mentioned earlier. The effective area and location of artificial leaks are generated using two independent uniform distributions to preserve the randomness. The same as the first case study, one of these test sets, together with the training set, is used to determine the leak zone parameters. The second test set is utilized to evaluate the performance of the model. In both test sets, the leak size is changed randomly with the precision of 0.001 mm². Considering $\Delta x = 0.5m$, wave speed=1205 m/s, a transient trace of 0.1593 s, and $\Delta t_{simulation} = 0.00042s$ each sample in datasets contains 382 head values.

6.3. Leak zone identification

In the first stage, based on the results of the first case study, a Cubic SVR is trained and tested to be applied as the leak zone identification module. Figure (9) shows the leak detection error on the train and the first test sets. Accordingly, the maximum value of the MLDE on the train and first test set (MLDE=2.74m) is applied to obtain the leaky zone radius. On this basis, the radius of the leaky zone is determined as 3 m, which at most includes 13 leak candidate locations for in-zone leak detection. In the second stage, each sample in the second test set is

evaluated using the trained SVR, and the closest point to the prediction is selected as the center of the zone.

6.4. In-zone leak detection

Finally, an LDA-based ensemble classifier with 40 component LDAs and a subspace of 50% is trained using the in-zone subset of samples in the training set. The leak's accurate location corresponds to the second test set sample predicted. Figure (10) presents the leak detection error of the model on the train and two test sets. As can be seen, the model has efficiently detected the location of leaks in all datasets.

6.5 Application to experimental data

To assess the model performance on experimental data, the length of the feature vector in the simulated and measured signals should exactly be the same. To do so, using the linear interpolation, the sample time step ($\Delta t_{measurement} = 0.001s$) is reduced to $\Delta t_{simulation} = 0.00042s$, so that both simulated and measured data include the same number of head values=382. Then, the trained zone identification module was applied to the measured data. The SVR suggested the points 29.41m and 50.23m as the center of the zone for the artificial leaks at 32 and 48 m, respectively. Therefore, the points 29.5m and 50m (the closest leak candidates) were selected as the zone centers. Accordingly, based on the zone radius=3, in each case, a subset of the training set, including 130 samples, is used to train the in-zone ensemble classifier. In the end, the classifiers are applied to the measured data. Table (7) presents the application of the proposed method on experimental data in detail. As can be seen, in both cases, the model reported the correct location of the leak.

6.6. Uncertainty analysis

The quality of the dataset can significantly affect the performance of the applied ML algorithms. In this study, it is assumed that the applied hydraulic model is calibrated before

applying it in the dataset generation stage. Nevertheless, in practice, utilizing a not-well-calibrated model affects the quality of the dataset by imposing some uncertainties in system parameters like pipe's roughness and initial flow. To investigate the impact of uncertainties in pipe roughness and initial flow on the proposed method, two uncertainty scenarios were considered here: 1) the pipe roughness has uncertainty; 2) the initial pipe flow has uncertainty. A test set with 200 samples was randomly generated for each scenario with uncertainties in the range ± 10 to $\pm 20\%$. Therefore, a total of 8 test sets with uncertainty were provided. The uncertain flow and pipe roughness were generated using a uniform statistical distribution with the mean value of 0.09 *litr/s* and 0.005 *mm* , respectively. Other simulation parameters are the same as the training set. These test sets with uncertainty can be considered as the real data that contain uncertainties in practice. Table (8) presents the performance of the proposed model on various scenarios and different levels of uncertainty. The results indicated that, even in the presence of 20% uncertainty in pipe roughness, the model could precisely detect more than 97% of cases. Also, the maximum leak detection error of the model on miss classified cases is limited to 4.69% of pipe length. The reason corresponds to the fact is that the classifier uses the first $3L/a$ of the transient wave and, the energy dissipation effects at this part of the response signal are not noticeable. On the other hand, uncertainties higher than 5% in initial flow have a considerable impact on the model's performance. The reason is that the initial flow is an important parameter in the applied transient excitation, and its variation can noticeably affect the response head values.

7. Conclusion

A novel framework for leak detection in pipelines has been introduced based on machine learning and transient flow analysis. The proposed method takes advantage of both transient

and machine learning to suggest a fast and reliable technique for leak detection in pipelines. The problem search space is modeled using a wide range of leak scenarios, and a two-stage Machine Learning approach is taken to discriminate the patterns resulting from a variety of leak states. The presented methodology is evaluated using a large-scale theoretical pipeline and an experimental Reservoir-Pipe-Valve model. The outcomes of this study are as follow.

