

## Experiments based comparative evaluations of machine learning techniques for leak detection in water distribution systems

Maryam Kammoun <sup>a,\*</sup>, Amina Kammoun<sup>a</sup> and Mohamed Abid<sup>a,b</sup>

<sup>a</sup> CES Research Lab, National Engineering School of Sfax, Sfax University, Sfax, Tunisia

<sup>b</sup> Digital Research Center (CRNS), Technopole of Sfax, Sfax, Tunisia

\*Corresponding author. E-mail: kammounmaryam@gmail.com

 MK, 0000-0001-5966-749X

### ABSTRACT

Leakage in water distribution systems is a significant long-standing problem due to the huge economic and ecological losses. Different leak detection studies have been examined in literature using different types of technologies and data. Currently, although machine learning techniques have achieved tremendous progress in outlier detection approaches, they are still limited in terms of water leak detection applications. This research aims to improve the leak detection performances by refining the choices of learning data and techniques. From this perspective, commonly used techniques for leak detection are assessed in this paper, and the characteristics of hydraulic data are investigated. Four intelligent algorithms are compared, namely k-nearest neighbors, support vector machines, logistic regression, and multi-layer perceptron. This study focuses on six experiments based on identifying outliers in various packages of pressure and flow data, yearly data, seasonal data, night data, and flow data difference to detect leakage in water distribution networks. Different scenarios of realistic water demand in two networks from the benchmark dataset LeakDB are used. Results demonstrate that the leak detection accuracy varies between 30% and 100% depending on the experiment and the choices of algorithms and data.

**Key words:** algorithms evaluation, comparative study, intelligent techniques, LeakDB, leak detection, machine learning

### HIGHLIGHTS

- This is the first work to use a common benchmark dataset to compare four machine learning techniques used for leak detection application.
- The choice of data has a significant effect on improving leak detection performances using learning algorithms.
- Six experiments are proposed to offer a comprehensive comparative study of algorithms behavior adopted for leak detection.

### INTRODUCTION

Water stress is a prominent issue across the globe. Some of the most plausible reasons include the critical conditions of water distribution networks, uncontrolled and untreated leaks, and breaks in the distribution pipelines all over the world. According to [Kanakoudis \*et al.\* \(2015\)](#), around 50% of the water volume entering a water distribution system is lost. Therefore, inspecting and monitoring water distribution networks can detect and repair leaks to reduce this huge water loss ([Kanakoudis & Tolikas 2001](#)). Leakage detection and location have become the central focus of most research works over the last years. Several software leak detection methods were presented in the literature. Some of them are based on transient analysis of pressure wave reflection using physical equations ([Srirangarajan \*et al.\* 2013](#)); others use statistical equations and algorithms grounded on pressure point analysis ([bin Md Akib \*et al.\* 2011](#)).

Further methods based on model conception encompassed all the hydraulic characteristics of a real distribution network using mathematical equations ([Adedeji \*et al.\* 2017](#)). Recently, with the emergence of big data and artificial intelligence algorithms, data-driven based methods have appeared, particularly prediction and classification methods which rest on machine learning-based outlier detection ([Wu & Liu 2017](#)). These methods rely upon data analysis to find abnormal features that do not conform to the expected behavior. In leak detection literature, these methods were developed to detect leaks as abnormal patterns in water distribution networks through analyzing hydraulic data, mainly pressure and flow data.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (<http://creativecommons.org/licenses/by/4.0/>).

Different data types and machine learning techniques were used to determine intelligent and efficient leak detection solutions. Some research works were based on historical datasets using test networks, while others worked with generated datasets using modeled real networks.

The efficient key of outlier detection methods, as a learning category, corresponds to the quality of input data and their compatibility with machine learning techniques (Ahmed *et al.* 2016). However, consistent noise in data tends to be similar to the actual outliers, which makes it difficult to be distinguished and classified. For example, in water distribution networks, the variations of pressure and flow data are closely related to changes in water consumption demands. These consumptions are periodic, according to seasons and days, and unplanned due to unpredictable demands or repair. Therefore, a leak can be considered as an unpredictable increase in demands, which makes it difficult to be classified as an outlier. Furthermore, an increase in demand which is not associated with leakage in training input data can deeply disturb the learning quality and subsequently the detection efficiency. Within this framework, this study unveils six experiments using, for each one, various pressure and flow training data, namely yearly data, seasonal or night data, flow difference data, and data with incipient or abrupt leaks. All these packages are trained separately to eliminate related noises and evaluate for each experiment the consequential effect. This work aims to compare and evaluate four machine learning algorithms for leak detection applications: artificial neural networks (ANN), k-nearest neighbors (KNN), support vector machine (SVM), and logistic regression (LR). This experimental comparison rests upon the benchmark dataset LeakDB using different scenarios for realistic water demand.

The chief objective of this investigation resides in enhancing the performances of leak detection methods by evaluating the behavior of learning techniques. Different experiments were conducted to examine the effect of learning data choice on improving the efficiency of machine learning algorithms. Additionally, this research represents the pioneering comparative work that uses a common benchmark dataset LeakDB. The results can contribute considerably to the development of robust leak detection methods in water distribution networks based on artificial intelligent techniques.

The next section displays the existing related literature works, then the general research methodology is identified, and the proposed dataset and algorithms are introduced. The paper then illustrates the comparative experiments and the obtained results, and finally concludes the work and provides new perspectives for future works.

## RELATED WORKS

Artificial intelligence methods are frequently applied in the leak detection field to improve the performance and efficiency of software solutions, which are mostly either very complex and expensive or inefficient. Numerous machine learning algorithms have been adopted. However, unsupervised methods are so as used more frequently than supervised ones because the latter require a labeled dataset for training, which is not always available and is particularly expensive.

Caputo & Pelagagge (2002) presented a prediction leak detection method based on generated data using ANN. They compared three architectures: probabilistic neural network (PNN), radial basis function (BRF), and multi-layer perceptron (MLP). They found that MLP presents the best results in predictive capacity and noise rejection. Mounce *et al.* (2011) applied SVM for anomaly detection using flow and pressure data. Kang *et al.* (2017) introduced a water leakage detection system using acoustic noise data. They combined a one-dimensional convolutional neural network (CNN) with SVM. The method achieved 99.3% of accuracy. Chen *et al.* (2004) highlighted a pressure transient wave-based method. They used SVM to detect negative pressure waves in the pressure curve. Their method offered better performance compared to other transient wave-based ones. Zhou *et al.* (2019) also developed a method of leak identification based on machine learning techniques analyzing real recorded transient pressure wave data. They used a CNN to extract the informative texture-features of transient wave samples and the deep neural network (DNN) to train the neural weights and produce a leak event indication. It has been reported that this method was able to achieve a failure rate that is lower than  $6 \times 10^4$  when the signal-to-noise ratio is 0 dB.

