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-support 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-support vector machine (SVM) (77.3%) and sequential forward selection-SVM (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 (NRW)
Transient reflection method (TRM)
Minimized classification error (MCE)
Support vector machine (SVM)
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 the developing nations, the old pipeline system is a significant cause of these losses. In 2023, Malaysia had 37.1% NRW, with a vast amount compared to the 15% international water association standard [1]. Over 70 percent of NRW is caused by physical leaks related to the failure of the pipeline, and the effects on the different regions could be very large. During the period 2015 to 2017, the Sabah state alone reported RM657 million in water leakage losses [2]. These problems are in line with the United Nations Sustainable Development Goal 6 which highlights the need to have universal access to clean water by 2030. The causes of the leaks in the pipes are varied and comprise ageing of infrastructure, movement of the ground, deterioration of materials used in construction, excessive pressure applied to the pipes, and defects during installation. Utilities are able to detect issues, as well as minimise loss of water, minimise the costs of repair, and achieve reliability of the services [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 that contain several pipe characteristics and different operational characteristics. The Transient Reflection Method (TRM) has some benefits in detecting leakage by observing the behavior of pressure waves within a pipeline [4]. When a transient pressure wave hits any discontinuities like a leak, elbow, or junction, there are characteristic reflections. TRM identifies the existence of these reflections to recognise the pipeline characteristics and determine the condition of their system. Although successful when applied in very simple pipeline systems, TRM interpretation is difficult when there are many discontinuities in a network, where overlapping reflections can become a problem in assessing the signal [5].

Machine learning offers possible remedies for the automated interpretation of signals used in 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].The Support Vector Machines (SVMs) have proved useful in the pattern recognition tasks that involve high-dimensional data that have very few samples [9, 10]. Past research studies by Qu et al. used SVMs on a wavelet packet decomposition pipeline for leak detection, and were able to achieve a classification accuracy of over 85% [11]. Xu et al. applied variation mode decomposition to SVM in leak identification, where 92% accuracy was obtained [12]. Yet, these studies have used manually chosen feature sets that were not optimised systematically, which might be restrictive to the classification results and practicality. In vibration and sensor-based monitoring systems, artificial neural networks and recurrent neural networks have demonstrated strong capability in recognising dynamic fault patterns and extracting temporal dependencies

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

from sequential data [13, 14]. Yet, these studies have used manually chosen feature sets that were not optimized systematically, which might be restrictive to 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 [15]. 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 [16]. Despite the theoretical advantages of MCE, limited studies have systematically compared its performance with other feature selection approaches in transient response monitoring (TRM)-based pipeline monitoring applications. A review of existing pipeline monitoring research shows that most studies work with relatively small feature sets, typically examining between four and ten characteristics, without thoroughly exploring the full range of time-domain features available. Moreover, the scholars did not make a systematic comparison of the various feature selection techniques to find out which one suits best when classifying TRM signals. The other significant factor that has to be put into consideration is that very few studies have investigated the implementation of the balance between the necessity to have a correct classification and whether systems need to execute in real time within a real-life pipeline network or not.

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, classification models based on support vector machines will be developed to differentiate between 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 possibility of real-time system monitoring when computational facilities are constrained. 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, revealing not only the presence of an anomaly but also its type.

2. Materials and Methods

2.1 Experimental Setup

A testing facility in the form of a pipeline was developed and built as a simulation of the conditions of operation of water distribution, so that it could be experimented with. The component specifications and system configuration are as given below.

