International Journal of Robotics and Control Systems
Vol. 5, No. 4, 2025, pp. 2218-2245
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

Enhance the Effectiveness of Leak Detection in Water Distribution Networks (WDNs) Using Machine Learning (ML) and Pressure Transducers

Chaima Chabira [a,1,*], Haddi Bakhti [b,2], Miloud Bentoumi [a,3], Ibraheem Shayea [c,4], Riyadh Benyettou [d,5]

[a] Laboratory of System Signal Analysis (LASS), Department of Electronics, Faculty of Technology, University of M'sila, University Pole, Road Bordj Bou Arreridj, M'sila 28000, Algeria
[b] Laboratory of Genie Electric (LGE), Department of Electronics, Faculty of Technology, University of M'sila, University Pole, Road Bordj Bou Arreridj, M'sila 28000, Algeria
[c] Department of Electronics and Communication Engineering, Istanbul Technical University (ITU), 34467 Istanbul, Turkey
[d] Laboratory of Materials and Mechanics of Structures (LMMS), Department of Electronics, Faculty of Technology, University of M'sila, University Pole, Road Bordj Bou Arreridj, M'sila 28000, Algeria
[1] chaima.chabira@univ-msila.dz; [2] elhadi.bakhti@univ-msila.dz; [3] miloud.bentoumi@univ-msila.dz; [4] shayea@itu.edu.tr; [5] riyadh.benyettou@univ-msila.dz
* Corresponding Author

ARTICLE INFO

Article history
Received May 17, 2025
Revised June 27, 2025
Accepted November 08, 2025

Keywords
WDNs;
Pressure Transducers;
Signal Processing;
Machine Learning;
Leak Detection

ABSTRACT

Water distribution systems (WDS) are critical infrastructure that supply water to communities, but leaks often occur in these systems and can lead to significant water losses and operational inefficiencies. Traditional leak detection methods often struggle with environmental noise and limited scalability. This research proposes a data-driven approach for leak detection using pressure transducers and machine learning (ML). The research contribution is the design of a low-cost, noise-resistant leak detection framework and a comparative analysis of multiple ML classifiers based on experimental data. To facilitate data acquisition, we constructed a prototype PEHD hydraulic circuit that measures 100 meters in length and 40 mm in diameter, on which two pressure transmitters were installed. Data were collected via a dSPACE acquisition system during both normal and leak-induced conditions. Six ML models—support vector machine (SVM), decision tree (DT), random forest (RF), naïve Bayes (NB), logistic regression (LR), and k-nearest neighbors (KNN)—were trained and evaluated using standard classification metrics. LR outperformed all other models, achieving 100% across accuracy, precision, recall, and F1-score. This result may be attributed to the linear separability of the leak signatures in the experimental setup. However, further validation is necessary to assess model generalizability under real-world conditions with varying pipe materials, flow rates, and noise levels. The study demonstrates that integrating pressure transducers with ML can enable reliable leak detection in WDNs, offering a scalable, hardware-efficient monitoring solution. Future work will focus on expanding the dataset and evaluating system performance in live environments.

1. Introduction

Water distribution networks (WDNs) are a vital part of urban infrastructure, overseen by government agencies and utilities to ensure a stable and safe water supply to consumers [1]. Their critical role involves supplying clean and reliable water to residential, commercial, and industrial users [2]. These networks encompass essential components such as water transportation pipelines, storage tanks for reserve capacity, pumping stations for pressure adjustment, and water quality monitoring systems for continuous safety evaluation [3]. The efficient administration and upkeep of these components are essential for public health and the seamless operation of urban environments [4].

A significant challenge in WDNs is the occurrence of leaks. A certain level of pressure is required to move water through pipelines, and a drop below a certain threshold, combined with a leak, can lead to contamination of the public water supply [5]. Leaks are detrimental to infrastructure due to the damage they inflict and may also result in public health hazards [6]. Therefore, effective water leak detection is crucial for preventing water loss, mitigating infrastructure damage, and ensuring the efficient use of resources [7]. Early detection of leaks can minimize operational expenses, avert water wastage, and preserve the integrity of the water supply. Leaks are traditionally identified using specialized monitoring instruments, including contemporary or conventional water leak detectors [8]. These devices measure various characteristics such as alterations in water pressure, vibrations, acoustic emissions, flow rate, and ultrasound, which can assist in identifying leaks [9]. Modern leak detection technologies often employ a mix of sensors that monitor multiple physical properties. For example, variations in water pressure can suggest leakage because a drop in pressure often occurs when there is an uncontrolled release of water [10]. Vibration sensors are also often utilized since leaks can create changes in the vibrations conveyed via pipes, which can be detected by sensitive equipment [11]. Acoustic emission sensors are useful for distinguishing sound waves created by water escaping from a pipe, allowing for detection of leaks [12], even in hard-to-reach or underground regions [13]. Additionally, assessing flow rates can assist in uncovering disparities between expected and actual water flows, which can suggest leaks [14].

ML algorithms have been widely used in recent research for anomaly detection, defect identification, and predictive maintenance [15]. ML algorithms excel in their capacity to process massive volumes of data and automatically find complicated patterns or abnormalities [16]. These techniques focus on the ability of a computer systems to collect and analyze data and develop accurate representations that best portray the conditions in the dataset [17]. The approach requires collecting diverse datasets representing both normal and unusual settings [18]. By examining these data, ML models learn to recognize the differentiating elements of abnormalities without requiring explicit programing or manual intervention. Over time, as the models are exposed to more data, they become more adept at accurately recognizing abnormal patterns. The literature generally shows that applying ML methods is effective for classifying with high accuracy [19], [20]. Specifically, they can classify data points into different categories such as normal or anomalous. The use of algorithms such as DT, SVM, RF, NB, LR, and KNN has been demonstrated to produce reliable results in various domains and for determining the presence or absence of leaks. These models can achieve high precision, recall, and F1 scores; thus, they are valuable tools for automating and optimizing anomaly detection tasks, leading to more efficient monitoring and quicker identification of issues.

While existing ML-based leak detection methods have shown promise, many are susceptible to environmental noise, such as that caused by road traffic and other vehicles, which can lead to false positives or missed detections. Furthermore, a critical assessment of the limitations of these methods often reveals a lack of comprehensive comparative analysis across different ML algorithms in real-world simulated environments and insufficient focus on robust, noise-immune sensing technologies.

In this study, we use ML techniques to confirm the presence of leaks. To overcome the problem of ambient noise, we specifically utilize pressure transmitters, which are demonstrably more robust to environmental disturbances compared to acoustic or vibration sensors alone. This combination of noise-resistant pressure data with advanced ML algorithms represents a significant step forward in reliable leak detection. To achieve this goal, we designed a new prototype pipeline of length 100 m and diameter 40 mm to simulate a real distribution network, on which two pressure transducers will be installed at previously known positions. A 12 mm circular hole was drilled at a distance of 14.5 m from the pump to simulate leakage. We installed a 220-volt solenoid valve with a power rating of 8 watts, which was controlled by a push button at the drilled location to stimulate leakage.

Based on the gathered data, we employed ML-based algorithms to accurately categorize the pressure transducer signals recorded in the presence and absence of leakage. The suggested approach uses the superior data processing and pattern recognition capabilities inherent in ML techniques to recognize small variations in pressure signals, hence providing reliable detection of leak events. This study makes a significant contribution to the field of leak detection in water distribution networks by combining precise data acquisition with state-of-the-art analytical methods. The main contributions of this study are summarized as follows:

- We introduce a novel approach that integrates highly noise-immune pressure transducers with a comprehensive comparative benchmarking of several high-performance machine learning algorithms. This work specifically addresses the critical research gap related to environmental noise interference in leak detection scenarios using an experimental pipeline network in the laboratory.
- The signals are collected from the WDNs using pressure transducers, and the pressure levels are continuously monitored and recorded. These transducers are highly sensitive sensors meant to measure and convert the pressure within the pipelines. Network monitoring can be performed remotely utilizing transducers connected with Wi-Fi transmitters positioned on crucial portions to cover the whole WDN. The Wi-Fi transmitters allow the collected data, including the pressure levels and fluctuations, to be transferred to centralized monitoring systems or cloud-based platforms.
- We design a system architecture for efficient leak detection using several high-performance ML algorithms and IoT. This system aims to ensure accurate, real-time detection of leaks within WDNs, minimizing water loss, reducing operational costs, and improving the overall reliability of urban infrastructure. While the paper focuses on the core leak detection methodology, we acknowledge that full-scale IoT integration would require further consideration of data transmission latency.
- We rigorously test and evaluate the performance of multiple well-known ML algorithms for leak detection, which is vital for assuring a system's robustness, adaptability, and efficiency in diverse settings. This includes a detailed analysis of their accuracy, efficiency, and suitability for real-world application, offering valuable insights into their comparative strengths and weaknesses.

The remainder of this paper is organized as follows: In Section 2, we examine the research background on leak detection in WDNs, focusing on structural and operational weaknesses. This strategy poses the issues of undetected or delayed leak identification. These risks, including old infrastructure and pressure changes, often lead to leaks that go undetected or are identified too late. This section illustrates the issues caused by such delayed identification, resulting in water loss, increased expenditures, and potential damage, notably to pressure transducers, in real-time data monitoring. The integration of ML and pressure transducers offers innovative solutions. In Section 3, we present related works that review previous research and studies related to the topic of this paper. this sets the context by summarizing existing knowledge and identifying gaps that the current study addresses. This section contextualizes the current study by explaining the system, solutions, techniques, advantages, and downsides. In Section 4, we cover the data collecting process and acquisition and detail how the data were obtained and acquired for the study. This section discusses the procedures, tools, and techniques used to obtain data. In addition, it describes the development of the ML models, outlining the steps required to build and train these models, including data preprocessing, feature selection, and model evaluation to ensure optimal leak detection performance. Section 5 describes the performance of ML models, their implications and limitations, and gives a comparison of different ML methods for water leak detection, discusses the implications of the results, highlights the strengths and weaknesses of each model, and evaluates their accuracy, efficiency, and suitability for leak detection. This study is summarized in a general conclusion that summarizes the obtained results.

2. Research Background

WDNs are essential infrastructures for delivering potable water to communities. However, these systems are prone to leaks, leading to significant water loss, high repair costs, and potential damage to the infrastructure [21]. Traditional leak detection methods are often inefficient and require considerable time and manual effort [22]. The advancement of ML offers an innovative approach to enhance leak detection by automating the process, improving accuracy, and reducing operational delays [23]. ML algorithms can analyze pressure fluctuations in WDNs and differentiate between leak and no-leak events, thereby enabling faster identification of problem areas [24]. By using pressure transducers for their precision and resistance to environmental noise, combined with advanced data acquisition tools, researchers can generate reliable data for effective model training. This study explores various ML algorithms to accurately classify leaks, demonstrating the potential of ML-based solutions to optimize leak detection processes and mitigate water loss in WDNs [25].

2.1. Water Distribution Network Systems

WDNs are complex infrastructures that deliver clean water from treatment plants or reservoirs to end-consumers such as residential, industrial, and commercial buildings. These networks consist of extensive underground pipelines, pumps, storage tanks, valves, and pressure sensors, and they work together to maintain the water pressure and flow across various points in the system. Effective management of WDNs ensures a consistent water supply to communities, making them vital to public health and economic stability.

The WDNs structure is typically hierarchical, with large main pipes branching into smaller distribution lines. The key components include pressure transducers and sensors, which monitor the water flow and detect fluctuations caused by leaks or breaks in the pipes. Leak detection is crucial for maintaining system integrity and reducing water loss, which has become a growing concern in aging infrastructures. Technological advancements, such as the integration of ML and smart sensors, offer promising solutions to improve leak detection and system performance [26], as shown in Fig. 1.

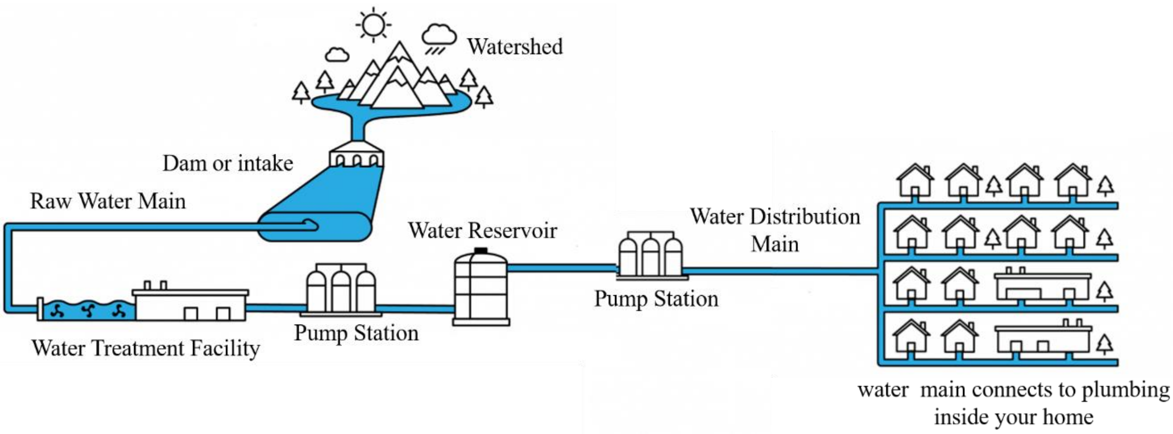

Fig. 1. Basic structure of a water distribution network (WDN)

2.2. Leaks in WDNs

Leaks in WDNs pose significant challenges, leading to water loss, infrastructure damage, and increased maintenance costs. Leaks can occur for various reasons, such as pipe aging, external damage, and pressure fluctuations. These issues often go undetected for long periods, worsening the problem by allowing water to escape unnoticed. As the leak worsens, it weakens the surrounding infrastructure, causing additional damage such as sinkholes and road collapses. Detecting leaks early is critical to minimize their impact. Conventional leak detection methods, such as manual inspections and acoustic sensors, are often time-consuming and ineffective for underground networks. The integration of modern technologies, such as pressure sensors and machine learning algorithms, has shown promise in improving leak detection speed and accuracy. These systems can detect changes in water flow and pressure, alerting operators of potential leaks before they cause extensive damage.

