INTERNATIONAL JOURNAL OF AUTOMOTIVE AND MECHANICAL ENGINEERING
ISSN: 2229-8649 e-ISSN: 2180-1606
VOLUME 23, ISSUE 2, 2026, 13649 - 13660
DOI: https://doi.org/10.15282/ijame.23.2.2026.13.1032

RESEARCH ARTICLE

Pipe feature classification using support vector machine based on time-domain feature extraction

Nor Azinee Said[1,2], Mohd Fairusham Ghazali1[1]*, Mohd Fadhlan Mohd Yusof[1], Nurul Hazwani Mokhtar[1], Muhammad Amsyar Azmil[1], Zairif Anas Mohammad Zahib[1]

[1]Advanced Structural Integrity and Vibration Research (ASIVR), Faculty of Mechanical and Automotive Engineering Technology, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia
[2] Faculty Engineering Technology, University College TATI (UC TATI), Jalan Panchor, Telok Kalong, 24000 Kemaman, Malaysia

Abstract - Water infrastructure deterioration threatens global water security, with utilities losing over $14 billion annually due to non-revenue water (NRW). Malaysia reported 37.1% NRW in 2023, significantly exceeding international benchmarks. This study develops an automated pipeline monitoring system integrating pressure transient analysis with machine learning for leak detection in water distribution networks. Three feature selection methods (minimum classification error, minimum redundancy maximum relevance, and sequential forward selection) were systematically evaluated to optimize classification performance. The proposed minimised classification error-surface vector machine model (MCE-SVM), which is based on two time-domain features slope sign change and integrated EMG to characterize three pipeline conditions, including elbow, junction, and leak, attained a test accuracy of 87.4%, substantially higher than those achieved by Minimum redundancy maximum relevance- surface vector machine (77.3%) and sequential forward selection-surface vector machine (65.8%), demonstrating the effectiveness of the minimised classification error based method in selecting highly discriminatory features with a minimal feature set. Moreover, the suggested system is capable of fast processing of the signal, which highlights the appropriateness of the proposed system to the report on near real-time monitoring of pipelines. The controlled laboratory experiments were conducted using 300 transient pressure signals recorded with one configuration under 4 bar in one pipe setup (63 mm medium-density polyethene).

Article History
Received : 31 July 2025
Revised : 22 September 2025
Accepted : 5 May 2026
Published : 30 June 2026

Keywords
Non-revenue water
Transient reflection method
Minimized classification error
Support vector machines
Feature selection

1. Introduction

The degradation of water infrastructure poses a challenge to global water security, and through non-revenue water (NRW), utilities have lost more than 14 billion dollars annually. In developing nations, the old pipeline system is a significant cause of these losses. In 2023, Malaysia had an NRW of 37.1%, a high level compared to the International Water Association standard of 15% [1]. Over 70 per cent of NRW is caused by physical leaks due to pipeline failures, and the effects across regions can be significant. During the period 2015 to 2017, the state of Sabah alone reported RM657 million in water leakage losses [2]. These problems align with the United Nations Sustainable Development Goal 6, which aims to achieve universal access to clean water by 2030. The causes of the leaks in the pipes are varied and include ageing of the infrastructure, ground movement, deterioration of construction materials, excessive pressure on the pipes, and defects during installation. Utilities can detect issues, minimise water loss, reduce repair costs, and achieve service reliability [3]. Conventional leak detection methods involve acoustic, ground-penetrating radar, slope, and infrared thermography, as well as manual inspection. Nonetheless, these methods are limited to complex distribution systems with multiple pipe and operational characteristics. The transient reflection ethod (TRM) offers benefits for detecting leakage by observing the behaviour of pressure waves within a pipeline [4]. When a transient pressure wave hits a discontinuity, such as a leak, elbow, or junction, characteristic reflections occur. TRM identifies these reflections to recognise pipeline characteristics and determine the condition of the system. Although successful when applied to very simple pipeline systems, TRM interpretation is difficult when a network contains many discontinuities, where overlapping reflections can complicate signal assessment [5].

Machine learning offers potential solutions for the automated interpretation of signals within a complex pipeline system. Machine learning approaches have also demonstrated excellent capability in solving engineering prediction problems involving nonlinear relationships and complex operating conditions, including friction-welded joint toughness prediction, refrigeration system performance estimation, and eco-safe driving behaviour modelling [6]-[8]. Support Vector Machines have proved useful in pattern recognition tasks involving high-dimensional data with very few samples. Past research by Qu et al. [9] used support vector machines (SVMs) within a wavelet packet decomposition pipeline for leak detection and achieved a classification accuracy of over 85%. Xu et al. [10] applied variational mode decomposition to SVM for leak identification, achieving 92% accuracy. Nevertheless, these studies have used manually selected feature sets that were not systematically optimised, which may limit the classification results and practicality. In vibration- and sensor-based monitoring systems, artificial and recurrent neural networks have demonstrated strong capabilities in recognising dynamic fault patterns and extracting temporal dependencies from sequential data. However, these studies

*CORRESPONDING AUTHOR | M. Fairusham | ✉ fairusham@umpsa.edu.my

have used manually selected feature sets that were not systematically optimised, which may limit the classification results and practicality.

Feature selection algorithms identify the most informative signal characteristics while reducing data dimensionality and computational burden. Generally, feature selection approaches can be categorised into three main groups. Filter methods evaluate features using statistical measures independently of the classifier, wrapper methods assess feature subsets based on classifier performance, and embedded methods incorporate feature selection directly within the learning algorithm. The minimised classification error (MCE) technique belongs to the embedded category and directly optimises classification accuracy during the feature selection process [11]. In addition, machine learning-assisted reliability and surrogate modelling approaches have attracted increasing attention in structural engineering applications due to their capability to enhance prediction efficiency while reducing computational complexity [12]. Despite the theoretical advantages of MCE, limited studies have systematically compared its performance with other feature selection approaches in transient response monitoring-based pipeline monitoring applications. A review of existing pipeline monitoring research shows that most studies use relatively small feature sets, typically examining between four and ten characteristics, without thoroughly exploring the full range of available time-domain features. Moreover, the scholars did not conduct a systematic comparison of feature selection techniques to determine which is best suited for classifying TRM signals. Another significant factor to consider is that very few studies have examined the balance between the need for accurate classification and the requirement for real-time execution in a real-world pipeline network.