2.1.1 Pipeline Configuration

The experimental pipeline was constructed using medium-density polyethene (MDPE) material with a diameter of 63 mm and an overall pipeline length of 158.27 m. To minimise the influence of overlapping wave reflections and simplify transient wave propagation analysis, the pipeline was configured in a straight-line layout throughout the experimental setup. This simple design resembles typical residential distribution network designs. The setup simplifies efforts needed in initial validation since it removes the level of complexity of multiple bends and directional changes, and remains of interest to real-world field applications. While this controlled configuration represents a deliberate simplification relative to real distribution networks, it provides a well-defined baseline for isolating and evaluating the discriminative capacity of time domain features. Controlled laboratory conditions allow the effect of feature type on classification accuracy to be studied without confounding variables such as varying pipe materials, branch topology, or ambient noise. Future work will extend the setup to multi-branch networks and diverse pipe materials to assess how the methodology generalises beyond this initial validation.

2.1.2 Operating Conditions

The experiments had a low pressure of 4 bar being maintained in the system. The significance of this level of pressure is to enhance the signal-to-noise ratio, and thus, it is easy to derive meaningful features in pressure measurements. Previous TRM investigations have adopted similar approaches, as stronger transient signals generally lead to more reliable analysis outcomes [17]-[18].

2.1.3 System Configuration

Water Supply System:
The water supply consisted of a 500-litre capacity storage tank connected to a centrifugal pump with a power rating of 1.5 kW operating at 50 Hz. This pump was capable of pumping up to 10 litres per second at a pressure of 0 to 4 bar.

Transient Generation:
An SLA 2-way pilot-operated piston solenoid valve was used to create pressure transients to a maximum pressure of 0 to 25 bar in its application to water. The valve was run on 24V DC.

Pressure Measurement:
The pressure signals were recorded on a piezoelectric pressure transducer (PCB Piezotronics model 113B27). It had a sensitivity of 7.25 mV/Kpa, was capable of measuring up to 6895 kPa of pressure, and the sensor had a resonant frequency of greater than 500 KHZ, rendering it able to capture high-pressure/low-pressure changes in a short time. Figure 1 presents the complete experimental schematic, including all components and their spatial arrangement.

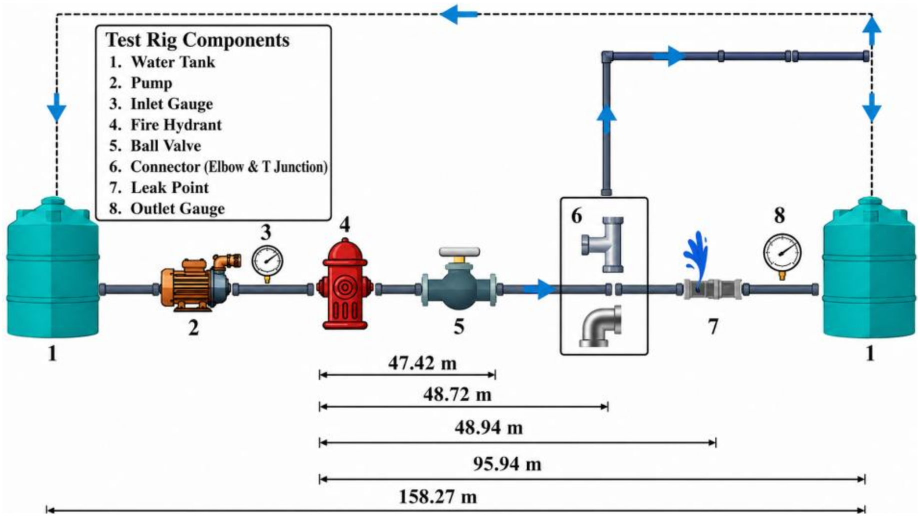

Figure 1. Schematic diagram of experimental pipeline system

2.2 Data Collection

2.2.1 Pipeline feature configuration

This study tested three distinct pipeline conditions to represent common distribution network scenarios, as shown in Table 1. All of the three configurations were put in a similar position, which was approximately 49 metres away from the transducer. This uniform position meant that any detected differences in signal could be limited to the true nature of the feature in question, and not the difference in time taken by 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 variable distances from sensors, and future work will investigate the sensitivity of the selected features to varying feature positions in order 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 offers a well-controlled and repeatable simulation of a perforation-type leak and is consistent with methodologies adopted 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