2.3. Leak Detection

Leak detection in WDNs faces several challenges. One of the main challenges is the early identification of leaks, particularly when they are small or hidden within underground pipes [27]. Leaks are often subtle and can go undetected for long periods, causing significant water loss and infrastructure damage. In large and complex WDNs, traditional methods such as acoustic monitoring or manual inspection may be inefficient due to the extensive labor and time. In addition, background noise and environmental factors can interfere with sensor readings, further complicating detection [28]. The limitations of conventional approaches highlight the need for more advanced techniques, such as integrating smart sensors or ML algorithms, that offer real-time monitoring capabilities [29]. However, even these newer methods face challenges in environments with high variability, which makes leak detection a persistent issue in ensuring efficient and sustainable water distribution [30]. As shown in Fig. 2.

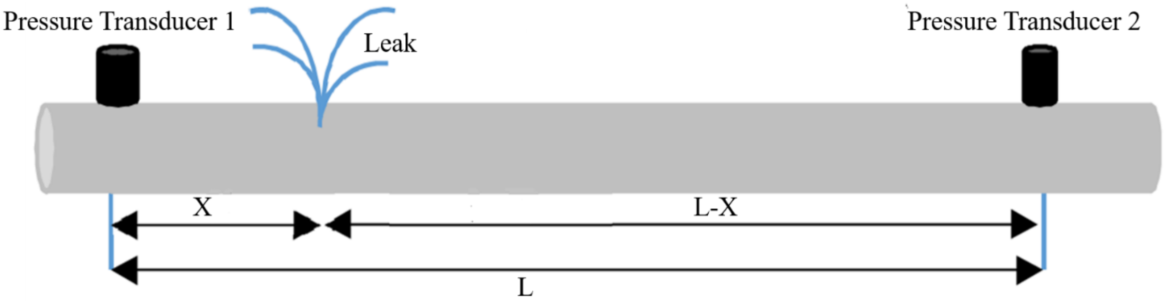

Fig. 2. Leak in a section of the pipe

2.4. Internet of Things (IoT)

IoT is defined as a collection of devices that are equipped with special sensors and microchips; thereby, the devices are capable of collecting and sharing by employing the extensive communication network that they form [31]. This situation creates additional optimization opportunities for the application and increases the number of automated operations in the application [32]. These consequences reduce costs and increase efficiency [33]. To achieve these capabilities, IoT uses several technologies, such as advanced sensors and actuators, Bluetooth, cellular data, WIFI, RFID, NFC, Ethernet, 5G, and cloud computing [34]. IoT technology has been described as a promising innovation [11]. The original concept of the IoT was centered on achieving connectivity among all things and, objects to create a digital and intelligent network system [35]. In this system, devices can communicate and collaborate, streamlining and accelerating various tasks and processes [36], as illustrated in Fig. 3. Within the IoT ecosystem, users interact with IoT devices via smartphones and computers, offering significant convenience in areas such as industry, home management, and healthcare [37].

2.5. ML and IoT Applications

ML models trained, on data from IoT devices have, significantly improved the efficiency of various activities, including healthcare, coordination, and urban planning. In areas requiring constant monitoring, small transducers of Wi-Fi technology continuously collect information such as weight, flow rate, and acoustic signals and transmit them in real time for analysis [38]. By leveraging verifiable data in near real-time, ML models can predict the likelihood of anomalies occurring, thereby improving support efforts and reducing the effort required with traditional methods [39]. However, despite these benefits, the massive amount of sensitive data collected by IoT devices raises security concerns. This information could be exploited for malicious purposes, such as controlling healthcare choices based on health tracking information or causing disruptions to the underlying infrastructure by hacking into IoT systems [40]. These dangers, if not addressed, can have serious consequences for people, businesses, and security. Therefore, protecting sensitive data used for ML training has become a fundamental issue, which has attracted increased attention from academia and industry partners [41]. In the WDNs context, it is essential to ensure data security and integrity to avoid unauthorized access or modification, which may compromise the productivity of leak detection and structure operations in general [42]. Adjusting the preferences of ML-IoT integration with the requirement of rigorous data protection is essential to maximize the exploitation potential in the field of WDNs and broader applications [43]. The integration of IoT with ML and transducers can transform the way data are collected and analyzed in different applications [44]. In this system, IoT devices equipped with transducers continuously collect data on changes in real time. The data are then transmitted to central systems, where ML algorithms analyze them to detect regularities or irregularities as well as potential deficiencies [45]. The combined use of IoT and ML improves efficiency, enables predictive support, reduces operational costs, and improves the overall management of systems in which monitoring is crucial.

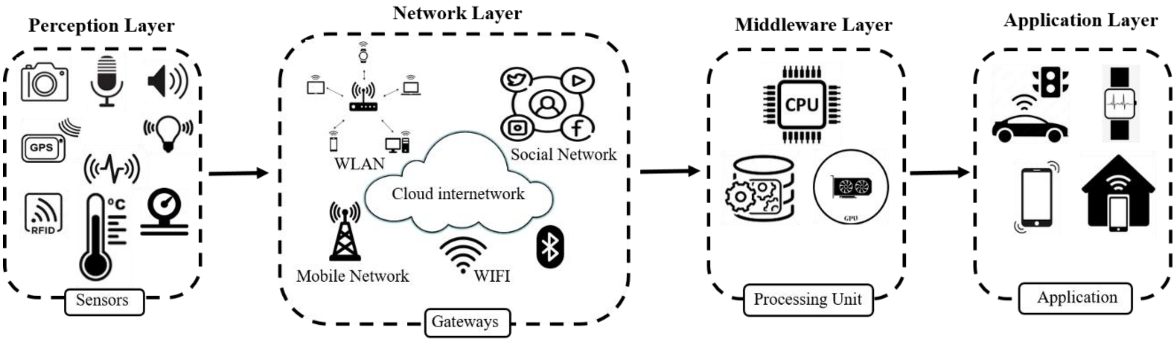

Fig. 3. Internet of things (IoT) architecture

2.6. Ongoing General Problems

An ongoing general problem in water distribution networks is the challenge of rapid and accurate leak detection. Traditional methods such as manual inspections and acoustic techniques are often insufficient due to their inefficiency in detecting small hidden leaks, especially in large or complex networks. These approaches are also time-consuming and labor-intensive, making them unsuitable for real-time monitoring of large-scale water distribution networks. In addition, ambient noise and pipe materials further reduce the accuracy of conventional methods. Given the importance of maintaining water infrastructure, there is a growing need for more advanced leak detection techniques. Integrating pressure transducers with ML algorithms can provide a more accurate and real-time solutions for identifying and localizing leaks in water distribution networks. This approach will help utilities respond more efficiently, save water, and minimize the economic impact of leak-related damage.

3. Related Works

This section provides a detailed analysis of recent state-of-the-art (SOTA) studies on leak detection in WDNs using ML techniques. This section summarizes the system, solution, methods, advantages and Disadvantages in each study to highlight their effectiveness and limitations, as presented in Table 1.

Tariq et al. investigated the use of MEMS-based accelerometers for leak detection in WDNs. Addressing significant water losses due to leakage; their study examines the performance of these accelerometers in metallic and non-metallic pipes over a 10-month test. They employed a stacked experimental design to accelerate the data, and they processed and developed ML models, such as random forest and AdaBoost to enhance the leak detection accuracy. Their findings showed that the random forest model achieved an accuracy rate of 100% for steel pipe and 94.93% for nonmetallic pipe and highlighted the potential of MEMS-based accelerometers in improving leak detection and contributing to water conservation measures [46].

Liu et al. introduced an advanced method for detecting leaks in WDNs using a Long Short-Term Memory Generative Adversarial Network (LSTM-GAN). This approach overcomes the limitations of traditional ML-based acoustic methods, which struggle with data scarcity and lack of diversity. The LSTM-GAN model generates synthetic leak signals to enhance the dataset, thereby improving leak detection performance. Evaluations using t-SNE analysis and acoustic characteristics show that LSTM-GAN outperforms other generative methods. The findings suggest that the propose approach significantly enhances the robustness of ML models for water leak detection, providing a novel solution to the challenges associated with limited real data [47].

Banjara et al. explore the efficacy of combining acoustic emission AE techniques with ML algorithms to detect and localize pipeline leaks. AE techniques are valued for their sensitivity to weak acoustic signals indicating faults. The study simulates pipeline leaks using pressure release valves and analyzes the AE signals through features such as AE counts, cumulative AE energy, and signal strength. These features are processed using SVM and relevance vector machine (RVM) algorithms to classify and localize leaks. Results show that both SVM and RVM effectively identify and classify leaks, with SVM demonstrating superior accuracy in binary and multiclass classifications [48].

Puning Xue et al. highlighted the gap in the application of ML methods for leakage detection in district heating networks. This study introduces a novel approach using ML to detect leakage faults by analyzing data from flowmeters and pressure sensors. Incorporating a delayed alert triggering algorithm and hydraulic simulation models, the proposed method achieves high leak detection accuracy. This proposed method provides a feasible solution for timely and accurate leakage detection in district heating networks [49].

Rodolfo Pinheiro da Cruz et al. presented a novel technique for detecting and locating leaks in low-pressure gas pipelines using acoustic sensors and ML algorithms. This study addresses the significant issues of environmental disasters, economic losses, and potential fatalities caused by pipeline leaks. Traditional methods often struggle to detect small leaks and reducer false alarms. This method uses acoustic signals captured by microphones and analyzed using ML models like LR, KNN, and SVM. The experimental results show high accuracy in leak detection (99.6%) and low false alarm rates (0.3%), with a maximum location error of 4.31%. This approach is a reliable and efficient alternative for monitoring low-pressure pipelines, enhancing safety, and reducing hazards [50].

Mashhadi et al. used ML techniques to detect and localize leaks in water distribution systems using, Lille University campus as a case study. They evaluated six ML methods: LR, DT, RF, hierarchical classification, PCA with K-means, and artificial neural networks ANN. Data were generated using the EPANET software to simulate various leakage scenarios. The results demonstrate that supervised methods, particularly LR, RF, and ANN, methods, demonstrated high accuracy in localizing leaks, whereas unsupervised methods struggled due to overlapping clusters. This research highlights the potential of ML to improve leak detection accuracy in complex water networks, although further validation with real-world data is needed [51].

Lee and Yoo developed a leak detection model for WDNs using RNN-LSTM algorithms. Their approach, which utilizes inflow meter data and a multi-threshold model, aimed to reduce false predictions. Data were generated using EPANET to simulate various scenarios. Evaluating with actual leakage data, the model showed over 90% leak detection accuracy, excluding singularities. This research underscores the effectiveness of DL in accurately and promptly identifying leaks in water distribution systems, thereby enhancing smart water infrastructure management [52].

Niamat Ullah, et al. The proposed an ML-based leakage detection system using AE sensor data. Features such as kurtosis, skewness, and RMS were extracted from the AE signals to train models such as neural networks, DT, RF, and KNN. The proposed system achieved a high classification accuracy of 99% for detecting water and gas pipeline leaks, including pinhole-sized leaks. This demonstrates the effectiveness of AE technology and ML in pipeline monitoring, with further potential for improving scalability, adaptability, and computational efficiency [53].

Coelho et al. developed an IoT-based system for real-time water leak detection using a wireless sensor network and ML. The system employs low-cost sensors to monitor water flow and uses RF algorithms for analysis. The proposed implementation demonstrated 75% accuracy in real-world settings. This cost-effective solution improves water management efficiency for both agricultural and household applications [54].