This study seeks to address such gaps by achieving several objectives. Thirty-seven time-domain-derived statistical features of TRM pressure transient signals will be extracted and evaluated. The study will compare three feature selection algorithms (MCE, mRMR, and SFS) to identify which features provide the most informative details. Next, support vector machine-based classification models will be developed to differentiate elbows, junctions, and leaks in pipeline components. Performance validation will use statistical measures such as precision, recall, and F1 scores. Finally, the research will analyse the feasibility of real-time system monitoring under computational resource constraints. The classifications (elbow, junction, and leak) were made based on operational needs. To plan maintenance, water utilities must identify normal pipeline features (elbows, junctions) and unusual ones (leaks). Permanent components like elbows and junctions need different considerations from leaks. The multiclass action-by-objective method yields field data that reveal not only the presence of an anomaly but also its type

2. Materials and Methods

2.1 Experimental Setup

A testing facility, designed as a pipeline, was developed and constructed to simulate water distribution operating conditions. The component specifications and system configuration are outlined as follows. The experimental pipeline was fabricated using medium-density polyethene (MDPE) with a diameter of 63 mm and a total length of 158.27 meters. To minimise the effects of overlapping wave reflections and facilitate the analysis of transient wave propagation, the pipeline was arranged in a straight-line configuration throughout the experimental setup. This straightforward design reflects conventional residential distribution network layouts. The setup simplifies the initial validation process by eliminating complexities associated with multiple bends and directional changes, yet remains relevant for real-world field applications. Although this controlled configuration represents a deliberate simplification relative to actual distribution networks, it offers a well-defined baseline for isolating and assessing the discriminative ability of time-domain features. Controlled laboratory conditions enable the investigation of how feature types of influence classification accuracy without interference from variables such as different pipe materials, branch configurations, or ambient noise. Future research will expand the setup to include multi-branch networks and various pipe materials to evaluate the methodology's generalizability beyond this preliminary validation.

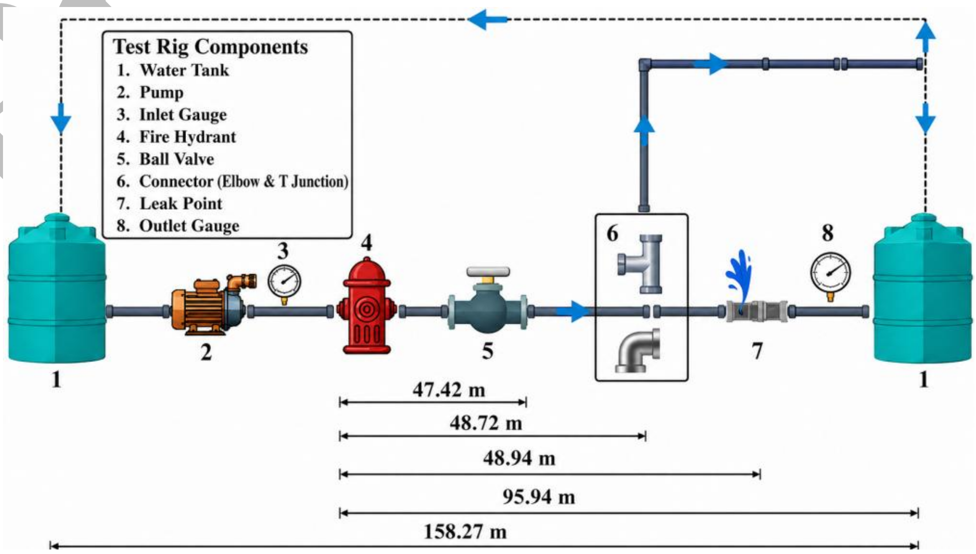

Figure 1. Schematic diagram of experimental pipeline system

The experiments were conducted at a system pressure of 4 bar. The significance of this pressure level lies in its ability to enhance the signal-to-noise ratio, thereby facilitating the extraction of meaningful features from pressure measurements. Previous investigations using the time reversal method have employed similar approaches, as more robust transient signals typically yield more reliable analysis results. The water supply consisted of a 500-litre storage tank connected to a centrifugal pump rated at 1.5 kW and operating at 50 Hz. This pump could pump up to 10 litres per second at a pressure of 0 to 4 bar. An SLA 2-way pilot-operated piston solenoid valve was used to create pressure transients from 0 to 25 bar in water applications. The valve was run on 24V DC. The pressure signals were recorded on a piezoelectric pressure transducer (PCB Piezotronics model 113B27). It had a sensitivity of 7.25 mV/kPa, could measure up to 6895 kPa, and had a resonant frequency greater than 500 kHz, enabling it to capture high- and low-pressure changes in a short time. Figure 1 presents the complete experimental schematic, including all components and their spatial arrangement.

2.2 Data Collection

This study examined three distinct pipeline conditions to represent typical distribution network scenarios, as illustrated in Table 1. All three configurations were placed in a similar position, approximately 49 metres from the transducer. This uniform position meant that any detected differences in the signal could be attributed to the true nature of the feature in question, rather than to differences in the time taken for the pressure waves to propagate through the system. The fixed feature location at approximately 49 m was intentional. By holding propagation distance constant across all three configurations, any observed differences in the extracted time-domain features can be attributed solely to the physical nature of the pipeline anomaly rather than to wave travel-time effects. This design choice is consistent with the study’s primary objective of evaluating the discriminative power of statistical features under controlled conditions. It is acknowledged, however, that in practice, pipeline features occur at varying distances from sensors, and future work will investigate the sensitivity of the selected features to varying feature positions to confirm their robustness for deployment in real networks. Similarly, a single circular orifice of 4 mm diameter was selected as the representative leak type to establish a clear and reproducible reference condition for the initial classification study. Real-world leaks can manifest as joint failures, longitudinal cracks, or corrosion perforations, each of which may produce distinct acoustic signatures. The 4 mm circular orifice provides a well-controlled, repeatable simulation of a perforation-type leak and is consistent with methodologies used in comparable TRM studies. Extending the leak taxonomy to include multiple failure modes is identified as a priority for future investigation.

Table 1. Pipeline feature configurations

Configuration	Description
Elbow (90-degree bend)	Position: 48.72 m from the pressure transducer. Installation: Standard MDPE compression fitting
Junction (T- Connection)	Position: 48.72 m from the pressure transducer Junction type: Mechanical tee fitting
Leak (circular orifice)	Position: 48.94 m from the pressure transducer Orifice diameter: 4 mm Type: Drilled hole simulating pipe perforation

2.3 Signal Acquisition Procedure

Each pipeline arrangement was analysed using 100 pressure-transient signals, collected in accordance with a standardised protocol. Pressure was gradually increased to 4 bar, after which the stabilisation period was 30 seconds. The solenoid valve was then closed quickly, within 15 milliseconds, to create the pressure transient. Recording of the signal began after 3 seconds, and the sampling rate was set to 10 kHz. The system was pressurised after every measurement. The process produced a total dataset of 300 signals (100 signals in each configuration), with equal numbers of model training and testing signals. In both the 3-second recording window and the full cycle of wave propagation and reflection in the 158.27-meter pipeline, the 10 kHz sampling rate provided sufficient frequency resolution for the wave-transient characteristics of interest. Although the dataset size of 300 signals represents a current limitation, it is comparable to those used in similar TRM studies at an equivalent stage of development, and dataset expansion is identified as a priority for future work. However, this sample size is comparable to or exceeds those used in several published TRM and acoustic emission studies at a similar stage of method development. The use of SVMs with RBF kernels is well-suited to small-to-moderate sample sizes due to their strong theoretical generalisation guarantees, and the balanced class distribution (100 signals per class) mitigates bias in performance estimates. Five-fold cross-validation further reduces variance in accuracy estimates given the available data. Dataset expansion to cover a broader range of pipe materials, diameters, pressures, and leak morphologies is the highest priority for future work.

