buildings

MDPI

Article

Application of Machine Learning for Leak Localization in Water Supply Networks

Abdul-Mugis Yussif [1,*], Haleh Sadeghi [2,*] and Tarek Zayed [1]

[1] Department of Building and Real Estate (BRE), Faculty of Construction and Environment (FCE), The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
[2] Department of Mechanical, Aerospace and Civil Engineering, School of Engineering, The University of Manchester, Manchester M13 9PL, UK
* Correspondence: mugis.yussif@connect.polyu.hk (A.-M.Y.); hsadeghiaa@connect.ust.hk (H.S.)

Abstract: Water distribution networks (WDNs) in urban areas are predominantly underground for seamless freshwater transmission. As a result, monitoring their health is often complicated, requiring expensive equipment and methodologies. This study proposes a low-cost approach to locating leakages in WDNs in an urban setting, leveraging acoustic signal behavior and machine learning. An inexpensive noise logger was used to collect acoustic signals from the water mains. The signals underwent empirical mode decomposition, feature extraction, and denoising to separate pure leak signals from background noises. Two regression machine learning algorithms, support vector machines (SVM) and ensemble k-nearest neighbors (k-NN), were then employed to predict the leak's location using the features as input. The SVM achieved a validation accuracy of 82.50%, while the k-NN achieved 83.75%. Since the study proposes using single noise loggers, classification k-NN and decision trees (DTs) were used to predict the leak's direction. The k-NN performed better than the DT, with a validation accuracy of 97.50%, while the latter achieved 78.75%. The models are able to predict leak locations in water mains in urban settings, as the study was conducted in a similar setting.

Keywords: leak localization; noise loggers; water distribution networks; machine learning; acoustic sensors

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

Academic Editor: Irem Dikmen

Received: 9 February 2023
Revised: 20 March 2023
Accepted: 21 March 2023
Published: 24 March 2023

1. Introduction

Based on recent statistics, the total coverage of constructed pipelines in 120 countries is approximately 2,175,000 miles, of which 65%, 8%, and 3% of the whole length of pipelines belong to the US, Russia, and Canada, respectively. Additionally, some other statistics show that since 2010, 27% of the world's population has been living in water-scarce areas, and it is predicted that this amount will go beyond 42% in 2050 [1]. Implementing effective management strategies should improve the current water distribution networks (WDNs). This improvement will help solve the global water scarcity problems and minimize non-revenue water (NRW). It will also strengthen the significant roles of pipelines in water collection, transportation, and distribution [2,3].

Considering the above, the NRW percentage goes beyond 50% for old WDNs. However, this amount is approximately 30% in most WDNs [4]. Notably, the main source of NRW is leakage in the water pipelines since it comprises 70% of the sources of NRW [5]. The leakage problems can lead to wasting natural resources and money and threaten public health as contaminants can enter the water networks through the leaks. The development of leak detection and localization technologies in WDNs is essential.

Approximately 50% of the WDNs in Hong Kong were constructed 40 years ago, resulting in the aging and degradation of pipes and, consequently, water loss through leaks [6]. The annual financial loss from pipeline leakages is approximately 173 million USD; therefore, the respective departments should take action to control it [7]. Although the water supplies

department (WSD) has started the rehabilitation and replacement scheme to improve the WDSs and prevent water loss through the pipelines, the water network in this country still needs further improvements since the number of reported leaks in pipelines was approximately 8512 in 2017. With this in mind, improving the leak detection and localization approaches is urgent for the concerned municipalities to minimize the damages. Therefore, several techniques have been utilized so far for leak localization on WDNs. However, using acoustic sensors for leak localization is more reliable and less costly [8]. The acoustic sensors that measure the acoustic waves in the pipes are classified into several groups, such as traditional accelerometers, fiber optic sensors, magnetostriction sensors, pulsed lasers, and noise loggers. Due to the high sensitivity of fiber optic sensors, magnetostriction sensors, and pulsed lasers, they are popular in laboratory environments. However, since noise loggers have a water-resistant design and wireless communication ability, they are more widely used on-site [7,8]. Additionally, based on the conducted studies in this area, the current literature is imbued with the following shortcomings:

1. There is a lack of studies that utilize one noise logger for leak localization on real water networks, which is less costly and time-consuming [9,10]. Most of the studies in this area have utilized two or more noise loggers to localize the leaks in WDNs, which increases the project's total cost and processing time.
2. Additionally, delving into the literature revealed that utilizing ML-based techniques for leak localization on real WDNs has not achieved significant success yet. Most recent studies have been conducted in laboratories under controlled conditions.

Considering the mentioned gaps, the objectives of this study are as follows:

1. To utilize one noise logger for localizing leaks on real water networks since the time and cost of the project can be reduced considerably.
2. From objective (1), develop ML-based techniques for leak localization in real WDNs in Hong Kong.

The primary goal of this study is to develop ML-based techniques for leak localization in real WDNs by utilizing one noise logger in Hong Kong.

Background

Conducting a literature review is essential to capture the current breakthroughs in the research industry [11]. Different steps are required for conducting a comprehensive literature review, as follows: (1) conducting a comprehensive bibliometric by searching the popular primary databases, like Scopus, Web of Science, and Google Scholar [10]; (2) conducting snowballing to increase the number of papers [12]; (3) filtering the unrelated papers using different filters [13]; and (4) conducting a quantitative and qualitative analysis [14].

For localizing the leaks in WDNs and overcoming the leakage issues in the pipelines, different leak detection and localization techniques have been considered by various researchers around the world. Additionally, several techniques have been utilized for leakage detection, such as acoustic leak detection, optical fiber acoustic sensing, analysis of the pressure point, ultrasound leak detection, infrared thermography, and electromagnetic methods [15–25]. Since acoustic sensors are inexpensive and highly reliable, the acoustic leak detection technique is one of the most reliable techniques for leak detection and localization [8]. The signal behavior can easily be studied under several factors, and the acoustic equipment is usually mechanically stable. This stability grants it the potential to be helpful in a wide range of environmental and physical conditions. Moreover, operating the noise loggers does not require demanding training because using them is not sophisticated. One of the acoustic-based techniques is cross-correlation, the most used technique for leak detection and localization. Although its success is remarkable, it might cause false alarms in the city networks due to background noises and water usage. Their high expenses are another drawback to consider [8]. Notably, leakage detection is comprised of two steps: detection of the leak, where the

sensors are deployed in the WDSs for acquiring the acoustic signals and predicting their states, and localization of the leaks, where their locations are pinpointed in the WDNs. Either one sensor or multiple sensors can be utilized for leak localization. Some studies suggest that using two sensors is more reliable than one. However, there are some limitations to using several sensors, such as suitable environments and financial issues [10]. For instance, Ref. [9] proposed new technology for localizing downhole tubing leaks using only one sensor. As mentioned by [26], the position of leaks can be localized in the WDNs using a listening stick for listening to each service connection, step testing, or noise loggers. Notably, using noise loggers for localizing leaks on the WDNs is preferable because step testing and listening stick surveys require night work and are time-consuming. Whereas noise loggers can be configured and deployed to record the acoustic signals on their own at any time.

Additionally, in studies where more than one feature is considered for leak localization, there is a need to consider sophisticated techniques for understanding the relationship between the leakage location and several features. Researchers have implemented machine learning (ML)-based models, such as artificial neural networks, k-nearest neighbors, deep neural networks, support vector machines, naïve Bayes, decision trees, and random forests, for detecting and localizing laboratory simulated leaks [25–31]. Another study proposed a leak identification technique based on transient frequency response and deep learning. Noise loggers are generally positioned at the meeting point of pipes during data collection for leak detection. Therefore, using a single noise logger without knowing the pipeline on which the leak is located poses a massive problem during repairs. The questions to answer here are: (1) Where is the exact location of the leak? (2) Which side of the valve or connection (location of the noise logger) is appropriate for excavation? (3) What is the most reliable method for determining the exact location of the leak? This research explores the solutions to these questions and presents models that tell in which direction excavation of the soil will be required, pivoting to the position of the noise logger. In this case, it is represented as "Right" or "Left". Theoretically, any of these labels could be in the water flow direction or against it. The possibility of the disturbance from the interaction between the acoustic leak waves and water flow waves is applied in this study. However, it should be noted that the actual properties of the waves are not studied. Instead, the effects of this interaction during feature extraction are utilized. The idea was motivated by the following findings: Mahmutoglu and Turk [32] used the receiver signal strength technique with two receivers to determine the location and direction of the leak point. The leak's distance was computed with one receiver's data, while the second receiver was used to estimate the direction. According to the authors, using multiple receivers to determine the leak distance increases simulation errors. Furthermore, a study indicated that when the water leaves the leak holes with turbulence, the vibrations transmit acoustic signals through the pipe walls and into the water [33]. The propagation of waves in water and other fluids is a complicated phenomenon. Still, it can be understood to occur through the water core and the pipe wall simultaneously, as if the two media act as a couple. Hence, the attenuation of the waves in both the water core and the pipe walls agrees to co-occur [34]. Therefore, the propagation of the acoustic leak signal is affected by the water flow and the pipe.