Table 1. Summary and comparative analysis of related work

Refs	System	Solution	Methods	Advantage	Disadvantage
[54]	• Sensor Network. • Data Aggregation and Transmission. • ML Analysis. • Low Power and Cost-Efficiency.	Real-time leak detection in WDNs using sensors and ML models and like SVM, DT, and NN, to identify leaks.	• SVM. • DT. • RF. • Neural Networks (NN). • XGBoost.	• Real-Time Leak Detection. • Efficiency and Precision. • Cost-Efficient Communication. • Improved ML Approach.	• Limited Dataset. • Lack of Real-World Validation. • Expensive Sensors. • Limited Sensor Integration.
[50]	• Real-time monitoring. • Acoustic sensing. • Signal processing. • Data acquisition. • ML models. • Performance optimization.	Improve the UNICAMP experiment with larger samples, noise cancelation, and benchmarking.	• LR. • KNN. • SVM-L. • SVM-RBF. • RF. • AdaBoost. • XG-Boost.	• Comprehensive Data Collection. • ML Implementation. • Acoustic Sensing Technology.	• Limited Experiment Size. • Dependency on Acoustic Data. • Lack of Real-World Validation. • Computational and Resource Constraints. • Benchmarking Issues.
[49]	• Leakage Simulation • Data Collection • ML Integration • Baseline Comparison • Challenges	DHN leaks were modeled using XG-Boost, hydraulic simulation, and optimized automation.	• SVM • XG-Boost.	• Simulated Leakage Scenarios. • Comprehensive Dataset • Baseline Data for Comparison. • ML Approach.	• Limited Training Data. • Manual Hyperparameter Specification. • Binary Classification Focus. • Generalization Challenges. • Reliance on ML Algorithms.
[48]	• Data collection. • Signal processing. • ML analysis. • Adjustable mode.	Improve WDN leak detection by automating sampling, optimizing ML, and generalizing.	• SVM. • RVM.	• Field-based experiments. • Strategic sensor placement. • Comprehensive dataset. • Use of advanced ML models.	• Model performance fluctuation. • Context-specific result. • Acoustic properties are not fully addressed. • Further experiments are required.
[51]	• Hydraulic Zone. • Data Collection & Simulation. • ML models. • Optimization of Detection.	Improve WDN leak detection by granularity, temporal analysis, and automation.	• LR. • DT. • RF. • Hierarchical Classification. • PCA with K-	• Detailed Simulation. • Large-scale Network. • Diverse Leak Scenarios	• Limited Granularity. • Temporal Limitations. • Complex Model Setup.

Refs	System	Solution	Methods	Advantage	Disadvantage
	• Real-time Alerts.		means.	• Multiple Models	• Potential Overfitting.
[52]	• RNN-LSTM Prediction. • Time-Series Data. • Real-Time Monitoring. • Data Input to Model. • Extended Data Sets. • Frequent Re-Evaluation.	Improve accuracy with diverse data, optimized patterns, and model comparisons.	• ANN. • RNN-LSTM-based flow prediction.	• Real Data Usage. • Flow Rate Prediction. • Daily Pattern Capture.	• Limited Data. • Shape Extraction Problems. • Need for Performance Evaluation. • Lack of Comparative Analysis.
[46]	• Acoustic Sensors & Accelerometers. • Midnight Data Collection. • ML. • Real-Time Monitoring. • Extended Experimentation.	Enhance leak detection via real-world tests, advanced sensors, and ML.	• RF • AdaBoost	• Acoustic Signal Collection. • Midnight Data Collection. • Increased Data Volume • Application of ML.	• Laboratory and Testbed Settings. • Multiple Leak Scenarios & Ambient Noise. • Simplified Pipe Geometries. • Challenges encountered by MEMS-Based Accelerometers.
[21]	• AE-Based Monitoring System • ML Algorithms for Leak Detection • Adaptive Threshold-Based Sliding Window • User Interface for Monitoring • Real-Time Alerts	Improve leakage detection using CNNs, expanded data, and real-time monitoring.	• LR • DT • KNN • RF	• High Accuracy: achieved of 99% • Comprehensive Feature Extraction • Classifiers • Real-Time Leak Detection • Diverse Dataset	• Limited Leak Sizes • Dataset Limitations • Algorithm Complexity • Pressure Variations
[15]	• LSTM-GAN • Analyze the captured acoustic signals • Predict leak occurrence • Enhanced dataset created from 1s audio samples • Continuous Improvement	Enhance leak detection with outlier detection, GANs, acoustic studies, and context.	• LSTM-GAN • Adversarial Network	• Comprehensive Data Collection • Increased Data Volume • Real-World Application	• Potential Outliers and Noise • Need for Conditional GANs • Insufficient Acoustic Analysis • Lack of Contextual Information

4. Proposed Solution Based on Pressure Transducers

The proposed leak detector uses pressure transducers to capture the pressure signals. In doing so, the following sections first describe the hydraulic and electrical plant prototypes. Two pressure transducers are mounted on a pipe at a distance of L from each other. Initially, the position of the leak is estimated to be 14m away from the pump. The positions of the transducers relative to the leak are predetermined: one transducer is fixed at a position 1.5m from the leak, while the other was moved along the pipe at intervals of 1.5m, up to a distance of 78m. Measurements were initiated by activating a push-button to simulate leakage. The recording times was set to 20s, which is sufficient to describe the transient phenomenon of the leak. We took the middle of this interval approximately 10s after the start of each recording to trigger the leak using a push button. The leak is detected by a sudden change in the pressure signals, which indicates a pressure drop in the pipe. This pressure drop is detected by the transducers. The pressure wave propagated in both directions from the leak and reached the transducers at different times, as demonstrated in Fig. 4.

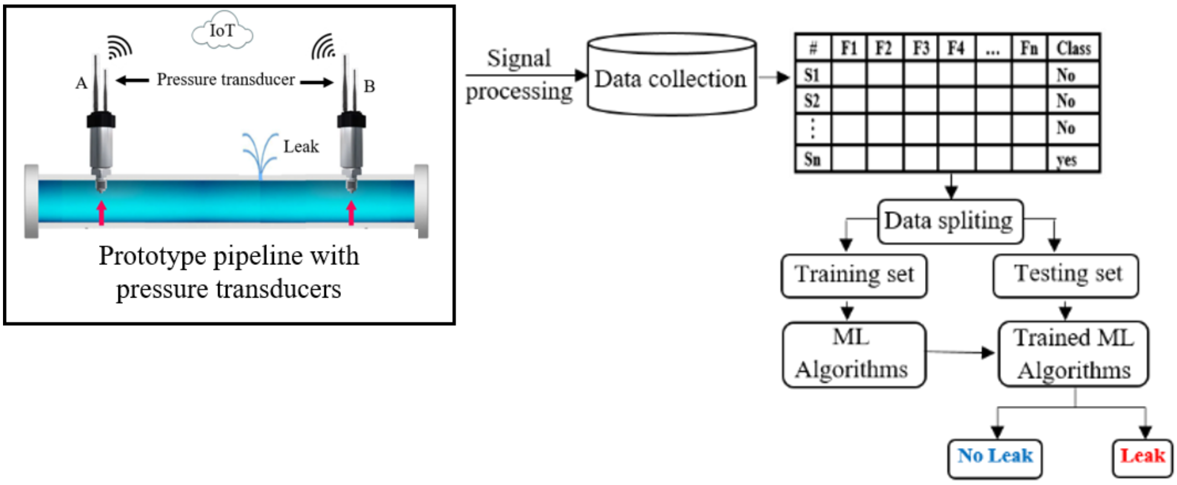

Fig. 4. Block diagram representing the implemented leak detection approach

4.1. Dataset Description

The dataset includes pressure transducer signal readings from a pipeline prototype designed and built in the laboratory to simulate distribution network leaks for detection to ensure the efficiency and safety of water distribution systems, as shown in Fig. 5. The proposed system is divided into two parts:

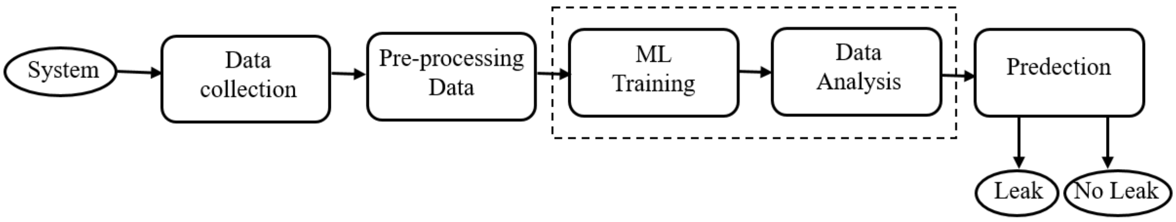

Fig. 5. Flowchart of leak detection system

- **Part 1: Hydraulic Circuit**

The hydraulic prototype circuit consisted of a closed HDPE pipe with a diameter of Ø=40 mm, length of 100 m, and thickness of 2.4 mm. The latter was manufactured in 2022 by PLAST TUBE in SETIF, as well as several elements such as the 150 L tank, the pump with (P = 0.8 Kw, flow rate = 100 L/min, and RPM = 2850), two highly sensitive pressure transducers, model 811FMA with a pressure range of 0-100 PSIG, with an accuracy of ±0.25% *FS* (Full Scale), and other accessories, as shown in Fig. 6.

- **Part 2: Acquisition system**

This part is an acquisition system comprising a conditioning card for exploiting the 4-20 mA analog signals coming from the two transducers. The real-time control of the continuous system was performed using a PC connected to a professional acquisition card (dSPACE DS1104). We took the sampling frequency equal to 1 kHz. The signals from the pressure transmitters, which are in the 4-20 mA range, are relatively weak and are accurately modeled using Equation (1) given by the manufacturer. The model of the pressure image as a function of the transducer current is given by the following function extracted from the technical sheet:

$$I(mA) = 0{,}16 * P + 4mA \tag{1}$$

The leak was simulated by a hole drilled in the pipe connected to a solenoid valve (8 watts, 0.7 bar, 220 V). The pressure in the pipeline is measured by the pressure transmitters, which provide a current proportional to the pressure at a given point. The current is converted into voltage by a conditioning card. The loop requires a 24 V power supply supplied by a stabilized power supply. The two pressure signals were recorded and visualized by a computer connected to a dSPACE professional acquisition card. The signals are recorded in a file with a CSV extension and processed using programs developed in MATLAB. As shown in Fig. 7.

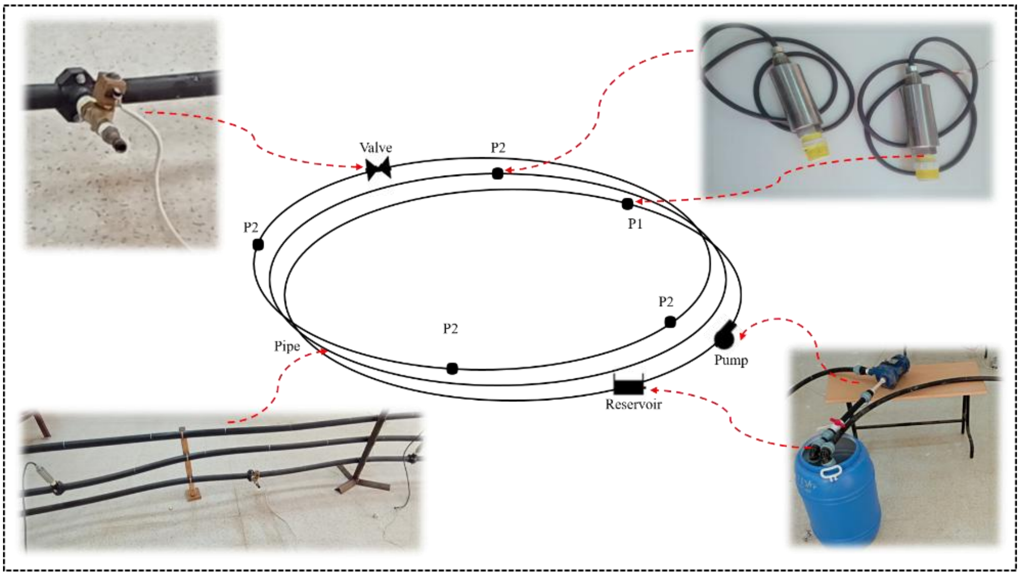

Fig. 6. Experimental hydraulic plant used for leak detection

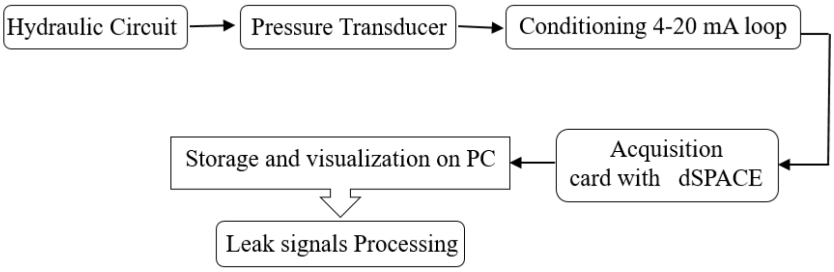

Fig. 7. Synoptic of the acquisition system

4.1.1. Data Collection

The signals from the pressure transducers are continuously monitored and recorded. These raw time-series pressure data are then processed to extract meaningful features that represent pressure variations over time. Data is collected in a tabular format, where each row represented a sample (S1, S2, ..., Sn) with corresponding features (F1, F2, ..., Fn) and a class label indicating whether a leak was detected ("Yes or No leak). The transducers were calibrated prior to deployment to ensure sensitivity to pressure deviations indicative of leakage while reducing the influence of environmental noise such as road traffic. Each sample corresponds to a time window during which transient pressure responses were captured. Sampling was conducted under both leak and no-leak conditions in a controlled lab setup, with a fixed leak size and location.