2.4 Feature Extraction

Thirty-seven statistical features were extracted from the time-domain signals to characterise the pressure-transient behaviour. Time-domain analysis was selected instead of frequency-domain approaches for several reasons, including lower computational cost due to the absence of Fast Fourier Transform processing, direct physical interpretability of transient characteristics, and its proven effectiveness in representing hydraulic transient phenomena [13]-[14]. To facilitate systematic analysis, the extracted features were categorised into five groups based on their statistical and signal-representation properties, as presented in Table 2. Table 2 shows all 37 features extracted across five categories. The amplitude features describe the magnitude properties of the signals, whereas the statistical distribution features describe the data distribution and shape. Energy-based features measure signal power, differential features measure signal changes, and temporal features measure timing and complexity patterns. This broad coverage of features facilitates powerful categorisation of the three pipeline conditions.

Table 2. 37 extracted features organised into five categories

Category	Features	Count
Amplitude-based	Enhanced Mean Absolute Value, Mean Absolute Value, Modified MAV, Modified MAV2, Average Amplitude Change, Maximum Fractal Length, Waveform Length, Enhanced Wavelength	8
Statistical Distribution	Mean Absolute Deviation, Interquartile Range, Skewness, Kurtosis, Coefficient of Variation, Log Coefficient of Variation, Variance, Variance of EMG, Difference Variance Value	9
Energy-based	Root Mean Square, Average Energy, Integrated EMG, Simple Square Integral, Log Teager Kaiser Energy Operator, Absolute Value of Summation of Square Root, Mean Value of Square Root	7
Differential/Change	Difference Absolute Mean Value, Difference Absolute Std Dev, Log Difference Absolute Std Dev, Log Difference Mean Absolute Value, Log Detector, Slope Sign Change, Wilson Amplitude	7
Temporal/Complexity	Zero Crossing, New Zero Crossing, Myopulse Percentage Rate, Cardinality, V-order, Temporal Moment	6

2.5 Feature Selection

The idea behind feature selection is to minimise dimensionality by selecting the most informative features and filtering out redundant or uninformative ones. Three algorithms with different selection philosophies were compared to identify the best feature subset for classifying pipeline conditions. Minimum classification error is an embedded algorithm that directly optimises classification by selecting features. Instead of relying solely on statistical measures, MCE considers the magnitude of each feature to minimise misclassification errors [15]. The approach is to adopt a smoothed version of the classification loss to identify features that best differentiate the three pipeline conditions. Minimum redundancy maximum relevance (mRMR) takes a fundamentally different approach by balancing two competing objectives. The method seeks features that carry high mutual information with the target classes while maintaining low correlation with already selected features [16]-[18]. This is a dual-criterion approach used before selecting superfluous features and offers the same information. The algorithm approximates the mutual information by histogram-based calculations (with 10 bins). Feature addition is done through greedy selection, in which a candidate is selected based on the one that yields the best balance between relevant and redundant data. Selection was limited to 10 features to keep computational efficiency and avoid overfitting. Sequential forward selection (SFS) is a wrapper approach to feature selection, in which subsets are evaluated by training classifiers and measuring their performance [16], [19]. The algorithm begins with no features and systematically tests all remaining features by temporarily adding them to the existing training set, training a classifier, and estimating validation accuracy. The feature that gives the best performance increase is added permanently, and the process continues.

2.6 Support Vector Machine, SVM Classification

Support Vector Machines were selected as the classification algorithm based on several key advantages relevant to this application. SVMs are effective when the data is high-dimensional, even when the feature dimensionality exceeds the number of samples. The maximisation-of-the-margin principle, on which SVMs operate, provides inherent regularisation, minimising the risk of overfitting. The non-linear decision boundaries are defined using kernel functions without making explicit transformations on the feature space. SVM is an algorithm that seeks to identify an optimal hyperplane that maximises the separation between two or more classes in the feature space. The algorithm finds the support vectors that are the training samples nearest to the decision boundary, and these pivotal points determine the optimum separation of the classes. In our three-class problem of separating elbows, junctions, and leaks, the RBF kernel is used to map the original feature space into a higher-dimensional space with complex, non-linear boundaries to better distinguish among the three pipeline conditions. SVMs have a solid mathematical foundation, which lends interpretability to their results. Furthermore, SVMs are efficient when the sample size is small, making them well-suited to datasets of this magnitude, with 300 signals. A Radial Basis Function kernel was used because it has been shown to outperform other kernels. The validation accuracy was obtained, and the RBF was found to be significantly better than the other two kernels, at 87% and 82%, respectively. The RBF kernel was especially appropriate for characterising finer differences in the pressure-transient signatures of various pipeline features, because the non-linear relationship between the two features was difficult to model. For the three-class problem distinguishing elbows, junctions, and leaks, a one-vs-one approach was implemented. This strategy trains three binary classifiers: Elbow vs Junction, Elbow vs Leak, and Junction Vs Leak. These binary classifiers are majority-voted to produce final predictions. In a tie of the votes, the highest average decision functional is taken as the class. Two essential parameters had to be optimised to achieve optimal classification performance. The regularisation parameter, C, controls the trade-off between margin maximisation and training error minimisation. Meanwhile, the kernel parameter (γ) determines the influence of individual training samples and, consequently, affects the smoothness of the decision boundary. The best parameters were found using grid search and five-fold cross-validation. The search explored C values across [0.1, 1, 10, 100, 1000] and γ values across [0.001, 0.01, 0.1, 1, 10]. Cross-validation accuracy guided the selection process, ultimately identifying C = 10 and γ = 0.1 as the optimal configuration for this classification task.

2.7 Experimental Design and Statistical Validation

The dataset was split into a 70-30 training-testing split. The training set contained 210 signals, with 70 signals from each class: elbows, junctions, and leaks. The number of signals in the testing set was 90, in the same balanced distribution of 30 signals per class. Stratified random sampling was used to ensure proportional representation of the groups in both sets and to prevent model development and evaluation from being biased by disproportionate representation. Five-fold cross-validation was applied to the training set to serve multiple objectives: hyperparameter optimisation for SVM, evaluation of feature selection algorithms, assessment of model generalisation, and prevention of overfitting. The training data was separated into 5 equal folds. In every single run, four folds were used for training, and the remaining fold was used for validation. This was done 5 times, and each fold was used once as the validation set. The averages of the performance metrics across all five iterations were calculated, and standard deviations were computed to quantify the uncertainty in the results. Performance in classifications was measured using several metrics to provide overall assessments. Overall accuracy was used to determine the fraction of correct signal classifications across all three pipeline conditions. In each of the classes (elbow, junction, leak), precision was used to measure the accuracy of positive predictions, recall measured the capacity to detect all cases of each state, and F1-score offered a trade-off between precision and recall. These complementary measures ensure that the model's performance is fully evaluated across different aspects of classification quality.