Ayadi *et al.* (2019) adopted a combined kernelized leak detection technique based on pressure data, using Kernel Fisher Discriminant Analysis as a reduction dimensionality technique. Additionally, they compared One Class Support Vector Machine (OCSVM) and KNN classifiers. Results revealed that OCSVM was more efficient than KNN. Oliveira *et al.* (2018) proposed a classification method for leak detection using the LR algorithm. The study was based on defining the threshold between anomalies and normal data. They found that the choice of the threshold is a compromise between improving accuracy and decreasing false alarm. Kayaalp *et al.* (2017) analyzed pressure data with KNN to real-time leakage

detection and location. Bjerke (2019) used generated data from the LeakDB dataset and compared two recurrent neural network architectures for leak detection. Results disclosed good accuracies according to the chosen scenarios. However, the research study did not use the original data from the LeakDB dataset. Moreover, incipient leaks, which are hard to detect, were neglected.

The previous works centered around improving the efficiency of detection and classification, and optimizing the parameters and functions of the proposed algorithms. Most of them neglected the difficulties related to the data type, flow behaviors of water, or disturbances of consumption demands. In fact, they uncommonly defined architectures and sizes of their water distribution networks, as well as sizes and types of injected leaks. Furthermore, they adopted an unsupervised learning mode, which further complicates the validation of their proposed models.

The present work displays a clear comparative study of different algorithms within a supervised learning using the benchmark dataset LeakDB. An experimental implementation using the same dataset and the same types of networks and leaks is used to compare performances and efficiency of intelligent techniques and to evaluate their behavior through applying them with hydraulic data. The methodology of the study and various experiments are detailed in the next sections.

## RESEARCH METHODOLOGY

Software methods for detecting leaks in water distribution networks are mainly based on analyzing hydraulic data, pressure, flow, or acoustic noise data. Variations of these data depend on the architecture of installations, their dimensions, shapes, position, and water runoff in pipelines, which relays on water demands. These demands vary periodically according to the usual consumption, and sometimes randomly according to unpredictable demands or repair. Leakage phenomenon can be regarded as a sudden or gradual increase in water demands, which complicates its classification as an outlier in certain cases. In fact, there are gradual increases in water demands which may refer to seasonal changes, intensive water demand, or the existence of large or progressive growing leaks. These assumptions in addition to ordinary measurement disturbances, distribution pipe faults, and even environmental changes yield a noisy flow and pressure data field. This can greatly affect the gait and performances of detection techniques, especially machine learning techniques.

These findings were so inspiring that they were an impetus for this research to propose an experimental study that compares and evaluates the most-used intelligent learning techniques, namely KNN, SVM, LR, MLP. These techniques affect a series of experiments, each of which uses a different package of pressure and flow data, as yearly, seasonal, night, flow difference data, or data with incipient or abrupt leaks. Each package choice allows us to eliminate the linked noise and to assess the consequent effect on detection results.

Figure 1 shows the general process of this evaluation. These packages are created from the LeakDB benchmark database generating pressure and flow data using a realistic demand scenario from two different networks: Water Distribution Network (WDN) Hanoi and Net-1. The central objective of this research work is to identify anomalies for each data package.

## EVALUATION SUPPORTS

To test the behavior of different types of intelligent techniques for leak detection application, we opted to compare four analytical techniques that belonged to different categories of learning (KNN, SVM, LR, and MLP) using the benchmark dataset LeakDB.

### LeakDB dataset

LeakDB corresponds to a benchmark dataset of simulated hydraulic data, created in 2018 by Vrachimis *et al.* (2018) using the python library Water Network Tool for Resilience (Klise *et al.* 2017) and EPANET which was widely used in the analysis of the hydraulic characteristics of water distribution networks (Cheung *et al.* 2005).

This dataset rests on two different architectures of water distribution networks: Net-1 and Hanoi. Net-1 is a simple, small, and virtual WDN architecture, used for small-scale testing. It consists of nine nodes, one reservoir, and one tank. Meanwhile, Hanoi is a simplified architecture of the real WDN from the city of Hanoi (Vietnam). It consists of one reservoir, 32 nodes, and 34 pipes. Figure 2 depicts these architectures. For each WDN, 1,000 scenarios were created with structural parameters that were modified from the original values based on an uncertainty value fixed at 25%. Demand nodes were simulated for 1 year based on a model of historical real data from water utilities. The model relies on three-signal components, namely yearly component, weekly periodic component, and random component (Eliades & Polycarpou 2012). The yearly component describes variations in water consumption referring to seasonal change over the course of a year. The weekly periodic

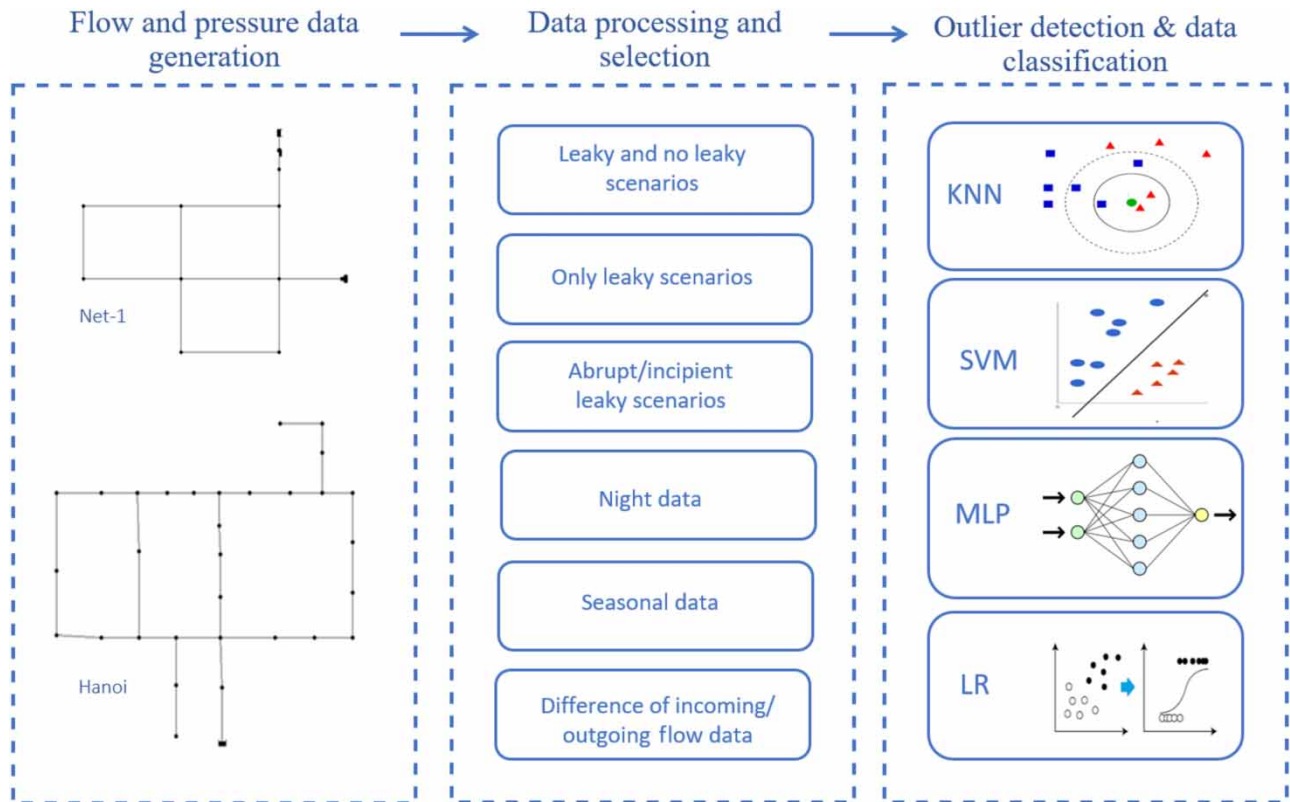


Figure 1 | Evaluation process.