4.1.2. Pre-Processing Data

Data splitting is a crucial step in ML because the collected data are divided into two distinct sets. The first set is the Training Set, which is used to train the ML algorithms. This set provides the model with the required information to learn patterns and make predictions. The testing set is used to evaluate the performance of the trained model. By assessing how well the model performs on this separate set of data, we can gauge its accuracy and effectiveness when making predictions on new, unseen data.

Prior to training, the data underwent preprocessing steps including outlier removal and normalization, Features were extracted using both time-domain. These features were selected based on domain knowledge and their separability across leak and non-leak conditions.

4.1.3. ML Training and Data Analysis

In this section, the major classification techniques are discussed along with their basic works. We used the following ML algorithms for modeling: SVM, DT, RF, NB, LR, KNN, and XGBoost. In the ML training process, various algorithms, including decision trees and neural networks, are trained on the features extracted from the training set. During this phase, the algorithms learn to recognize the patterns in the data that correspond to the conditions of leakage or no leakage, the next step involves data analysis, where the trained ML models analyze the testing data. These models are equipped to make predictions based on the features provided, effectively identifying leak versus no leak scenarios. This systematic approach allows the models to apply their learned knowledge to new data, thus enhancing their predictive capabilities. We selected a diverse sets of well-known and widely used machine learning algorithms for this study to provide a comprehensive comparative analysis of their effectiveness in leak detection. The chosen algorithms include:

- Support Vector Machine (SVM)

SVM a powerful algorithm known for its ability to find optimal hyperplanes for classification, effective in high-dimensional spaces and for non-linear decision boundaries through various kernel functions [55]. As shown in Fig. 8.

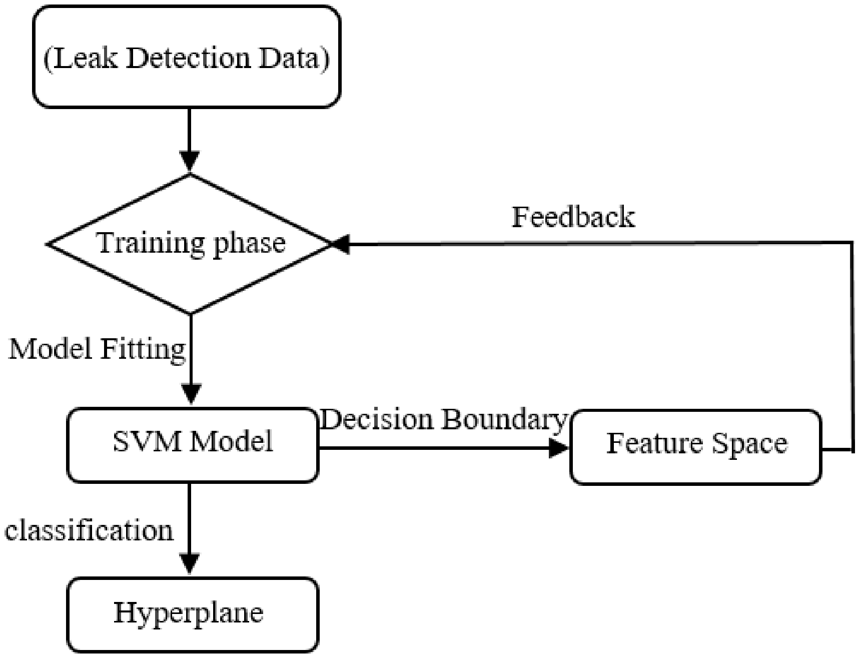

Fig. 8. Flowchart support vector machine classification

- Decision Tree (DT)

DT is a simple, interpretable model that partitions the data based on feature values, useful for understanding decision rules. The optimal tree division is typically determined using entropy Equation (2) or Classification error, resulting in feature need scores for interpretation [56]. As illustrated in Fig. 9.

$$Entropy = -\sum_{j=1}^{m} p_{ij} log_2 pij \quad (2)$$

- Random Forest (RF)

Vol. 5, No. 4, 2025, pp. 2218-2245

RF is an ensemble method that builds multiple decision trees and merges their predictions to improve accuracy and control overfitting. It is robust to noise and highly effective for complex datasets [57], as illustrated in Fig. 10.

- Naive Bayes (NB)

The proposed classifier is a probabilistic classifier based on Bayes' theorem, assuming independence between features. It is simple and computationally efficient, as shown in Equation (3), despite strong perceptions of simplicity and layer independence, NB performs remarkably well in a variety of classification tasks due to its efficiency and scalability [58].

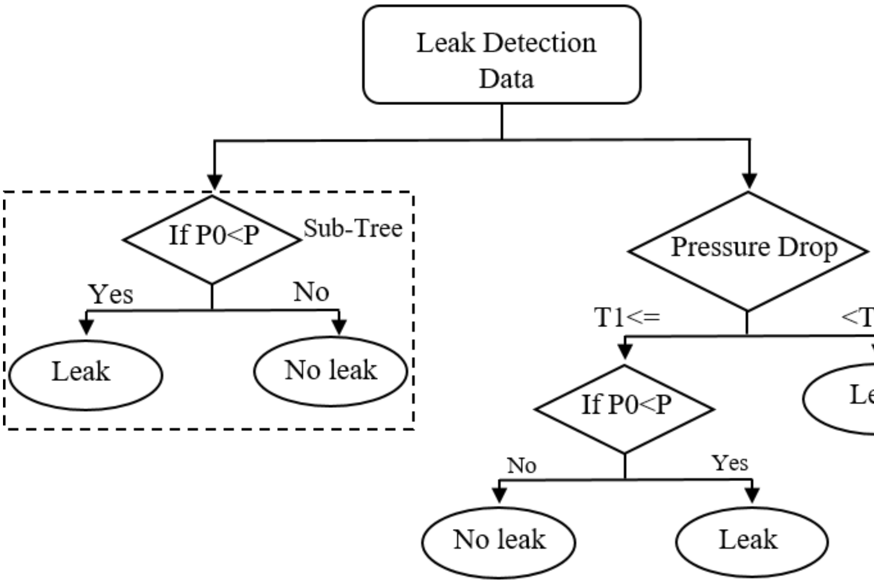

Fig. 9. Flowchart of the decision tree structure in machine learning

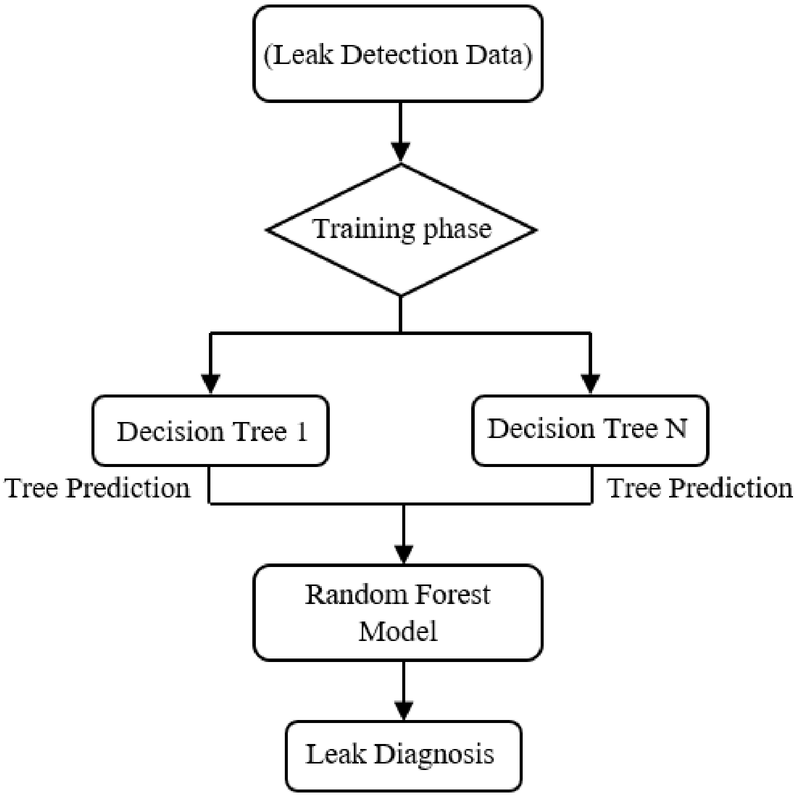

Fig. 10. Flowchart of the random forest in machine learning

$$P(Y = y|X = x) = \frac{P(Y = y)\prod_{i=1}^{n} P(X = x_i|Y = y)}{\sum_i^C P(y_i, X = x)} \tag{3}$$

- Logistic Regression (LR)

A linear classification model often effective for binary classification tasks where classes are linearly separable. The following Equation (4)-(6) illustrate how to determine P(X). LR is valued for its efficiency, especially when dealing with large datasets [59].

$$logistic(p) = ln\left[\frac{p(X)}{1 - p(X)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{4}$$

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k} \tag{5}$$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}} \tag{6}$$

- K-Nearest Neighbor (KNN)

The KNN algorithm is a non-parametric, instance-based learning algorithm that classifies data points based on the majority class of their nearest neighbors [60]. As demonstrated in Fig. 11.

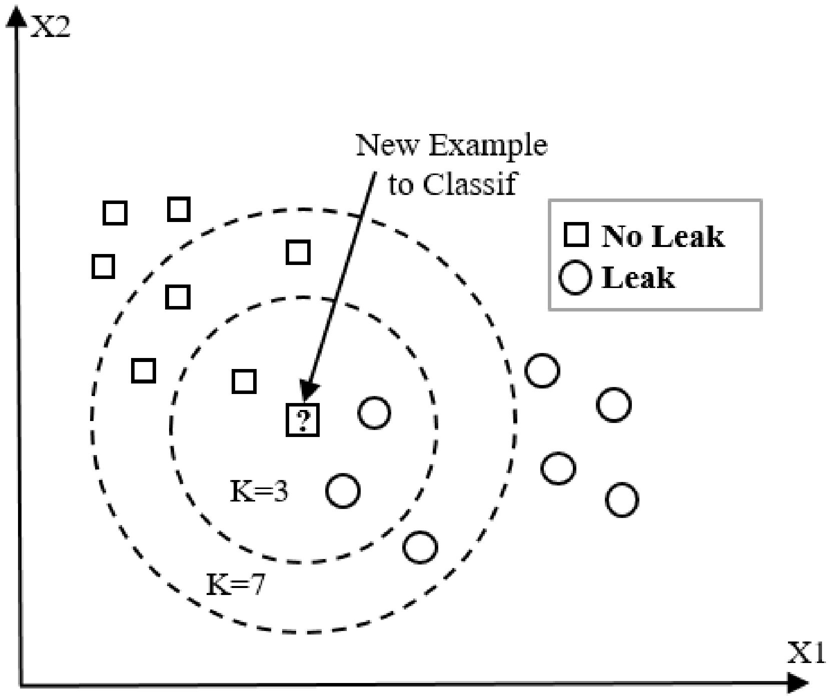

Fig. 11. K-Nearest neighbor in machine learning

- Extreme Gradient Boosting (XGBoost)

Shorten for XGBoost is an ML model that is a highly efficient and robust gradient boosting framework known for its superior performance in various prediction tasks due to its optimized boosting technique. That combines predictions from simple learners typically DT, to generate complex models. The latter method uses a gradient growth framework to optimize the model by reducing a specific loss function. The objective function $Obj^{(t)}$in step, t is determined by Equation (7), which is recognized for its speed, scalability, and regularization methods for handling overfitting and missing values [61]. As shown in Fig. 12.

Vol. 5, No. 4, 2025, pp. 2218-2245

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{t} \Omega(f_t) \quad (7)$$

4.2. Key Performance Evaluation

In fact, the confusion matrix is a valuable tool for evaluating the performance of classifiers in binary classification tasks. This subsection summarizes the predictions made by the classifier compared to the actual labels. The four possible prediction outcomes in a binary classification confusion matrix are:

- True Positive (TP): The classifier correctly predicted the positive class.
- True Negative (TN): The classifier correctly predicted the negative class.
- False Positive (FP) (Type I error): The classifier incorrectly predicted the positive class when the actual class was negative.
- False Negative (FN) (Type II error): The classifier incorrectly predicted the negative class when the actual class was positive.

From these four outcomes, several key performance metrics can be calculated such as Accuracy, Sensitivity, Specificity, Precision, and F1-score.