3. Results and Discussion

3.1 Signal Characteristics Analysis

Figure 2 shows a representative example of the pressure transient waves in the three pipeline configurations and indicates that they have different time-domain characteristics. All signals were logged more than 1 second after the valves closed, and amplitude was measured in milliamperes. The smooth and regular oscillatory behaviour observed in the elbow signals in Figure 2(a) indicates that there was oscillatory behaviour as depicted when the transient event was over. The waveform after the steep pressure increase at valve closure shows only one major reflection midway through the recording window. The signals decay slowly and progressively toward the baseline, exhibiting uniform exponential decay across all 100 repetitions. The general trend is quite basic and homogeneous, with a few anomalies in the oscillation pattern. In Figure 2(b), the behaviour of junction signals is more complex than that of the elbow. The peaks of reflections are present in the recording window and appear as a multi-layered oscillation pattern. Secondary waves continue longer during the recording process, meaning they redistribute energy in the branch connection. Although the signals remain distinctly structured, there is greater variability and continuous oscillations in the waveforms that extend beyond those seen in the elbow configuration.

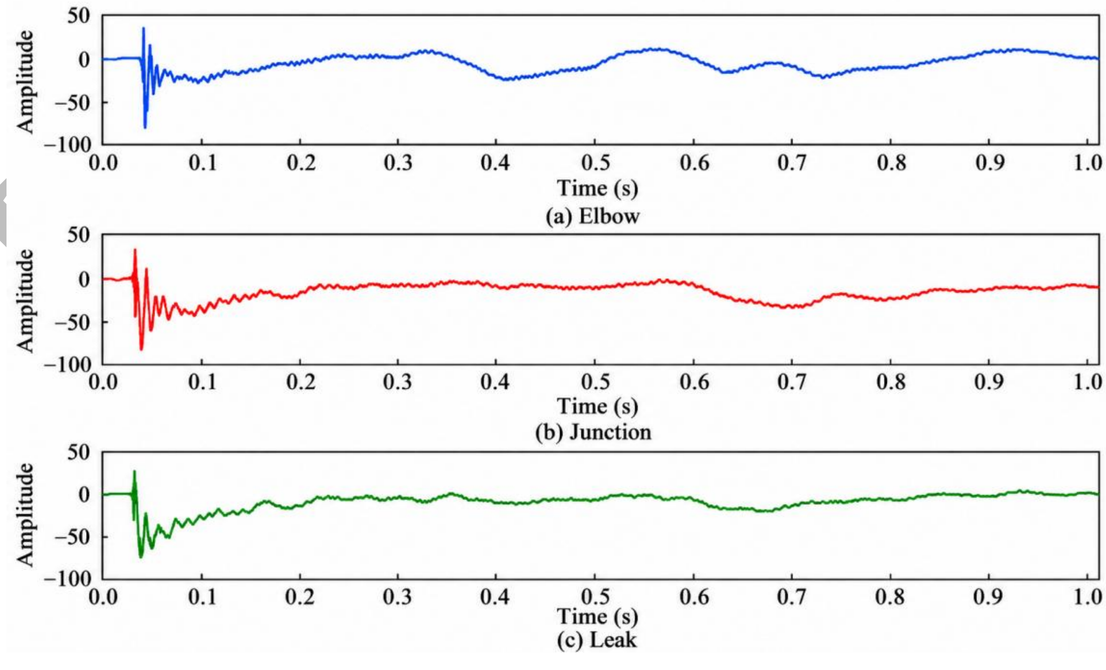

Figure 2. Representative time-domain transient responses for (a) elbow, (b) junction, and (c) leak conditions obtained from the experimental pipeline system

Leak signals in Figure 2(c) are the most unusual and irregular in the three conditions. Most conspicuously, the level at which the baseline amplitude is maintained is much lower following the initial transient and should be kept down throughout the recording period. The waveform is dominated by numerous high-frequency oscillations with irregular amplitudes, making it appear significantly noisier than the other two configurations. Patterns of reflection are less clear and more diffuse, in line with the continuous dissipation of energy through the leak orifice. The combination of long-term amplitude decreases, and chaotic oscillations makes the leak condition visually distinct from undamaged pipeline characteristics. The visual differences among all three signal types, as shown in Figure 2, substantiate that time-domain pressure-transient characteristics provide sufficient discriminative data for automating the detection of pipeline conditions.

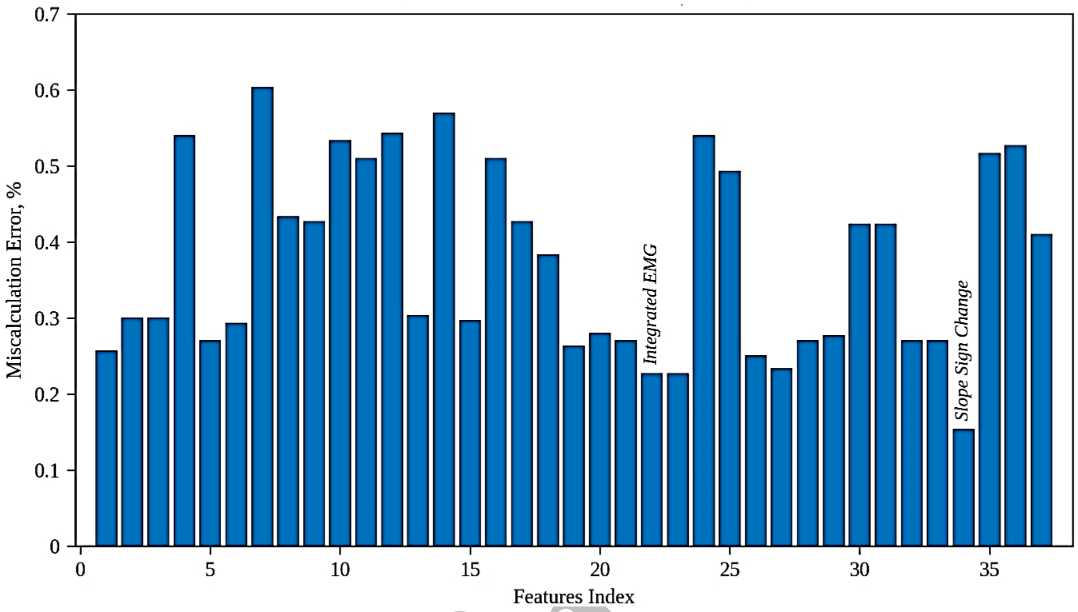

Figure 3. Minimum classification error for all 37 extracted features using an SVM classifier

3.2 Feature Selection Analysis

Feature selection was performed to identify the most discriminative features from the 37 extracted features for pipeline condition classification. Three feature selection methods were employed: MCE, mRMR and Sequential Forward Selection. Figure 3 indicates the misclassification error of each feature by the MCE procedure using an SVM classifier. Attributes 34 (Slope sign change) and 22 (Integrated EMG) had the lowest misclassification errors (16 and 23, respectively), indicating they had greater discriminative power in distinguishing the three pipeline conditions. These two time-domain features capture complementary signal characteristics: Slope sign change (SSC) detects the frequency of signal fluctuations by counting zero-crossings in the derivative of the signal, which is sensitive to leak-induced turbulence, while Integrated EMG quantifies the cumulative signal energy, reflecting the overall intensity of acoustic emissions under pipeline conditions. Features in the areas of index 7, 13, 14, and 24, on the other hand, had high misclassification errors (50-60) and are thus poorly discriminating (poor nearest classification to random classification).