Considering the review of the current literature, this study aims to grapple with the following gaps:

- The current literature lacks a study that considers the application of ML-based techniques for localizing leaks in real water networks. Most studies in this area are conducted in labs, which lack the conditions WDNs are subjected to in the field.
- Most studies in this area have also used multiple noise loggers for localizing leaks in water networks, which have two main shortcomings.
    (i) A significant increase in the total expenses incurred by the project;

(ii) Increased time spent in data processing from multiple acoustic devices due to correlation.

Therefore, the current body of literature lacks a study that utilizes one noise logger for leak localization on real water networks, which is less costly and time-consuming.

The rest of the paper is organized as follows: Section 2 explains the study research methodology in detail, including the data acquisition, the theoretical background of the adopted signal decomposition technique, and the feature selection methods. The ML models' development and their optimized parameters are also included in this section. Section 3 presents the performances and discussions of the developed models. The effectiveness of the feature selection method (PCA) used in the study are highlighted in this section. Finally, Section 4 summarizes the achieved study objectives and proposes relevant future research directions in the study domain.

2. Methodology

Figure 1 illustrates the detailed research framework of this study, comprising four phases: (1) conducting a comprehensive literature review, (2) acoustic signal acquisition and leak detection, (3) signal processing and data preparation, and (4) leak localization. Regarding the first phase, a literature review of noise-logger-based studies and machine learning-based techniques was conducted. This review was conducted to ascertain the need for the study. In the second phase, the noise loggers were deployed at midnight to acquire acoustic signals. It took 12 months to acquire enough acoustic signals for the study. The third phase, analyzing the dataset, was conducted simultaneously with the second phase and was concerned with signal processing, feature extraction, and data preparation. In the fourth phase, several ML-based techniques were applied to calculate the leak distance and predict the direction of the leak.

The concept utilizes the features extracted from the acoustic signals generated and recorded with noise loggers connected to the WDNs. The detailed explanations of the different phases (see Figure 1) involved in this study are as follows:

2.1. PHASE 1: Conducting a Comprehensive Literature Review

The first phase reviewed the noise logger-based studies and machine learning-based techniques. A detailed systematic review was conducted to find suitable techniques for leak localization in WDNs. Several steps were taken to conduct a systematic review. As the initial step, a comprehensive bibliometric search, including the key terms, was conducted in three main databases: Scopus, Web of Science, and Google Scholar. After which, to increase the number of papers, snowballing was conducted. The next step is the filtration of irrelevant articles using several filters, such as those published in peer-reviewed journals or the past fifteen years. After shortlisting the appropriate papers, a quantitative and qualitative analysis was conducted.

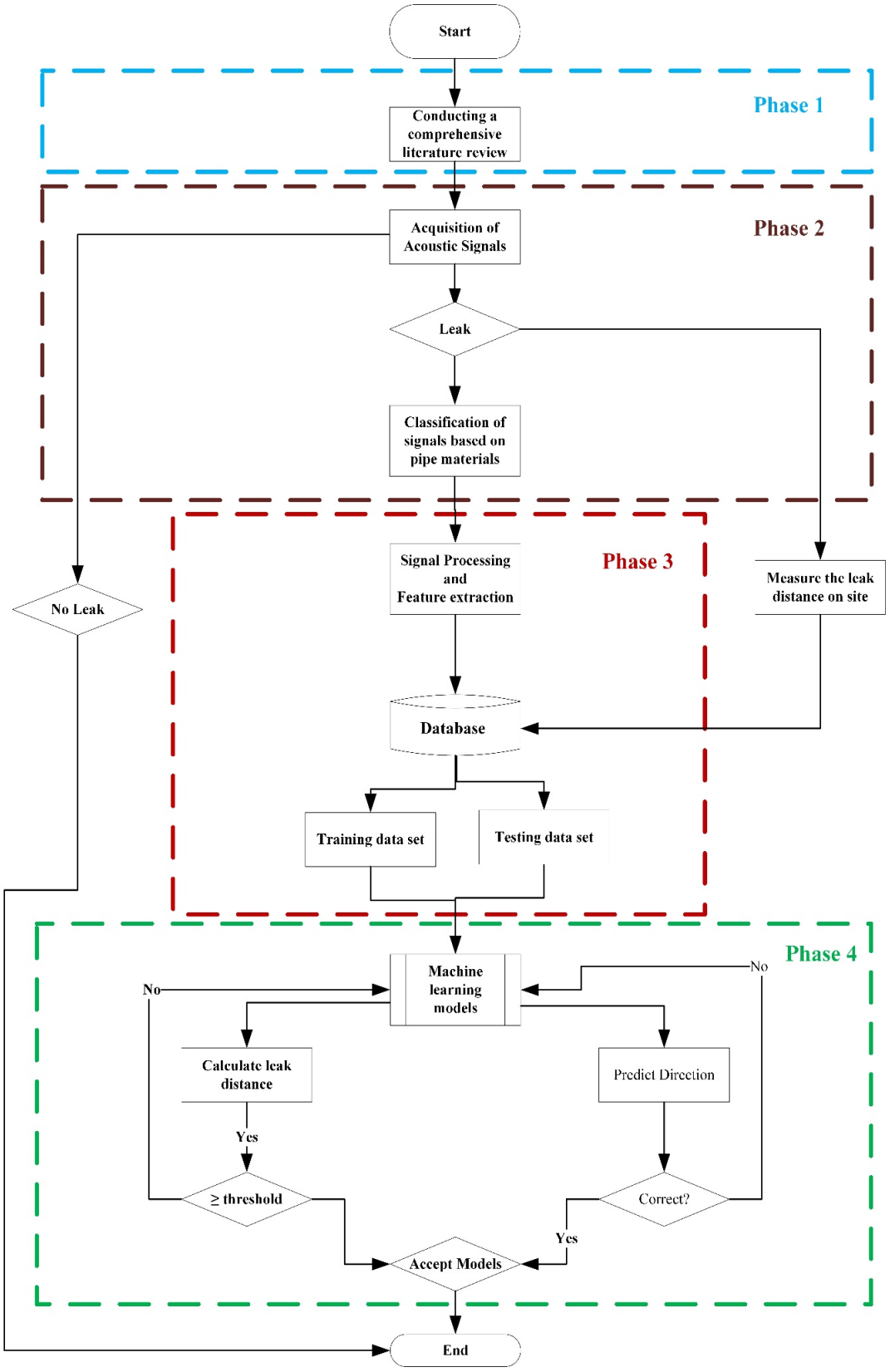

Figure 1. Research Framework.

2.2. PHASE 2: Signal Acquisition and Leak Detection

Most leak detection and localization studies were conducted in laboratories under controlled conditions. However, localization of leakages in the WDNs in the cities could be challenging since the background noise, and other ambient conditions are difficult to control in real WDNs that supply many households and companies. The Hong Kong Water Supplies Department (WSD) provided information about suspected leaks and their location maps to the research team. The research team then deployed the noise loggers and configured them for time interval midnight recordings. The loggers automatically record the signals and transfer them to the cloud for convenient recovery. They were configured to take four (4) recordings every night from 3:30 a.m. to 4:15 a.m. at 15-min intervals. The choice of operation time is to minimize the interference of noise from the surrounding environment [27] in recording the actual acoustic leak signal. It is, however, necessary to note that the distance from which noise loggers can detect acoustic signals is determined by physical characteristics, such as pipe material, thickness, diameter, etc. The signals travel farther in narrow and metallic pipes than wider pipes composed of other materials, such as asbestos. Ideally, the signals could travel approximately 300 m maximum in narrow ductile pipelines [26].

A sampling frequency of 4096 Hz was used with a recording length of 10 s for acquiring acoustic vibrations each time. The frequency of the leak signal in the WDN typically ranges from a few hundred Hz to a few kHz [27]. According to the Nyquist–Shannon sampling theorem, the sampling rate of the equipment must be at least twice that of the highest frequency component of the signal to capture its content accurately. Moreover, a very high sampling rate was avoided to prevent a larger dataset generation, which would cause storage burdens. Further, the distance of the proposed leak location from the sensor is measured for localization model development. The deployed noise loggers were left on the site for approximately 3–5 days to ensure that the data obtained was devoid of significant biases. The recorded signals were downloaded from the cloud at the end of the deployment period. It took the research team 12 months to acquire sufficient data to develop and validate the localization models. A discrete wavelet transform was applied to the collected signals to study their frequency patterns and changes further. The signals then underwent feature extraction to obtain numeric data formats for the machine learning techniques application. These features are level, root-mean-square (RMS), spread, frequency spread, kurtosis, autocorrelation kurtosis, skewness, maximum amplitude, peak amplitude, time-domain average amplitude, frequency-domain average amplitude, peak frequency, frequency centroid, crest factor, energy, maximum Lyapunov exponent (MLE), and autocorrelation MLE. Refer to [8,28] for further information on features and leak detection. More information about the physical characteristics of the features and their derivations is also found in those studies. The acoustic signals from the no-leak cases were separated, and only those from the leak cases were used in the subsequent stages of localization.

It was required to study the behavior of the signals generated in a non-leak condition to reduce the influence of background noise on the pure leak signal. Due to this, the team deployed noise loggers to record signals at locations in the WDNs that were initially identified as leak points but subsequently repaired by the water supplies department. In this process, the signal properties from the non-leak conditions were studied and used to reinforce the application of the machine learning models used in the study.