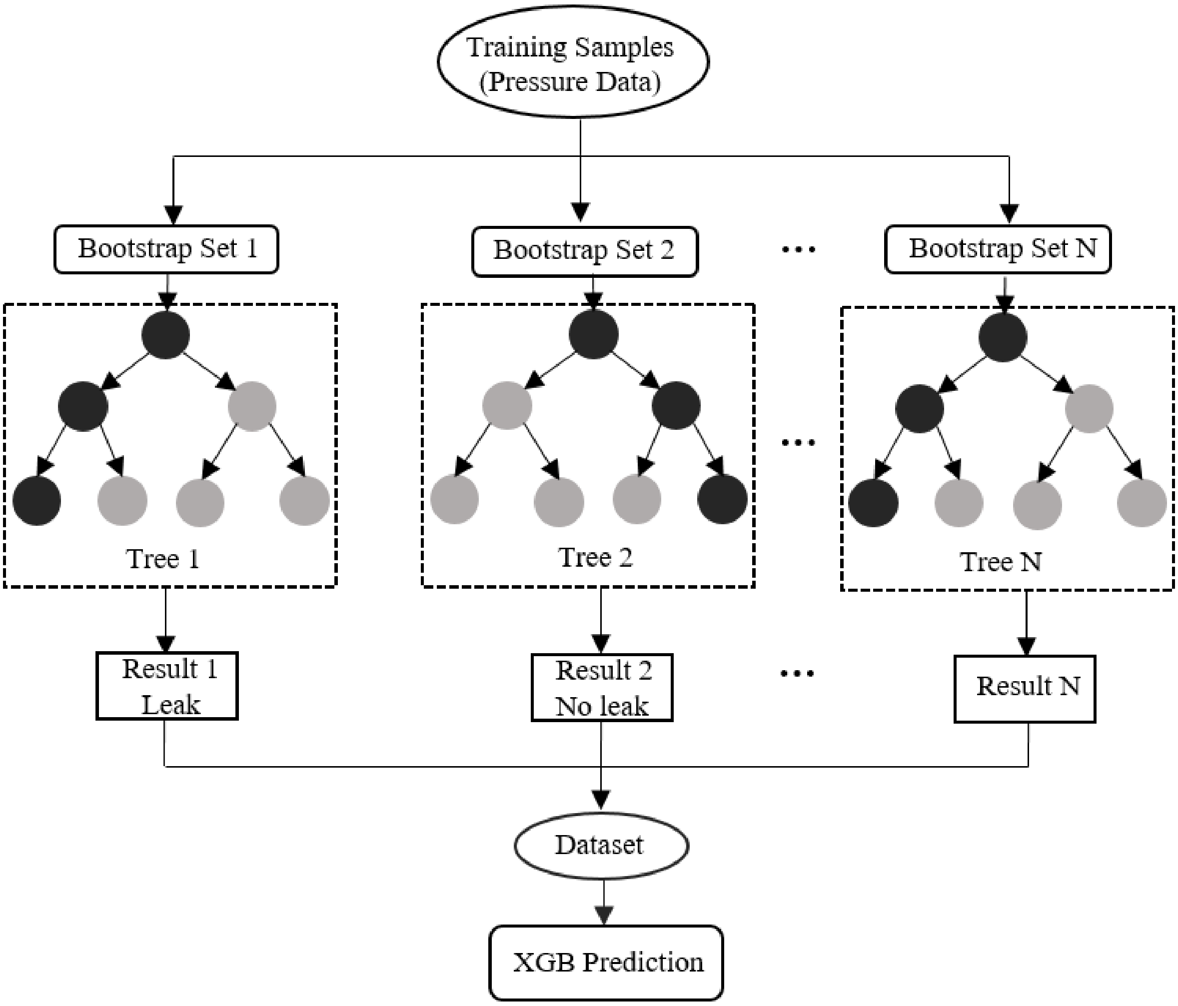

Fig. 12. Flowchart of the extreme gradient boosting in machine learning

4.2.1. Accuracy

Accuracy is a statistic used in ML to assess a classification model's performance. True positives and true negatives represent the proportion of accurately anticipated instances to the total number of cases [62], as illustrated in Fig. 13. The accuracy formula is determined by Equation (8).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \tag{8}$$

4.2.2. Sensitivity

Sensitivity is an ML variable that measures a classification model's efficiency in identifying positive cases. The proposed method focuses on the model's ability to separate true positives from all actual positives [63], as presented in Fig. 14. The sensitivity formula is derived from Equation (9).

$$Sensitivity = \frac{TP}{TP + FN} * 100\% \tag{9}$$

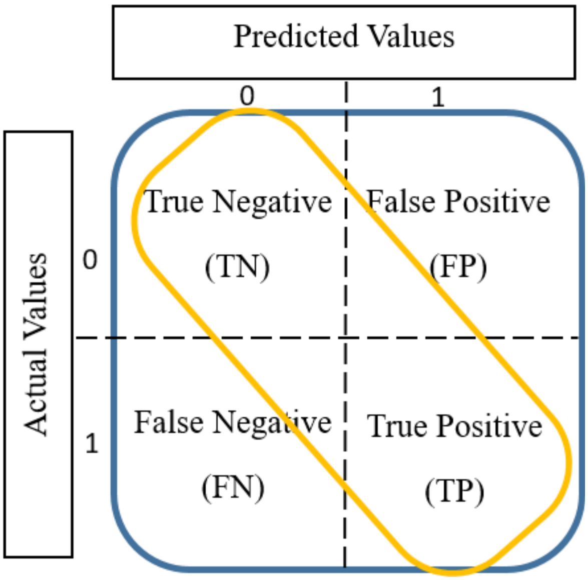

Fig. 13. Confusion matrix for binary classification and accuracy formula illustration

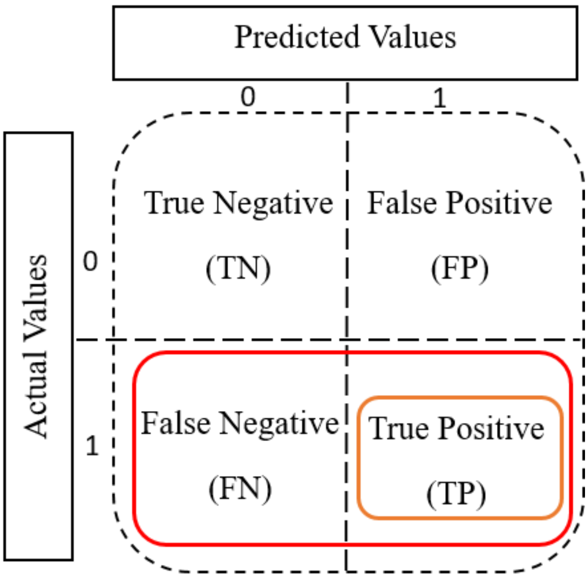

Fig. 14. Binary classification confusion matrix and sensitivity calculation

4.2.3. Specificity

In ML, specificity is a metric that measures a model's ability to correctly identify negative instances. This section focuses on how well the model detects true negatives out of all actual negative cases [64], as represented in Fig. 15. The formula for specificity is expressed in Equation (10).

$$Specificity = \frac{TN}{TN + FP} * 100\% \tag{10}$$

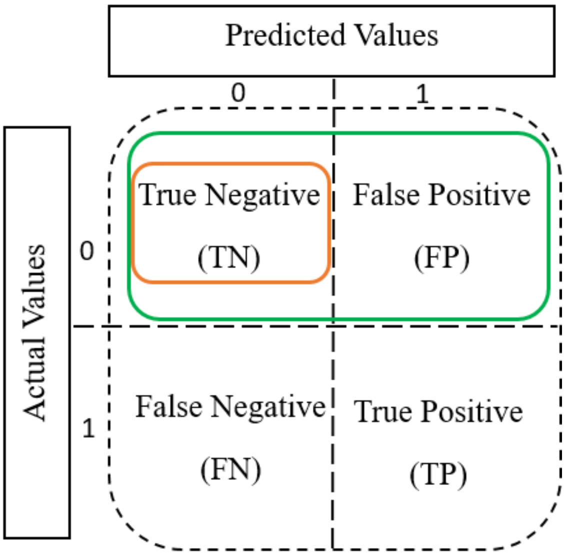

Fig. 15. Confusion matrix and specificity representation for binary classification tasks

4.2.4. Precision

In ML, precision measures the accuracy of positive predictions. This gives tells us the proportion of instances predicted as positive that are actually positive. Precision focuses on the number of predicted positives that are true positives, which is important in cases in which false positives carry significant consequences [64], as shown in Fig. 16. The formula for precision follows Equation (11).

$$Precision = \frac{TP}{TP + FP} * 100\% \tag{11}$$

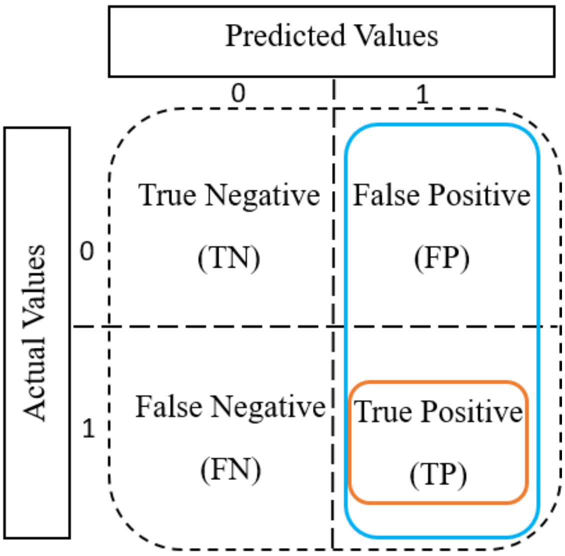

Fig. 16. Visual representation of precision and confusion matrix in binary classification

4.2.5. F1-Score

The F1 score is a metric used in ML to evaluate the performance of a classification model, particularly in cases where the class distribution is imbalanced. The proposed method combines precision and recall into a single score, providing a balance between precision and recall [65]. The F1 score is particularly useful to find an optimal balance between false positives and false negatives. Fig. 17, shows a confusion matrix for problems involving binary classification tasks.

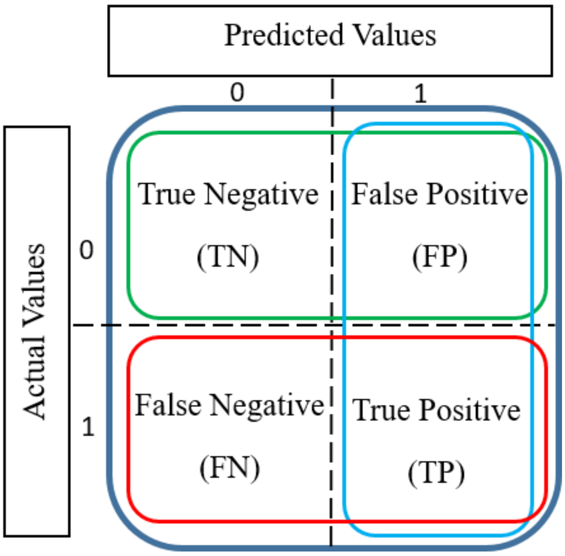

Fig. 17. F1-score evaluation and confusion matrix for binary classification

5. Results and Discussions

This section presents a comprehensive analysis of the performance of various ML algorithms applied to leak detection in our prototype circuit. We discuss the empirical results obtained from our experimental setup, evaluate the strengths and weaknesses of each model, and contextualize our findings within the broader literature.

5.1. Acquisition Signals

The signals from the previously read pressure transducers are analyzed. Detection is achieved by analyzing the two acquired signals. The leak location is determined by applying the equations obtained from the developed mathematical models. For this purpose, we recorded data at different positions along the pipeline with two pressure transducers placed on either side of the leak. The experimental results obtained using the proposed method are presented in this section, with the data curves shown in Fig. 18 and Fig. 19. It is clear that there are fluctuations in the pressure loss lines, indicating that the pressure loss curve is not smooth. These interferences can be effectively minimized using an adaptive filter. We have presented in Fig. 18 the signals without leaks for a randomly chosen position. The x-axis represents time in seconds, and the y-axis represents pressure in PSI. Fig. 18. Compares the noised and denoised pressure signals over time without leakage. The noise signal (red line) fluctuates around 164 PSI, showing rapid noise variations. The de-noised signal (blue line) stabilized at 166.5 PSI with a smoother, steady trend, indicating successful noise removal. Both signals remained consistent throughout the experiment, suggesting stable conditions and without leakage.

In Table 2, based on the provided data, pressure sensor P1 is fixed at a distance of 1.5 m, whereas pressure sensor P2 changes at different distances. The delay between the signals denoted as ΔT, varies with each P2 value. The relationship between the distance and the time delays could imply the fluid pressure dynamics in a system where P2's distance affects the time taken for the sensor to detect pressure changes. The greater the distance of P2, the longer the delay, suggesting that as the distance between the sensor and the variable pressure points increases, the system takes more time to stabilize or register a pressure change. Fig. 18 shows that the signal closest to the pump has a significantly higher noise level than the signal recorded farther away. This is due to fluctuations in the pumping system. As the distance between the transducer and pump increases, the magnitudes of the pressure signal decrease. Additionally, we observed a distinct signal pattern at the moment the leak occurred: a sharp pressure drop followed by a rise, followed by stabilization at an average value lower than the pre-leak state. Fig. 19 presents the signals captured after the leak at distances of 9m, 38m, 66m, and 76m.