2.2.2 Signal acquisition procedure

Each pipeline arrangement was analysed using 100 pressure transient signals, which were taken in accordance with a standardized protocol. Pressure was introduced gradually to 4 bar, after which the period of stabilisation was 30 seconds. The solenoid valve was then closed quickly within 15 milliseconds in order 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 resulted in the production of a total dataset of 300 signals (100 signals in each configuration) with equal classes of model training and testing. In both the 3-second recording window, the full cycle of wave propagation and reflect was recorded in the 158.27-meter pipeline, whereas the 10 kHz sampling rate allowed a reasonable frequency resolution of 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 specifically 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 estimation. 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.

2.3 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 (FFT) processing, direct physical interpretability of transient characteristics, and its proven effectiveness in representing hydraulic transient phenomena [19, 20]. To facilitate systematic analysis, the extracted features were categorised into five groups according to their statistical and signal representation properties, as presented in Table 2.

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

Table 2 shows all 37 features that were extracted in five categories. The features of amplitude describe the magnitude properties of the signals, whereas the features of statistical distribution describe the data distribution and shape. Energy based features are used to measure signal power, differential features are used to measure signal changes, and temporal features are used to measure timing and complexity patterns. This broad covering feature facilitates powerful categorising the three pipeline conditions.

2.4 Feature Selection

The idea behind feature selection is to minimise the dimensions by determining the most informative characteristics, filtering out the redundant or uninformative ones. Three algorithms that have different selection philosophies were compared to identify the best subset of features whose joints could be used to classify the pipeline conditions.

2.4.1 Minimum classification error

Minimum Classification Error (MCE), is an embedded algorithm that directly optimises classification on the feature selection. Instead of using just the statistical measures, MCE considers the magnitude of each feature in minimising the misclassification errors [21]. The approach is to adopt a smoothed version of the classification loss with the aim of identifying features that can best differentiate the three conditions of the pipeline.

2.4.2 Minimum redundancy maximum relevance

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 [22-24]. This is a dual-criterion approach before the superfluous features are selected and offers the same information. The algorithm approximates the mutual information by the histogram-based calculations (with 10 bins). The addition of features is done through greedy selection, in which a specific candidate is selected based on which candidate would yield the best balance of relevant and redundant data. Selection was limited to 10 features to keep the computational efficiency and avoid overfitting.

2.4.3 Sequential forward selection

Sequential Forward Selection (SFS) represents the wrapper approach to feature selection, where subsets are evaluated by actually training classifiers and measuring their performance [22, 25]. The algorithm begins with no features in it and systematically tests all remaining features by putting them temporarily in 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.

2.5 Support Vector Machine, SVM Classification

2.5.1 Rationale for SVM selection

Support Vector Machines (SVMs) were selected as the classification algorithm based on several key advantages relevant to this application. SVMs are effective when the data is in high dimensions, even when the dimensions of the features are higher than the number of samples. The maximisation of the margin principle, on which SVMs operate on, delivers inherent regularisation, minimizing the chances 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 tries to identify an ideal hyperplane that performs the maximum 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 the case of our three-class problem of separating elbows, junctions, and leaks, the RBF kernel is used to map the original feature space into a space of a higher dimension with complex and non-linear boundaries to well distinguish among the three pipeline conditions. SVMs have a good mathematical background, which gives interpretability to their results. Furthermore, SVMs are efficient when the size of the sample is small, and thus they are suited to datasets of such magnitude, having 300 signals.

2.5.2 Kernel selection

Radial Basis Function kernel was used because it was shown that its performance is better compared to other kernels. The accuracy of the validation was obtained, and RBF was found to be significantly better than the other two kernels, which were 87% and 82%, respectively. The RBF kernel was especially appropriate in characterising finer differences in the pressure transient signatures of the various features in a pipeline due to the fact that the non-linear relationship between the two features was difficult to model.

