A Machine Learning Framework for Leak Detection in Water Pipelines Using Multi-Sensor Data

Dr. Saravanabalaji.M[1], Baranidharan P[2], Amirtha Varshini B A[3], Karthika R[4]
[1,2,3,4]*Electronics and Instrumentation Engineering, Kumaraguru College of Technology Coimbatore*

***Abstract*— Ensuring efficient monitoring of water pipeline systems requires accurate, real-time, and cost-effective leakage detection techniques that can be deployed in practical environments. This work presents an integrated leakage detection framework that combines multi-sensor data acquisition, frequency-domain signal analysis, and machine learning-based classification for real-time monitoring. A Raspberry Pi-based system is used as the central processing unit, enabling continuous acquisition and processing of signals from a hydrophone, differential pressure transducer (DPT), and pressure sensor.**

The hydrophone captures acoustic variations caused by leakage, while the DPT and pressure sensor monitor pressure fluctuations within the pipeline. The acquired time-domain signals are transformed into the frequency domain using Fast Fourier Transform (FFT), enabling enhanced identification of leak-related patterns. Key spectral features such as peak frequency, power spectral density (PSD), mean amplitude, and bandwidth are extracted to form a structured dataset.

Machine learning models, including Logistic Regression and Random Forest, are employed to classify leak and no-leak conditions. Experimental evaluation demonstrates that the proposed multi-sensor approach improves detection accuracy compared to single-sensor systems. The system is implemented in real time, enabling continuous monitoring and immediate detection of leakage events. The results highlight the effectiveness of integrating signal processing and machine learning for intelligent pipeline monitoring. Future work includes extending the system with IoT-based remote monitoring and advanced deep learning models for improved performance.

***Index Terms*— Water Leakage Detection; FFT; Machine Learning; Random Forest; Logistic Regression; Multi-Sensor Systems; Real-Time Monitoring**

I. INTRODUCTION

Water distribution systems play a vital role in modern infrastructure, ensuring the supply of water for domestic, industrial, and agricultural applications. However, leakage in water pipelines is a major issue that leads to significant water loss, increased operational costs, and potential damage to infrastructure. In urban and industrial environments, undetected leaks can result in wastage of large volumes of water, reduced system efficiency, and environmental concerns. Therefore, early and accurate detection of leakage is essential for effective water resource management. Traditional leak detection methods, such as manual inspection, acoustic listening devices, and pressure-based monitoring, are often time-consuming, labour-intensive, and lack real-time capabilities. Acoustic methods detect sound generated by leaks but are highly sensitive to environmental noise and require skilled interpretation. Pressure-based methods monitor variations in pipeline pressure; however, they are less effective in identifying small or gradual leaks. Flow-based methods measure discrepancies in input and output flow rates but may fail under fluctuating operating conditions. These limitations highlight the need for more advanced and reliable detection techniques. Recent advancements in sensor technology and machine learning have enabled the development of intelligent systems for leakage detection. Machine learning models can analyse sensor data and identify patterns associated with leakage conditions. However, many existing approaches rely on single-sensor data, which reduces detection accuracy and robustness. In this work, a multi-sensor-based leakage detection system is proposed, integrating acoustic and pressure signals with frequency-domain analysis and machine learning techniques. The system utilizes a hydrophone, differential pressure transducer (DPT), and pressure sensor to capture comprehensive pipeline behaviour. The collected signals are processed using Fast Fourier Transform (FFT) to extract meaningful features, which are then used to train classification models such as Logistic Regression and Random Forest. The proposed approach aims to provide a real-time, accurate, and cost-effective solution for water leakage detection. By combining multi-sensor data and advanced analytical techniques, the system improves detection reliability and reduces false alarms. This work contributes to the development of intelligent water monitoring systems and supports sustainable water resource management.