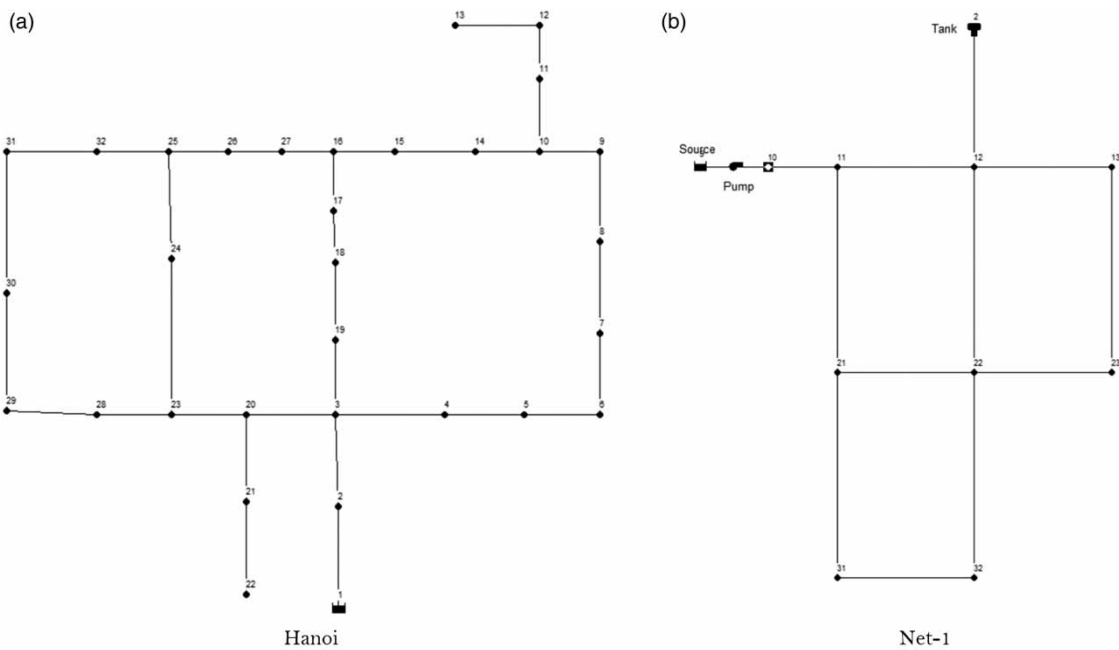


Figure 2 | The architectures of water distribution networks used from LeakDB: (a) Hanoi network; (b) Net-1 network.

component displays changes in demand signals throughout a week based on the social and economic characteristics of consumers. Both of them are approximated using Fourier series (Yamauchi & Huang 1977). The random component describes

random variations due to unpredictable demands or repair. It was created using a normal distribution with zero mean and a standard deviation fixed at 0.33 in this dataset. Figure 3 illustrates an example of the three components. Flow and pressure data were generated every 30 min and stored as .csv files. Leaks occurred randomly in terms of date, duration, types, positions, and size. A scenario may contain 0 to 2 leaks, incipient or abrupt, with a diameter varying in the range of [2 cm, 20 cm]. The leak can last from a few hours to several months.

### Machine learning algorithms

KNN is a classification algorithm categorized as a lazy learning algorithm (Zhang 2016). It directly classifies new input data according to their distances from previously classified data without building a predictive model. In our case, there are two classes, namely data related to normal consumption or abnormal data due to a leak. KNN calculates the distance and stores the KNN ( $k = 3$ ) for each sample of learning pressure or flow input, and then deduces the class of the new input by looking for the class which constitutes the majority neighbors of the latter.

SVM is an algorithm belonging to the category of linear classifiers. The process resides in creating a separation among the different input data class with a hyper-plane, called support vector, maximizing the margin between them. SVM plots each sample of pressure and flow data as a point in  $n$ -dimensional space. Then, it calculates and estimates the most plausible location of the hyper-plane between classes. Therefore, the algorithms are trained to learn to predict the class of the new input. SVM is based on its kernel function, which transforms the input vectors from low-dimensional space to a high-dimensional space. In this study, the radial basis function kernel is chosen from the best results reported in literature (Bernhard Schölkopf 2018).

MLP is a category of ANN which models a mathematical relationship between input training data and their output labels. MLP consists of input and output layers, in addition to one or more non-linear hidden layers. This study introduces an architecture based on one hidden layer with 100 units and a Rectified Linear activation function.

LR is a linear statistical algorithm based on the estimation of a linear relationship between input data and their output labels to predict the classification of test data. For the binary classification presented in this study, the sigmoid activation function in logistic regression was more appropriate (Rymarczyk *et al.* 2019).

### Metrics

Capacity of detection, false alarms rate, and accuracy were calculated in every node using the standard classification metrics of the confusion matrix: true positive rate (TPR), false positive rate (FPR), and accuracy (ACC):

$$TPR = \frac{TP}{(TP + FN)} \quad (1)$$

$$FPR = \frac{FP}{(FP + TN)} \quad (2)$$

$$ACC = \frac{(TP + FP)}{(FP + FN + TN + TP)} \quad (3)$$

TPR, also called sensitivity, measures the proportion of data labeled as leaks that are correctly identified. It demonstrates clearly the capacity of detection. FPR, also called specificity, measures the proportion of data labeled as non-leaks which are identified as leaks. It indicates the false alarms rate. ACC measures the proportion of correctly labeled data; where true positive (TP) indicates the number of correctly identified labeled leaky data; false positive (FP) corresponds to the number of leaky data incorrectly identified; false negative (FN) stands for the number of non-leaky data incorrectly identified, and true negative (TN) refers to the number of non-leaky data correctly identified.

## EMPIRICAL VALIDATION & RESULTS

Considering the big size of the dataset LeakDB and the similarity of events, this study did not use all scenarios. In fact, unrealistic scenarios, whose leaks last for a few hours or several months, were ignored. The set of selected scenarios includes the different categories of leaks and demands in each network. Two reduced datasets were defined. The first one consists of 54 leakage scenarios from Hanoi network, containing 34 incipient leaks and 35 abrupt leaks. The second consists of 13 scenarios from Net-1 network, containing nine incipient leaks and nine abrupt leaks. Tables 1 and 2 show all used scenarios



and the characteristics of their leaks, which are leak nodes, types, durations, and leak demands range in cubic meter per hour (CMH). The advocated choice ensures the existence of the two types of leaks with different sizes in all nodes positions, with a duration extending from 7 days to 8 months.

Subsequently, a series of implementation was applied using a kit of seven scenarios for training and three scenarios for testing from each reduced dataset. In this section, the undertaken experiments as well as the obtained results are shown in details.