2.5.3 Multi-class classification

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 come up with final predictions. In a tie of the votes the highest average decision functional is taken as the class.

2.5.4 Hyperparameter optimisation

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 $\gamma = 0.1$ as the optimal configuration for this classification task.

2.6 Experimental Design and Statistical Validation

2.6.1 Data partitioning

The dataset was split into a 70-30 training and 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. The stratified random sampling was used to guarantee proportional representation of the groups in both sets to exclude the development and evaluation of models biased due to disproportionate representation.

2.6.2 Cross-validation strategy

Five-fold cross-validation was applied to the training set to serve multiple objectives: hyper parameter 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 employed to train, and the remaining fold was employed as the validation. This was done 5 times and every fold was used as the validation set once. The averages of the performance metrics of all five iterations were calculated, and standard deviations were calculated to measure the uncertainty in the results.

2.6.3 Performance metrics

Performance in classifications was measured through several metrics in order to offer overall measurements. The overall accuracy was used to determine the fraction of the correct signal classification of 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 tell all cases of each state, and F1-score offered a trade-off between precision and recall. These complementary measures make sure that the model performance is completely evaluated in different elements of classification quality.

3. Results and Discussion

3.1 Signal Characteristics Analysis

Figure 2 shows the representative 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 closing of valves, and amplitude was measured in milli-amperes.

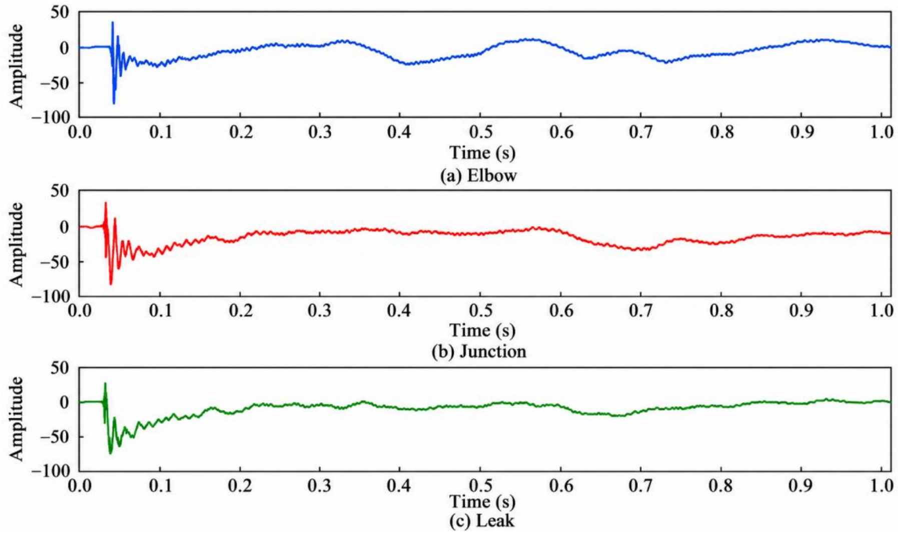

Figure 2. Representative time-domain transient responses for (a) elbow, (b) junction, and (c) leak conditions obtained from the experimental pipeline system

3.1.1 Elbow configuration

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 increase in pressure at the valve closure shows only one major reflection at the middle of the recording window. The signals decay slowly and progressively to the baseline, displaying uniform exponential drying throughout and across all 100 repetitions. The general trend is quite basic and homogeneous with a little few anomalies in the oscillation pattern.

3.1.2 Junction configuration

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 as well as in the form of a multi-layered oscillation pattern. Secondary waves continue during the recording process longer, meaning that they redistribute the energy in the branch connection. Although the signals remain distinctly structured, there is more variability and continuous oscillations in the waveforms that go further than the those seen in the elbow configuration.