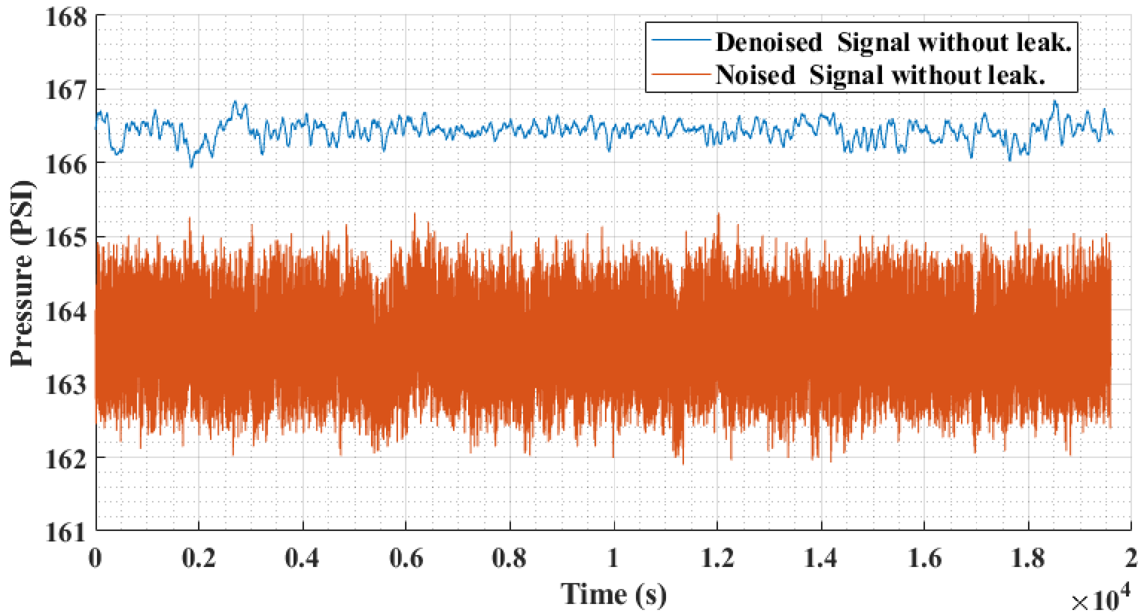

Fig. 18. Noised and denoised pressure signals without leaks were at distances of 1 m and 9 m from the pumping system, respectively

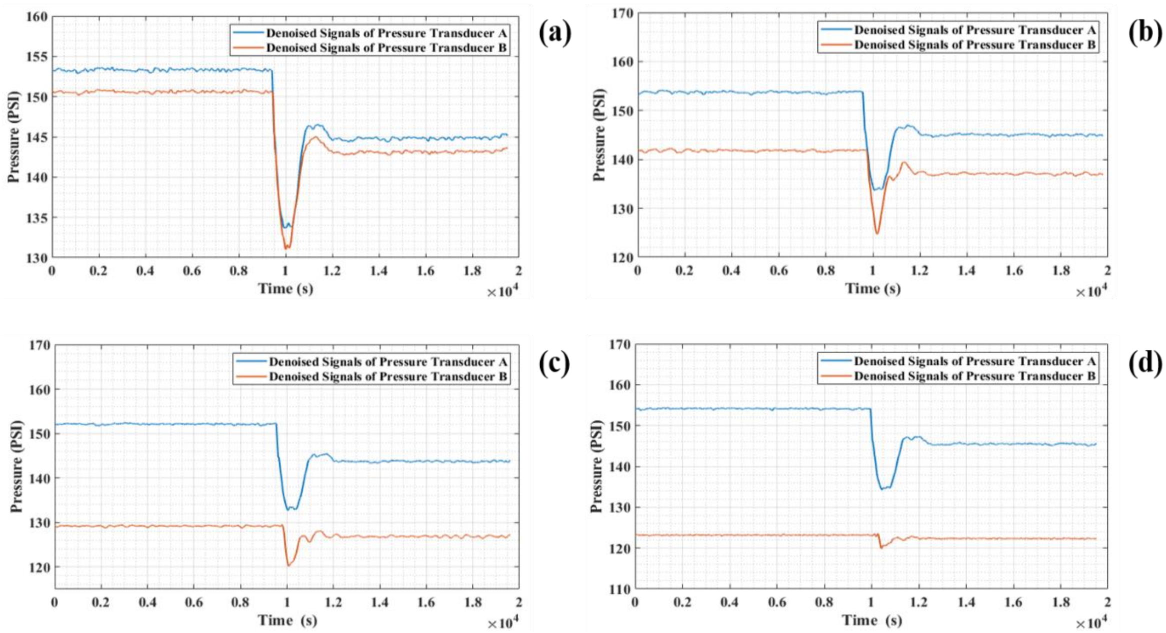

Fig. 19. Denoised pressure signals at various distances. (a) 9 m (b) 38 m (c) 66 m and (d) 76 m

5.2. Performance Assessment

In this section, the dataset used in this study was split into training and testing sets, with 80% of the dataset allocated to training the ML models and the remaining 20% was reserved for testing their performance on unseen data. The model's performance was evaluated using standard metrics, including accuracy, precision, recall, F1-score, and specificity. These metrics provide a comprehensive assessment of the model's ability to correctly classify water supply pressure data. Table 3 presents the performance assessment of the employed ML algorithms for water supply pressure data based on commonly used evaluation criteria. The results indicate a varied performance across the different machine learning models, with LR achieving a perfect score of 100% across all metrics (Accuracy, Precision, Recall, F1-Score, and Specificity). This outstanding performance suggests that LR was exceptionally effective in distinguishing between leak and non-leak conditions within our experimental dataset. In contrast, NB exhibited the lowest overall performance, particularly in Specificity (33.33%), indicating a higher rate of false positives compared to other models. This suggests that NB struggled to accurately identify non-leak instances. SVM achieved perfect Recall (100%), meaning it successfully identified all leak instances, but its Precision (76.19%) and Specificity (16.67%) were lower, leading to more false positives. RF, DT, and KNN demonstrated robust and similar performance, with F1-scores around 88.24% and Accuracy of 81.82%. XGBoost showed strong performance, with an Accuracy of 90.91% and perfect Recall (100%), making it a highly reliable model for leak detection in this context, only slightly trailing LR. The exceptional performance of Logistic Regression, achieving 100% accuracy, precision, recall, F1-score, and specificity, is a significant finding of this study. This suggests that the relationship between the pressure transducer readings and the presence or absence of a leak is highly linear and separable in our controlled experimental environment. The consistent pressure drop observed when the leak was active, combined with the low ambient noise environment ensured by our experimental setup and the robustness of pressure transducers, likely enabled LR to establish a clear decision boundary. This indicates that for situations where leak-induced pressure changes present a distinct and consistent pattern, simpler linear models like LR can be highly effective, potentially offering a computationally efficient solution for real-time monitoring. Conversely, the relatively lower performance of models like NB and SVM (in terms of specificity for NB and precision for SVM) indicates their sensitivity to the dataset characteristics. NB's assumption of feature independence might not perfectly align with the subtle correlations present in pressure variations, while SVM's performance can be sensitive to the choice of kernel and regularization parameters, which might not have been optimally tuned for our specific data distribution, leading to some misclassifications despite good recall.

Table 2. Pressure sensor distance dependence on leak and signal delay (ΔT)

Distance (Meters)	P1= 1.5 P2= 9	P1= 1.5 P2=38	P1= 1.5 P2=66	P1= 1.5 P2=76
Delay between signals ΔT(Second)	0.1	0.56	0.8	0.9

Table 3. Performance metrics for various leak detection models

Model	**Metrics**				**Performance (%)**				
	TN	TP	FN	FP	Accuracy	Precision	Recall	F1-Score	Specificity
NB	2	13	4	3	68.18	76.47	81.25	78.79	33.33
SVM	1	16	5	0	77.27	76.19	100	86.49	16.67
RF	3	15	3	1	81.82	83.33	93.75	88.24	50
DT	3	15	3	1	81.82	83.33	93.75	88.24	50
KNN	3	15	3	1	81.82	83.33	93.75	88.24	50
XGBoost	4	16	2	0	90.91	88.89	100	94.12	66.67
LR	**6**	**16**	**0**	**0**	**100**	**100**	**100**	**100**	**100**

5.3. The Top-Performing Models

Fig. 20 presents depicts a comprehensive overview of the performance of the five best-performing ML models RF, DT, KNN, XGBoost, and LR, evaluated using several important performance metrics, including Accuracy, Precision, Recall, F1-Score, and Specificity. These metrics are critical for assessing the effectiveness of each model in terms of addressing the leak detection problem. In this study, LR demonstrated optimal performance with perfect metrics across all evaluation criteria, achieving perfect results across all evaluation criteria. The model consistently provided accurate classifications, with no false positives (non-leaks predicted as leaks) or false negatives (leaks missed). This makes LR the most reliable choice for this application, particularly in scenarios requiring both high sensitivity and specificity. XGBoost exhibits robust performance, particularly excelling in Recall and F1-Score, making it an excellent choice for imbalanced datasets where maximizing Recall is essential. The RF, DT, and KNN models generally demonstrate good performance; however, they require enhancement in specificity to reduce the occurrence of false positives. In summary, LR emerged as the optimal model for this task due to its perfect evaluation metrics, and XGBoost was found to be a strong alternative, particularly for imbalanced datasets. The RF, DT, and KNN models, although effective in many respects, require further refinement to reduce the number of false positives and achieve higher specificity for better practical application.

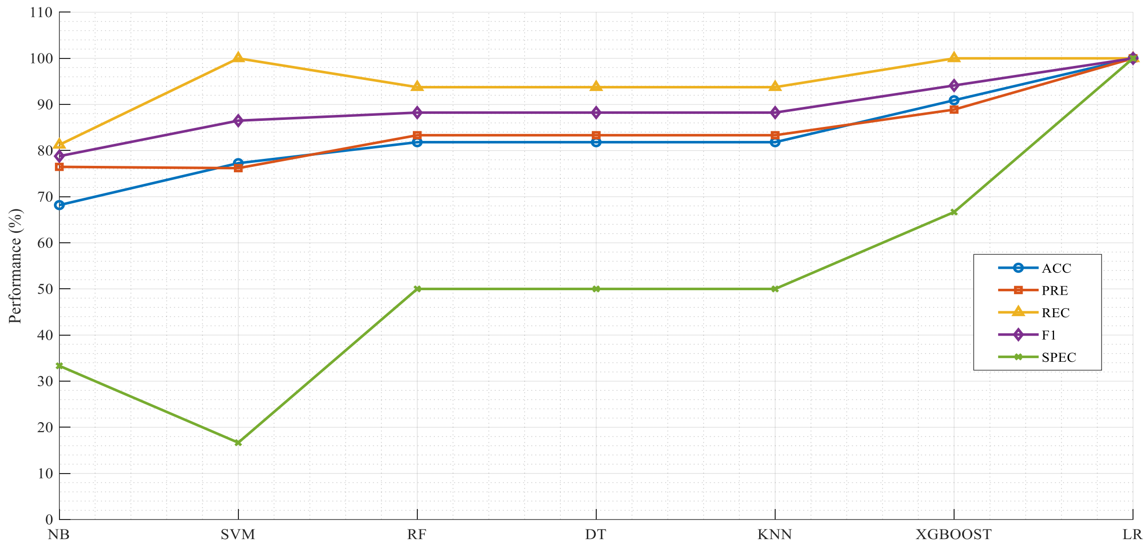

Fig. 20. Comparison of the performances of the top five models' performances

Table 4 presents the confusion matrix for the logistic regression model applied to the leak detection dataset. The confusion matrix is a powerful tool for evaluating and visualizing the performance of classification algorithms. This section presents a detailed comparison of the true labels (actual classes) and the predicted labels (classes predicted by the model). This comparison is critical for understanding the performance of the model across various categories. In this case, the LR model achieves the following perfect classification:

- True Negatives (No_Leak correctly identified): 6
- True Positives (Leak correctly detected): 16
- False Positives (No_Leak wrongly classified as Leak): 0
- False Negatives (Leak missed by the model): 0

These results indicate that the model achieved an ideal balance between sensitivity (ability to detect leaks) and specificity (avoiding false alarms). The absence of false positives and false negatives highlights the model's capability to classify data with 100% accuracy, which makes it highly reliable for real-world leak detection scenarios. By visualizing this relationship in the confusion matrix, we can easily interpret and validate the performance of the classification algorithm, providing clear insights into its strengths and weaknesses. For the LR model in this study, the matrix demonstrated flawless performance, thereby strengthening its suitability for deployment in applications where precise and dependable leak detection is crucial.

5.4. Comparison with Other Studies

Table 5 summarizes the results of other studies and their performance in various leak detection analyses in WDNs. This comparison highlights the reliability and effectiveness of different ML methods under different conditions and pipe materials. Each study addresses specific scenarios, demonstrating the adaptability of ML techniques to achieve high leak detection accuracy.

Our study's finding of 100% accuracy with Logistic Regression for PEHD pipes aligns with similar high-accuracy results reported in the literature, particularly for controlled environments or simulated data. For instance, studies using data generated by EPANET [20] also reported 100% accuracy with LR and RF, suggesting that highly accurate classification is achievable when data quality is high and noise is minimal. However, it is crucial to highlight the methodological differences that impact direct comparability. While [23] and [15] report strong accuracies (84.79% and up to 100% for metal pipes, 94.93% for non-metal pipes) using RF, these studies often involve different pipe materials, network complexities, sensor types, and noise environments. Our experimental setup, using pressure transducers, aims to mitigate ambient noise, which contributes to the clarity of the pressure signals. Studies that rely on acoustic or vibration sensors, for example, may encounter higher noise levels from the environment, leading to slightly lower overall accuracies even for robust models. Specifically, our experimental setup involves a controlled prototype pipeline with a known leak location and size, allowing for precise data acquisition under defined conditions. This differs from studies using large-scale real-world networks or purely simulated data. The high accuracy achieved by LR in our study underscores its potential for highly distinct leak signatures, which may not always be present in more complex, noisy, and uncontrolled real-world scenarios. The choice of pressure transducers, known for their robustness against external acoustic noise, played a pivotal role in maintaining the integrity of our sensor readings, thereby facilitating the observed high performance of the ML models.