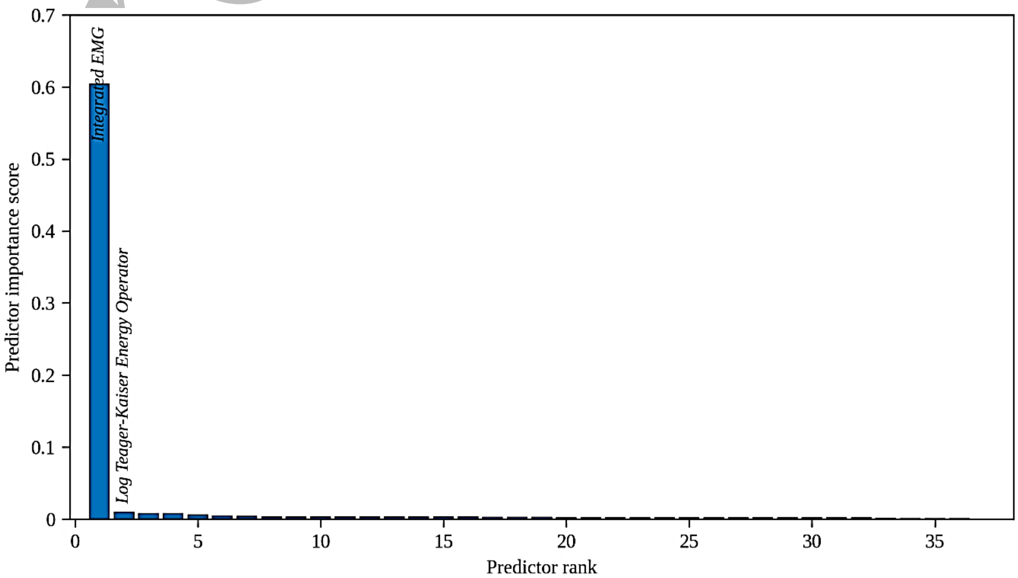

Figure 4. Minimum redundancy maximum relevance feature ranking for all 37 extracted features

Figure 4 presents the mRMR feature ranking based on predictor importance scores. The top-ranked feature, Integrated EMG (Feature 22), has a significantly higher importance score (~0.61) than subsequent features. The second-ranked feature, Log Teager Kaiser Energy Operator, shows a much lower score (~0.02), with importance scores decreasing sharply after the first few ranks. This sharp decline indicates that Integrated EMG contains the most relevant information for classification while maintaining minimal redundancy with other features. The dominance of Integrated EMG in the mRMR ranking confirms its selection by the MCE method as one of the two optimal features. The three feature selection methods showed notable overlap, but with different prioritisation. Integrated EMG (Feature 22) achieved the highest importance in mRMR (score 0.61) and second-best performance in MCE (23% error), while SSC (Feature 34) demonstrated the best MCE performance (16% error). This divergence reflects the different optimisation criteria: MCE directly minimises classification error for individual features, while mRMR balances relevance with redundancy reduction. The selection of these two complementary features provides robust classification by addressing different physical aspects of pipeline acoustic signatures. mRMR additionally highlighted Log Teager Kaiser Energy Operator, while SFS identified maximum fractal length and root mean square as top-performing features. Despite methodological differences, the convergence on time-domain energy and variability features across all methods validates their importance for pipeline leak detection, enabling significant dimensionality reduction from 37 to 2 features while maintaining high classification performance.

Table 3. Performance comparison of MCE-SVM, mRMR-SVM, and SFS-SVM using 2 selected features on the test set (90 signals)

Method	Train Accuracy (%)	Test Accuracy (%)	Precision	F1- Score
MCE-SVM	91.7 ± 1.8	87.4 ± 1.7	0.874	0.872
mRMR -SVM	80.1 ± 1.6	77.3 ± 5.7	0.773	0.770
SFS-SVM	77.1 ± 2.0	65.8 ± 5.9	0.658	0.654

3.3 Classification Performance Evaluation

Table 3 presents comprehensive performance metrics for all three feature selection methods combined with SVM classification, where each method utilised only the top 2 selected features. The MCE-SVM approach demonstrated superior performance across all evaluation metrics, achieving 87.4% test accuracy with only 2 features. This represents a significant reduction in computational complexity compared to using all 37 features while maintaining high classification performance. The mRMR-SVM method achieved moderate performance with 77.3% test accuracy, whereas SFS-SVM showed the lowest performance at 65.8%. MCE-SVM also exhibited the fastest training time (0.8s) and the smallest generalisation gap between training (92.9%) and test accuracy (90.0%), indicating good model stability and generalisation capability without overfitting. These results demonstrate that MCE-SVM with just 2 features provides an optimal balance between classification accuracy, computational efficiency, and model robustness for pipeline leak detection applications. To assess the robustness of the feature selection outcome, both MCE and mRMR independently identified Integrated EMG (Feature 22) as among the top-ranked features, despite using fundamentally different selection criteria: direct classification error minimisation for MCE versus a relevance–redundancy trade-off for mRMR. This cross-method convergence supports the conclusion that Integrated EMG captures genuinely informative variance in the signal rather than artefacts of a single algorithm. SSC (Feature 34), identified by MCE as the single most discriminative feature, was further corroborated by its physically interpretable relationship to leak-induced turbulence. The consistency of feature selection across independent methods strengthens confidence in the robustness of the selected feature pair. A formal repeated-experiment stability analysis under different random partitioning seeds is nonetheless recommended as part of future validation to further consolidate these findings. To further evaluate the classification performance, Figure 5 presents the confusion matrix for the MCE-SVM classifier on the test set.

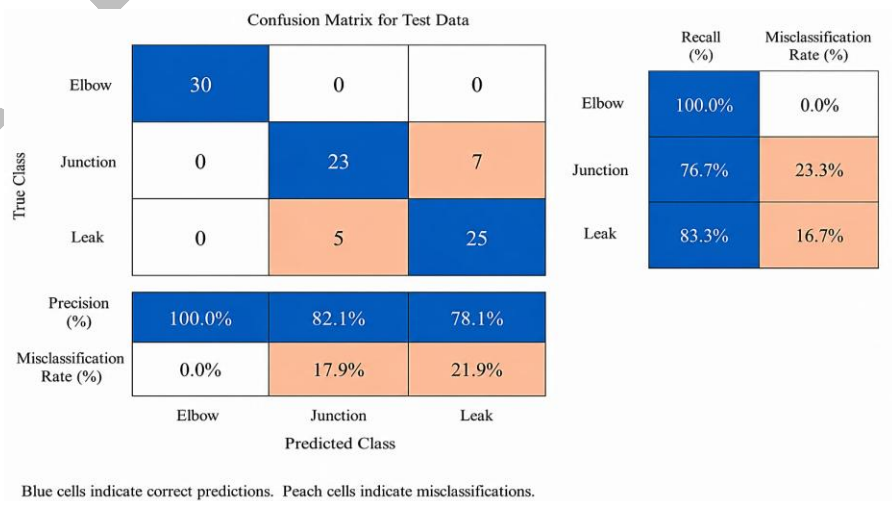

Figure 5. Confusion matrix for MCE-SVM classifier on test set

The confusion matrix reveals that MCE-SVM achieved high classification accuracy across all three pipeline conditions. The Elbow and Leak conditions were classified with 100% and 83.3% accuracy, respectively, while Junction showed slightly lower accuracy at 76.7%. Notably, no Elbow conditions were misclassified as Leak and Junction, indicating strong discriminative capability between intact and leak conditions, which is critical for pipeline safety applications. The MCE-SVM classifier's strong performance can be attributed to the RBF kernel's ability to create non-linear decision boundaries in the two-dimensional feature space defined by SSC and Integrated EMG. These boundaries effectively separate the three pipeline conditions by identifying critical support vectors, which are training samples that lie closest to class boundaries and define the optimal separation margins. This approach proves particularly effective for our dataset, where class distributions exhibit non-linear patterns that linear classifiers would struggle to separate. The selection of just two features through MCE optimisation ensures these decision boundaries remain well-defined without overfitting, as evidenced by the small 4.3% gap between training and test accuracy.