- The introduced model combines regression and classification approaches to address the issue of many classes in high-resolution leak detection of large-scale problems. In the first stage, the problem search space is reduced to restrict the maximum leak detection error using the Support Vector Regression. In the second stage, an Ensemble classifier based on Linear Discriminant components is trained and utilized to determine the exact location of the leak within the identified zone. The applied Majority Voting technique provides most likely correct predictions with higher confidence based on the outputs of the individual component classifiers.
- In practice, due to uncertainties, the roughness of the pipeline can change by aging and are not crisp values. In most conventional model-based methods, the accuracy and uniqueness of results are affected by uncertainties; thus, in methods like ITA, real-time simultaneous calibration and leak detection are required to enhance the reliability of the results. Such an aspect can increase the dimension of the corresponding optimization problem and make it computationally expensive. In this study, two techniques are applied to enhance the robustness of the model against pipe roughness uncertainties. First, only the initial part of the transient wave is used as the system's feature vector. This selection corresponds to the fact that the first $2L/a$ of the transient waves contain information on the complete pipeline with negligible energy dissipation effects. Second, to obtain diverse classifiers in the ensemble model, the random subspace method is applied. In this method, different subsets of pressure heads in each sample

data are used to train individual linear discriminant classifiers. Compared to a single classifier, this approach, reduce the chance of selecting inaccurate features through the random selection of small subsets of features. The uncertainty analysis confirmed that the model is stable and reliable against uncertainties in pipe roughness.

- In the proposed method, the computational effort is concentrated in the training of the applied Machine Learning algorithms. When these stages are complete, the processing of new measurements and predicting new leaks can be performed almost immediately. On this basis, compared to the model-based methods like the Inverse Transient Analysis, no complex and time-consuming optimization is required in real field applications. Thus, the method is much faster than the inverse transient analysis.
- Previous investigations on inverse methods such as Inverse Transient Analysis have shown that raised optimization problems can be highly multimodal and ill-posed with non-unique solutions. The introduced model is a data-driven technique for direct leak detection. Both algorithms applied in this method give unique predictions. Thus, the final output of the model is a unique location for the leak.

This paper tested the proposed method on an experimental pipeline and a large-scale theoretical case study. The results demonstrated the promising potential of the proposed framework for real field application. Nevertheless, further evaluation of the method in real field application is required to uncover the limitations and potentials of the method. Also, the method can be extended to more complex situations with more than one leak in the presence of other anomalies like the blockage in piping systems.

8. Abbreviations

ML	Machine Learning
AI	Artificial Intelligence
TDM	Transient Damping Methods

692 TRM Transient Reflection Methods
693 SRM System Response Methods
694 ITM Inverse Transient Methods
695 ITA Inverse Transient Analysis
696 ELT Experience Learning Theory
697 SLT Statistical Learning Theory
698 ANN Artificial Neural Networks
699 SVM Support Vector Machine
700 SVC Support Vector Classification
701 SVR Support Vector Regression
702 LDA Linear Discriminant Analysis
703 K-NN K Nearest Neighbor
704 RPV Reservoir-Pipe-Valve
705 MOC Method of Characteristics
706

707 **9. Statements and Declarations**

708

709 **9.1. Ethical Approval**

710 Not applicable.

711

712 **9.2. Consent to Participate**

713 Not applicable.

714

715 **9.3. Consent to Publish**

716 Not applicable.

717

718 **9.4. Authors Contributions**

719 All authors contributed to the study conception and design. Model development, data
720 generation and analysis were performed by Amir Houshang Ayati. The first draft of the
721 manuscript was written by Amir Houshang Ayati and all authors commented on previous
722 versions of the manuscript. All authors read and approved the final manuscript.

723

724 **9.5. Funding**

725 The authors declare that no funds, grants, or other support were received during the preparation
726 of this manuscript.

727

728 **9.6. Competing Interests**

729 The authors have no relevant financial or non-financial interests to disclose.

730

731 **9.7. Availability of data and materials**

732 The simulation and experimental data as used during the study are available from the
733 corresponding author upon reasonable request.