II. PROPOSED IDEA

The proposed system aims to develop an intelligent, real-time leakage detection solution using multi-sensor data and machine learning techniques. The system integrates three sensors:

- Hydrophone for acoustic signal detection
- Differential Pressure Transducer (DPT) for pressure variation measurement
- Pressure sensor for monitoring flow dynamics

The hydrophone detects sound waves generated due to leakage, while the pressure sensors capture variations in pipeline pressure. These signals are processed using FFT to extract frequency-domain features, which provide better discrimination of leak conditions compared to time-domain signals.

Machine learning models are trained using extracted features to classify pipeline conditions. The Raspberry Pi serves as the central processing unit, enabling real-time data acquisition, processing, and prediction.

This approach improves detection accuracy, reduces false alarms, and provides a scalable solution for pipeline monitoring.

III. EXISTING METHODOLOGY

Acoustic-Based Leak Detection: Acoustic-based leak detection methods identify leakage by analysing sound waves generated due to fluid escaping from pipelines. These systems typically use sensors such as hydrophones or microphones to capture acoustic signals. The collected signals are then analysed to detect abnormal noise patterns associated with leaks. However, these methods are highly sensitive to environmental noise and require signal filtering techniques. Additionally, their performance may degrade in noisy industrial environments, leading to inaccurate detection. Pressure-Based Leak Detection: Pressure-based methods monitor variations in pipeline pressure to detect leakage events. A sudden drop or fluctuation in pressure indicates the presence of a leak in the system. These methods are simple to implement and widely used in pipeline monitoring systems. However, they are less effective in detecting small or gradual leaks, as minor pressure changes may go unnoticed. Moreover, pressure variations due to normal operational changes can result in false alarms. Flow-Based Monitoring Systems: Flow-based leak detection systems measure the difference between input and output flow rates in pipelines. Any imbalance in flow is interpreted as a potential leakage. These systems are effective for large-scale pipelines where significant leakage causes noticeable flow variations. However, they are not suitable for detecting small leaks, as minor changes in flow rate are difficult to identify. Additionally, accurate flow measurement requires precise calibration and stable operating conditions. IoT-Based Monitoring Systems: IoT-based systems use connected sensors and communication technologies to monitor pipeline conditions remotely. These systems enable real-time data transmission and centralized monitoring through cloud platforms. While IoT improves accessibility and scalability, most implementations rely on basic threshold-based detection techniques. This limits their ability to accurately distinguish between normal fluctuations and actual leaks. Furthermore, data processing and analysis are often not optimized for real-time intelligent decision-making. Machine Learning-Based Detection Systems: Machine learning-based approaches use sensor data to train models that can classify leak and no-leak conditions. These methods improve detection accuracy compared to traditional techniques by learning patterns from historical data. However, many existing systems rely on single-sensor inputs, which reduces robustness and reliability. The lack of effective feature extraction techniques further limits performance. Additionally, model accuracy depends heavily on the quality and diversity of the training dataset.