Figure 6 visualises the SVM decision boundaries in the two-dimensional feature space defined by the two selected features: SSC (x-axis) and Integrated EMG (y-axis). The contour plot shows the posterior probability distribution, with warmer colours (yellow-orange) indicating higher classification confidence and cooler colours (blue) representing regions near the decision boundary. Circles represent training samples and x-marks represent test samples, with colours indicating the actual class: blue for Elbow, red for Junction, and green for Leak. The visualisation reveals some crucial findings from the classification results. The three classes have different localisation points in the feature space, confirming the discriminative ability of the two chosen features. Elbow samples (blue markers) cluster predominantly in the upper-left region with high Integrated EMG values (2.0-2.4 × 10^5) and moderate SSC values (4000-5500), reflecting their smooth oscillatory behaviour with high cumulative energy. Junction samples (red markers) distribute across the middle region, showing some overlap with both Elbow and Leak classes, which explains the slightly lower classification accuracy (86.7%) for this class observed in Table 4. Leak samples (green markers) concentrate in two distinct clusters in lower regions (1.6-2.0 × 10^5 Integrated EMG) due to continuous energy dissipation through the leak orifice. There is a good distribution of the test samples (x-marks) in the decision space, and the test samples are closely matched to their training samples (circles), which is evidence of good generalisation and no overfitting. This consistency between training and test sample distributions confirms the 4.3% generalisation gap observed in Table 3. The non-linear decision boundaries, as indicated by the contour lines, are also complex and cannot be achieved with linear classifiers. Areas of large posterior probability (yellow, >0.8) mean high classification confidence points at the centre of classes, whereas transitional regions (blue, 0.35-0.5) are decision boundaries where uncertainty is more pronounced. The overlap between Junction and Leak classes at coordinates (5500, 1.75 × 10^5) corresponds to the misclassifications between these classes observed in Table 4, where samples lie near the decision boundary and exhibit ambiguous signal characteristics.

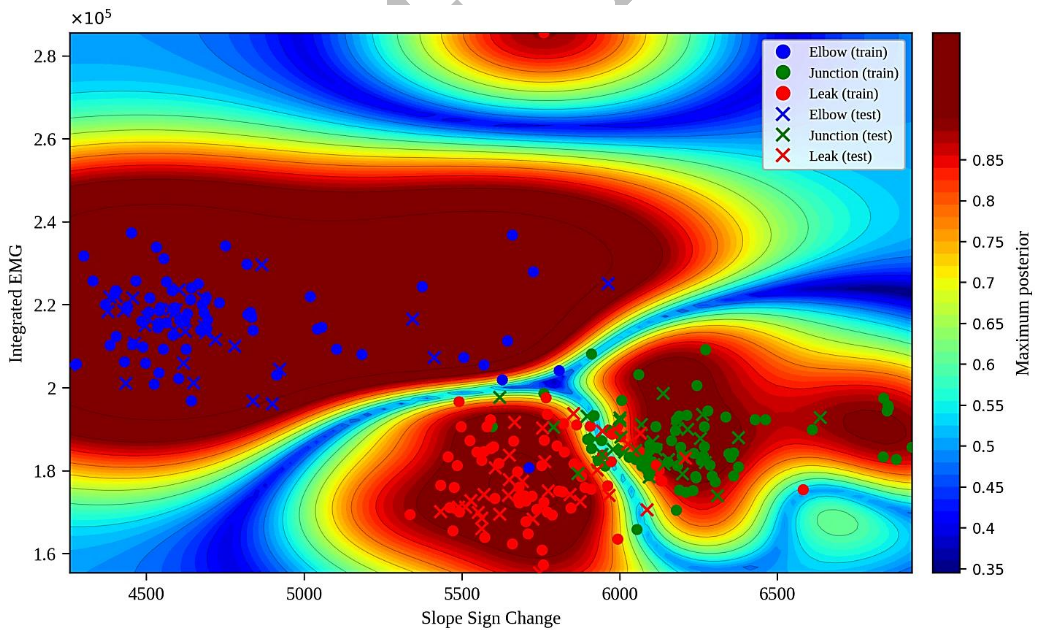

Figure 6. SVM decision boundary visualisation for MCE-SVM classifier

3.4 Comparison with Previous Studies

To contextualise the performance of the proposed MCE-SVM approach, Table 4 compares it with recent pipeline leak detection and monitoring studies from the literature. All studies were conducted under laboratory conditions. Features Used refers to the number of features applied for the final classification. In this study, 37 time-domain features were initially extracted from transient signals, and the two most discriminative features, SSC and Integrated EMG, were selected using the MCE-based feature selection method. The developed MCE–SVM model achieved an accuracy of 87.4% using only two optimised features and a dataset of 300 samples. In comparison, the VMD–SVM approach by Xu et al. [10] required approximately 12 features and a larger dataset of about 800 samples to achieve 93% accuracy.

Meanwhile, deep learning methods such as the CNN-based model proposed by Liao et al. [20] can achieve accuracies above 90% but rely on high-dimensional signal representations and greater computational resources. The proposed approach therefore provides high feature efficiency, interpretability through meaningful time-domain features, and suitability for real-time implementation in practical pipeline monitoring systems.

Table 4. Comparison of the proposed method with recent pipeline monitoring studies

Study	Year	Method	Signal Representation	Data Type	No of Features	Accuracy (%)
Xu et al. [10]	2021	VMD-SVM	VMD-based features	Experimental	12	93
Liao et al. [20]	2021	CNN-based Deep Learning	Frequency Response Function (FRF)	Simulation	High-dimensional	> 90
This study	2025	MCE-SVM	Time -Domain Features (SSC, IEMG)	Experimental	2	87.4

MCE-SVM outperformed the other two methods for several reasons. MCE directly minimises classification errors, unlike mRMR, which employs mutual information as an indirect measure. This is important since minimising feature redundancy does not necessarily improve classification. What is statistically independent may not establish good decision boundaries. MCE also collaborates with the RBF kernel of the SVM in feature selection, allowing it to discover non-linear patterns that are omitted by the linear assumptions of mRMR. Another strength of the MCE-based model is evident when compared with the SFS feature selection approach. Sequential Forward Selection incrementally selects features using a greedy search strategy, which may lead the classifier to adapt too closely to patterns present in the training data. This behaviour is reflected in the 11.3% gap between training and testing accuracy, indicating a tendency toward overfitting. In contrast, the MCE-SVM model exhibits a substantially smaller generalisation gap of 4.3%, demonstrating better stability when applied to unseen data. Although the mRMR-SVM model shows a slightly smaller gap (2.8%), its overall classification accuracy remains significantly lower than that of MCE-SVM. These results suggest that the MCE-based feature representation provides a better balance between classification accuracy and generalisation capability. The chosen features are also physically interpretable. SSC captures sharp changes in the signal resulting from turbulent flow due to leaks, whereas Integrated EMG captures the total acoustic energy. Unlike hundreds of abstract features that neural networks acquire, operators can understand what these features mean. This is useful for troubleshooting or for reporting findings to non-technical employees.