10. References

- Abouhamad, M., Zayed, T., & Moselhi, O. (2016). Leak Detection in Buried Pipes Using Ground Penetrating Radar; A Comparative Study. In *Pipelines 2016* (pp. 417-424). <https://ascelibrary.org/doi/10.1061/9780784479957.039>
- Ayati, A. H., Haghighi, A., & Ghafouri, H. R. (Forthcoming). Machine Learning Assisted Model for Leak Detection in Water Distribution Networks Using Hydraulic Transient Flows. *Journal of Water Resources Planning and Management (ASCE)*. doi:10.1061/(ASCE)WR.1943-5452.0001508
- Ayati, A. H., Haghighi, A., & Lee, P. (2019). Statistical Review of Major Standpoints in Hydraulic Transient-Based Leak Detection. *Journal of Hydraulic Structures*, 5(1), 1-26. doi:10.22055/jhs.2019.27926.1095
- Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2005). Ensemble diversity measures and their application to thinning. *Information Fusion*, 6(1), 49-62. doi:<https://doi.org/10.1016/j.inffus.2004.04.005>
- Bohorquez, J., Alexander, B., Simpson Angus, R., & Lambert Martin, F. (2020). Leak Detection and Topology Identification in Pipelines Using Fluid Transients and Artificial Neural Networks. *Journal of Water Resources Planning and Management*, 146(6), 04020040. doi:10.1061/(ASCE)WR.1943-5452.0001187
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers*. Paper presented at the 5th Annual Workshop Computational Learning Theory, ACM, Pittsburgh.
- Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1), 5-20. doi:<https://doi.org/10.1016/j.inffus.2004.04.004>
- Brunone, B. (1999). Transient test-based technique for leak detection in outfall pipes. *Journal of Water Resources Planning and Management*, 125(5), 302-306.
- Brunone, B., Golia, U. M., & Greco, M. (1995). Effects of Two-Dimensionality on Pipe Transients Modeling. *Journal of Hydraulic Engineering*, 121(12), 906-912. doi:10.1061/(ASCE)0733-9429(1995)121:12(906)
- Butterworth, S. (1930). On the Theory of Filter Amplifiers. *Experimental Wireless and the Wireless Engineer*, 7, 536-541.
- Caputo, A. C., & Pelagagge, P. M. (2003). Using neural networks to monitor piping systems. *Process Safety Progress*, 22(2), 119-127. doi:10.1002/prs.680220208
- Carreño-Alvarado, E. P., Reynoso-Meza, G., Montalvo, I., & Izquierdo, J. I. (2017). *A comparison of machine learning classifiers for leak detection and isolation in urban networks*. Paper presented at the Congress on Numerical Methods in Engineering, Valencia, Spain.
- Chamasemani, F. F., & Singh, Y. P. (2011, 27-29 Sept. 2011). *Multi-class Support Vector Machine (SVM) Classifiers -- An Application in Hypothyroid Detection and Classification*. Paper presented at the 2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications.
- Chaudhry, M. H. (2016). *Applied Hydraulic Transients*: Springer New York.
- Chunli, Fan, Sun Fengrui, and Yang Li (2005). Investigation on nondestructive evaluation of pipelines using infrared thermography. 2005 Joint 30th International Conference on Infrared and Millimeter Waves and 13th International Conference on Terahertz Electronics, IEEE. Vol. 2, pp. 339-340. Doi: 10.1109/ICIMW.2005.1572551
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Mach. Learn*, 20(3), 273-297.
- Covas, D., & Ramos, H. (2010). Case Studies of Leak Detection and Location in Water Pipe Systems by Inverse Transient Analysis. *Journal of Water Resources Planning and Management*, 136(2), 248-257. doi:10.1061/(ASCE)0733-9496(2010)136:2(248)
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. In: Mozer M.C., Jordan M.I., and Petsche T. (Eds.), *Advances in Neural Information*