2.2.1. Nature of the Deployment of Noise Loggers for Data Collection

Diverse leak sizes for different deployment distances were used to collect signals for the training dataset. Further, three (3) different leak sizes, i.e., large, medium, and small, were used to validate the developed localization models. The first of the noise loggers was 3.556 m away from the large leak, the second was 0.9 m away from the medium-sized leak, and the third was 1.314 m away from the small-sized leak. These distances were not pre-simulated for data acquisition since it was field data. Distances here refer to the stretch

between the leak spot and the pipeline valve through which the researchers could access the pipelines. The noise logger deployment is shown in Figure 2.

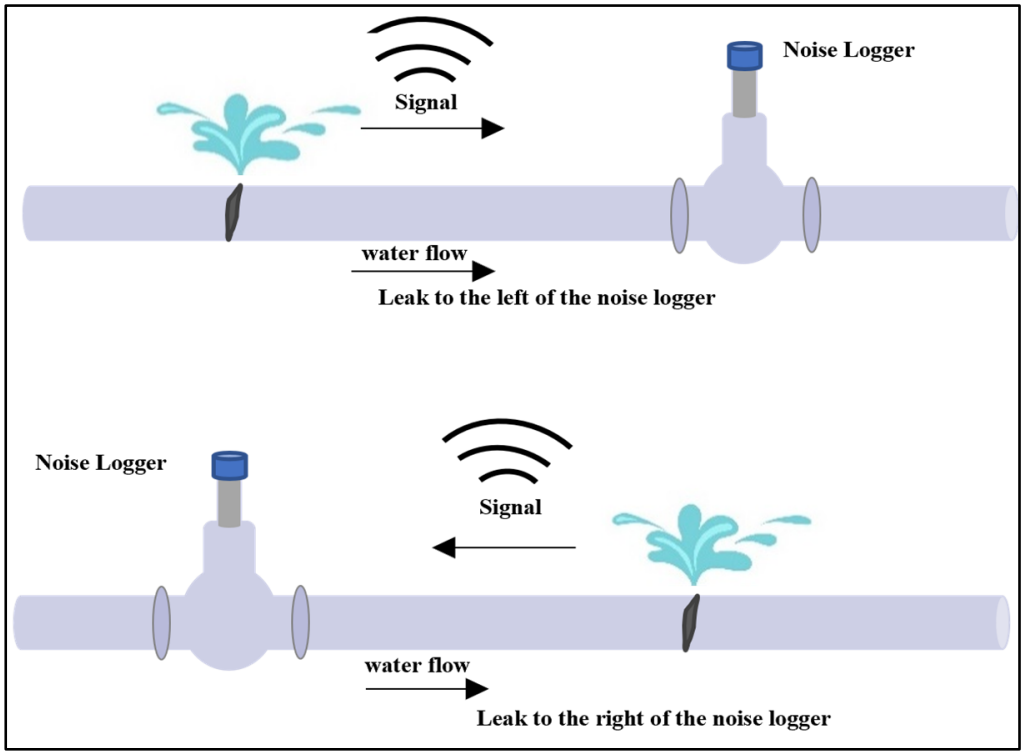

Figure 2. Leak direction to the Right and Left of the noise logger.

2.3. PHASE 3: Signal Processing and Feature Extraction

2.3.1. Signal Processing

Working with these signals requires prior knowledge of their behavior in a leak or non-leak situation. Signal processing has been an integral part of this study since everything is based on the behavior of the signals. It is imperative to obtain a clear pattern for the signals for feature extraction. Tijani et al. [8] used correlation analysis to show how significantly better data generated from processed signals performed than raw data. This study was conducted with the same dataset used in this study.

Several signal processing techniques have previously been used to denoise leak signals from acoustic noise loggers. Among these techniques are linear prediction (LP), singular value decomposition (SVD), dual-tree complex wavelet transform (DTCWT), empirical mode decomposition (EMD), a method based on the received signal strength (RSS) [10,28], variational mode decomposition (VMD), [29] and fast Fourier transform (FFT). FFT is implemented to compute the discrete Fourier transform since its direct computation is inefficient [30]. The RSS method was proposed by [31] to denoise signals for leak detection and localization. The authors went further with the application of the method to determine the sizes of the leaks, which will be of great importance should it successfully differentiate between small and large leak sizes. With this method, they indicated that distant leakages could be located with minimal average errors. The main drawback of this approach is the requirement of prior knowledge about the emanating area of the strength and the wavelet propagation. A study by Liu et al. [29] used a denoising technique based on VMD for leak localization in water supply pipes. The method achieved an average accuracy of 93.58%, which is a good performance. LP is efficient in analyzing wavelets that are propagated from sounds and voices. It has largely been used in speech recognition and speech synthesis. The technique requires simple computational demands in its application, coupled with its efficiency in sound

signal processing. It has been used to extract the significant features of a leak signal recorded by acoustic noise loggers. The method works on the principle that the resonant frequencies of a linear system can be captured using discrete-time signal outputs. It considers a short time range to apply this discretization but has been able to obtain the essential features required. It is then used to obtain the main features of the leak-induced wavelet excitation and the system in decomposition [32,33]. Cui et al. [34] applied EMD in dealing with the problems of wave attenuation and dispersion in a pipe during leak detection. Propagated waves face these problems when they have lower frequencies and have to travel long distances.

Moreover, undecomposed wavelets with these characteristics lead to low localization accuracy. EMD identifies the main parts of the leak wavelet and facilitates its separation from background noise. It is also effective in processing nonlinear and nonstationary signals from acoustic emissions and is used to decompose signals in both time and frequency domains. With an adaptive noise cancellation mechanism [35], high-fidelity acoustic leak signals can be extracted without needing prior knowledge of the behavior of either background noise or leak noise. The authors realized they could identify the features of nonstationary, white, and color noises. This data-driven mechanism effectively removes all external noise unrelated to the actual leak signal.

It should be noted that most of these studies were performed in the laboratory, and the leaks were simulated. In this study, we used data from the field subjected to different conditions from those obtained in laboratories, usually under controlled conditions. The EMD method was used to extract the features for the ML models' development, given its effectiveness against background noise.

2.3.2. Feature Extraction

Feature extraction entails the production of the data-driven values of certain characteristics of the acoustic waves using mathematical expressions. MATLAB/Octave codes were used for the extraction of these features. The features were extracted using time-domain and frequency-domain dependencies. Both domain attributes were used because vibroacoustic emissions have been shown to exhibit disparate features under each of them [36]. The extracted features are presented in Table 1.

Table 1. EMD feature extraction results.

Feature	Maximum Value	Average Value	Minimum Value
Level	8.00×10^1	6.65×10^1	5.07×10^1
Spread	5.07×10^1	1.10×10^1	2.76×10^0
Root Mean Square	2.37×10^1	6.16×10^2	8.33×10^3
Time-domain Average Amplitude	1.63×10^1	4.72×10^2	6.40×10^3
Peak Amplitude	1.89×10^0	2.80×10^1	5.17×10^2
Crest Factor	1.72×10^1	4.94×10^0	3.06×10^0
Energy	5.63×10^1	6.15×10^2	6.94×10^4
Maximum Lyapunov Exponent (MLE)	3.00×10^4	2.05×10^4	8.41×10^3
Autocorrelation Kurtosis	1.44×10^4	1.52×10^3	4.65×10^0
Autocorrelation MLE	2.87×10^4	2.21×10^4	8.34×10^3
Frequency-Domain Average Amplitude	1.34×10^3	3.04×10^4	2.00×10^5
Peak Frequency	1.96×10^3	6.98×10^2	1.06×10^1
Maximum Amplitude	4.46×10^2	9.03×10^3	4.40×10^4
Frequency centroid	1.38×10^3	6.62×10^2	2.84×10^1
Skewness	2.18×10^1	5.38×10^0	9.26×10^1
Kurtosis	8.49×10^2	8.91×10^1	3.78×10^0
Frequency Spread	1.33×10^0	1.01×10^1	4.88×10^4

The EMD decomposes the signal $x(t)$ in successive intrinsic mode functions (IMFs), c_i. It is a data-adaptive method that iterates to fine-tune the signal and minimize its residue r_m, thereby eliminating the white noise from the pure signal. Thus, the iteration continues for m IMFs until r_m reaches a monotonic function from which IMFs cannot be further extracted. As shown in Equation (1), the residue r_m converges to stop the iterations [37].

$$x(t) = \sum_{i-1}^{m} c_i + r_m \tag{1}$$

The residue is calculated at every time t [38] as:

$$r(t) = x(t) - c(t) \tag{2}$$

The results of the extracted features in EMD are presented in Table 1. The table shows the range of the values for each feature.

2.3.3. Data Preparation: Feature Selection by the Principal Component Analysis (PCA)

Classical machine learning algorithms can be susceptible to overfitting if the dataset size and its features are not well balanced. Processes were taken to study the effects of the feature combinations on the performance of the models. The principal component analysis (PCA) was used since the attributes have significant variabilities. The combination was set to select a group of features whose cumulative variability equals or exceeds 95% of the total variability. The selection was achieved by preferring features with more significant variabilities over those with smaller variabilities. Since PCA is sensitive to the data scale, normalization with range transformation, that is, (0–1), was applied to the dataset before running the analysis.