3.1.3 Leak configuration

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 was maintained is much lower following the initial transient, and it is worth keeping down during the whole recording period. The waveform is dominated by a lot of high-frequency oscillations with irregular amplitudes, which make it look significantly noisier than the other two configurations. Patterns of reflection are less clear and are more diffused, in line with the continuous dissipation of energy via the leak orifice. The combination of long-term amplitude decreases, and chaotic oscillations causes the leak condition to be visually conspicuously different from undamaged pipeline characteristics. The visual differences in all three signal types, as shown in Figure 2, substantiate the fact that time-domain pressure transient characteristics have adequate discriminating data in automating the detection of pipeline conditions.

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: Minimum Classification Error (MCE), Minimum Redundancy Maximum Relevance (mRMR), and Sequential Forward Selection (SFS). 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 least misclassification error (16 and 23, respectively), which showed that they had higher discriminative potential to give the correct three pipeline conditions. These two time-domain features capture complementary signal characteristics: Slope Sign Change detects the frequency of signal fluctuations by counting zero-crossings in the signal derivative, which is sensitive to leak-induced turbulence, while Integrated EMG quantifies the cumulative signal energy, reflecting the overall intensity of acoustic emissions from 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).

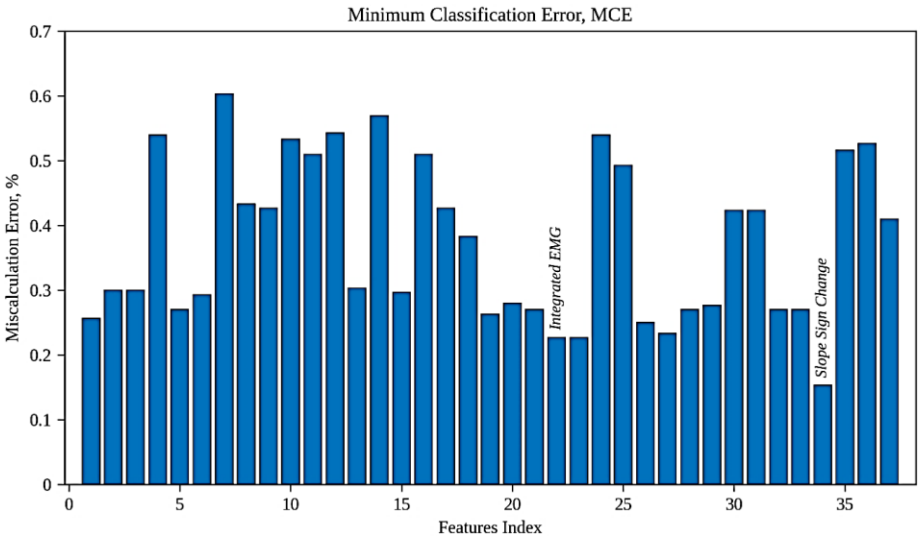

Figure 3. Minimum Classification Error (MCE) for all 37 extracted features using an SVM classifier

Figure 4 presents the mRMR feature ranking based on predictor importance scores. The top-ranked feature, Integrated EMG (Feature 22), demonstrates significantly higher importance (score ~0.61) compared to 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.