- Processing Systems 9, MIT Press, Cambridge, MA. *Advances in Neural Information Processing Systems*, 155-161.
- Duan, H. F., & Lee, P. J. (2016). Transient-Based Frequency Domain Method for Dead-End Side Branch Detection in Reservoir Pipeline-Valve Systems. *Journal of Hydraulic Engineering*, 142(2), 04015042. doi:10.1061/(ASCE)HY.1943-7900.0001070
- Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd ed.). New York: Wiley.
- Farley, B., Mounce Stephen, R., & Boxall Joby, B. (2011). *Field Validation of "Optimal" Instrumentation Methodology for Burst/Leak Detection and Location*. Paper presented at the Water Distribution Systems Analysis Conf, Reston, VA.
- Feng, J., & Zhang, H. (2006, 2006//). *Algorithm of Pipeline Leak Detection Based on Discrete Incremental Clustering Method*. Paper presented at the Computational Intelligence, Berlin, Heidelberg.
- Ferrante, M., & Brunone, B. (2003). Pipe system diagnosis and leak detection by unsteady-state tests. 1. Harmonic analysis. *Advances in Water Resources*, 26(1), 95-105. doi:10.1016/S0309-1708(02)00101-X
- Ferrante, M., Brunone, B., & Meniconi, S. (2009). Leak-edge detection. *Journal of Hydraulic Research*, 47(2), 233-241. doi:10.3826/jhr.2009.3220
- Gong, J., Zecchin Aaron, C., Simpson Angus, R., & Lambert Martin, F. (2014). Frequency Response Diagram for Pipeline Leak Detection: Comparing the Odd and Even Harmonics. *Journal of Water Resources Planning and Management*, 140(1), 65-74. doi:10.1061/(ASCE)WR.1943-5452.0000298
- Goulet, J.-A., Coutu, S., & Smith, I. F. C. (2013). Model falsification diagnosis and sensor placement for leak detection in pressurized pipe networks. *Advanced Engineering Informatics*, 27(2), 261-269. doi:https://doi.org/10.1016/j.aei.2013.01.001
- Haghighi, A., Covas, D., & Ramos, H. (2012). Direct backward transient analysis for leak detection in pressurized pipelines: from theory to real application. *Journal of Water Supply: Research and Technology - Aqua*, 61(3), 189-200. doi:10.2166/aqua.2012.032
- Haghighi, A., & Shamloo, H. (2011). Transient generation in pipe networks for leak detection. *Proceedings of the Institution of Civil Engineers - Water Management*, 164(6), 311-318. doi:10.1680/wama.2011.164.6.311
- Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832-844. doi:10.1109/34.709601
- Huang, Y.-C., Lin, C.-C., & Yeh, H.-D. (2015). An Optimization Approach to Leak Detection in Pipe Networks Using Simulated Annealing. *Water Resources Management*, 29(11), 4185-4201. doi:10.1007/s11269-015-1053-4
- Jönsson, L., & Larson, M. (1992). Leak Detection through Hydraulic Transient Analysis. In B. Coulbeck & E. P. Evans (Eds.), *Pipeline Systems* (pp. 273-286). Dordrecht: Springer Netherlands.
- Jung, B. S., & Karney, B. W. (2008). Systematic exploration of pipeline network calibration using transients. *Journal of Hydraulic Research*, 46(sup1), 129-137. doi:10.1080/00221686.2008.9521947
- Kim, S. H. (2018). Development of Multiple Leakage Detection Method for a Reservoir Pipeline Valve System. *Water Resources Management*, 32(6), 2099-2112. doi:10.1007/s11269-018-1920-x
- Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, 51(2), 181-207. doi:10.1023/A:1022859003006
- Kwok, J. T.-Y. (1998, 20-20 Aug. 1998). *Support vector mixture for classification and regression problems*. Paper presented at the Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170).
- Lee, P. J., Duan, H.-F., Tuck, J., & Ghidaoui, M. (2015). Numerical and Experimental Study on the Effect of Signal Bandwidth on Pipe Assessment Using Fluid Transients. *Journal of Hydraulic Engineering*, 141(2), 04014074. doi:10.1061/(ASCE)HY.1943-7900.0000961