RapidMiner 9.10 was used for the PCA attribute reduction. The "dimensionality reduction" was set to "keep variance", and the "variance threshold" was 0.95. This threshold was set to ensure that the process stops adding new attributes when a cumulative variance of 95% is achieved. The results of the PCA include the standard deviation of the individual features, the proportion of variance, and the cumulative variance of the entire set of features in the dataset, provided with the eigenvalues. The eigenvectors are provided so that users can identify the features comprising the variance that might be considered. The following steps were followed in choosing the best feature combination.

1. The combined most variable features are used for modeling to observe the performance.
2. The next most variable feature is added to the features used in step 1, the model is revised, and the performance is observed again.
3. If the performance from step 2 is better than step 1, step 2 is repeated until there is a reduction in accuracy or no increase in the subsequent accuracies of the predictions.
4. The features from step 3 are chosen as the optimal feature combination.

Out of the 17 initially extracted features, six emerged as the most suitable combination for the best performance of the models. The dataset consists of 265 data points, of which 80% were used for training and the remaining 20% for testing. Another leak dataset consisting of 80 data points was collected to validate the developed algorithms. The resulting features are presented in Table 2.

Table 2. The cumulative variance of selected features.

Attribute	Standard Deviation	Proportion of Variance	Cumulative Variance
Frequency Centroid	0.567	0.382	0.382
Maximum Amplitude	0.488	0.283	0.665
FD Average Amplitude	0.359	0.154	0.819
Crest Factor	0.247	0.073	0.892
Level	0.168	0.034	0.926
Spread	0.156	0.029	0.955

2.4. PHASE 4: Developing ML-Based Techniques for Leak Localization

In this study, the ML-based techniques for leak localization were developed in RapidMiner Studio 9.10. After selecting the operators from the RapidMiner software library, their parameters were configured based on our preferred definitions. A total of four ML techniques were developed in this study for the prediction of leak direction and distance. As shown in Figure 2, two arbitrary options were considered in predicting the leak directions, i.e., left and right. Moreover, two ML-based techniques, decision trees (DT) and k-nearest neighbors (k-NN), were utilized to predict leak direction. A regression-based ensemble k-NN and a support vector machine (SVM) model were also used for leak distance prediction. Notably, the same features and datasets were used for the training and testing of the ML-based models since a comprehensive comparison could be conducted among the developed models. Different parameters were employed for developing the models in the RapidMiner Studio, such as multiply, select attribute, set role, apply model, etc. By using the "apply model" function, the ML-based algorithms can be applied to the dataset. The output of the models will be presented when the process is executed.

2.4.1. Cross-Validation

The models were trained, tested, and validated on the datasets using cross-validation. The cross-validation method is usually used when the dataset is small, as in this case. It comprises two subprocesses: the training subprocess and the testing subprocesses. The former houses the core operator of the model, while the latter applies the model to test its performance. The parameters used for cross-validation in all the developed models (i.e., DT, k-NN, and SVM) were the same and are as follows: split on batch attribute: no, leave on out: no, number of folds: 10 (i.e., the dataset is divided into ten equal subsets, from which one is used for testing and the remaining are used for training), sampling type: automatic, use local random seed: yes (value = 1992), enable parallel execution: yes. Most of these parameters were left default, except for the "number of folds", configured to create a suitable number of subsets for the dataset. Automatic sampling uses stratified sampling since it is a nominal classification problem. This sampling type ensures that the distribution of the target variable amongst the subsets is even. Specifying the local random seed allows the same subsets to be created whenever the model is executed.

The explanations of the utilized techniques for the prediction of leak direction are as follows:

2.4.2. Development of Regression Models for Leak Distance

Support Vector Machines (LibSVM)

After the PCA analysis of the dataset to obtain the most variable features, a special form of support vector regression (LibSVM tool) was applied to predict the leak distances with a 1 m accuracy. During its invention and development, experiments were used to show its possession of superior performance in many cases for SVM applications [39]. The parameter settings for LibSVM used in this study are listed in Table 3. The last three were seen as expert parameters. The RBF can easily handle both linear and

nonlinear values and has a good way of handling the complexity of the models by using hyperparameters.

Table 3. Parameters used for tuning the ML techniques for leak direction and distance.

ML Techniques	Parameters
LibSVM	kernel type: radial basis kernel function (RBF), gamma (γ): 0, complexity constant (C): 7.5, cache size: 80, epsilon: 0.1, tolerance of loss function of epsilon-SVR (p): 0, shrinking: calculate confidences: confidence for multiclass: False.
Ensemble k-NN	transformation regression: transformation method = exp, z-scale = False. k-NN operator: k = 5, weighted vote = True, measure types = mixed measures, and mixed measure = mixed Euclidean distance.
Decision Tree	criterion: Gini index, maximal depth: 10, apply pruning: yes, confidence: 0.1, apply prepruning: yes, minimal gain: 0.01, minimal leaf size: 2, minimal size for split: 10, number of prepruning alternatives: 3
k-NN	k = 3, weighted vote: yes, measure types: mixed measures, mixed measures: mixed Euclidean distance

SVM requires a solution to the optimization problem.

$$\begin{aligned} \underset{\mathrm{b,w}}{\text{minimize}}: \quad & \tfrac{1}{2} w^T w \\ \textit{subject to}: \quad & \mathrm{y_n}\left(w^T \S_\mathrm{n} + b\right) \geq 1 \end{aligned} \tag{3}$$

where the linear equation $y_n\left(w^T x_n + b\right)$ is the equation of the plane [40].

The regularization parameter C in Equation (5) is used to control the algorithm's complexity, which is essential because of the source of the data. Even though robust signal processing and feature selection were performed, there is no guarantee that they will be completely noise-free. The other parameters were left at their defaults.

The choice of a small C (7.5) was considered for regularization and controlling the sensitivity of the outliers. The C was carefully chosen with the RBF kernel and the other properties to provide a soft-margin SVM. Unlike a hard-margin SVM, the soft-margin widens the boundary of the support vector while allowing some of the data points into it, thereby violating the boundary. However, in return, it will achieve a better score, as seen in Section 3.

Considering the training data $(x_n, y_n)\ in\ \mathbb{R}^n, n = 1, 2, \ldots, N$, the SVM allows a slack $\xi_n \geq 0$ for the number of violations in the margin allowed to the data points.

Considering the slack, the second part of Equation (3) [40] becomes flexible.

$$y_n\left(w^T x_n + \mathrm{b}\right) \geq 1 - \xi_\mathrm{n} \tag{4}$$

The optimization then becomes:

$$\begin{aligned} \underset{\mathrm{b,w},\xi}{\text{minimize}}: \quad & \tfrac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n \\ \textit{subject to}: \quad & \mathrm{y}_n\left(w^T \S_n + b\right) \geq 1 - \xi_n,\ \xi_n \geq 0,\ \nabla_n = 1, 2, ..., N \end{aligned} \tag{5}$$

The RBF kernel is helpful when the dataset is complex with high nonlinearity. There is no doubt that multicollinearity in a dataset affects the learning and leads to overfitting in most cases. Additionally, extreme complexity makes it tedious for the ML algorithm to draw a simple pattern for the dataset. The RBF kernel [39] is defined as:

$$k(x, x') = \exp\left(-\gamma \left\| x - x' \right\|^2\right) \tag{6}$$

where $\|x - x'\|$ is the Euclidean distance between the two examples. It can be seen from Equation (6) that γ directly affects the width of the kernel. A small γ = 0 was chosen in this

work because of the dataset complexity to make the kernel as wide as possible. However, it should be noted that the integrity of the support vectors was not over-compromised.

Ensemble Learner with a k-NN

A k-NN was used in an ensemble learner using the transformed regression operator in the RapidMiner software. The transformation method used for the nested operator was "exp", which is exponential, and the "z scale" was set to false. Table 3 shows the parameters utilized for ensemble k-NN in this study.

The choice of transformation specifies which type of transformation is to be performed on the label variable. The process creates a linear relationship between the dependent and the independent variables. A typical simple exponential transformation is shown in Equation (7) [41].

$$y = B^{a_0 + a_1 x} \rightarrow \log(y) = a_0 + a_1 x \bullet \log(B) \quad (7)$$

A total of five nearest neighbors gave a better score for the training and validation of the distance model. Voting based on the majority was used to classify the unknown point.

2.4.3. Development of Classification Models for Direction

Decision Tree (DT)

A decision tree technique is a tree-like model based on supervised learning and is widely used in many machine learning problems. The DT structure comprises several internal nodes, branches, and leaves. Notably, each internal node of the trees shows a test on an attribute. Since the decision tree produces a visual flowchart as an outcome, it can be easily interpreted and understood. Another advantage of the decision tree is that it requires less effort and computation time than other techniques [42–44]. The tree structure is presented in Figure 3, and Table 3 shows the parameters used to develop the decision tree model in this study.