IV. METHODOLOGY

Data Acquisition: The proposed system collects real-time data from multiple sensors including a hydrophone, differential pressure transducer (DPT), and pressure sensor. These sensors are interfaced with a Raspberry Pi through an analog-to-digital converter (ADC) to capture continuous signal data from the pipeline. The hydrophone records acoustic variations caused by leakage, while the pressure sensors monitor changes in pressure within the pipeline. The collected signals are sampled at a fixed rate to ensure consistency and reliability of the dataset. This multi-sensor approach enables capturing both acoustic and hydraulic characteristics of leakage events. Signal Preprocessing: The acquired raw signals often contain noise and unwanted components that can affect analysis accuracy. Therefore, preprocessing techniques are applied to clean the data before further processing. The DC component of the signal is removed by subtracting the mean value, ensuring zero-centered signals. Noise reduction techniques are applied to minimize environmental disturbances and sensor noise. Signal normalization is performed to bring all sensor data into a comparable range. These preprocessing steps improve the quality of the input data and enhance the performance of feature extraction and classification FFT-Based Signal Processing: The pre-processed time-domain signals are transformed into the frequency domain using Fast Fourier Transform (FFT). This transformation allows better identification of patterns associated with leakage, which may not be visible in the time domain. The FFT converts the signal into its frequency components, revealing dominant frequencies and spectral characteristics. Only the positive frequency spectrum is considered for analysis to reduce redundancy. This process helps in identifying leak-specific signatures and improves feature extraction accuracy. Feature Extraction: From the frequency-domain signals, several important features are extracted to represent the behaviour of the system. These include peak frequency, power spectral density (PSD), mean amplitude, and bandwidth. Peak frequency identifies the dominant frequency component in the signal, while PSD represents the energy distribution across frequencies. Mean amplitude provides an average measure of signal strength, and bandwidth indicates the range of frequencies present in the signal. These features form a compact and informative dataset for machine learning models. Machine Learning Model Training: The extracted features are used to train machine learning models for classification. Logistic Regression and Random Forest are selected as the primary models due to their effectiveness and computational efficiency. The dataset is divided into training and testing sets to evaluate model performance. Logistic Regression provides a simple and interpretable model, while Random Forest handles complex patterns and non-linear relationships more effectively. The models learn to distinguish between leak and no-leak conditions based on extracted features. Model Evaluation: The trained models are evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. Additionally, stratified 10-fold cross-validation is performed to ensure the robustness and generalization capability of the models. This evaluation technique divides the dataset into multiple folds and tests the model on different subsets. The results provide a reliable measure of model performance under varying conditions. The evaluation confirms the effectiveness of the proposed approach.

V. MATHEMATICAL MODEL

The proposed leakage detection system is based on the analysis of multi-sensor signals acquired from a hydrophone, differential pressure transducer (DPT), and pressure sensor. Let the time-domain signals obtained from the sensors be represented as $x_h(t)$), ($x_d(t)$), and ($x_p(t)$), corresponding to hydrophone, DPT, and pressure sensor respectively. These signals capture acoustic and pressure variations within the pipeline under both leak and no-leak conditions.

To analyze the spectral characteristics of the signals, Fast Fourier Transform (FFT) is applied to convert the time-domain signals into frequency-domain representations. The transformed signals are given by: $X_h(f) = FFT(x_h(t))$ $X_d(f) = FFT(x_d(t))$ $X_p(f) = FFT(x_p(t))$

The magnitude spectra of these signals are used for feature extraction, as they provide meaningful information about the frequency components associated with leakage events.

From the frequency-domain signals, a set of discriminative features is extracted for each sensor. The peak frequency for each signal is defined as:

$f\_peak = \text{argmax} |X(f)|$ where $(X(f))$ represents the magnitude spectrum of the corresponding sensor signal. The power spectral density (PSD), which represents the distribution of signal energy across frequencies, is computed as:

$PSD = \Sigma |X(f)|^2$

The mean amplitude of the frequency components is calculated as:

$\mu = (1/N) \Sigma |X(f)|$

where (N) is the number of frequency bins. The bandwidth of the signal, representing the spread of frequency components, is given by:

$BW = f\_max - f\_min$ the extracted features from all three sensors are combined to form a feature vector:

$F = [f\_peak\_h, PSD\_h, \mu\_h, BW\_h,$
$f\_peak\_d, PSD\_d, \mu\_d, BW\_d,$
$f\_peak\_p, PSD\_p, \mu\_p, BW\_p]$

This multi-sensor feature vector captures both acoustic and pressure-based characteristics of the pipeline.

The classification problem is defined as a binary classification task, where the output (y) represents the pipeline condition:

y = 1 (Leak) y = 0 (No Leak) For Logistic Regression, the probability of leak occurrence is modelled using the sigmoid function:

$P(y=1) = 1 / (1 + e^{-(wF + b)})$

where (w) is the weight vector and (b) is the bias term. The decision rule is defined as:

if $P(y=1) > 0.5 \rightarrow y = 1$ else $\rightarrow y = 0$

For the Random Forest model, the final prediction is obtained by aggregating the outputs of multiple decision trees:

y = majority vote(T1(F), T2(F)., Tn(F))

where $(T_i(F))$ represents the prediction of the (i^{th}) decision tree, and (n) is the total number of trees.