835 Lee, P. J., Lambert, M. F., Simpson, A. R., Vítkovsky, J. P., & Misiunas, D. (2007). Leak location in single
 836 pipelines using transient reflections. *Australasian Journal of Water Resources*, 11(1), 53-65.
 837 doi:10.1080/13241583.2007.11465311
 838 Lee, P. J., Vítkovský, J. P., Lambert, M. F., Simpson, A. R., & Liggett, J. A. (2005a). Leak location using
 839 the pattern of the frequency response diagram in pipelines: a numerical study. *Journal of*
 840 *Sound and Vibration*, 284(3), 1051-1073. doi:https://doi.org/10.1016/j.jsv.2004.07.023
 841 Leu, S.-S., & Bui, Q.-N. (2016). Leak Prediction Model for Water Distribution Networks Created Using
 842 a Bayesian Network Learning Approach. *Water Resources Management*, 30(8), 2719-2733.
 843 doi:10.1007/s11269-016-1316-8
 844 Liggett, A., & Chen, L. C. (1994). Inverse Transient Analysis in Pipe Networks. *Journal of Hydraulic*
 845 *Engineering*, 120(8), 934-955. doi:10.1061/(ASCE)0733-9429(1994)120:8(934)
 846 Liou, C. P. (1998). Pipeline Leak Detection by Impulse Response Extraction. *Journal of Fluids*
 847 *Engineering*, 120(4), 833-838. doi:10.1115/1.2820746
 848 Mashford, J., Silva, D. D., Marney, D., & Burn, S. (2009). An Approach to Leak Detection in Pipe
 849 Networks Using Analysis of Monitored Pressure Values by Support Vector Machine. Third
 850 International Conference on Network and System Security, IEEE, pp. 534-539.
 851 Mashford, J., De Silva, D., Burn, S., & Marney, D. (2012). LEAK DETECTION IN SIMULATED WATER PIPE
 852 NETWORKS USING SVM. *Applied Artificial Intelligence*, 26(5), 429-444.
 853 doi:10.1080/08839514.2012.670974
 854 Mitchell, T. M. (1997). *Machine Learning*: McGraw-Hill, Inc.
 855 Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of Machine Learning* (F. Bach Ed. 2
 856 ed.). Cambridge, Massachusetts, London, England: MIT Press.
 857 Mounce, S. R., Mounce, R. B., & Boxall, J. B. (2011). Novelty detection for time series data analysis in
 858 water distribution systems using support vector machines. *Journal of Hydroinformatics*, 13(4),
 859 672-686. doi:10.2166/hydro.2010.144
 860 Muhlbaier, M., Topalis, A., & Polikar, R. (2005, 2005//). *Ensemble Confidence Estimates Posterior*
 861 *Probability*. Paper presented at the Multiple Classifier Systems, Berlin, Heidelberg.
 862 Mukherjee, S., Osuna, E., & Girosi, F. (1997, 24-26 Sept. 1997). *Nonlinear prediction of chaotic time*
 863 *series using support vector machines*. Paper presented at the Neural Networks for Signal
 864 Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop.
 865 Nash, G. A., & Karney, B. W. (1999). Efficient inverse transient analysis in series pipe systems. *Journal*
 866 *of Hydraulic Engineering*, 125(7), 761-764.
 867 Nixon, W., Ghidaoui, M. S., & Kolyshkin, A. A. (2006). Range of Validity of the Transient Damping
 868 Leakage Detection Method. *Journal of Hydraulic Engineering*, 132(9), 944-957.
 869 doi:10.1061/(ASCE)0733-9429(2006)132:9(944)
 870 Ozevin, Didem, & Harding, James (2012). Novel leak localization in pressurized pipeline networks using
 871 acoustic emission and geometric connectivity. *International Journal of Pressure Vessels and*
 872 *Piping*, 63-69. doi:10.1016/j.ijpvp.2012.01.001
 873 Plath, M., Mathias, E., & Knut, W. (2014). Energy efficiency and energy saving in the German water
 874 industry. *Water Pract Technol*, 9(2), 256-263.
 875 Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*,
 876 6(3), 21-45. doi:10.1109/MCAS.2006.1688199
 877 Poulakis, Z., Valougeorgis, D., & Papadimitriou, C. (2003). Leakage detection in water pipe networks
 878 using a Bayesian probabilistic framework. *Probabilistic Engineering Mechanics*, 18(4), 315-
 879 327. doi:10.1016/s0266-8920(03)00045-6
 880 Puust, R., Kapelan, Z., Savic, D. A., & Koppel, T. (2010). A review of methods for leakage management
 881 in pipe networks. *Urban Water Journal*, 7(1), 25-45. doi:10.1080/15730621003610878
 882 Ranginkaman M H, Haghighi A, Vali Samani H M. Inverse Frequency Response Analysis For Pipelines
 883 Leak Detection Using The Particle Swarm Optimization. *International Journal of Optimization*
 884 *in Civil Engineering*. 2016; 6 (1) :1-12URL: http://ijoce.iust.ac.ir/article-1-234-en.html
 885 Refaeilzadeh, P., Tang, L., & Liu, H. (2009). *Cross-validation*. Springer, Berlin.