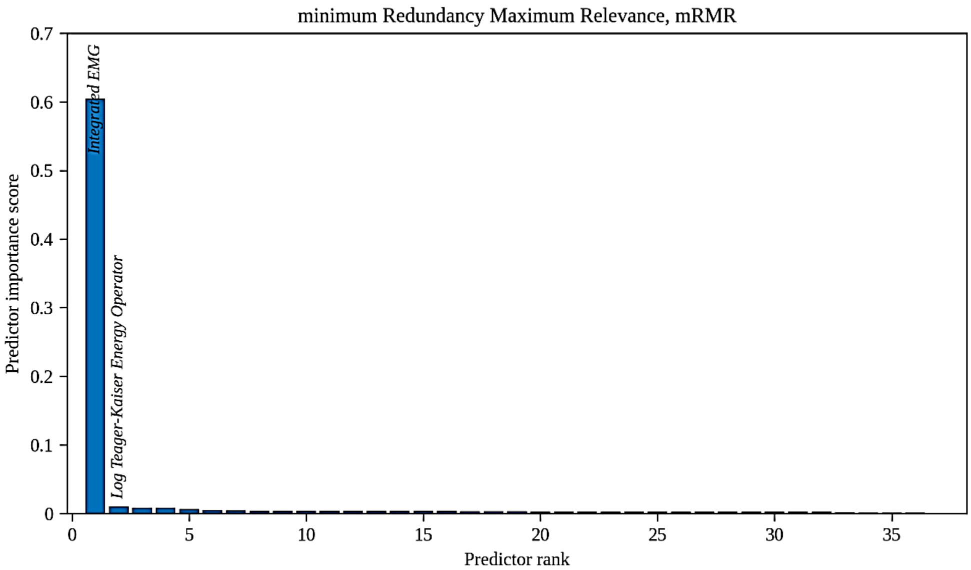

Figure 4. Minimum Redundancy Maximum Relevance (MRMR) feature ranking for all 37 extracted features

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 Slope Sign Change (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.

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, while SFS-SVM showed the lowest performance at 65.8% test accuracy.

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

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, namely direct classification error minimisation in the case of MCE versus relevance–redundancy trade-off in the case of mRMR. This cross-method convergence supports the conclusion that Integrated EMG captures genuinely informative signal variance rather than artefacts of a single algorithm. Slope Sign Change (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.

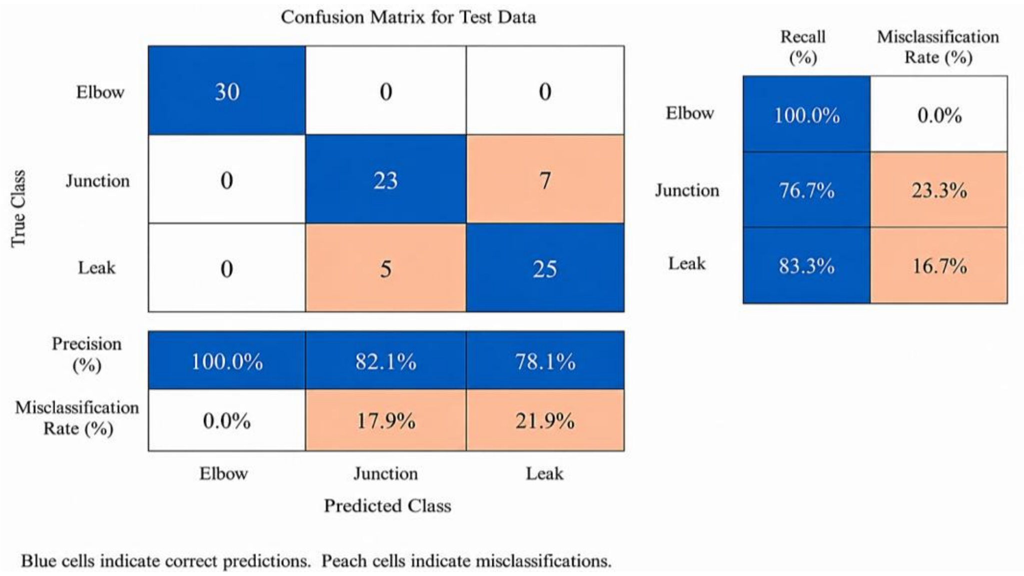

Figure 5. Confusion matrix for MCE-SVM classifier on test set

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 Slope Sign Change 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 optimization ensures this 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: Slope Sign Change (x-axis) and Integrated EMG (y-axis). The contour plot shows the posterior probability distribution, with warmer colors (yellow orange) indicating higher classification confidence and cooler colors (blue) representing decision boundary regions. Circles represent training samples and x-marks represent test samples, with colors indicating the actual class: blue for Elbow, red for Junction, and green for Leak. The visualization proves some crucial findings of the results of the classification. The three classes have different localization points in the feature space and it confirms 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 Slope Sign Change values (4000-5500), reflecting their smooth oscillatory behavior 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.