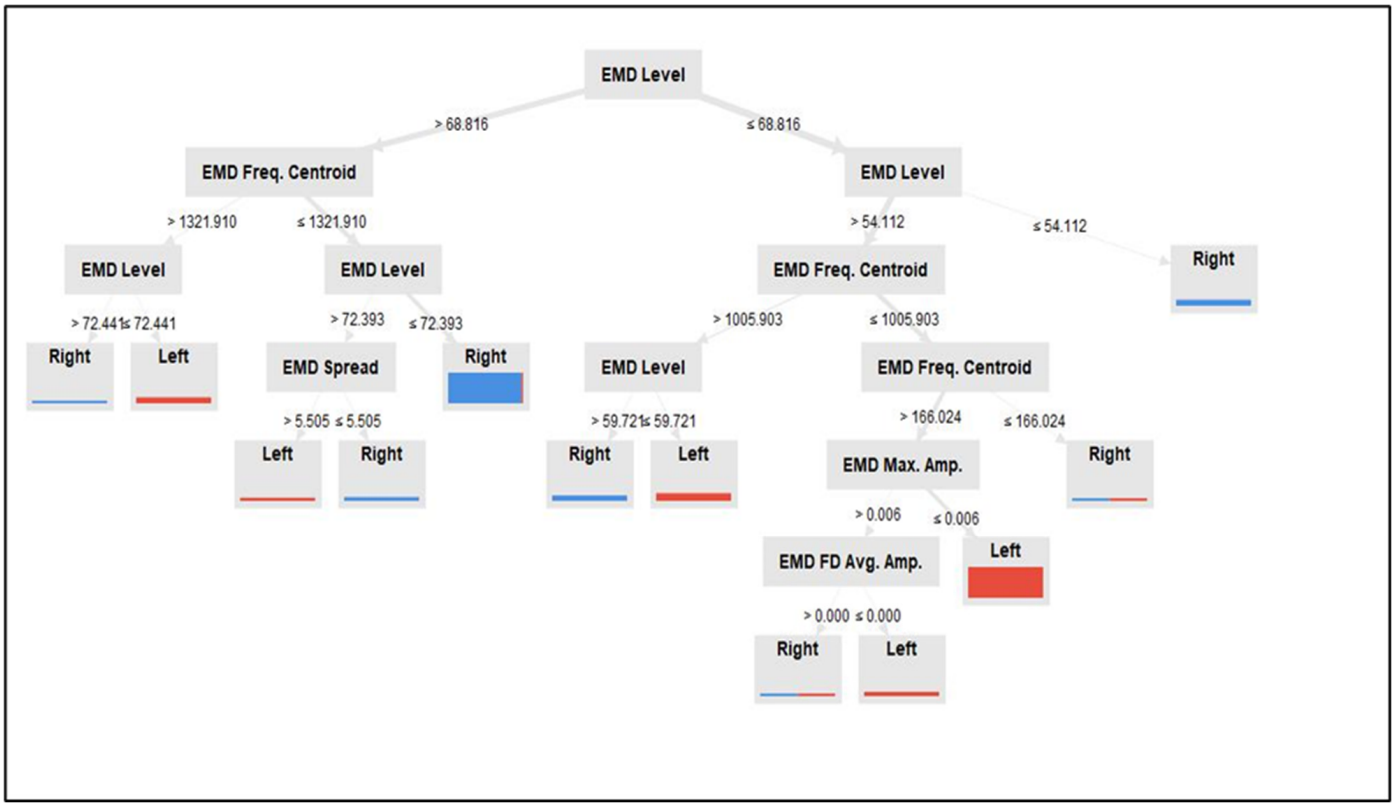

Figure 3. Decision tree model for leak direction prediction.

The Gini index measures the inequalities in the distribution of the label characteristics, and it is used to create partitions in the dataset. Low values of the Gini index show high purity and stable node for a decision to be made, usually requiring no split anymore. It is also determined which feature to use as the root or a node for the split. The wide diversity of the attribute values resulted in the large distribution of the tree, requiring a parameter that must sufficiently support the performance of the tree algorithm. Equation (8) [45] indicates the mathematical description of the Gini index.

$$G = 1 - \sum_{k=1}^{m} p_k^2 \tag{8}$$

where $k = 1, 2, \ldots, m$ and p_k is the proportion of samples belonging to class K.

The depth of the tree, pruning, minimal gain, and size limitations were set to combat overfitting issues. It is necessary to control and curb the decision tree's growth to ensure that memorization and overlearning do not occur during training. Otherwise, it will result in poor performance when validated with unseen data.

k-Nearest Neighbor (k-NN)

k-NN is a nonparametric supervised learning algorithm utilized for classification purposes in this study. This algorithm can predict the new dataset based on the similarity measure. In doing so, the k-NN algorithm assumes the similarity between the new data and the k-training dataset and puts the new case into the class nearest to the available classes. In other words, it classifies the data based on how its neighbors are classified and put the new data point in the category with the highest probability of housing it [46]. One of the advantages of the k-NN algorithm is that it can provide good results with high accuracies for large and small datasets in case they are labeled and noise-free. However, finding a suitable k-number with a high-accuracy model is challenging [47]. The k-NN technique is suitable for classification purposes. This study used it to predict the direction of leaks in real water networks. The parameters for the developed k-NN model can be seen in Table 3.

3. Model Implementation and Validation

3.1. Performance Measures

The performance of the leakage distance prediction was measured using a 1-m maximum distance rule. The purpose of pinpointing the leak location is to make the excavation and repair easier for the Water Supplies Department or any associated organization concerned. The successful location of the position of a leak is performed within a 1-m radius of the actual leakage point. Likewise, class recall, class precision, and classification accuracy were used for assessing the leakage direction models. The computations of these metrics are shown in Equations (9)–(11), respectively.

Given the following definitions [45]:

- TR = true Right: when both the predicted class and the actual class are R;
- TL = true Left: when both the predicted class and the actual class are L;
- FR = false Right: when the predicted class is R, but the actual class is L; and
- FL = false Left: when the predicted class is L, but the actual class is R.

$$\text{Class recall} = \frac{TR}{TR + FL} \; or \; \frac{TL}{TL + FR} \tag{9}$$

$$\text{Class precision} = \frac{TR}{TR + FR} \; or \; \frac{TL}{TL + FL} \tag{10}$$

$$\text{Prediction accuracy} = \frac{TR + TL}{TR + TL + FL + FR} \tag{11}$$

3.2. The Analysis of ML-Based Models for Leak Distance

3.2.1. LibSVM

The LibSVM yielded a bias of −4.196 for the prediction, similar to the intercept c in the well-known line equation $y = mx + c$. Other salient outcomes to consider about the model are the weights of the individual features, like the coefficients of the terms in a line equation. These parameters are essential because the SVM algorithms use them to create the hyperplane to achieve optimized predictions. They also reflect the actual effects carried by the attributes in the dataset. The weights of the features are presented in Table 4.

Table 4. Weights of attributes for LibSVM.

Attribute	Weight
Level	4.89×10^5
Spread	8.14×10^4
Crest Factor	4.45×10^4
Avg. Amp.	1.43×10^0
Max. Amp.	2.87×10^1
Freq. Centroid	6.02×10^6

The LibSVM model achieved a training accuracy of 85.28% and a validation accuracy of 82.50% using a maximum of 1m radius measure as the acceptable limit. The errors were uniformly distributed. Thus, the model performs well. A mean error of 0.036 and a standard deviation of 0.980 were achieved. Therefore, the errors were not overspread about their mean value.

3.2.2. Ensemble k-NN

The k-NN achieved an accuracy of 80.0% for training and 83.75% for validation, maintaining the 1 m as the acceptable accuracy limit. The validation yielded a mean error of 0.864 and a standard deviation of 4.381. These values were high because the few wrongly calculated distances were strangely large compared to the actual. However, the mean error and deviation of the correctly calculated distances were insignificant and as low as 0.067 and 0.260, respectively. Conspicuously different from the combined results, showing how significantly the few wrongly predicted distances affected the entire model. Theoretically, normalization is usually required when using k-NN because of its sensitivity to the scale of the attributes. Outlier detection could also be used to minimize the effects of these impurities, though none of them were considered in this study. For a pragmatic approach, it was realized that these data cleansing mechanisms were not inapplicable to this model and the given dataset as they resulted in poor performances. Subsequently, PCA was used, and it proved useful and successful. The training and validation errors are presented in Table 5.

Table 5. Training and validation of RMSEs and MAEs.

Model	Parameter	Training	Validation
LibSVM	Root mean square error	2.888	0.981
	Mean Absolute error	0.976	0.689
Ensemble k-NN	Root mean square error	4.589	4.465
	Mean Absolute error	1.879	1.561

In Table 5, the validation RMSE and MAE values supersede those of the training sets in terms of optimality. Here, the RMSE and MAE are agreed upon as performance measures because the behaviors of the models follow an acceptable trend. This result is a reliable performance indication of the models because the validation dataset was obtained from a different field location and was never used in training.

The predicted leakage and actual leakage distances can be visualized in the plots shown in Figure 4. The three noise loggers were stationed at three different distances

(0.9 m, 1.314 m, and 3.556 m) from the three leakage spots, see Section 2.2.1. Most points recorded validation errors of less than 1 m, with a few data points having high variations. K-NN has six data points with wider marginal errors, and it is noteworthy to recall that normalization and outlier detection were not considered during the data processing stage.