- Sarkamaryan, S., Haghighi, A., Ashrafi, S. M., & Samani, H. M. V. (2020). Surrogate-Assisted Inverse Transient Analysis (SAITA) for Leakage Detection in Pressurized Piping Systems. *Iranian Journal of Science and Technology, Transactions of Civil Engineering*. doi:10.1007/s40996-020-00516-4
- Sattar, A. M., & Chaudhry, M. H. (2008). Leak detection in pipelines by frequency response method. *Journal of Hydraulic Research*, 46(sup1), 138-151. doi:10.1080/00221686.2008.9521948
- Shamloo, H., & Haghighi, A. (2009). Leak detection in pipelines by inverse backward transient analysis. *Journal of Hydraulic Research*, 47(3), 311-318. doi:10.3826/jhr.2009.3428
- Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press.
- Soldevila, A., Fernandez-Canti, R. M., Blesa, J., Tornil-Sin, S., & Puig, V. (2017). Leak localization in water distribution networks using Bayesian classifiers. *Journal of Process Control*, 55, 1-9. doi:https://doi.org/10.1016/j.jprocont.2017.03.015
- Sophocleous, S., Savić, D., & Kapelan, Z. (2019). Leak Localization in a Real Water Distribution Network Based on Search-Space Reduction. *Journal of Water Resources Planning and Management*, 145(7). doi:10.1061/(asce)wr.1943-5452.0001079
- Tang, X., Liu, Y., Zheng, L., Ma, C., & Wang, H. (2009, 4-5 July 2009). *Leak Detection of Water Pipeline Using Wavelet Transform Method*. Paper presented at the 2009 International Conference on Environmental Science and Information Application Technology.
- Tao, T., Huang, H., Li, F., & Xin, K. (2014). Burst Detection Using an Artificial Immune Network in Water-Distribution Systems. 140(10), 04014027. doi:doi:10.1061/(ASCE)WR.1943-5452.0000405
- Vapnik, V. N., & Kotz, S. (1982). *Estimation of dependences based on empirical data: Springer series in statistics (Springer series in statistics)*. Berlin: Springer.
- Vitkovsky, J., Lambert, M., Simpson, A. (2000). Advances in unsteady friction modeling in transient pipe flow. Proc.8th Int. Conf. On Pressure Surges, Professional Engineering Publishing Limited, London, UK
- Vitkovsky John, P., Lambert Martin, F., Simpson Angus, R., & Liggett James, A. (2007). Experimental Observation and Analysis of Inverse Transients for Pipeline Leak Detection. *Journal of Water Resources Planning and Management*, 133(6), 519-530. doi:10.1061/(ASCE)0733-9496(2007)133:6(519)
- Vitkovsky , J. P., Liggett , J. A., Simpson , A. R., & Lambert , M. F. (2003). Optimal Measurement Site Locations for Inverse Transient Analysis in Pipe Networks. *Journal of Water Resources Planning and Management*, 129(6), 480-492. doi:10.1061/(ASCE)0733-9496(2003)129:6(480)
- Walt, J. C., Heyns, P. S., & Wilke, D. N. (2019). Pipe network leak detection: comparison between statistical and machine learning techniques. *Urban Water Journal*, 15(10), 953-960. doi:10.1080/1573062x.2019.1597375
- Wang, X.-J., Lambert Martin, F., Simpson Angus, R., Liggett James, A., & Vítkovský John, P. (2002). Leak Detection in Pipelines using the Damping of Fluid Transients. *Journal of Hydraulic Engineering*, 128(7), 697-711. doi:10.1061/(ASCE)0733-9429(2002)128:7(697)
- Wu Zheng, Y., & Sage, P. (2006). *Water Loss Detection via Genetic Algorithm Optimization-based Model Calibration*. Paper presented at the 8th Annual Int. Symp. on Water Distribution System Analysis, Reston, VA.
- Wu Zheng, Y., Sage, P., & Turtle, D. (2010). Pressure-Dependent Leak Detection Model and Its Application to a District Water System. *Journal of Water Resources Planning and Management*, 136(1), 116-128. doi:10.1061/(ASCE)0733-9496(2010)136:1(116)
- Xu, X., & Karney, B. (2017). An Overview of Transient Fault Detection Techniques. In C. Verde & L. Torres (Eds.), *Modeling and Monitoring of Pipelines and Networks: Advanced Tools for Automatic Monitoring and Supervision of Pipelines* (pp. 13-37). Cham: Springer International Publishing.
- Yu, H., & Kim, S. (2012). SVM Tutorial — Classification, Regression and Ranking. In *Handbook of Natural Computing* (pp. 479-506).

- Zhang, Q., Wu Zheng, Y., Zhao, M., Qi, J., Huang, Y., & Zhao, H. (2016). Leakage Zone Identification in Large-Scale Water Distribution Systems Using Multiclass Support Vector Machines. *Journal of Water Resources Planning and Management*, 142(11), 04016042. doi:10.1061/(ASCE)WR.1943-5452.0000661
- Zhang, X., Srinivasan, R., & Van Liew, M. (2009). Approximating SWAT Model Using Artificial Neural Network and Support Vector Machine1. *JAWRA Journal of the American Water Resources Association*, 45(2), 460-474. doi:https://doi.org/10.1111/j.1752-1688.2009.00302.x

Figure 1. Step by step flow of the proposed method

Figure 2. The core idea of SVM regression (SVR)

Figure 3. First case study

Figure 4. Leak detection error of the Cubic SVR on the train and first test datasets (first case study).

Figure 5. Percentage exceedance associated with the absolute error in the leak location prediction for train and first datasets (first case study).