Fig. 21 provides a visual comparison of the leak detection accuracy achieved in our study versus the findings from other studies. By presenting these results side by side, the figure highlights the advancements made in this work and underscores the high reliability of the proposed method, particularly for PEHD pipes. The comparison demonstrates the superiority of the proposed method in terms of leveraging ML techniques to achieve flawless accuracy in leak detection. These findings demonstrate that the selection of ML algorithms must consider the specific characteristics of the dataset, type of pipe material, and operational conditions of the water distribution network. The proposed methods also reinforce the potential for further improvement in leak detection methods by integrating advanced algorithms and optimizing feature selection to cater to diverse practical scenarios.

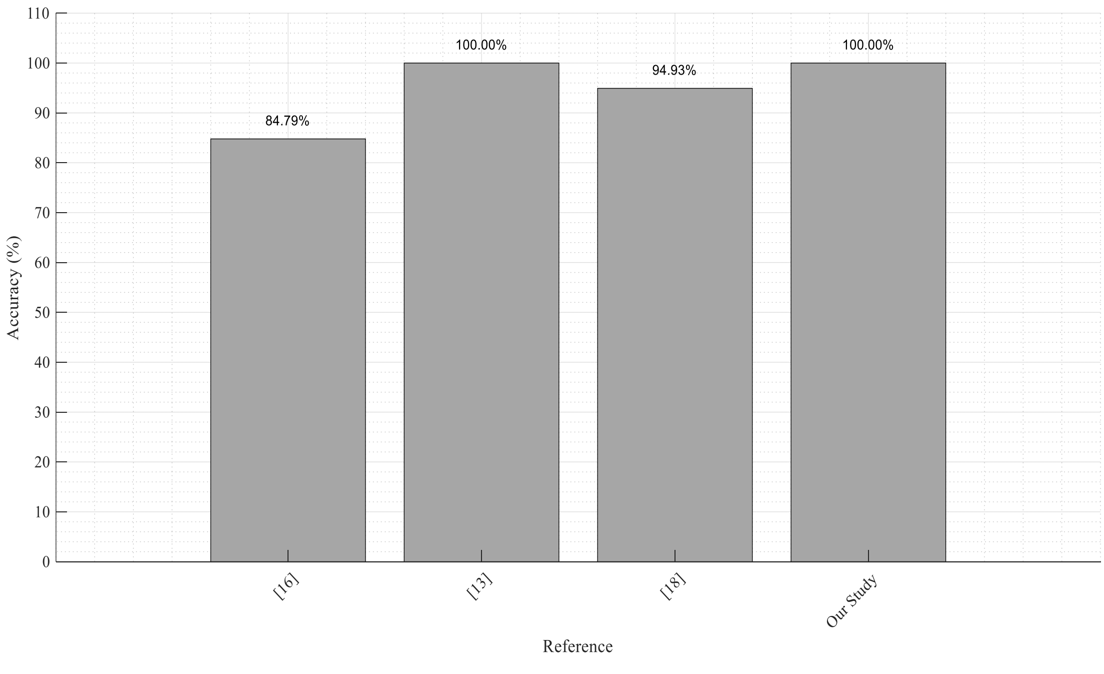

Fig. 21. Comparison of leak detection accuracy with findings from other studies

Table 4. Confusion matrix for LR leak detection model

Total N. of samples		Predicted Labels	
		No_leak	Leak
True Labels	No_Leak	6	0
	Leak	0	16

Table 5. Comparison accuracy in leak detection with other studies

Best Methods	Performance	References
RF	For non-metal pipes (Accuracy = 84.79 %)	[23]
LR RF	For data generated by EPANET (Accuracy = 100%)	[20]
RF	For a metal pipe (Accuracy=100%) For non-metal pipes (Accuracy=94.93%)	[15]
LR	**For PEHD pipes (Accuracy = 100%)**	**Our Study**

6. Conclusion

This study underscores the profound impact of advanced ML techniques on improving leakage detection in urban infrastructure. We successfully developed an ML-based leak detection model utilizing pressure signals meticulously collected from a simulated WDN via two strategically calibrated pressure transducers. These transducers proved highly effective in mitigating the influence of ambient noise and road traffic, ensuring high-quality signal acquisition critical for robust model training. The dataset was carefully curated and processed to accurately represent both leakage and non-leakage signals, ensuring data integrity. Several ML algorithms were employed to classify the data into two distinct categories: leak and no-leak. Among the diverse models tested, LR emerged as the best-performing algorithm, achieving a remarkable 100% accuracy, precision, recall, F1-score, and specificity in our experimental setup. This result underscores the significant potential of ML techniques in improving the accuracy and reliability of leak detection systems in water distribution networks.

This research contributes new knowledge to the domain by demonstrating a practical and reliable ML-based leak detection framework that integrates noise-resistant pressure transducers with a Wi-Fi-enabled IoT system, validated through a custom-built experimental setup. Our approach provides a foundation for developing scalable and efficient solutions for infrastructure monitoring, especially in environments where ambient noise is a challenge. However, it is crucial to acknowledge the limitations of the present study to ensure a balanced perspective on the applicability of our findings. This research was conducted on a relatively small-scale prototype pipeline (100 m length, 40 mm diameter) with a single, fixed leak location and size, under controlled laboratory conditions. Consequently, the observed 100% accuracy for Logistic Regression, while robust within this specific context, may not directly generalize to the inherent complexities of large-scale, real-world WDNs, which involve diverse pipe materials, multiple and dynamic leak scenarios, varying environmental conditions, and intricate network topologies. The dataset, while experimentally derived, is also relatively small, and future work will require extensive data collection to ensure broader applicability.

Future work will focus on expanding the dataset to incorporate a wider variety of scenarios, including different pipe materials, multiple leak types and sizes, and diverse environmental conditions to assess model generalizability. We also aim to explore innovative methodologies, such as DL models, to handle more complex and multi-leak datasets and to investigate strategies for improving model interpretability and computational efficiency for real-time deployment. Furthermore, we intend to pursue field validation with utility partners to test the system in operational urban water distribution networks, moving beyond the current laboratory-scale prototype.

Author Contribution: All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding: This research received no external funding.

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

Abbreviations and Acronyms

This section provides a list of all abbreviations and acronyms used throughout the manuscript to enhance clarity and ensure consistent understanding for readers. Each term is defined upon first use in the text and listed in Table 6 for quick reference.

Table 6. Abbreviations and acronyms used in this study

Abbreviation	Explanation
ML	Machine Learning
IoT	Internet of Things
WDNs	Water distribution network systems
PEHD	Polyethylene High-Density
RFID	Radio Frequency Identification
NFC	Near-field communication
SVM	Support Vector Machines
DT	Decision Trees
RF	Random Forests
NB	Naïve Bayes
LR	Logistic Regression
KNN	K-Nearest Neighbor
XGBoost	Extreme Gradient Boosting
PSI	Pounds per Square Inch
MEMS	Micro-Electro-Mechanical Systems
AdaBoost	Adaptive Boosting
LSTM-GAN	Long Short-Term Memory Generative Adversarial Network
t-SNE	through-Stochastic Neighbor Embedding
AE	Acoustic Emission
RVM	Relevance Vector Machine
PCA	Principal Component Analysis
ANN	Artificial Neural Networks
RNN-LSTM	Recurrent Neural Networks for Long Short-Term Memory
DL	Deep Learning
FS	Full Scale
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
CSV	Comma-Separated Values

References

[1] P. Verma and R. P. Parouha, "Solving systems of nonlinear equations using an innovative hybrid algorithm," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 46, no. 4, pp. 1005-1027, 2022, https://doi.org/10.1007/s40998-022-00527-z.

[2] I. A. Tijani, S. Abdelmageed, A. Fares, K. H. Fan, Z. Y. Hu, and T. Zayed, “Improving the leak detection efficiency in water distribution networks using noise loggers,” *Science of The Total Environment*, vol. 821, p. 153530, 2022, https://doi.org/10.1016/j.scitotenv.2022.153530.

[3] T. Luna, J. Ribau, D. Figueiredo, and R. Alves, “Improving energy efficiency in water supply systems with pump scheduling optimization,” *Journal of Cleaner Production*, vol. 213, pp. 342-356, 2019, https://doi.org/10.1016/j.jclepro.2018.12.190.

[4] B. Duan, J. Gao, H. Cao, and S. Hu, “Energy-Efficient Management of Urban Water Distribution Networks Under Hydraulic Anomalies: A Review of Technologies and Challenges,” *Energies*, vol. 18, no. 11, p. 2877, 2025, https://doi.org/10.3390/en18112877.

[5] A. V. Serafeim, N. T. Fourniotis, R. Deidda, G. Kokosalakis, and A. Langousis, “Leakages in Water Distribution Networks: Estimation Methods, Influential Factors, and Mitigation Strategies—A Comprehensive Review,” *Water*, vol. 16, no. 11, p. 1534, 2024, https://doi.org/10.3390/w16111534.

[6] C. Chabira, H. Bakhti, M. Bentoumi and S. Meftah, "RNL used for the verification of the efficiency of a localization model in a real WDNs," *2022 International Conference of Advanced Technology in Electronic and Electrical Engineering (ICATEEE)*, pp. 1-4, 2022, https://doi.org/10.1109/ICATEEE57445.2022.10093092.

[7] E. Farah and I. Shahrour, “Water Leak Detection: A Comprehensive Review of Methods, Challenges, and Future Directions,” *Water*, vol. 16, no. 20, p. 2975, 2024, https://doi.org/10.3390/w16202975.

[8] A. Alshami, E. Ali, M. Elsayed, A. E. E. Eltoukhy and T. Zayed, "IoT Innovations in Sustainable Water and Wastewater Management and Water Quality Monitoring: A Comprehensive Review of Advancements, Implications, and Future Directions," *IEEE Access*, vol. 12, pp. 58427-58453, 2024, https://doi.org/10.1109/ACCESS.2024.3392573.

[9] S. Meftah, M. Bentoumi, D. H. Burhannuddin, H. Bakhti, and C. Chabira, “Novel Leak Detector Based on DWT an Experimental Study,” *International Journal of Robotics and Control Systems*, vol. 4, no. 3, pp. 1109-1134, 2024, https://doi.org/10.31763/ijrcs.v4i3.1458.

[10] S. Zhao, Q. Lu, T. Zhang, S. He, P. Shi and J. Li, "Leak Detection of Underground Water Pipelines Using Acoustic Feature Extraction," *IEEE Internet of Things Journal*, vol. 12, no. 16, pp. 33411-33420, 2025, https://doi.org/10.1109/JIOT.2025.3575465.

[11] T. Yu, X. Chen, W. Yan, Z. Xu, and M. Ye, “Leak detection in water distribution systems by classifying vibration signals,” *Mechanical Systems and Signal Processing*, vol. 185, p. 109810, 2023, https://doi.org/10.1016/j.ymssp.2022.109810.

[12] R. Vanijjirattikhan *et al*., “AI-based acoustic leak detection in water distribution systems,” *Results in Engineering*, vol. 15, p. 100557, 2022, https://doi.org/10.1016/j.rineng.2022.100557.

[13] S. Lee and B. Kim, “Machine Learning Model for Leak Detection Using Water Pipeline Vibration Sensor,” *Sensors*, vol. 23, no. 21, p. 8935, 2023, https://doi.org/10.3390/s23218935.

[14] L. Romero-Ben, D. Alves, J. Blesa, G. Cembrano, V. Puig, and E. Duviella, “Leak detection and localization in water distribution networks: Review and perspective,” *Annual Reviews in Control*, vol. 55, pp. 392-419, 2023, https://doi.org/10.1016/j.arcontrol.2023.03.012.

[15] N. M. Mahdi, A. H. Jassim, S. A. Abulqasim, A. Basem, A. A. F. Ogaili, and L. A. Al-Haddad, “Leak detection and localization in water distribution systems using advanced feature analysis and an Artificial Neural Network,” *Desalination and Water Treatment*, vol. 320, p. 100685, 2024, https://doi.org/10.1016/j.dwt.2024.100685.

[16] P. Pandey, N. T. Mücke, S. Jain, P. Ramachandran, S. M. Bohté, and C. W. Oosterlee, “Machine Learning-Based Digital Twin for Water Distribution Network Anomaly Detection and Localization,” *Engineering Proceedings*, vol. 69, no. 1, p. 201, 2024, https://doi.org/10.3390/engproc2024069201.

[17] D. P. Sousa, R. Du, J. Mairton Barros da Silva Jr, C. C. Cavalcante, and C. Fischione, “Leakage detection in water distribution networks using machine-learning strategies,” *Water Supply*, vol. 23, no. 3, pp. 1115-1126, 2023, https://doi.org/10.2166/ws.2023.054.