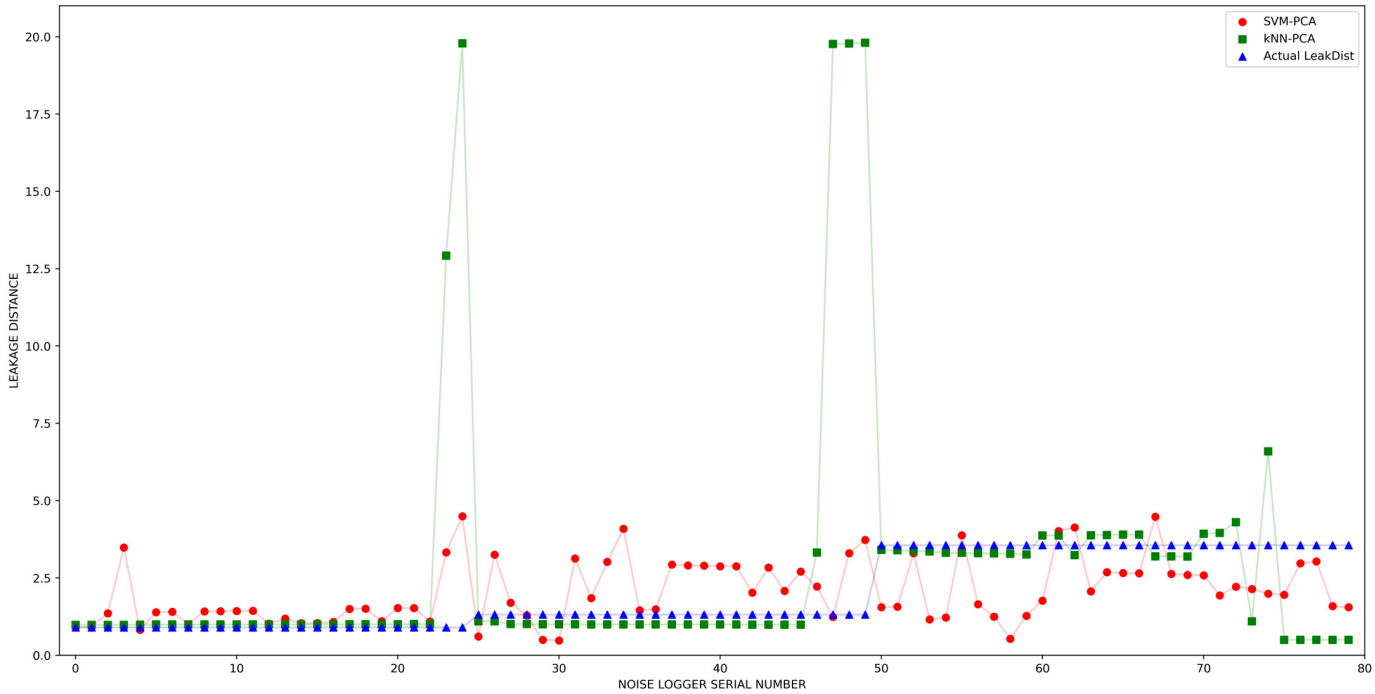

Figure 4. Visualization of the predicted and actual leakage distances for ensemble k-NN and LibSVM with PCA.

Accuracy Table

The training and validation accuracies for all the models are presented in Table 6. A maximum of 1 m is considered in all cases. The accuracies are calculated by Equation (12).

$$\text{Accuracy} = \frac{\text{Number of correctly predicted datapoints}}{\text{Total number of predictions}} \quad (12)$$

Table 6. The training and validation performances of the models.

Model	Training	Validation
LibSVM	85.28%	82.50%
Ensemble k-NN	80.00%	83.75%

3.3. The Analysis of ML-Based Models for Leak Direction

3.3.1. Decision Tree (DT)

The DT achieved a training and validation accuracy of 90.08% and 78.75%, respectively. Also, the DT classification structure is shown in Figure 3. A confusion matrix in Tables 7 and 8 shows the training and validation accuracy of the model.

Table 7. The confusion matrices for the training of the direction machine-learning models.

k-NN			
Accuracy = 92.08%	true Right	true Left	class precision
predicted Right	68	7	90.67%
predicted Left	5	74	93.67%
class recall	93.15%	91.36%	92.06%
Decision Tree			
Accuracy = 90.08%	true Right	true Left	class precision
predicted Right	67	9	88.16%
predicted Left	6	72	92.31%
class recall	91.78%	88.89%	90.08%

Table 8. The confusion matrices for validation of the direction of machine-learning models.

k-NN			
Accuracy = 97.50%	true Right	true Left	class precision
predicted Right	29	1	96.67%
predicted Left	1	49	98.00%
class recall	96.67%	98.00%	97.50%
Decision Tree			
Accuracy = 78.75%	true Right	true Left	class precision
predicted Right	29	16	64.44%
predicted Left	1	34	97.14%
class recall	96.67%	68.00%	78.75%

3.3.2. k-NN

Tables 7 and 8 contain the training and validation results of the developed k-NN model. It is seen that the k-NN model achieved a training accuracy of 92.08% and a validation accuracy of 97.50%. Both class precision and class recall in all cases were also 90% or more, showing the reliability of the model and its performance. The class precision and recall indicate the relevance of the classifier. A low validation error, i.e., 2.50% from the model, is also a good performance indicator.

3.4. Analyzing the Effectiveness of Adopting the PCA

All 17 extracted features were used as input attributes to show the cogency of adopting the PCA to remove multicollinearity. The same parameters applied to the previous models were used. The features were (1) level, (2) spread, (3) crest factor, (4) FD average amplitude, (5) maximum amplitude, (6) frequency centroid, (7) RMS, (8) TD average amplitude, (9) peak amplitude, (10) crest factor, (11) Energy, (12) MLE, (13) autocorrelation kurtosis, (14) autocorrelation MLE, (15) peak frequency, (16) skewness, and (17) kurtosis.

The LibSVM model achieved a prediction accuracy of 37.50% (see Table 9), which is lower than the accuracy of the same model when PCA analysis was done to generate only the most variable features. Likewise, the ensemble k-NN also achieved an accuracy of 62.50%.

Table 9. Training and validation accuracies of 17 features.

Model	**Training**	**Validation**
LibSVM	82.64%	37.50%
Ensemble k-NN	90.94%	62.50%

The addition of redundant features increases computational costs and induces overfitting and complexity. To avoid these difficulties in the study, the brute force method, which involves feeding all the input variables into the model while relinquishing the responsibility of obtaining the best combination for achieving the maximum accuracy [48],

is eliminated. The authors conducted a leak detection study in which fewer attributes from feature selection techniques produced better results than the brute force practice. The effects are seen when all 17 features are used in modeling; the validation accuracy is low, but the errors' standard deviation, RMSE, and MAE look good (see Tables 9 and 10). This uneven performance shows the presence of overfitting. The actual and predicted results are shown in Figure 5.

Table 10. Training and validation of RMSEs and MAEs for 17 features.

Model	Parameter	Training	Validation
LibSVM	Root mean square error	2.812	1.966
	Mean Absolute error	1.103	1.560
Ensemble k-NN	Root mean square error	2.950	2.515
	Mean Absolute error	0.679	1.713

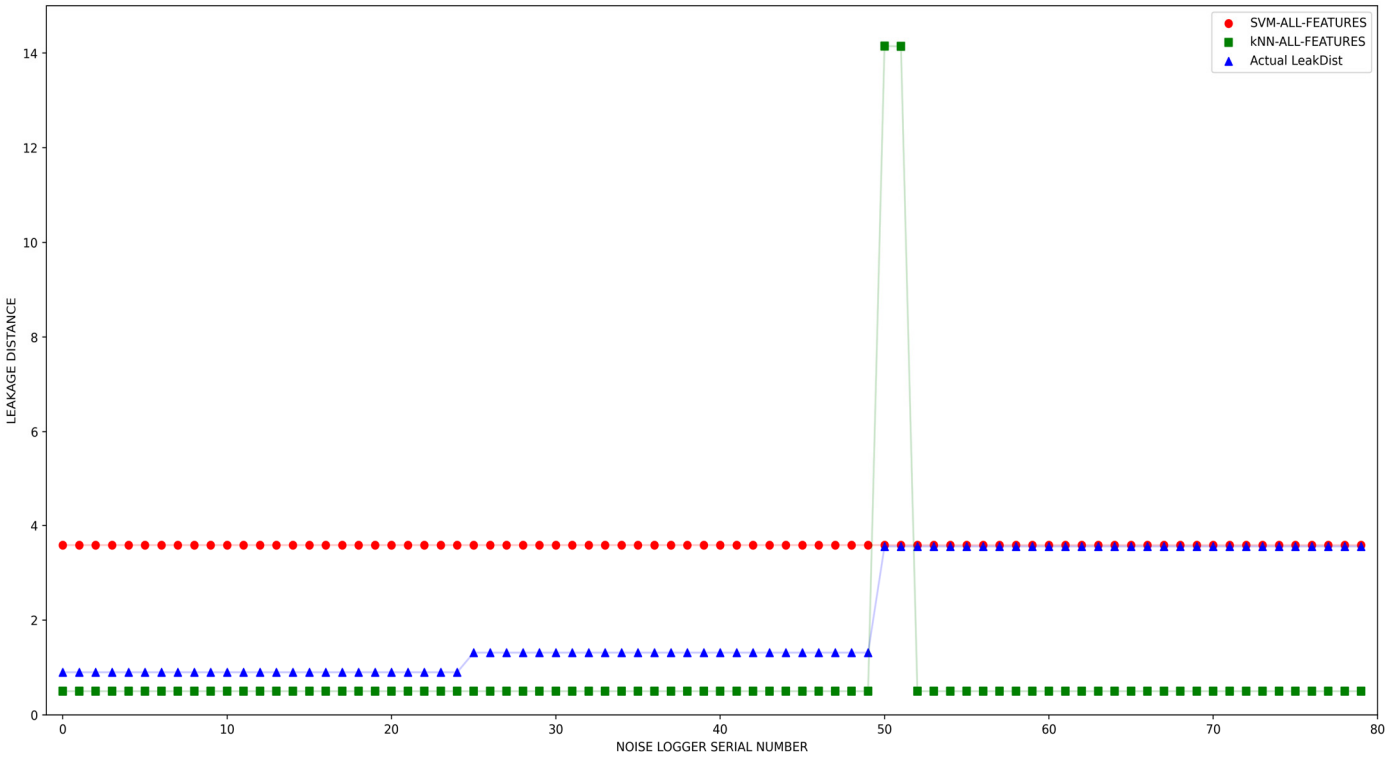

Figure 5. Visualization of the predicted and actual leakage distances for Ensemble k-NN LibSVM with all features.