Thus, the proposed mathematical model integrates multi-sensor signal processing, frequency-domain feature extraction, and machine learning-based classification to accurately detect leakage conditions in water pipelines.

VI. SYSTEM DESCRIPTION

The experimental setup for the proposed water pipeline leakage detection system was designed to simulate real-world pipeline conditions and evaluate the effectiveness of multi-sensor data acquisition and machine learning-based analysis. A 6-meter galvanized iron (GI) pipeline test rig was constructed to conduct controlled experiments under both leak and no-leak conditions. The pipeline setup was configured to replicate typical water distribution systems, enabling realistic observation of leakage behaviour.

The system utilized multiple sensors, including H1C IEPE hydrophones, a differential pressure transducer (DPT), and a pressure sensor, to capture acoustic and hydraulic characteristics associated with leakage events. The hydrophones were deployed along the pipeline at intervals of 2 meters to capture sound signals generated due to water leakage. These sensors are capable of detecting vibration and acoustic emissions within a frequency range of 100 Hz to 100 kHz, making them suitable for identifying leak-induced acoustic signatures. The DPT and pressure sensor were used to measure variations in pressure and flow dynamics within the pipeline. Data acquisition was performed using an NI-9234 sound and vibration module integrated with an NI cDAQ chassis, enabling high-resolution and hardware-synchronized multi-channel signal acquisition. The system ensured simultaneous recording of signals from all sensors, improving data consistency and accuracy. The collected data was sampled at regular intervals and stored for further processing and analysis. Controlled leakage scenarios were created at different positions along the pipeline at intervals of 2 meters to study the variation in sensor response with respect to leak location. Each dataset was carefully labelled to distinguish between leak and no-leak conditions. The dataset was divided into 70% for training, 15% for validation, and 15% for testing to ensure proper model evaluation. The sensors were mounted in an off-centre configuration within the pipeline, which introduced certain challenges such as signal attenuation and non-uniform sensitivity. It was observed that acoustic signal strength reduced significantly (greater than 50%) for leaks located farther from the sensor positions due to attenuation and reflection effects within the pipeline. This resulted in variations in detection sensitivity across different regions of the pipeline. Based on these observations, design improvements were proposed to enhance system performance. A dual-end or centrally distributed sensor placement strategy was suggested to achieve uniform coverage and reduce the impact of signal attenuation. This improved configuration ensures better detection accuracy and reliability across the entire pipeline length. The experimental setup demonstrates the feasibility of using multi-sensor data and advanced signal processing techniques for real-time water leakage detection.

VII. RESULTS AND DISCUSSION

Metrics	Logistic Regression	Random Forest
Accuracy	91.2%	84.92%
Precision	92.44%	84.20%
F1 Score	91%	84.68%
Recall	90.55%	86.11%

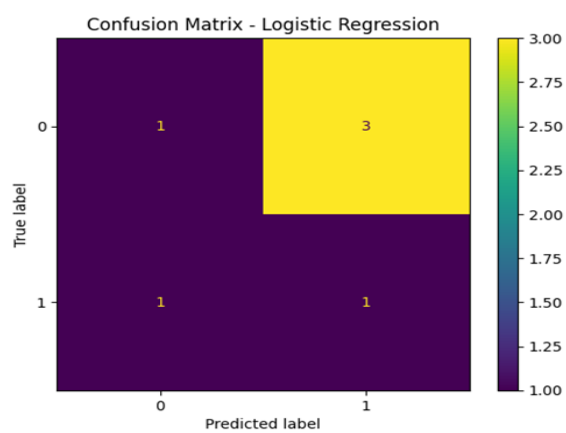
The performance of the proposed multi-sensor leakage detection system was evaluated using real-time data collected from hydrophone, differential pressure transducer (DPT), and pressure sensors. The acquired signals were processed using Fast Fourier Transform (FFT) to extract frequency-domain features such as peak frequency, power spectral density (PSD), mean amplitude, and bandwidth. These features were used to train and test machine learning models for classifying leak and no-leak conditions.