[18] R. J. L. Taloma, F. Cuomo, D. Comminiello, and P. Pisani, “Machine learning for smart water distribution systems: exploring applications, challenges and future perspectives,” *Artificial Intelligence Review*, vol. 58, no. 4, p. 120, 2025, https://doi.org/10.1007/s10462-024-11093-7.

[19] U. Rajasekaran and M. Kothandaraman, "Comparative analysis of machine learning and deep learning based water pipeline leak detection using EDFL sensor," *Journal of Pipeline Systems Engineering and Practice*, vol. 14, no. 4, p. 04023026, 2023, https://doi.org/10.1061/JPSEA2.PSENG-1439.

[20] Z. Rehman, N. Tariq, S. A. Moqurrab, J. Yoo, and G. Srivastava, "Machine learning and internet of things applications in enterprise architectures: Solutions, challenges, and open issues," *Expert Systems*, vol. 41, no. 1, p. e13467, 2024, https://doi.org/10.1111/exsy.13467.

[21] S. Zuluaga, S. Saxe, and B. Karney, “Roles of Value in the Evaluation and Modeling of Decision Strategies for Pipe Maintenance in Water Distribution Networks,” *Journal of Water Resources Planning and Management*, vol. 150, no. 4, p. 04024006, 2024, https://doi.org/10.1061/JWRMD5.WRENG-6171.

[22] M. R. Islam, S. Azam, B. Shanmugam and D. Mathur, "A Review on Current Technologies and Future Direction of Water Leakage Detection in Water Distribution Network," *IEEE Access*, vol. 10, pp. 107177-107201, 2022, https://doi.org/10.1109/ACCESS.2022.3212769.

[23] D. Alves, J. Blesa, E. Duviella, and L. Rajaoarisoa, “Leak Detection in Water Distribution Networks Based on Water Demand Analysis,” *IFAC-PapersOnLine*, vol. 55, no. 6, pp. 679-684, 2022, https://doi.org/10.1016/j.ifacol.2022.07.206.

[24] G. Rajan and S. Li, "A Systematic Literature Review on Flow Data-Based Techniques for Automated Leak Management in Water Distribution Systems," *Smart Cities*, vol. 8, no. 3, p. 78, 2025, https://doi.org/10.3390/smartcities8030078.

[25] Y. L. Tsai, H. C. Chang, S. N. Lin, A. H. Chiou, and T. L. Lee, "Using convolutional neural networks in the development of a water pipe leakage and location identification system," *Applied Sciences*, vol. 12, no. 16, p. 8034, 2022, https://doi.org/10.3390/app12168034.

[26] N. Hartem, “Etude des pertes d’eau dans les réseaux d’eau potable (cas réseau de la commune de Bendjerrah),” *Working Paper*, 2020, https://dspace.univ-guelma.dz/jspui/handle/123456789/10112.

[27] N. A. M. Yussof and H. W. Ho, “Review of Water Leak Detection Methods in Smart Building Applications,” *Buildings*, vol. 12, no. 10, p. 1535, 2022, https://doi.org/10.3390/buildings12101535.

[28] A. Negm, X. Ma, and G. Aggidis, “Review of leakage detection in water distribution networks,” *IOP Conference Series: Earth and Environmental Science*, vol. 1136, no. 1, p. 012052, 2023, https://doi.org/10.1088/1755-1315/1136/1/012052.

[29] P. Nagapurkar, N. Sharma, S. Garcia, and S. Nimbalkar, "Evaluating Acoustic vs. AI-Based Satellite Leak Detection in Aging US Water Infrastructure: A Cost and Energy Savings Analysis," *Smart Cities*, vol. 8, no. 4, p. 122, 2025, https://doi.org/10.3390/smartcities8040122.

[30] L. Bykerk and J. V. Miro, "Detection of Water Leaks in Suburban Distribution Mains with Lift and Shift Vibro-Acoustic Sensors," *Vibration*, vol. 5, no. 2, pp. 370-382, 2022, https://doi.org/10.3390/vibration5020021.

[31] M. Soori, B. Arezoo, and R. Dastres, “Internet of things for smart factories in industry 4.0, a review,” *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 192-204, 2023, https://doi.org/10.1016/j.iotcps.2023.04.006.

[32] M. R. Islam, S. Azam, B. Shanmugam and D. Mathur, "An Intelligent IoT and ML-Based Water Leakage Detection System," *IEEE Access*, vol. 11, pp. 123625-123649, 2023, https://doi.org/10.1109/ACCESS.2023.3329467.

[33] M. Boujelben, Z. Benmessaoud, M. Abid, and M. Elleuchi, “An efficient system for water leak detection and localization based on IoT and lightweight deep learning,” *Internet of Things*, vol. 24, p. 100995, 2023, https://doi.org/10.1016/j.iot.2023.100995.

[34] M. Rahmati, H. Yazdizadeh and A. Yazdizadeh, "Leakage detection in a gas pipeline using artificial neural networks based on wireless sensor network and Internet of Things," *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, pp. 659-664, 2017, https://doi.org/10.1109/INDIN.2017.8104850.

[35] X. -L. Li, P. Wan, H. -T. Zhang, M. Li and Y. -H. Jiang, "The Application Research of Internet of Things to Oil Pipeline Leak Detection," *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 211-214, 2018, https://doi.org/10.1109/ICCWAMTIP.2018.8632561.

[36] C. Liu, B. Chen, W. Shao, C. Zhang, K. K. L. Wong and Y. Zhang, "Unraveling Attacks to Machine-Learning-Based IoT Systems: A Survey and the Open Libraries Behind Them," *IEEE Internet of Things Journal*, vol. 11, no. 11, pp. 19232-19255, 2024, https://doi.org/10.1109/JIOT.2024.3377730.

[37] R. Jalinder Jadhav, P. Radhakrishnan, D. Arun Jadhav, B. Ashreetha, J. Divya and S. Mukherjee, "Internet of Things Enabled Gas Leakage Detection Over Industrial Areas using Powerful MQ Series Sensor and

Controller," *2024 International Conference on Inventive Computation Technologies (ICICT)*, pp. 1679-1686, 2024, https://doi.org/10.1109/ICICT60155.2024.10544961.

[38] A. A. Chaudhry, R. Mumtaz, S. M. Hassan Zaidi, M. A. Tahir and S. H. Muzammil School, "Internet of Things (IoT) and Machine Learning (ML) enabled Livestock Monitoring," *2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET)*, pp. 151-155, 2020, https://doi.org/10.1109/HONET50430.2020.9322666.

[39] K. Choi, J. Yi, C. Park and S. Yoon, "Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines," *IEEE Access*, vol. 9, pp. 120043-120065, 2021, https://doi.org/10.1109/ACCESS.2021.3107975.

[40] J. K. Samriya, C. Chakraborty, A. Sharma, M. Kumar and S. K. Ramakuri, "Adversarial ML-Based Secured Cloud Architecture for Consumer Internet of Things of Smart Healthcare," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2058-2065, 2024, https://doi.org/10.1109/TCE.2023.3341696.

[41] A. Mitra, B. Bera, A. K. Das, S. S. Jamal, and I. You, "Impact on blockchain-based AI/ML-enabled big data analytics for Cognitive Internet of Things environment," *Computer Communications*, vol. 197, pp. 173-185, 2023, https://doi.org/10.1016/j.comcom.2022.10.010.

[42] P. Shukla, "ML-IDS: A machine learning approach to detect wormhole attacks in Internet of Things," *2017 Intelligent Systems Conference (IntelliSys)*, pp. 234-240, 2017, https://doi.org/10.1109/IntelliSys.2017.8324298.

[43] M. S. Rahman, T. Ghosh, N. F. Aurna, M. S. Kaiser, M. Anannya, and A. S. M. S. Hosen, "Machine learning and internet of things in industry 4.0: A review," *Measurement: Sensors*, vol. 28, p. 100822, 2023, https://doi.org/10.1016/j.measen.2023.100822.

[44] R. O. Ogundokun, M. O. Arowolo, S. Misra, and J. B. Awotunde, "Machine Learning, IoT, and Blockchain Integration for Improving Process Management Application Security," *Blockchain Applications in the Smart Era*, pp. 237-252, 2022, https://doi.org/10.1007/978-3-030-89546-4_12.

[45] M. A. M. Ali, A. Basahr, M. R. Rabbani and Y. Abdulla, "Transforming Business Decision Making with Internet of Things (IoT) and Machine Learning (ML)," *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pp. 674-679, 2020, https://doi.org/10.1109/DASA51403.2020.9317174.

[46] S. Tariq, B. Bakhtawar, and T. Zayed, "Data-driven application of MEMS-based accelerometers for leak detection in water distribution networks," *Science of The Total Environment*, vol. 809, p. 151110, 2022, https://doi.org/10.1016/j.scitotenv.2021.151110.

[47] R. Liu, T. Zayed, and R. Xiao, "Advanced acoustic leak detection in water distribution networks using integrated generative model," *Water Research*, vol. 254, p. 121434, 2024, https://doi.org/10.1016/j.watres.2024.121434.

[48] N. K. Banjara, S. Sasmal, and 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.

[49] P. Xue, Y. Jiang, Z. Zhou, X. Chen, X. Fang, and J. Liu, "Machine learning-based leakage fault detection for district heating networks," *Energy and Buildings*, vol. 223, p. 110161, 2020, https://doi.org/10.1016/j.enbuild.2020.110161.

[50] R. P. da Cruz, F. V. da Silva, and A. M. F. Fileti, "Machine learning and acoustic method applied to leak detection and location in low-pressure gas pipelines," *Clean Technologies and Environmental Policy*, vol. 22, no. 3, pp. 627-638, 2020, https://doi.org/10.1007/s10098-019-01805-x.

[51] N. Mashhadi, I. Shahrour, N. Attoue, J. El Khattabi, and A. Aljer, "Use of Machine Learning for Leak Detection and Localization in Water Distribution Systems," *Smart Cities*, vol. 4, no. 4, pp. 1293-1315, 2021, https://doi.org/10.3390/smartcities4040069.

[52] C.-W. Lee and D.-G. Yoo, "Development of Leakage Detection Model and Its Application for Water Distribution Networks Using RNN-LSTM," *Sustainability*, vol. 13, no. 16, p. 9262, 2021, https://doi.org/10.3390/su13169262.

[53] N. Ullah, Z. Ahmed, and J.-M. Kim, "Pipeline Leakage Detection Using Acoustic Emission and Machine Learning Algorithms," *Sensors*, vol. 23, no. 6, p. 3226, 2023, https://doi.org/10.3390/s23063226.

[54] J. A. Coelho, A. Glória, and P. Sebastião, "Precise water leak detection using machine learning and real-time sensor data," *IoT*, vol. 1, no. 2, pp. 474-493, 2020, https://doi.org/10.3390/iot1020026.

[55] R. Xiao, Q. Hu, and J. Li, "Leak detection of gas pipelines using acoustic signals based on wavelet transform and Support Vector Machine," *Measurement*, vol. 146, pp. 479-489, 2019, https://doi.org/10.1016/j.measurement.2019.06.050.

[56] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decision Analytics Journal*, vol. 3, p. 100071, 2022, https://doi.org/10.1016/j.dajour.2022.100071.

[57] Y. Ao, H. Li, L. Zhu, S. Ali, and Z. Yang, "The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling," *Journal of Petroleum Science and Engineering*, vol. 174, pp. 776-789, 2019, https://doi.org/10.1016/j.petrol.2018.11.067.

[58] E. M. K. Reddy, A. Gurrala, V. B. Hasitha, and K. V. R. Kumar, "Introduction to Naive Bayes and a review on its subtypes with applications," *Bayesian reasoning and gaussian processes for machine learning applications*, pp. 1–14, 2022, https://www.taylorfrancis.com/chapters/edit/10.1201/9781003164265-1.

[59] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *Journal of Clinical Epidemiology*, vol. 122, pp. 56-69, 2020, https://doi.org/10.1016/j.jclinepi.2020.03.002.

[60] M. Suyal and P. Goyal, "A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning," *International Journal of Engineering Trends and Technology*, vol. 70, no. 70, pp. 43-48, 2022, https://doi.org/10.14445/22315381/IJETT-V70I7P205.

[61] G. H. Coffie and S. K. F. Cudjoe, "Using extreme gradient boosting (XGBoost) machine learning to predict construction cost overruns," *International Journal of Construction Management*, vol. 24, no. 16, pp. 1742–1750, 2024, https://doi.org/10.1080/15623599.2023.2289754.

[62] A.-M. Yussif, H. Sadeghi, and T. Zayed, "Application of Machine Learning for Leak Localization in Water Supply Networks," *Buildings*, vol. 13, no. 4, p. 849, 2023, https://doi.org/10.3390/buildings13040849.

[63] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168-192, 2021, https://doi.org/10.1016/j.aci.2018.08.003.

[64] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," *Scientific reports*, vol. 12, no. 1, p. 5979, 2022, https://doi.org/10.1038/s41598-022-09954-8.

[65] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, p. 6, 2020, https://doi.org/10.1186/s12864-019-6413-7.