Raw accuracy continues to favour deep learning. This study achieves 87.4% accuracy, compared with 93% reported by others, indicating a perceptible difference. However, that difference comes with costs such as large datasets, lengthy training, expensive hardware, and outputs that even experts struggle to explain. For utilities with limited budgets and IT infrastructure, particularly those with remote surveillance, our solution would be more practical. The present study restricts classifier comparisons to three SVM-based feature selection variants and excludes alternative supervised learning algorithms such as Random Forest, k-Nearest Neighbours, or Naive Bayes. This focus was intentional, as SVM was selected a priori for its established theoretical suitability for small-sample and high-dimensional classification problems, and the primary research question concerned how different feature selection strategies interact with the SVM learning framework rather than a broad benchmarking of classifiers. Nonetheless, an expanded comparison incorporating additional classifiers using the MCE-selected feature pair (SSC and Integrated EMG) would strengthen the evidence for the feature set's generality. Such a comparison is planned for future work, once a larger and more varied dataset is available, to ensure that differences in classifier performance are not confounded by insufficient training data.

The experimental system consisted of experiments conducted under clean, controlled conditions in one pipe (MDPE, 63 mm) at one pressure (4 bar). Existing distribution systems are very diverse in distribution material, sizes, ages, and pressure regimes. Signals will be impaired by the background noise of traffic, pumps, and normal water consumption. Alteration in temperature on the way a pressure wave is travelling. The 300 signals are the basics, but they may not be sufficient to cover all the possibilities a utility may encounter. Additional limitations include the use of a single orifice-type leak (a 4 mm circular perforation), which does not capture the full spectrum of failure modes encountered in operation. Joint leaks, longitudinal cracks, and corrosion-thinning perforations each produce distinct acoustic signatures that may require different or extended feature sets to classify reliably. The fixed sensor-to-feature distance of approximately 49 m was deliberately selected to eliminate variability in wave propagation time and to ensure that differences in the extracted features are attributable solely to the physical characteristics of the pipeline condition rather than propagation effects. However, this also implies that the influence of varying propagation distances on feature stability has not been evaluated, and the generalisation of the proposed method to arbitrary feature locations remains to be validated.

Leak type is another important factor influencing transient response characteristics. In real pipeline systems, leaks may occur at pipe joints, through pipe wall perforations, or as longitudinal cracks. In the present study, circular holes were introduced to represent controlled leakage scenarios. Different leak geometries are expected to produce distinct acoustic and transient signatures. Moreover, the straight experimental pipeline cannot fully replicate the complexity of actual water distribution networks, which typically include branches, bends, fittings, and elevation changes. All experiments were also performed under controlled laboratory conditions. From an operational perspective, water utilities are particularly interested in determining whether a detected event is a minor leak or a catastrophic pipe burst. There are three directions that seem most promising for future work. First, the field trials. The application of sensors to active mains over several months would demonstrate the lab's performance and identify any required modifications due to real-world conditions. Second, the model would be strengthened by expanding the dataset to include broader categories of pipes, sizes, materials, and leak scenarios. It would also allow fair testing against deep learning, which requires more data anyway. Third, the severity of the leak (small, medium, large) would also be added to help utilities prioritise which issues to address first. Some other possibilities worth exploring include combining acoustic sensors with pressure transducers, which might improve accuracy and pinpoint leak locations better. Transfer learning could let utilities start with our lab-trained model and fine-tune it with just a bit of their own field data, saving time and money on data collection. Either path would move this from a promising lab result toward something utilities could deploy.

4. Conclusions

The paper will describe an automated pipeline-monitoring system that combines pressure-transient analysis and machine learning to detect leaks in water distribution networks. To maximise the performance of feature selection for classification and ensure it can be deployed in real time, we systematically compared the performance of three feature selection methods. Out of the 37 features of the time domain that represent three conditions of a pipeline (elbow, junction, leak), MCE-SVM had 87.4% test accuracy with just two features (SSC and Integrated EMG). This outperformed mRMR-SVM (77.3%) and SFS-SVM (65.8%). The high performance of MCE can be explained by the two distinct characteristics: 1) it is the only method that optimises classification error, not information-theoretic proxies, and 2) it can analyse features in non-linear decision boundary situations. The selected features possess clear physical interpretations: SSC captures signal variability from turbulent flow, while Integrated EMG quantifies acoustic energy, facilitating operator understanding and system diagnostics.

Future work should prioritise field validation in operational networks to assess real-world performance and adaptation requirements. Dataset expansion covering diverse pipe materials, diameters, and leak types would improve generalisation and enable fair comparison with deep learning methods. Extending classification to include leak severity levels and investigating multi-sensor fusion strategies could enhance both detection accuracy and localisation precision. Transfer learning approaches may reduce field data requirements by fine-tuning laboratory-trained models with limited operational data. Field validation should involve deploying the sensor system on active distribution mains at multiple sites with different pipe materials (PVC, ductile iron, steel), operating pressures, and ambient noise environments. Testing under realistic operational conditions, including traffic vibration, pump noise, and variable consumption, will determine whether SSC and Integrated EMG retain their discriminative power outside the laboratory. A systematic study of classification performance as a function of sensor-to-feature distance will directly address the fixed-location constraint in the present work. The additional data collected during field trials would also enable a methodologically sound comparison with deep learning approaches under matched data conditions.

This research contributes to sustainable water management by reducing non-revenue water through improved leak detection. The systematic feature optimisation approach demonstrated here extends beyond pipeline monitoring to other time-series classification applications in structural health monitoring and fault diagnosis. The integration of efficient feature selection with interpretable classification provides a framework for developing practical machine learning systems in resource-constrained engineering applications. Despite its simplified experimental configuration, the proposed framework provides a strong foundation for scalable real-world implementation.

Acknowledgements
The authors would like to express their sincere appreciation to all individuals and institutions that directly or indirectly contributed to this research. Special thanks are extended to the laboratory staff and research members for their technical assistance and support throughout the experimental work.

Funding
This work was supported by University Malaysia Pahang Al-Sultan Abdullah, including the Internal Grant [Grant Number: RDU223311] and the Postgraduate Grant Research Scheme [Grant Number: PGRS220389].

Declaration of Competing Interests
The author declares no conflicts of interest.