Figure 6. The performance of the model on the train and test sets (first case study).

Figure 7. Experimental setup of the second case study. (a) schematic; (b) photograph

Figure 8. Experimental via Simulation response of the second case study a) no-leak condition; b) Leak (1): 2.3 mm² leak at 32 m; (c) Leak (2): 1.88 mm² leak at 48 m.

Figure 9. The leak detection error of Cubic SVR on the train and the first test sets (second case study).

Figure 10. the leak detection error of the model on the train and two test sets (second case study).

Table 1. SVR performance in leak zone identification (data without down-sampling)

Regression Algorithm	MAE		MLDE		MER %
	train	test	train	test	
SVR	Linear	36.9	31.9	366.0	36.60
	Quadratic	19.7	18.8	76.3	7.63
	Qubic	22.0	19.7	40.2	4.22
	RBF Fine	27.5	26.3	210.8	21.08
	RBF Medium	23.8	21.9	83.8	8.38
	RBF Coarse	49.1	41.1	412.3	41.23

Note: MAE is Mean absolute error, MLDE is Maximum Leak Detection Error, MER is Maximum Error Ratio. MER is calculated by dividing maximum error of train and test by the whole pipe length

Table 2. Comparison of different regression algorithms in leak zone identification (down-sampled data)

Regression Algorithm		MAE (m)		MLDE (m)		MER %
		train	test	train	test	
Tree	Fine (leaf size=4)	3.3	10.1	619.5	572.1	61.95
	Medium (leaf size=12)	4.7	10.7	837.5	572.1	83.75
	Coarse (leaf size=36)	10.4	14.6	911.8	575.0	91.18
SVR	Linear	38.2	33.1	373.5	319.0	37.35
	Quadratic	22.3	21.4	104.1	107.0	10.70
	Qubic	22.4	20.6	41.2	43.5	4.35
	RBF Fine	27.3	25.9	285.6	139.9	28.56
	RBF Medium	24.0	22.0	685.1	215.3	68.51
	RBF Coarse	48.6	40.4	584.1	343.0	58.41
Ensemble	Boosted Trees	23.4	25.3	707.7	168.2	70.77
	Bagged Trees	2.0	6.3	628.5	183.4	62.85
Linear Regression		24.5	22.3	310.9	235.0	31.09

Note: MAE is Mean absolute error, MLDE is Maximum Leak Detection Error, MER is Maximum Error Ratio. MER is calculated by dividing maximum error of train and test by the whole pipe length

1205

Table 3. Sensitivity analysis of the distance between leak candidates

Distance between leak candidates(m)*	No. Leak Candidates	MAE		MLDE		MER %	No. in-zone leak candidates
		train	test	train	test		
2	499	22.4	20.6	41.2	43.5	4.35	45
4	249	22.7	21.1	42.5	92.1	9.21	49
8	124	24.3	24.1	75.8	160.9	16.09	43
10	99	23.4	22.0	38.1	185.3	18.53	39
20	49	22.9	22.8	35.6	418.7	41.87	43
25	39	24.9	32.9	37.1	1079.3	107.93	89
40	24	23.3	28.9	35.6	1004.8	100.48	53
50	19	24.9	33.2	37.1	1120.8	112.08	47
100	9	20.8	79.8	29.7	4412.0	441.20	91
200	4	22.0	146.3	29.7	6636.2	663.62	69

1206 *The distance between leak candidates=leak detection resolution.

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

Table 4. Comparison of different classifiers for in-zone leak detection

	Classifier	Training time (s)	Accuracy %		MAE (m)		MLDE (m)		MER %
			train	test	train	test	train	test	
SVM	Linear	198	92.6	88.2	0.1	0.2	2	6	6.8
	Quadratic	634	77.8	70.2	1.1	1.2	6	64	72.7
	Qubic	811	2.7	2.7	30.2	28.8	80	80	90.9
	RBF Fine	11	73.8	16.9	1.8	7.9	30	72	81.8
	RBF Medium	7	13.1	10.0	14.2	14.6	66	68	77.3
	RBF Coarse	7	3.3	4.0	31.8	31.5	80	78	90.9
KNN	Fine	2	12.8	11.3	7.1	7.6	50	56	63.6
	Medium	3	2.2	4.4	9.3	11.0	18	48	54.5
	Coarse	4	1.8	4.2	19.2	22.8	88	86	100.0
	Cosine	5	7.8	8.9	7.6	8.4	28	34	38.6
	Cubic	17	2.0	3.3	9.4	9.5	20	28	31.8
	Weighted	18	15.7	11.3	6.3	7.6	44	56	63.6
Ensemble	Boosted Trees	115	32.7	28.4	3.3	3.3	26	24	29.5
	Bagged Trees	86	48.4	32.2	2.1	2.7	2	28	31.8
	Subspace LDA	94	100.0	100.0	0.0	0.0	0	0	0.0
	Subspace KNN	110	10.3	6.4	11.8	15.9	42	82	93.2
	RUS Boosted Trees	209	23.6	20.0	4.6	4.6	38	38	43.2
	LDA	3	100.0	96.2	0.0	0.7	0	88	100.0