### Leaky and non-leaky scenarios

In the first validation, all types of scenarios were tested. The 64 leaky scenarios were mixed with 40 non-leaky scenarios that were chosen randomly from the Hanoi network. The 22 leaky scenarios were mixed with 15 non-leaky scenarios that were chosen randomly from the Net-1 network. The mixture between leaky and non-leaky scenarios, and the use of different kits of scenarios instead of only one scenario, guaranteed the evaluation of different types of demands and leaks with different sizes. Table 3 summarizes the results of the artificial intelligence (AI) algorithms application for leak detection using pressures and flow data from Net-1 and Hanoi.

### Only leaky scenarios

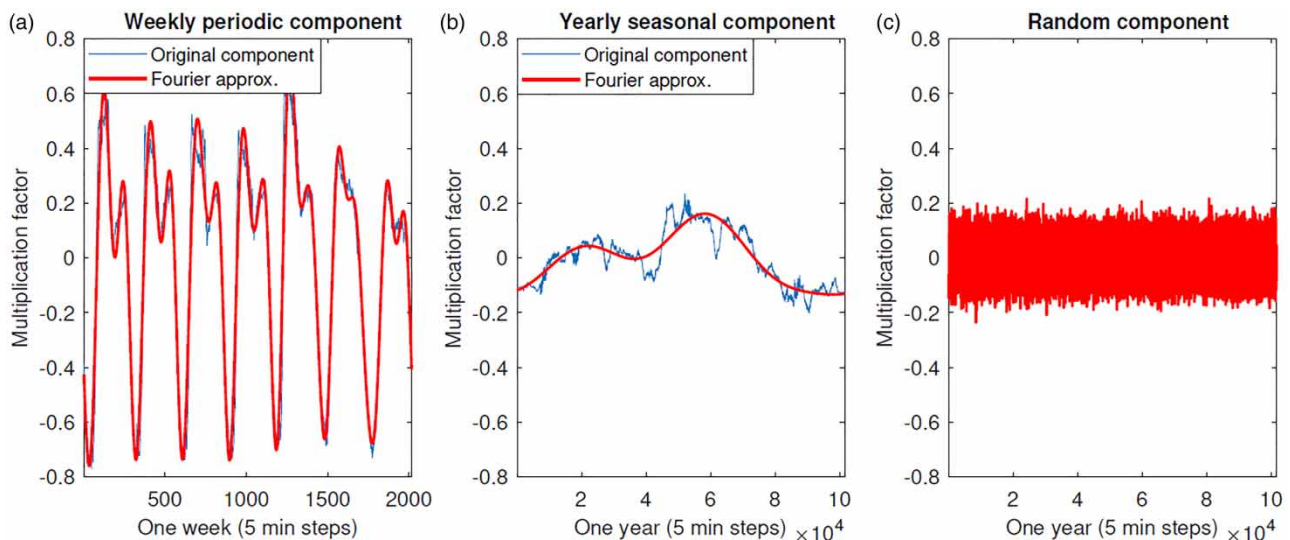
In this experiment, to evaluate the effect of labeled data on the leak detection learning, all non-leaky scenarios were eliminated, and the training and testing kits were constructed only with the reduced datasets of scenarios from Net-1 and Hanoi. The results are outlined in Table 4.

### Abrupt and incipient leaks

In water distribution networks, leaks are categorized into two types: abrupt leaks and incipient leaks. Figure 4 displays an example of the flow and pressure variation for the two types of leaks. According to Vrachimis *et al.* (2018), incipient leaks are small and increase gradually. They are the hardest to detect. However, abrupt leaks are large and start with peaks. This experiment (the third) assesses the performance of AI algorithms according to the type of leaks. Two types of kits are used. One contains only incipient leaky scenarios for training and testing, while the other contains only abrupt leaky scenarios. The results are shown in Tables 5 and 6.

### Night data

Among the methods used for improving leak detection performance, the overnight measurement of pressure and flow parameters minimizes disturbances related to the consumption. In this experiment (fourth experience), pressure and flow data from 00:00 to 04:00 over 1 year were used for each scenario in the train and test sets. Table 7 shows the results obtained for Net-1 and Hanoi.



**Figure 3** | The three-signal component of demands model indicating the original signal and Fourier approximation (Vrachimis *et al.* 2018).

**Table 1** | Dataset characteristics I

Scenarios	Leak information			
	Node	Type	Duration	Leak demand (CMH)
	<b>Hanoi network</b>			
Scenario 2	6	Incipient	17 days	[43.2; 46.8]
Scenario 3	19	Abrupt	36 days	[90; 93.6]
	26	Incipient	100 days	[86.4; 97.2]
Scenario 5	21	Incipient	124 days	[1,512; 1,911]
Scenario 6	18	Incipient	34 days	[600; 655]
Scenario 7	14	Abrupt	7 days	[1,537.2; 1,742.4]
Scenario 8	17	Incipient	61 days	[370.8; 486]
	21	Abrupt	74 days	[28.8; 36]
Scenario 10	12	Abrupt	5 days	[2,450; 2,750]
	28	Abrupt	150 days	[1,640; 1,890]
Scenario 11	12	Abrupt	28 days	[367.2; 396]
Scenario 16	3	Abrupt	76 days	[219; 237.6]
Scenario 22	2	Abrupt	36 days	[1,512; 1,519.2]
	9	Abrupt	45 days	[79; 90]
Scenario 101	11	Incipient	21 days	[838.8; 872]
	14	Abrupt	34 days	[1,760; 1,991]
Scenario 102	5	Abrupt	37 days	[2,200; 2,550]
	27	Abrupt	18 days	[1,270; 1,500]
Scenario 110	7	Incipient	10 days	[590; 630]
Scenario 111	4	Abrupt	79 days	[870; 990]
Scenario 124	3	Incipient	85 days	[470; 500]
	30	Incipient	61 days	[187.2; 219.6]
Scenario 127	29	Abrupt	73 days	[226.8; 244]
Scenario 128	13	Abrupt	17 days	[1,860; 2,090]
Scenario 133	17	Abrupt	155 days	[1,350; 1,440]
	26	Incipient	94 days	[1,700; 1,995.8]
Scenario 134	22	Incipient	69 days	[392.4; 424.8]
Scenario 144	6	Abrupt	117 days	[1,300; 1,515]
Scenario 151	4	Incipient	12 days	[2,000; 2,200]
Scenario 152	7	Abrupt	37 days	[799; 835]
Scenario 153	29	Incipient	141 days	[1,450; 1,800]
	31	Abrupt	36 days	[104.4; 111.6]
Scenario 156	22	Abrupt	33 days	[140.4; 151.2]
	29	Incipient	41 days	[72; 82.8]
Scenario 162	25	Abrupt	18 days	[266.4; 285]
Scenario 165	23	Abrupt	105 days	[331; 352]
Scenario 169	15	Incipient	18 days	[1,170; 1,371.6]
Scenario 172	3	Incipient	64 days	[594; 612]
	26	Abrupt	13 days	[460.8; 500.4]

*(Continued.)*

**Table 1** | Continued

Scenarios	Leak information			
	Node	Type	Duration	Leak demand (CMH)
	<b>Hanoi network</b>			
Scenario 175	20	Abrupt	69 days	[428.4; 483]
Scenario 193	10	Abrupt	106 days	[637.2; 690]
Scenario 208	18	Incipient	85 days	[1,360; 1,450]
Scenario 212	25	Incipient	20 days	[570; 630]
Scenario 213	14	Incipient	29 days	[720; 830]

### Seasonal data

To eliminate the disturbances due to seasonal changes in a year, in the fifth experiment the used pressure and flow data were taken from leaky scenarios for a period ranging from 1 to 2 months. The leaky scenario was divided into two parts, one for training and the other for testing. In the first test, the algorithms were applied for the whole node pressure and flow data in Net-1 and Hanoi networks. In the second test, algorithms were applied only for the leaky zone in the Hanoi network. The obtained results are shown in [Tables 8 and 9](#).