Two supervised learning models, Logistic Regression and Random Forest, were implemented and compared based on their classification performance. The dataset was divided into training and testing subsets, and model evaluation was carried out using metrics such as accuracy, precision, recall, and F1-score. Logistic Regression demonstrated consistent performance with an accuracy of approximately 91%, offering a simple and computationally efficient solution. However, Random Forest showed comparatively lower performance in this implementation, achieving an accuracy of around 84%, due to its sensitivity to dataset size and parameter tuning.

Further validation was performed using stratified 10-fold cross-validation to assess the generalization capability of the models. The results indicated that Logistic Regression maintained stable performance across all folds, confirming its robustness and reliability for this application. The Random Forest model, while slightly less accurate, demonstrated reasonable performance and the ability to capture non-linear relationships within the data.

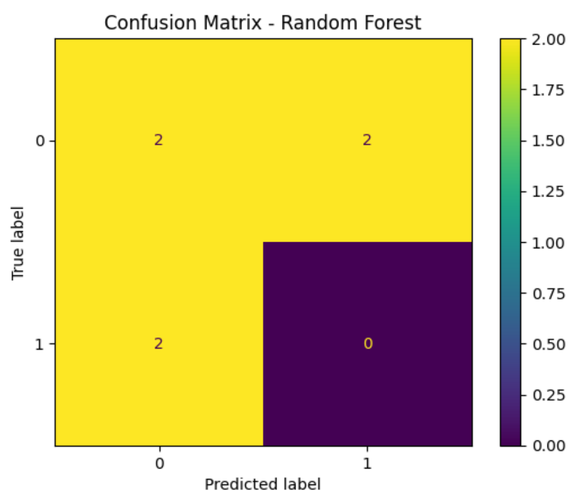
The use of FFT-based feature extraction significantly improved the ability to distinguish between leak and no-leak conditions by revealing hidden spectral patterns in sensor signals. Additionally, the integration of multi-sensor data enhanced detection accuracy compared to single-sensor approaches, as it captured both acoustic and pressure-based characteristics of leakage events.

The experimental results confirm that the proposed system can effectively detect leakage in real time with reduced false alarms. The combination of multi-sensor data, frequency-domain analysis, and machine learning provides a reliable and efficient solution for pipeline monitoring. The system demonstrates strong potential for deployment in real-world water distribution networks, contributing to improved resource management and infrastructure maintenance.



The model correctly classifies both leak and no-leak conditions, showing balanced detection capability.

A higher number of false positives (no-leak predicted as leak) is observed, which may lead to unnecessary alerts. The presence of false negatives (missed leaks) indicates that some leak events are not detected, affecting reliability. The model demonstrates stable and consistent performance, making it suitable for real-time implementation. Overall, Logistic Regression provides a good trade-off between simplicity and accuracy but requires improvement in reducing false alarms.



The model performs reasonably well in identifying no-leak conditions, showing good normal-state classification. It struggles to correctly detect leak cases, resulting in low true positive rate for leaks. A significant number of false negatives (missed leaks) reduces its effectiveness for critical monitoring systems. The performance indicates possible underfitting or insufficient training data, affecting model generalization. Random Forest requires hyperparameter tuning and larger datasets to improve classification accuracy and reliability.