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

1242

1243

1244

1245

1246

1247

Table 5. Sensitivity analysis of subspace size in LDA-based Ensemble classifier

No. LDA models	Subspace size	Training time (s)	Accuracy %		MAE (m)		MLDE (m)		MER %
			train	test	train	test	train	test	
30	10%	28.6	98.2	97.3	0.1	0.1	8	4	9.1
30	20%	32.7	100.0	99.6	0.0	0.0	0	2	2.3
30	30%	57.2	100.0	100.0	0.0	0.0	0	0	0.0
30	40%	77.7	100.0	100.0	0.0	0.0	0	0	0.0
30	50%	94.0	100.0	100.0	0.0	0.0	0	0	0.0
30	60%	118.5	100.0	100.0	0.0	0.0	0	0	0.0
30	70%	143.0	100.0	100.0	0.0	0.0	0	0	0.0
30	80%	204.3	100.0	100.0	0.0	0.0	0	0	0.0
30	90%	228.9	100.0	100.0	0.0	0.0	0	0	0.0

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

Table 6. Sensitivity analysis of the number of models in LDA-based Ensemble classifier

No. LDA models	Subspace size	Training time (s)	Accuracy %		MAE (m)		MLDE (m)		MER %
			train	test	train	test	train	test	
10	20%	20.4	99.8	99.6	0.0	0.0	2	2	2.3
20	20%	28.6	99.8	99.6	0.0	0.0	2	2	2.3
30	20%	32.7	99.8	99.8	0.0	0.0	2	2	2.3
40	20%	45.0	100.0	100.0	0.0	0.0	0	0	0.0
50	20%	57.2	100.0	100.0	0.0	0.0	0	0	0.0
60	20%	65.4	100.0	100.0	0.0	0.0	0	0	0.0
70	20%	73.6	100.0	100.0	0.0	0.0	0	0	0.0
80	20%	85.8	100.0	100.0	0.0	0.0	0	0	0.0
90	20%	89.9	100.0	100.0	0.0	0.0	0	0	0.0
100	20%	98.1	100.0	100.0	0.0	0.0	0	0	0.0

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

1296

Table 7. Application of the proposed method on Experimental Leaks (second case study)

		Experimental Leak (1)	Experimental Leak (2)
Leak parameters	Real Location	32m	48m
	Leak effective area	2.30 mm ²	1.88 mm ²
Zone identification	Regression model	Cubic SVR	Cubic SVR
	Identified zone radius	2.74m	2.74m
	Applied zone radius	3m	3m
	Identified zone center	29.41m	50.23m
	Applied zone center	29.5m	50m
	No. in-zone candidates	13	13
	In-zone leak candidates	26.5, 27, 27.5, 28, 28.5, 29, 29.5, 30, 30.5, 31, 31.5, 32, 32.5 m	47, 47.5, 48, 48.5, 49, 49.5, 50, 50.5, 51, 51.5, 52, 52.5, 53 m
	Size of zone dataset	130 samples	130 samples
In-zone leak detection	No. Ensemble models	40	40
	Component model type	LDA	LDA
	Subspace size	50%	50%
	Ensemble train accuracy	100%	100%
	Ensemble train MAE	0	0
	Ensemble train MLDE	0	0
	Percentage of Ensemble correct votes on experimental data	82.5%	87.5%
	Detected location	32m	48m

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311 Table 8. The performance of the proposed model considering different levels of uncertainty in
1312 pipe roughness and initial flow.

Uncertainty%	Uncertainty in pipe roughness				Uncertainty in Initial flow			
	Accuracy%	MAE (m)	MLDE (m)	MER%	Accuracy%	MAE (m)	MLDE (m)	MER%
5	99.58	0.015	0.5	0.8	94	0.125	1	1.6
10	99.23	0.015	0.5	1.6	77.5	1.245	7	10.9
15	98.47	0.02	1	3.1	67.5	2.74	18.5	28.9
20	97.64	0.02	1	3.1	60	3.83	26	40.6

1313

1314