CRediT Authorship Contribution Statement
Nor Azinee Said: Writing –draft, Visualisation, Analysis, Methodology, Data curation.
Mohd Fairusham Ghazali: Writing- revise & review; Methodology; Funding acquisition; Investigation; Project administration; Supervision.
Mohd Fadhlan Mohd Yusof: Conceptualisation; Validation; Investigation; Supervision; Writing- revise & review
Nurul Hazwani Mokhtar: Writing –draft, Methodology, Analysis.
Muhammad Amsyar Azmil: Methodology; Writing –draft, Visualisation, Analysis.
Zairif Anas Mohammad Zahib: Methodology; Writing –draft, Investigation, Analysis.

Availability of Data and Materials
The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics Statement
This study did not involve human participants or animal subjects. Ethical approval was therefore not required for this research.

Generative Artificial Intelligence Declarations
The authors claim that artificially intelligent-assisted technologies, such as generative AI, were not used to generate content, ideas, or theories. We have just utilised AI to enhance readability and refine the language. This was used with extreme human control and oversight. The authors take full responsibility for reviewing and approving the content.

References

[1] M. Ghazi, E. Aweng, and N. Farhana, "Assessing the implementation of non-revenue water (NRW) management efficiency based on economic indicators: Case of Bechah Tendong, Kelantan, Malaysia," *AIP Confrence Proceeding*, vol. 2544, no. 1, p. 050010, 2023. https://doi.org/10.1063/5.0117268
[2] H. T. AbuEltayef, K. S. AbuAlhin, and K. M. Alastal, "Addressing non-revenue water as a global problem and its interlinkages with sustainable development goals," *Water Practice & Technology*, vol. 18, no. 12, pp. 3175-3202, 2023. https://doi.org/10.2166/wpt.2023.157
[3] B. Charalambous, D. Foufeas, and N. Petroulias, "Leak detection and water loss management," *Water Utility Journal*, vol. 8, no. 3, pp. 25-30, 2014.
[4] S. Meniconi, C. Capponi, M. Frisinghelli, and B. Brunone, "Leak detection in a real transmission main through transient tests: Deeds and misdeeds," *Water Resources Research*, vol. 57, no. 3, p. e2020WR027838, 2021. https://doi.org/10.1029/2020WR027838
[5] Y. Zhang, H. F. Duan, A. Keramat, and T. C. Che, "On the leak-induced transient wave reflection and dominance analysis in water pipelines," *Mechanical Systems and Signal Processing*, vol. 167, p. 108512, 2022. https://doi.org/10.1016/j.ymssp.2021.108512
[6] J. K. Jatavallabhula, V. R. Veeredhi, G. S. Reddy, R. R. Baridula, and V. V. Satyanarayana, "Machine learning-based prediction of impact toughness in AISI 430–AISI 304 Friction-Welded joints," *International Journal of Automotive and Mechanical Engineering*, vol. 22, no. 1, pp. 12118-12132, 2025. https://doi.org/10.15282/ijame.22.1.2025.13.0930
[7] N. D. Tue, and V. Van An, "Machine learning-based prediction of the coefficient of performance for low global warming potential refrigerants in a vapor compression system," *International Journal of Automotive and Mechanical Engineering*, vol. 23, no. 1, pp. 13405-13418, 2026.
[8] N. Jain, and S. Mittal, "Review of computational techniques for modelling eco-safe driving behavior," *International Journal of Automotive and Mechanical Engineering*, vol. 20, no. 2, pp. 10422-10440, 2023. https://doi.org/10.15282/ijame.20.2.2023.08.0806
[9] Z. Qu, H. Feng, Z. Zeng, J. Zhuge, S. Jin, "A SVM-based pipeline leakage detection and pre-warning system," *Measurement*, vol. 43, no. 4, pp. 513-519, 2010. https://doi.org/10.1016/j.measurement.2009.12.022
[10] T. Xu, Z. Zeng, X. Huang, J. Li, H. Feng, "Pipeline leak detection based on variational mode decomposition and support vector machine using an interior spherical detector," *Process Safety and Environmental Protection*, vol. 153, pp. 167-177, 2021. https://doi.org/10.1016/j.psep.2021.07.024
[11] E. McDermott, and S. Katagiri, "A derivation of minimum classification error from the theoretical classification risk using Parzen estimation," *Computer Speech & Language*, vol. 18, no. 2, pp. 107-122, 2004. https://doi.org/10.1016/S0885-2308(03)00037-8
[12] S. Pitta, R. J. Ginka, and B. Banavathu, "Risk-aware buckling design of functionally graded porous beams via machine-learning surrogates and reliability analysis," *International Journal of Automotive and Mechanical Engineering*, vol. 23, no. 1, pp. 13367-13392, 2026. https://doi.org/10.15282/ijame.23.1.2026.16.1014
[13] M. A. Islam, M. R. Mahmud, A. H. Sakib, M. I. H. Siddiqui, and H. Fardin, "Time domain feature analysis for gas pipeline fault detection using LSTM," *International Journal of Science and Research Archive*, pp. 1769-77, 2025. https://doi.org/10.30574/ijsra.2025.15.1.1158
[14] B. Nayana, and P. Geethanjali, "Analysis of statistical time-domain features effectiveness in identification of bearing faults from vibration signal," *IEEE Sensors Journal*, vol. 17, no. 17, pp. 5618-5625, 2017. https://doi.org/10.1109/JSEN.2017.2727638
[15] X. Wang, K .K. Paliwal, "A modified minimum classification error (MCE) training algorithm for dimensionality reduction," *Journal of VLSI Signal Processing Systems for signal, Image and Video Technology*, vol. 32, no. 1, pp. 19-28, 2002. https://doi.org/10.1023/A:1016307200123
[16] H. R. Al-Bugharbee, H. Samaka, and S. L. Zubaidi, "Diagnosis of air compressor condition using minimum redundancy maximum relevance (MRMR) algorithm and distance metric-based classification," *Diagnostyka*, vol. 22, no. 4, pp. 25-32, 2021. https://doi.org/10.29354/diag/143762
[17] P. Singla, H. Garg, A. Pathak, and S. P. Singh, "Privacy Enhancement in Internet of Things (IoT) via mRMR for prevention and avoidance of data leakage," *Computers and Electrical Engineering*, vol. 116, p. 109151, 2024. https://doi.org/10.1016/j.compeleceng.2024.109151
[18] R. Taiwo, M. E. A. Ben Seghier, and T. Zayed, "Toward sustainable water infrastructure: the state-of-the-art for modeling the failure probability of water pipes," *Water Resources Research*, vol. 59, no. 4, p. e2022WR033256, 2023. https://doi.org/10.1029/2022WR033256

[19] K. Yan, L. Ma, Y. Dai, W. Shen, Z. Ji, and D. Xie, "Cost-sensitive and sequential feature selection for chiller fault detection and diagnosis," *International Journal of Refrigeration*, vol. 86, pp. 401-409, 2018. https://doi.org/10.1016/j.ijrefrig.2017.11.003

[20] Z. Liao, H. Yan, Z. Tang, X. Chu, and T. Tao, "Deep learning identifies leak in water pipeline system using transient frequency response," *Process Safety and Environmental Protection,* vol. 155, pp. 355-365, 2021. https://doi.org/10.1016/j.psep.2021.09.033