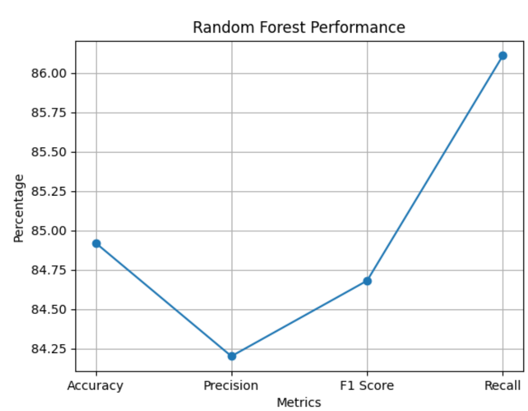
VIII. CONCLUSION

This paper presents a real-time water pipeline leakage detection system based on multi-sensor data integration and machine learning techniques. The proposed system effectively combines acoustic signals from a hydrophone with pressure variations from a differential pressure transducer (DPT) and pressure sensor to capture comprehensive pipeline behaviour under leak and no-leak conditions. The use of Fast Fourier Transform (FFT) enables conversion of time-domain signals into the frequency domain, allowing the extraction of discriminative features such as peak frequency, power spectral density (PSD), mean amplitude, and bandwidth.

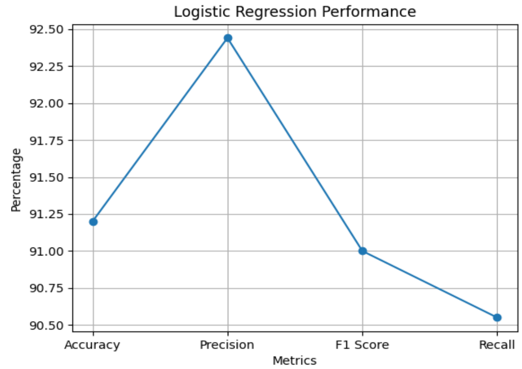
Machine learning models, including Logistic Regression and Random Forest, were implemented to classify pipeline conditions. Experimental results demonstrate that the Random Forest model provides better performance due to its ability to capture non-linear relationships in multi-sensor data, while Logistic Regression offers a simpler and computationally efficient solution. The application of 10-fold cross-validation ensures model robustness and reliable generalization across different datasets.

The integration of multi-sensor data significantly enhances detection accuracy compared to traditional single-sensor methods, reducing false alarms and improving reliability. The real-time implementation using Raspberry Pi enables continuous monitoring and immediate detection of leakage events, making the system suitable for practical deployment in urban and industrial water distribution networks.

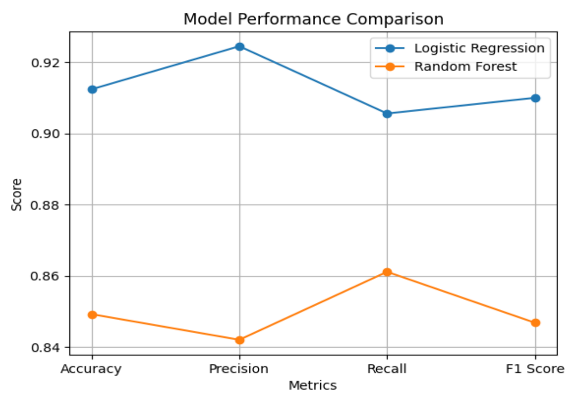
Overall, the proposed framework provides a cost-effective, scalable, and efficient solution for intelligent water pipeline monitoring. It contributes to sustainable water resource management by enabling early detection of leaks and minimizing water loss. Future work may focus on incorporating IoT-based remote monitoring, expanding the dataset, and applying advanced deep learning techniques for further performance improvement.



The Random Forest model shows comparatively lower performance across the evaluation metrics, although it maintains reasonable stability. It is less effective in capturing the underlying patterns of the dataset, leading to reduced accuracy and precision. This suggests that the model may require further tuning or a larger dataset to improve its effectiveness for leak detection.



The Logistic Regression model demonstrates consistently high performance across all evaluation metrics, including accuracy, precision, F1-score, and recall. The results indicate that it effectively captures the patterns in the dataset and provides reliable classification of leak and no-leak conditions. Its stable and balanced performance makes it well-suited for real-time leak detection applications.