The graphical representation of the modeling results with all 17 features is presented in Figure 5. It is seen that both models are merely copying one value closer to one of the actual leakage distances throughout the predictions. This constancy is why they perform better in the RMSE and the MAE, but the results show unreliable models.

4. Conclusions

Most studies on localizing pipe leakages have been undertaken in labs under controlled conditions, and only a few have been tested in the field. In addition, deploying several noise loggers on the WDNs is time-consuming and requires big budgets. Therefore, this study aims to overcome the mentioned shortcomings for the first time in the body of relevant literature using a novel methodological approach. The methodology employed in this study is based on deploying a single noise logger on real WDNs and utilizing advanced ML-based techniques. Because there would be confusion about the leak's location on the water mains in a blind deployment, this study predicts the leak's distance and direction.

Therefore, the noise logger is the reference point relative to the source of the acoustic leak signal. Water supply departments can use the developed models to determine the locations of leakages in water mains when detected. They will be required to obtain and input the acoustic leak signals into the models. The signals are decomposed, and the locations are predicted using the ML techniques discussed in the paper.

The data for this study was collected from 60 leaks and 32 non-leak sites in Hong Kong, while the validation dataset was obtained from three (3) leak spots at another leak site. Several experiments were conducted by placing the noise logger on pipelines in the WDNs. The study was conducted in an urban city with rife noise and disturbances. The adopted feature extraction and denoising technique is robust against background noises. It separates a high-fidelity leak signal from any external noise that might contaminate the leak signal. Therefore, the models can be applied to locate leaks in water distribution networks from various noise disturbances. If no background noise is recorded, the pure leak signal is maintained after the decomposition.

Using a single noise logger for leak localization is feasible, as shown in this study. The empirical mode decomposition (EMD) signal processing effectively extracts high-fidelity signal features from the acoustic leak signal for leak localization. The EMD was used to extract high-fidelity signal features by fine-tuning the leak wavelets and separating them from the background noises. Since obtaining leak signals devoid of noise from an operating WDN is inconceivable, this decomposition technique was used. It is very effective in dealing with the problems of white noise, which commonly plagues pressurized water transport pipelines.

This study establishes that the SVM and the transformed regression ensemble with k-NN techniques are effective for leak localization. Both models achieved closed accuracy, but the SVM demonstrated a more reliable performance due to its immunity against overfitting. Likewise, the classification techniques DT and k-NN are effective in determining the directions of the location of the leaks when the noise logger is used as the reference point. Out of the two ML algorithms employed, the k-NN performed better and is therefore preferred for direction determination.

5. Future Studies

We encourage future research on using a single noise logger for leak localization to observe the direction of the water flow during signal acquisition, as that was one of the limitations of this study. The observed flow direction can be used as a reference for stating the direction since it can be challenging to interpret arbitrary directions on the WDN when using maps. Moreover, metal pipe materials with diameters of approximately 600 mm were considered in this study, which caused the early attenuation of the acoustic signals. Therefore, the localization models were built on a dataset with a maximum distance of 25 m. A thorough signal analysis of the acoustic leak signals should be studied in two ways: (1) when the signals propagate in the same direction as the water flow and (2) when the signal propagates in the opposite direction to the flow of the water. Even though we tried our best to acquire enough data for our models' developments, we still believe that the quantity we used was insufficient for generalizability, especially when considering different cities and varieties of environmental conditions. Further studies involving a large amount of data with diverse distance ranges should be carried out to avoid the biases of insufficient data. The proposed research can also not locate multiple leaks simultaneously in the same pipeline. The noise logger deployment for individual leaks should be completed since the method relies on acoustic signal propagation. It is not yet clear how to differentiate between two or more leak signals traveling together due to interference. The noise logger records the signals it encounters as one, whether generated from a single leak or multiple leak points. When two leak signals interfere constructively, the recordings from the noise logger might be interpreted by the leak distance prediction models to be farther away than they actually are. This prediction results from increased pressure, amplitude, frequency, etc., due to two compressions

or rarefactions meeting. The models might also predict a distance shorter than the actual leak distance location when one leak signal compression component meets with a rarefaction component of another signal, thereby causing pressure reduction than in typical situations. The prediction will not yield a correct value if the noise logger records the signal in any of these situations.

The developed models did not also consider the network complexity in the modeling. The water network complexity is another factor that can hinder the performance of the developed models. The signals were acquired on pipelines whose layout arrangements could be determined using the available maps. In the cases where multiple networks conglomerate or there are bends, the models will face significant challenges in pinpointing the exact leak location. They will only provide the approximate distance from the noise logger, not including whether the pipeline is straight or arched.

More validation cases are required to establish the complete reliability of the models. The validation dataset only consisted of signals from short distances, i.e., 0.9 m, 1.314 m, and 3.556 m. Therefore, even though the models were trained to predict leak positions up to 20 m, there were no validation cases beyond 4 m. However, this only applies to the held-out testing dataset, but the training-validation steps were adequately completed using the 80% training and 20% testing practice. Future research will be targeted toward solving the above limitations.

Author Contributions: Conceptualization, H.S.; methodology, A.-M.Y. and H.S.; software, A.-M.Y. and H.S.; validation, A.-M.Y. and H.S.; formal analysis, A.-M.Y. and H.S.; investigation, A.-M.Y.; resources, T.Z.; data curation, A.-M.Y. and H.S.; writing—original draft preparation, A.-M.Y. and H.S.; writing—review and editing, A.-M.Y., H.S. and T.Z.; visualization, A.-M.Y. and H.S.; supervision, T.Z.; project administration, T.Z.; funding acquisition, T.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Innovation and Technology Fund [Innovation and Technology Support Programme (ITSP)] and the Water Supplies Department of Hong Kong under the grant number ITS/067/19FP.

Data Availability Statement: All data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. UN-WWAP (United Nations World Water Assessment Programme). *The United Nations World Water Development Report 2018: Nature-Based Solutions for Water*; UN-WWAP: Paris, France, 2018.
2. El-Zahab, S.; Zayed, T. Leak Detection in Water Distribution Networks: An Introductory Overview. *Smart Water* **2019**, *4*, 5. [CrossRef]
3. Taiwo, R.; Shaban, I.A.; Zayed, T. Development of Sustainable Water Infrastructure: A Proper Understanding of Water Pipe Failure. *J. Clean. Prod.* **2023**, *398*, 136653. [CrossRef]
4. Kanakoudis, V.; Muhammetoglu, H. Urban Water Pipe Networks Management towards Non-revenue Water Reduction: Two Case Studies from Greece and Turkey. *Clean-Soil Air Water* **2014**, *42*, 880–892. [CrossRef]
5. Van Zyl, J.; Clayton, C.R.I. The Effect of Pressure on Leakage in Water Distribution Systems. In *Institution of Civil Engineers-Water Management*; Thomas Telford Ltd.: London, UK, 2007; Volume 160, pp. 109–114.
6. Yue, D.P.T.; Tang, S.L. Sustainable Strategies on Water Supply Management in Hong Kong. *Water Environ. J.* **2011**, *25*, 192–199. [CrossRef]
7. Gupta, A. Hong Kong Is Wasting a Third of Its Water. Available online: https://chinadialogue.net/en/cities/9803-hong-kong-is-wasting-a-third-of-its-water/ (accessed on 20 October 2022).
8. Tijani, I.A.; Abdelmageed, S.; Fares, A.; Fan, K.H.; Hu, Z.Y.; Zayed, T. Improving the Leak Detection Efficiency in Water Distribution Networks Using Noise Loggers. *Sci. Total Environ.* **2022**, *821*, 153530. [CrossRef]
9. Liu, D.; Fan, J.; Wu, S. Acoustic Wave-Based Method of Locating Tubing Leakage for Offshore Gas Wells. *Energies* **2018**, *11*, 3454. [CrossRef]
10. Hu, Z.; Tariq, S.; Zayed, T. A Comprehensive Review of Acoustic Based Leak Localization Method in Pressurized Pipelines. *Mech. Syst. Signal Process.* **2021**, *161*, 107994. [CrossRef]