### Flow data difference

The study of LeakDB dataset by tracing flow and pressure data curves showed that in both cases when leaks were very small or the demands were unpredictable, the variations in pressure and flow during a leak were not remarkable. However, the difference between incoming and outgoing flows demonstrated large variations. [Figure 5](#) reports an example of flow/pressure data variations, and flows difference variation during a leakage period. The proposed experiment (the sixth) used the flow difference data to detect leaks. The differences between the incoming and outgoing flows were computed in each node of each chosen scenario from the Hanoi network. Some scenarios were used for training and others for testing. The results are shown in [Table 10](#).

## RESULTS DISCUSSION

In all experiments, the application of AI algorithms for leak detection in the Net-1 network provides better results compared to the Hanoi network. This can be accounted for through comparing the architectures and the length of the two networks. It is obvious that the disruption of data due to unforeseen demands in the Hanoi network is more intense than that in the Net-1 network. This also justifies the higher performance with the use of pressure data from Net-1 and flow data from Hanoi. Pressure data is the best parameter for leak detection, but it is the more sensitive to disturbance compared to flow data. This finding can be corroborated further by comparing the fourth experiment to the second one. When using night pressure and flow data, no improvement was recorded in performance compared to using pressure data from the Net-1 network. However, for the Hanoi network, the accuracy improved substantially, especially with pressure data. Furthermore, the four algorithms yielded almost the same ACC results for both experiments, except for LR which provided the minimum TPR, and for KNN which gave the maximum FPR.

The first and second experiments demonstrated that by using only leakage scenarios, the percentage of TPR decreased by 10% for all algorithms. When using leaky and non-leaky scenarios together for training and testing, the leak event became more distinguished by AI algorithms. In addition, the results of the third experiment were indicative that the type and size of leaks have the greatest effect on leak detection efficiency. Moreover, incipient leaks were more difficult to detect compared to abrupt leaks. When using abrupt leaky scenarios from Net-1, all TPRs for the different algorithms were greater than 90%. However, when using incipient leaky scenarios, TPRs did not exceed 45%. The TPR results of abrupt leaky scenarios from the Hanoi network were suggestive that SVM and LR are the most affected algorithms by noisy data.

With the fifth experiment, by eliminating the disturbances due to seasonal change and using the same scenario for testing and training, the performances for the different algorithms improved in the two networks compared to the second experiment. Indeed,



**Table 2** | Dataset characteristics II

Scenarios	Leak information			
	Node	Type	Duration	Leak demand (CMH)
<b>Hanoi network</b>				
Scenario 214	10	Incipient	32 days	[330; 400]
Scenario 233	11	Abrupt	184 days	[1,630.8; 1,900]
	31	Incipient	147 days	[810; 970]
Scenario 235	12	Incipient	97 days	[360; 432]
Scenario 258	13	Incipient	81 days	[1,400; 1,700]
Scenario 271	5	Incipient	82 days	[1,177; 1,230]
	30	Abrupt	61 days	[780; 870]
Scenario 285	8	Abrupt	136 days	[698; 760]
Scenario 293	2	Incipient	36 days	[1,447.2; 1,454.4]
Scenario 304	32	Incipient	11 days	[1,700; 1,950]
Scenario 327	20	Incipient	99 days	[1,370; 1,526.4]
Scenario 353	23	Incipient	130 days	[1,850; 2,000]
Scenario 369	8	Incipient	260 days	[495; 536.4]
Scenario 384	15	Abrupt	54 days	[1,375; 1,700]
Scenario 417	27	Incipient	145 days	[1,630.8; 1,990]
Scenario 420	18	Abrupt	26 days	[1,100; 1,200]
Scenario 485	11	Abrupt	86 days	[1,800; 2,150]
	32	Abrupt	131 days	[890; 1,100]
Scenario 611	16	Abrupt	58 days	[39.6; 43.2]
Scenario 612	19	Incipient	80 days	[480; 530]
Scenario 654	16	Incipient	61 days	[1,044; 1,200]
	20	Abrupt	18 days	[2,100; 2,380]
Scenario 813	9	Incipient	97 days	[68.4; 72]
Scenario 897	2	Abrupt	21 days	[205.2; 208.8]
	28	Incipient	19 days	[370; 440]
<b>NET-1 NETWORK</b>				
Scenario 2	31	Incipient	18 days	[320; 378]
Scenario 29	31	Abrupt	90 days	[150; 190]
Scenario 73	10	Abrupt	47 days	[61.2; 68.4]
	32	Incipient	51 days	[223.2; 201.6]
Scenario 74	2	Incipient	56 days	[560; 815]
Scenario 81	21	Incipient	46 days	[400; 1,050]
Scenario 91	21	Abrupt	16 days	[380; 900]
	31	Abrupt	121 days	[230; 350]
Scenario 163	12	Incipient	48 days	[158.4; 172.8]
	23	Abrupt	55 days	[410; 561]
Scenario 475	10	Incipient	125 days	[550; 1,400]
	22	Abrupt	13 days	[500; 1,480]
Scenario 492	12	Abrupt	14 days	[410; 1,200]
Scenario 498	2	Abrupt	10 days	[220; 280]

(Continued.)

**Table 2** | Continued

Scenarios	Leak information			
	Node	Type	Duration	Leak demand (CMH)
	<b>Hanoi network</b>			
Scenario 552	22	Incipient	13 days	[460; 1,330]
Scenario 722	13	Incipient	38 days	[320; 566]
	23	Incipient	16 days	[104.4; 118.8]
Scenario 829	13	Abrupt	57 days	[340; 840]

**Table 3** | Results of AI algorithms application using mixed scenarios

Net-1					Hanoi				
Data type	Algorithm	TPR %	FPR %	ACC %	Data type	Algorithm	TPR %	FPR %	ACC %
Pressures	KNN	79.09	0.28	97.45	Pressures	KNN	40.68	17.91	61.28
	SVM	71.72	0.1	96.47		SVM	16.33	2.21	73.19
	MLP	76.49	0.07	98.3		MLP	19.38	4.17	65.01
	LR	76.11	0.04	97.49		LR	17.59	3.6	65.77
Flow	KNN	36.5	1.92	90.2	Flow	KNN	37.54	14.55	61.38
	SVM	47.07	1.25	92.6		SVM	25.53	54.23	1.67
	MLP	42.11	1.52	90.98		MLP	27.17	5.38	60.73
	LR	30.17	0.65	91.9		LR	19.97	0.87	59.35