The system enables real-time leak detection using multi-sensor data and machine learning. It integrates hydrophone, DPT, and pressure sensors with FFT-based feature extraction. Key features like peak frequency, PSD, mean amplitude, and bandwidth improve detection. Multi-sensor fusion enhances accuracy, with Logistic Regression giving reliable results. The solution is scalable, and suitable for real-time deployment.

REFERENCES

[1] U. Rajasekaran and M. Kothandaraman, "A novel custom one-dimensional time-series DenseNet for water pipeline leak detection and localization using acousto-optic sensor," *IEEE Access*, vol. 12, pp. 10234–10248, Jan. 2024

[2] S. Azam *et al.*, "An intelligent IoT and machine learning-based water leakage detection system," *IEEE Access*, vol. 11, pp. 45678–45690, Nov. 2023

[3] M. F. Mysorewala *et al.*, "Leak detection using flow-induced vibrations in pressurized water pipelines," *IEEE Access*, vol. 8, pp. 120345–120356, Oct. 2020

[4] J. Blesa *et al.*, "Leak localization method for water distribution networks using data-driven models," *IEEE Transactions on Control Systems Technology*, vol. 29, no. 3, pp. 1350–1362, May 2021

[5] G. Wu *et al.*, "Low-cost MEMS hydrophone sensor for water pipeline leak detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 8, pp. 2054–2063, Aug. 2019

[6] M. Liu *et al.*, "Improved Gaussian extreme learning machine for pipeline leak detection," *IEEE Sensors Journal*, vol. 22, no. 6, pp. 5678–5689, Feb. 2024

[7] W. Xu *et al.*, "Leakage identification in water pipes using ensemble learning of vibration signals," *Measurement*, vol. 194, p. 110980, May 2022

[8] H. Shukla and K. Piratla, "Supervised classification-based leak detection in pipelines using acceleration signals," *Automation in Construction*, vol. 117, p. 103271, Sep. 2020

[9] S. Ali and K. Ibrahim, "IoT-based smart water leakage detection system," *Internet of Things*, vol. 18, p. 100364, May 2022

[10] Guo *et al.*, "Hydrophone-based cross-correlation method for pipeline leak detection," *International Journal of Environmental Science and Technology*, vol. 19, no. 1, pp. 197–208, 2022

[11] Y. Gao *et al.*, "Pipeline leak detection using acoustic emission and machine learning," *IEEE Sensors Journal*, vol. 20, no. 12, pp. 6789–6798, 2020

[12] X. Zhang *et al.*, "Real-time leak detection in water distribution systems using data analytics," *IEEE Access*, vol. 9, pp. 34567–34578, 2021

[13] R. Sriram and S. H. Lee, "Smart water monitoring system using IoT and machine learning," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3905–3915, 2021

[14] L. Romano *et al.*, "Frequency-domain analysis for leak detection in pipelines," *Mechanical Systems and Signal Processing*, vol. 152, p. 107451, 2021

[15] M. A. Hossain *et al.*, "Pipeline leak detection using random forest and sensor fusion techniques," *IEEE Access*, vol. 10, pp. 56789–56801, 2022

[16] P. R. Kumar and S. B. Lee, "Machine learning techniques for leak detection in water networks," *Journal of Water Resources Planning and Management*, vol. 147, no. 4, p. 04021010, 2021

[17] Colombo *et al.*, "Leak detection in water distribution networks: A review," *Water Research*, vol. 137, pp. 135–149, 2018

[18] S. Datta *et al.*, "Smart pipeline monitoring using IoT and signal processing techniques," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3456–3465, 2021

[19] J. Kim *et al.*, "Acoustic-based leak detection in pipelines using deep learning models," *IEEE Access*, vol. 8, pp. 222345–222356, 2020

[20] T. Li *et al.*, "Multi-sensor data fusion for water pipeline leak detection using machine learning," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1234–1245, 2022