11. Mohandes, S.R.; Sadeghi, H.; Fazeli, A.; Mahdiyar, A.; Hosseini, M.R.; Arashpour, M.; Zayed, T. Causal Analysis of Accidents on Construction Sites: A Hybrid Fuzzy Delphi and DEMATEL Approach. *Saf. Sci.* **2022**, *151*, 105730. [CrossRef]
12. Nguyen, P.H.D.; Fayek, A.R. Applications of Fuzzy Hybrid Techniques in Construction Engineering and Management Research. *Autom. Constr.* **2022**, *134*, 104064. [CrossRef]
13. Mohandes, S.R.; Zhang, X.; Mahdiyar, A. A Comprehensive Review on the Application of Artificial Neural Networks in Building Energy Analysis. *Neurocomputing* **2019**, *340*, 55–75. [CrossRef]
14. Siraj, N.B.; Fayek, A.R. Risk Identification and Common Risks in Construction: Literature Review and Content Analysis. *J. Constr. Eng. Manag.* **2019**, *145*, 3119004. [CrossRef]
15. Martini, A.; Troncossi, M.; Rivola, A. Leak Detection in Water-Filled Small-Diameter Polyethylene Pipes by Means of Acoustic Emission Measurements. *Appl. Sci.* **2017**, *7*, 2. [CrossRef]
16. Brunone, B.; Capponi, C.; Meniconi, S. Design Criteria and Performance Analysis of a Smart Portable Device for Leak Detection in Water Transmission Mains. *Measurement* **2021**, *183*, 109844. [CrossRef]
17. Habib, A.; Akram, S.; Ali, M.R.; Muhammad, T.; Zainab, S.; Jehangir, S. Radio Frequency Identification Temperature/CO_2 Sensor Using Carbon Nanotubes. *Nanomaterials* **2023**, *13*, 273. [CrossRef] [PubMed]
18. El-Zahab, S.; Asaad, A.; Abdelkader, E.M.; Zayed, T. Development of a Clustering-Based Model for Enhancing Acoustic Leak Detection. *Can. J. Civ. Eng.* **2019**, *46*, 278–286. [CrossRef]
19. Liu, Z.; Kleiner, Y. State of the Art Review of Inspection Technologies for Condition Assessment of Water Pipes. *Measurement* **2013**, *46*, 1–15. [CrossRef]
20. Liu, Y.; Ma, X.; Li, Y.; Tie, Y.; Zhang, Y.; Gao, J. Water Pipeline Leakage Detection Based on Machine Learning and Wireless Sensor Networks. *Sensors* **2019**, *19*, 5086. [CrossRef]
21. Meniconi, S.; Capponi, C.; Frisinghelli, M.; Brunone, B. Leak Detection in a Real Transmission Main through Transient Tests: Deeds and Misdeeds. *Water Resour. Res.* **2021**, *57*, e2020WR027838. [CrossRef]
22. Xu, X.; Karney, B. An Overview of Transient Fault Detection Techniques. *Model. Monit. Pipelines Netw.* **2017**, *7*, 13–37.
23. Stajanca, P.; Chruscicki, S.; Homann, T.; Seifert, S.; Schmidt, D.; Habib, A. Detection of Leak-Induced Pipeline Vibrations Using Fiber—Optic Distributed Acoustic Sensing. *Sensors* **2018**, *18*, 2841. [CrossRef]
24. Zuo, J.; Zhang, Y.; Xu, H.; Zhu, X.; Zhao, Z.; Wei, X.; Wang, X. Pipeline Leak Detection Technology Based on Distributed Optical Fiber Acoustic Sensing System. *IEEE Access* **2020**, *8*, 30789–30796. [CrossRef]
25. Yazdekhasti, S.; Piratla, K.R.; Atamturktur, S.; Khan, A. Experimental Evaluation of a Vibration-Based Leak Detection Technique for Water Pipelines. *Struct. Infrastruct. Eng.* **2018**, *14*, 46–55. [CrossRef]
26. El-Abbasy, M.S.; Mosleh, F.; Senouci, A.; Zayed, T.; Al-Derham, H. Locating Leaks in Water Mains Using Noise Loggers. *J. Infrastruct. Syst.* **2016**, *22*, 04016012. [CrossRef]
27. Tariq, S.; Bakhtawar, B.; Zayed, T. Data-Driven Application of MEMS-Based Accelerometers for Leak Detection in Water Distribution Networks. *Sci. Total Environ.* **2022**, *809*, 151110. [CrossRef]
28. Tijani, I.A.; Zayed, T. Gene Expression Programming Based Mathematical Modeling for Leak Detection of Water Distribution Networks. *Meas. J. Int. Meas. Confed.* **2022**, *188*, 110611. [CrossRef]
29. Liu, B.; Jiang, Z.; Nie, W.; Ran, Y.; Lin, H. Research on Leak Location Method of Water Supply Pipeline Based on Negative Pressure Wave Technology and VMD Algorithm. *Meas. J. Int. Meas. Confed.* **2021**, *186*, 110235. [CrossRef]
30. Sanchez-Gendriz, I. Signal Processing Basics Applied to Ecoacoustics. *Ecol. Inform.* **2021**, *66*, 101445. [CrossRef]
31. Mahmutoglu, Y.; Turk, K. Received Signal Strength Difference Based Leakage Localization for the Underwater Natural Gas Pipelines. *Appl. Acoust.* **2019**, *153*, 14–19. [CrossRef]
32. Cody, R.A.; Dey, P.; Narasimhan, S. Linear Prediction for Leak Detection in Water Distribution Networks. *J. Pipeline Syst. Eng. Pract.* **2020**, *11*, 04019043. [CrossRef]
33. Ebrahimkhanlou, A.; Salamone, S. Single-Sensor Acoustic Emission Source Localization in Plate-like Structures Using Deep Learning. *Aerospace* **2018**, *5*, 50. [CrossRef]
34. Cui, X.; Yan, Y.; Ma, Y.; Ma, L.; Han, X. Localization of CO2 Leakage from Transportation Pipelines through Low Frequency Acoustic Emission Detection. *Sens. Actuators A Phys.* **2016**, *237*, 107–118. [CrossRef]
35. Guo, C.; Wen, Y.; Li, P.; Wen, J. Adaptive Noise Cancellation Based on EMD in Water-Supply Pipeline Leak Detection. *Meas. J. Int. Meas. Confed.* **2016**, *79*, 188–197. [CrossRef]
36. Butterfield, J.D.; Meyers, G.; Meruane, V.; Collins, R.P.; Beck, S.B.M. Experimental Investigation into Techniques to Predict Leak Shapes in Water Distribution Systems Using Vibration Measurements. *J. Hydroinform.* **2018**, *20*, 815–828. [CrossRef]
37. Wu, Z.; Huang, N.E. Ensemble Empirical Mode Decomposition: A Noise-Assited. *Biomed. Tech.* **2010**, *55*, 193–201.
38. Wang, G.; Chen, X.Y.; Qiao, F.L.; Wu, Z.; Huang, N.E. On Intrinsic Mode Function. *Adv. Adapt. Data Anal.* **2010**, *2*, 277–293. [CrossRef]
39. Hsu, C.W.; Chang, C.C.; Lin, C.J. A Practical Guide to Support Vector Classification. *BJU Int.* **2016**, *101*, 1396–1400.
40. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
41. Nurmuhammad, A.; Muhammad, M.; Mori, M. Numerical Solution of Initial Value Problems Based on the Double Exponential Transformation. *Publ. Res. Inst. Math. Sci.* **2005**, *41*, 937–948. [CrossRef]
42. Sharma, H.; Kumar, S. A Survey on Decision Tree Algorithms of Classification in Data Mining. *Int. J. Sci. Res.* **2016**, *5*, 2094–2097.

43. El-zahab, S.; Abdelkader, E.M.; Zayed, T. An Accelerometer-Based Leak Detection System. *Mech. Syst. Signal Process.* **2018**, *108*, 276–291. [CrossRef]
44. Taiwo, R.; Ben Seghier, M.E.A.; Zayed, T. Towards Sustainable Water Infrastructure: The State-of-the-Art for Modeling the Failure Probability of Water Pipes. *Water Resour. Res.* **2023**, e2022WR033256. [CrossRef]
45. Kotu, V.; Deshpande, B. Classification. In *Data Science*; Elsevier: Amsterdam, The Netherlands, 2019; Volume 51, pp. 65–163. ISBN 9781493985791.
46. Quy, T.B.; Muhammad, S.; Kim, J.-M. A Reliable Acoustic EMISSION Based Technique for the Detection of a Small Leak in a Pipeline System. *Energies* **2019**, *12*, 1472. [CrossRef]
47. Fereidooni, Z.; Tahayori, H.; Bahadori-Jahromi, A. A Hybrid Model-Based Method for Leak Detection in Large Scale Water Distribution Networks. *J. Ambient Intell. Humaniz. Comput.* **2021**, *12*, 1613–1629. [CrossRef]
48. Butterfield, J.D.; Meruane, V.; Collins, R.P.; Meyers, G.; Beck, S.B.M. Prediction of Leak Flow Rate in Plastic Water Distribution Pipes Using Vibro-Acoustic Measurements. *Struct. Health Monit.* **2018**, *17*, 959–970. [CrossRef]