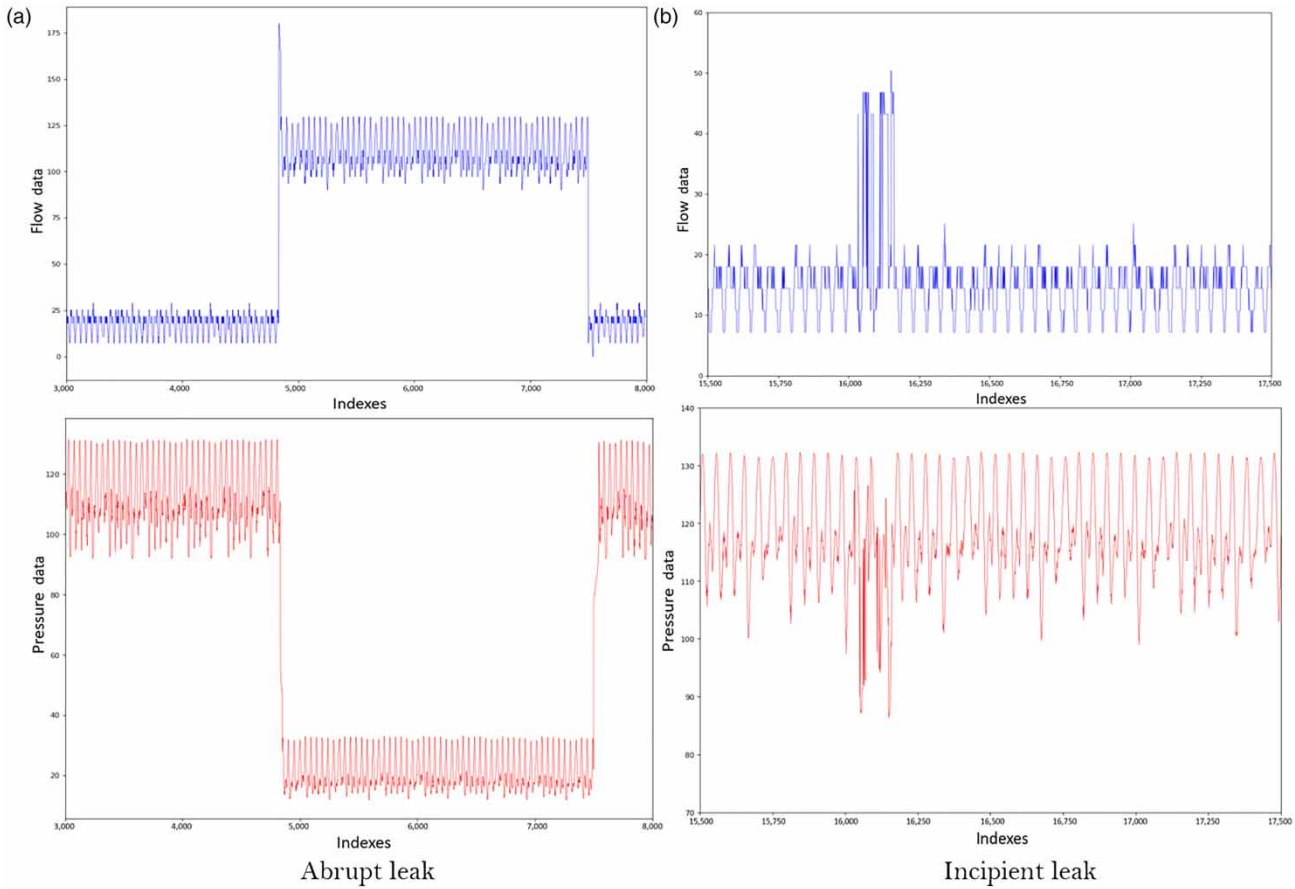
**Table 4** | Results of AI algorithms application using only leaky scenarios

Net-1					Hanoi				
Data type	Algorithm	TPR %	FPR %	ACC %	Data type	Algorithm	TPR %	FPR %	ACC %
Pressures	KNN	67.21	3.35	92.73	Pressures	KNN	18.27	10.93	57.67
	SVM	66.32	0.36	95.2		SVM	9.23	0.43	55.64
	MLP	61.32	0.32	94.56		MLP	5.44	0.29	57.89
	LR	60.38	0.03	94.69		LR	4.65	0.16	57.62
Flow	KNN	50.13	3.35	90.45	Flow	KNN	34.78	19.61	60.16
	SVM	48.13	0.32	92.81		SVM	22.11	2.51	62.13
	MLP	65.37	0.82	92.29		MLP	27.29	8.6	62.97
	LR	17.75	0.01	89.03		LR	11.6	0.64	60.43

better accuracy, more true labeled detection, and less false alarm were reported especially with SVM. Additionally, performances were improved in the Hanoi network when the algorithms were applied only in the leak areas.

In the sixth experiment, a new parameter was defined for leak detection, which is the flow difference. The proposed experiment yielded different results according to the test and train scenarios. As portrayed in Table 8, the best performance was achieved in scenario 111 with 100% TPR, 100% ACC, and 0% FPR for KNN, LR, and MLP, but for other scenarios, almost 50% ACC and 75% FPR were obtained. Furthermore, the variation triggered by the leak can be observed only on the leaky node when computing the flow difference between the incoming and the outgoing flow. The obtained results for the two last experiments can help identify leak localization using AI algorithms in future research.

It is noteworthy that the classification algorithms KNN and SVM provide the best performances compared to LR and MLP algorithms. Moreover, using night data and seasonal data KNN offers the highest TPR, but also the highest rate of false alarms, which makes SVM results more efficient and more reliable.



**Figure 4** | Variation of flow and pressure data for the two types of leaks: (a) abrupt leak in node 23, scenario 722, Net-1; and (b) incipient leak in the same node, scenario 163, Net-1.

**Table 5** | Results of AI algorithms application using abrupt leaky scenarios

Net-1					Hanoi				
Data type	Algorithm	TPR %	FPR %	ACC %	Data type	Algorithm	TPR %	FPR %	ACC %
Pressures	KNN	89.91	0.16	97.83	Pressures	KNN	30.72	12.82	82.05
	SVM	90.73	0.13	98.02		SVM	4.55	1.21	90.22
	MLP	90.64	0.19	97.95		MLP	12.42	0.11	77.78
	LR	90.18	0.07	97.96		LR	19.82	0.52	92.24
Flow	KNN	66.38	3.57	90.34	Flow	KNN	23.58	10.14	83.84
	SVM	68.62	2.74	91.46		SVM	18.0	1.15	91.5
	MLP	66.13	2.63	91.05		MLP	22.34	9.85	83.98
	LR	56.37	1.8	89.73		LR	6.45	0.0	91.49

## CONCLUSION

This paper examines an experimental comparative study of the most-used AI algorithms for leak detection and location in water distribution networks. Different experiments were carried out using the LeakDB dataset. Departing from the obtained results, it is deduced that the types of data and the choices of the training sets are significant in terms of enhancing performances in detecting different leakage data (TPR), reducing false alarms (FPR), and improving the accuracy (ACC).