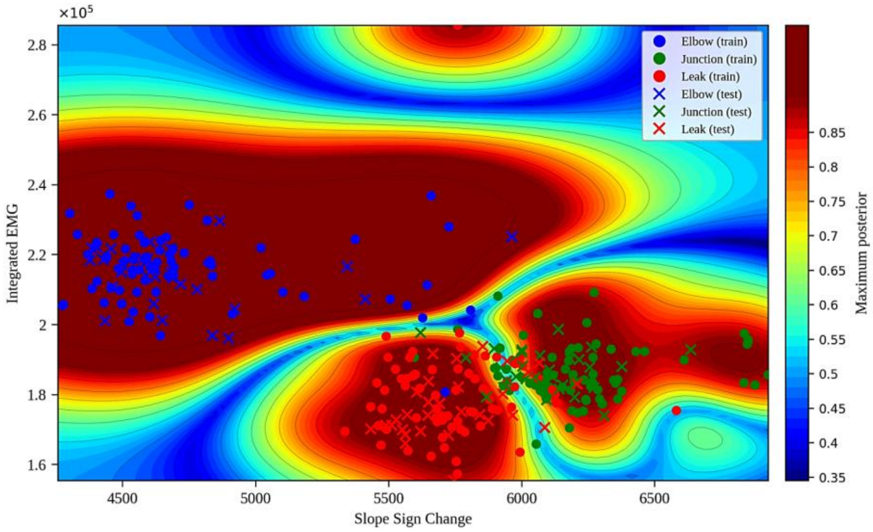

Figure 6. SVM decision boundary visualization for MCE-SVM classifier

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 determined in the contour lines, are also complex and cannot be attained using the 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 around coordinates (5500, 1.75 × 10^5) corresponds to the misclassifications between these classes observed in Table 4, where samples fall near the decision boundary with ambiguous signal characteristics.

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 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, Slope Sign Change (SSC) and Integrated EMG (IEMG), 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. (2021) 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. (2021) 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. [12]	2021	VMD-SVM	VMD-based features	Experimental	12	93%
Liao et al. [26]	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%

3.5 Discussion

MCE-SVM was the best among the other two methods for a number of reasons. MCE directly minimises classification errors, unlike mRMR, which employs mutual information as an indirect measure. This is important since minimising the redundancy of features does not necessarily correspond to improved 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 can be observed when compared with the SFS feature selection approach. Sequential Forward Selection (SFS) 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. Slope Sign Change captures sharp changes in the signal that occur as a result of turbulent flow from 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 in troubleshooting or reporting the findings to non-technical employees.

Raw accuracy continues to favour deep learning. This study achieves 87.4% accuracy compared to 93% reported by others, showing a perceptible difference. However, that difference comes with costs such as large datasets, lengthy training experiences, costly hardware, and outputs that lack explanation even by experts. In the case of 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 does not include alternative supervised learning algorithms such as Random Forest, k-Nearest Neighbours, or Naive Bayes. This focus was intentional as SVM was selected a priori based on 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 (Slope Sign Change 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, ensuring that differences in classifier performance are not confounded by insufficient training data.

3.6 Limitations and Future Work

The experimental system consisted of experiments conducted with clean and controlled conditions in one pipe (MDPE, 63 mm) and 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 (4 mm circular perforation), which does not represent the full spectrum of failure modes encountered operationally. 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 generalization 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 corresponds to a minor leakage or a catastrophic pipe burst.