**Table 6** | Results of AI algorithms application using incipient leaky scenarios

Net-1					Hanoi				
Data type	Algorithm	TPR %	FPR %	ACC %	Data type	Algorithm	TPR %	FPR %	ACC %
Pressures	KNN	44.8	1.55	91.72	Pressures	KNN	17.2	18.19	76.68
	SVM	42.64	0.07	92.75		SVM	4.55	1.21	90.22
	MLP	31.18	0.01	91.35		MLP	0.15	0.38	91.74
	LR	23.59	0.0	90.41		LR	0.15	0.04	91.94
Flow	KNN	33.08	5.83	86.51	Flow	KNN	18.88	19.5	75.62
	SVM	30.61	0.25	91.08		SVM	0.64	0.12	91.55
	MLP	22.08	0.77	89.55		MLP	13.61	14.67	79.64
	LR	10.48	0.01	88.76		LR	0.15	0.0	92.08

**Table 7** | Results of AI algorithms application using night data

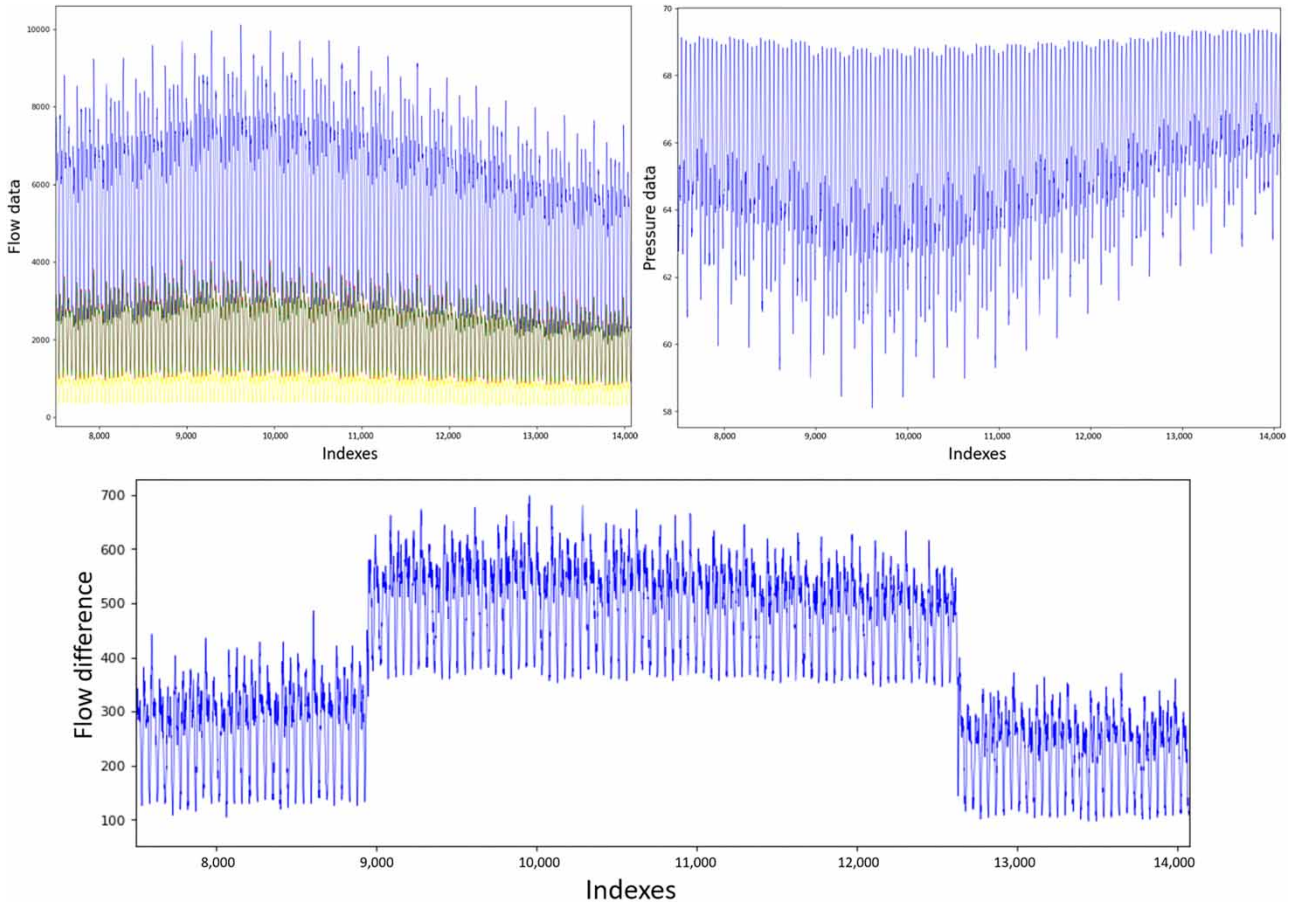
Net-1					Hanoi				
Data type	Algorithm	TPR %	FPR %	ACC %	Data type	Algorithm	TPR %	FPR %	ACC %
Pressures	KNN	67.82	2.99	93.13	Pressures	KNN	27.84	7.91	63.51
	SVM	67.5	0.48	95.27		SVM	19.17	0.02	64.04
	MLP	67.26	0.44	95.27		MLP	11.52	0.11	60.59
	LR	67.21	0.44	95.2		LR	11.22	0.0	60.52
Flow	KNN	62.8	1.69	93.5	Flow	KNN	37.47	18.34	62.01
	SVM	63.32	0.9	94.35		SVM	29.61	1.28	67.99
	MLP	63.73	0.44	94.8		MLP	34.9	10.69	65.11
	LR	48.95	0.23	93.02		LR	17.88	1.0	62.92

**Table 8** | Results of AI algorithms application using seasonal data from Hanoi network

Hanoi					Whole network				
Leak area	Algorithm	TPR %	FPR %	ACC %	Data type	Algorithm	TPR %	FPR %	ACC %
Pressures	KNN	72.46	24.76	74.6	Pressures	KNN	57.22	32.66	65.01
	SVM	62.89	0.00	91.46		SVM	41.11	6.1	81.76
	MLP	63.18	6.40	86.6		MLP	4.59	0.0	78.06
	LR	60.89	1.86	89.8		LR	16.68	0.61	80.37
Flow	KNN	72.90	22.55	76.4	Flow	KNN	58.92	31.54	66.27
	SVM	62.91	0.00	91.47		SVM	39.97	2.85	84.0
	MLP	62.89	1.9	90.01		MLP	31.39	16.03	71.88
	LR	62.99	0.1	91.4		LR	39.28	5.55	81.76

**Table 9** | Results of AI algorithms application using seasonal data from Net-1 network

Net-1					Net-1				
Network	Algorithm	TPR %	FPR %	ACC %	Data type	Algorithm	TPR %	FPR %	ACC %
Pressures	KNN	99.97	21.01	85.81	Flow	KNN	72.46	12.42	82.6
	SVM	90.87	2.68	95.23		SVM	71.8	2.96	88.8
	MLP	89.76	11.36	89.0		MLP	72.37	9.37	84.69
	LR	90.84	6.78	92.44		LR	69.11	10.33	82.99



**Figure 5** | Variation of flow data during leak period on pipe 2, pipe 19, pipe 3, and pipe 20, pressure data, and flow difference data on node 3 in scenario 16, Hanoi network.

**Table 10** | Results of AI algorithms application using flow data difference

Data type	Flow difference				Flow difference	Flow difference			
	Algorithm	TPR %	FPR %	ACC %		Algorithm	TPR %	FPR %	ACC %
Scenario 111	KNN	100.0	0.00	100.0	Scenario 162	KNN	100.0	73.04	48.09
	SVM	36.54	26.3	58.7		SVM	27.37	73.04	27.07
	MLP	100.0	0.00	100.0		MLP	100.0	75.87	46.08
	LR	100.0	0.00	100.0		LR	100.0	75.87	46.08

As a supervised comparison study, the major problem with the LeakDB dataset is defining the same label for each scenario for all the network nodes. This may be plausible for a large leak that disturbs the whole distribution systems; however, for small or incipient leaks, the nodes in the leak area are individually affected and by different degrees in large networks like Hanoi.

This work, to our knowledge, is the first to use a common benchmark dataset for different AI algorithms. It offers a clear comparison of the behavior of different algorithms in leak detection using flow and pressure data, and highlights the role of choosing learning data in terms of enhancing detection performance.

The obtained observations and results can shed light on the behavior of supervised learning techniques, allowing deeper insight to better explore leak detection applications. This research work is promising and valuable in terms of opening further lines of investigation and paving the way for constructive future research directions. This study can be taken further because it



lays the ground for future research into novel solutions to resolve the issue of leak detection and localization through the use of other state-of-the-art unsupervised techniques and deep-learning architectures.

## ACKNOWLEDGEMENT

This research work was accomplished in collaboration between Sofia Technologies Company and CES-laboratory in the National Engineering School of Sfax. This project was carried out under the MOBIDOC scheme, funded by the EU through the EMORI program, and managed by the ANPR.

## DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories. <https://goo.gl/zLJpuD>.

## REFERENCES

- Adedeji, K., Hamam, Y., Abe, B. & Abu-Mahfouz, A. M. 2017 leakage detection algorithm integrating water distribution networks hydraulic model. SimHydro 2017 Conference: Choosing the right model in applied hydraulics, Sophia Antipolis, Nice, France.
- Ahmed, M., Mahmood, A. N. & Hu, J. 2016 A survey of network anomaly detection techniques. *Journal of Network and Computer Applications* **60**, 19–31.
- Ayadi, A., Ghorbel, O., BenSalah, M. S. & Abid, M. 2019 Kernelized technique for outliers detection to monitoring water pipeline based on WSNS. *Computer Networks* **150**, 179–189.
- Bernhard Schölkopf, A. J. S. 2018 *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press Cambridge, MA, USA.
- Bin Md Akib, A., Bin Saad, N. & Asirvadam, V. 2011 Pressure point analysis for the early detection system. In: *7th Ieee International Symposium 2011 on Signal Processing and its Applications*, pp. 103–107.
- Bjerke, M. 2019 *leak detection in water distribution networks using gated recurrent neural networks*. Unpublished master's thesis, Norwegian University of Science and Technology, Trondheim, Norway.
- Caputo, A. C. & Pelagagge, P. M. 2002 An inverse approach for piping networks monitoring. *Journal of Loss Prevention in the Process Industries* **15** (6), 497–505.
- Chen, H., Ye, H., Chen, L. & Su, H. 2004 Application of support vector machine learning to leak detection and location in pipelines. In: *Proceedings of the 21st IEEE Instrumentation and Measurement Technology Conference (IEEE cat. no. 04ch37510)*, Vol. 3, pp. 2273–2277.
- Cheung, P., Van Zyl, J. & Reis, L. 2005 Extension of epanet for pressure driven demand modeling in water distribution system. *Computing and Control for the Water Industry* **1**, 311–316.
- Eliades, D. & Polycarpou, M. M. 2012 Leakage fault detection in district metered areas of water distribution systems. *Journal of Hydroinformatics* **14** (4), 992–1005.
- Kanakoudis, V. & Tolikas, D. 2001 The role of leaks and breaks in water networks: technical and economical solutions. *Journal of Water Supply: Research and Technology – AQUA* **50** (5), 301–311.
- Kanakoudis, V., Tsitsifli, S., Cerk, M., Banovec, P., Samaras, P. & Zouboulis, A. I. 2015 Basic principles of a DSS tool developed to prioritize NRW reduction measures in water pipe networks. *Water Quality, Exposure and Health* **7** (1), 39–51.
- Kang, J., Park, Y.-J., Lee, J., Wang, S.-H. & Eom, D.-S. 2017 Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems. *IEEE Transactions on Industrial Electronics* **65** (5), 4279–4289.
- Kayaalp, F., Zengin, A., Kara, R. & Zavrak, S. 2017 Leakage detection and localization on water transportation pipelines: a multi-label classification approach. *Neural Computing and Applications* **28** (10), 2905–2914.
- Klise, K. A., Bynum, M., Moriarty, D. & Murray, R. 2017 A software framework for assessing the resilience of drinking water systems to disasters with an example earthquake case study. *Environmental Modelling & Software* **95**, 420–431.
- Mounce, S. R., Mounce, R. B. & Boxall, J. B. 2011 Novelty detection for time series data analysis in water distribution systems using support vector machines. *Journal of Hydroinformatics* **13** (4), 672–686.
- Oliveira, E., Fonseca, M., Kappes, D., Medeiros, A. & Stefanini, I. 2018 Leak detection system using machine learning techniques. In: *6th International Congress on Automation in Mining*, Santiago, Chile.
- Rymarczyk, T., Kozłowski, E., Kłosowski, G. & Niderla, K. 2019 Logistic regression for machine learning in process tomography. *Sensors* **19** (15), 3400.
- Srirangarajan, S., Allen, M., Preis, A., Iqbal, M., Lim, H. B. & Whittle, A. J. 2013 Detection and localization of wavelet-based burst events in water distribution networks. *Journal of Signal Processing Systems* **72** (1), 1–16.
- Vrachimis, S. G., Kyriakou, M. S., Eliades, D. G. & Polycarpou, M. M. 2018 LeakDB: A benchmark dataset for leakage diagnosis in water distribution networks. In: *WDSA/CCWI Joint Conference Proceedings*, Vol. 1.
- Wu, Y. & Liu, S. 2017 A review of data-driven approaches for burst detection in water distribution systems. *Urban Water Journal* **14** (9), 972–983.

- Yamauchi, H. & Huang, W.-y. 1977 Alternative models for estimating the time series components of water consumption data 1. *Journal of the American Water Resources Association* **13** (3), 599–610.
- Zhang, Z. 2016 Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine* **4** (11), 218–229.
- Zhou, B., Lau, V. & Wang, X. 2019 Machine-learning-based leakage-event identification for smart water supply systems. *IEEE Internet of Things Journal* **7** (3), 2277–2292.

First received 26 March 2021; accepted in revised form 19 July 2021. Available online 17 August 2021