There are three directions that seem the most promising in terms of 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 increasing the dataset, which would include larger categories of pipes, sizes, materials, and leak situations. 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 allow utilities to pay more attention to which issues should be solved initially. 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 actually deploy.

4. Conclusions

The paper will describe an automated pipeline monitoring system that combines pressure transient analysis and machine learning to detect leaks within water distribution networks. To maximise the performance of feature selection to classify and ensure that it can be deployed in real-time, we systematically compared the performance of three methods of feature selection. 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 (Slope Sign Change 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: Slope Sign Change 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. In particular, 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 Slope Sign Change 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 of 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 Postgraduates Grant Research Scheme [Grant Number: PGRS220389].

Declaration of Competing Interest

The author declares no conflicts of interest.

CRediT Authorship Contribution Statement

Mohd Fairusham Ghazali (Writing – review & editing, Methodology, Funding acquisition, Investigation, Project administration, Supervision).
Mohd Fadhlan Mohd Yusof (Conceptualisation, Formal analysis, Validation, Investigation, Resources, Supervision).
Nor Azinee Said (Writing – original draft, Visualisation, Formal analysis, Methodology, Data curation).
Nurul Hazwani Mokhtar (Writing – original draft, Methodology, Formal analysis).
Muhammad Amsyar Azmil (Writing – original draft, Visualisation, Formal analysis).
Zairif Anas Mohammad Zahib (Writing – original draft, Investigation, Formal 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, 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, 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, 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, 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, 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, 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, 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, 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] S. Salcedo-Sanz, J.L. Rojo-Álvarez, M. Martínez-Ramón, G. Camps-Valls, "Support vector machines in engineering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 4, no. 3, pp. 234-267, 2014. https://doi.org/10.1002/widm.1125

[10] N.K. Banjara, S. Sasmal, S. Voggu, "Machine learning supported acoustic emission technique for leakage detection in pipelines," *International Journal of Pressure Vessels and Piping,* vol. 188, p. 104243, 2020. https://doi.org/10.1016/j.ijpvp.2020.104243

[11] 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

[12] 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

[13] M.F.F.M. Ruslan, M.F. Hassan, "Unbalance failure recognition using recurrent neural network," *International Journal of Automotive and Mechanical Engineering,* vol. 19, no. 2, pp. 9668-9680, 2022. https://doi.org/10.15282/ijame.19.2.2022.04.0746

[14] M.Z. Nuawi, A.R. Ismail, M. Nor, M.M. Rahman, "Comparative study of whole-body vibration exposure between train and car passengers: A case study in Malaysia," *International Journal of Automotive and Mechanical Engineering,* vol. 4, pp. 490-503, 2011. https://doi.org/10.15282/ijame.4.2011.10.0040

[15] E. McDermott, 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

[16] S. Pitta, R. J. Ginka, 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

[17] A. Navarro-Navarro, J.E. Ruiz-Sarrio, V. Biot-Monterde, J.A. Antonino-Daviu, V. Becker, S. Urschel, "Application of transient analysis techniques to fault diagnosis in low-and medium-power synchronous machines," *Machines,* vol. 11, no. 2, p. 288, 2023. https://doi.org/10.3390/machines11020288

[18] U.E. Edeh, T.T. Lie, M.A. Mahmud, "Assessment of transmission reliability margin: Existing methods and challenges and future prospects," *Energies,* vol. 18, no. 9, p. 2267, 2025. https://doi.org/10.3390/en18092267

[19] M.A. Islam, M.R. Mahmud, A.H. Sakib, M.I.H. Siddiqui, 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

[20] B. Nayana, 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

[21] 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

[22] H.R. Al-Bugharbee, H. Samaka, 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

[23] P. Singla, H. Garg, A. Pathak, 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

[24] R. Taiwo, M.E.A. Ben Seghier, 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

[25] K. Yan, L. Ma, Y. Dai, W. Shen, Z. Ji, 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

[26] Z. Liao, H. Yan, Z. Tang, X. Chu, 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