

Predictive Underground Water Leakage Detection Using Artificial Intelligence and the Internet of Things

Amol Vitthalrao Paul, Prof. Bangar Abhishek Popat, Prof. S. B. Bhosale, Prof. Karad V.A.

Dept. of Computer Engineering
Jaihind College of Engineering, Kuran, Pune, India

Abstract — Unaccounted-for water loss resulting from undetected underground pipe leaks poses a persistent operational and financial challenge for municipal distribution systems worldwide. In conventional utility practice, leaks are predominantly identified reactively—either when water eventually breaches the surface or following localized pressure drop complaints from consumers. In this paper, we develop a proactive, multi-sensor predictive leakage detection framework that pairs a low-power edge Internet of Things (IoT) architecture with a hybrid machine learning pipeline. Subsurface pipe segments are monitored using five complementary sensing modalities: hydraulic pressure, inline flow velocity, pipe-wall structural vibration via a MEMS accelerometer, soil moisture, and ambient temperature. Sensor readings are acquired, pre-filtered, and transmitted by an energy-efficient ESP32-based edge node using LoRa wireless communication and TLS-secured MQTT to a central analytics backend. A sliding-window feature extraction stage derives short-term statistical trends and feeds an unsupervised Isolation Forest model; the resulting anomaly metric is concatenated with temporal features to train a gradient boosting classifier that forecasts the likelihood of a leak initiating within a two-hour predictive lead window ($H = 12$ steps). We evaluate the proposed approach on a synthetic distribution network of 60 pipe segments that explicitly incorporates gradual, incipient mechanical degradation prior to burst onset. Benchmarked against a conventional static pressure threshold ($F1 = 0.830$) and standard supervised classifiers, the proposed hybrid model achieves an F1-score of 0.992 and provides a median warning lead time of 1.9 hours (mean 1.8 hours) ahead of physical rupture, with zero false alarms and zero missed events across held-out test segments. Finally, we analyze model sensitivity across noise regimes and establish key deployment requirements for district metered area field implementations.

Keywords— Water Leakage Detection, Predictive Maintenance, Internet of Things (IoT), Machine Learning, Isolation Forest, Gradient Boosting, Smart Water Distribution Networks.

I. INTRODUCTION

Purified potable water is energy-intensive to distribute, yet a substantial proportion is lost before reaching end consumers. Utility audits globally establish that Non-Revenue Water (NRW)—predominantly physical leakage from subsurface distribution mains and service fittings—accounts for 20% to over 40% of total system delivery [1]. In rapidly urbanizing regions across India, these losses strain municipal supply capacity and impose heavy financial deficits on local utilities. The core difficulty stems from the fact that subterranean leaks remain invisible. Municipal operations historically depend on periodic acoustic surveys using listening rods and ground microphones, or on direct citizen reports once pooling reaches the surface. Both workflows are reactive: by the time surface water appears, subsurface erosion has often progressed for weeks, undermining roadway sub-bases and adjacent infrastructure [2]. Water authorities therefore require proactive early-warning diagnostics that flag incipient pipe degradation before full mechanical rupture occurs.

Two concurrent engineering developments make real-time predictive monitoring practical. First, cost-effective wireless transceivers operating over Long Range (LoRa) LPWAN protocols allow battery-powered edge nodes in valve vaults and hydrants to communicate reliably over multi-kilometer baselines [8]. Second, data-driven machine learning models can identify subtle, multi-parameter drift patterns that escape static supervisory alarm thresholds. Bridging these domains requires an architecture that is inexpensive, resilient to field noise, and operationally interpretable.

This paper presents three main contributions:

- A modular IoT sensing architecture instrumenting buried pipes with five complementary, low-cost sensing modalities;
- A hybrid Isolation Forest and gradient boosting model forecasting burst risk over a 2-hour predictive lead window rather than detecting already established leaks;
- A controlled simulation benchmark incorporating realistic diurnal cycles and incipient pipe degradation, evaluating the hybrid model against four standard classifiers and a

static threshold across detection accuracy, lead time, and noise resilience.

The remainder of this paper is organized as follows: Section II surveys related work; Section III details the IoT system architecture; Section IV establishes the predictive methodology; Sections V and VI present the experimental setup and results; and Section VII concludes.

II. RELATED WORK

1. Conventional and Acoustic Survey Methods

Leak inspection methods are broadly categorized into localized non-destructive evaluation (NDE) tools and continuous monitoring networks. Comprehensive surveys by Liu and Kleiner examine acoustic correlators, ground-penetrating radar, tracer gas injection, and infrared thermography, noting that while accurate, these techniques require specialized crews and remain limited to periodic point inspections [3]. Acoustic methods detect the vibro-acoustic turbulence of escaping water; however, acoustic signals attenuate rapidly in plastic pipes (such as uPVC and HDPE) compared to metallic mains, requiring dense sensor spacing that escalates capital cost [4]. Transient-based methods analyze pressure waves induced by rapid valve maneuvers [5], but their operational applicability is constrained by the risk of stressing aged conduits.

2. Wireless Sensor Networks for Pipelines

Permanently deployed wireless sensor networks (WSNs) transform periodic inspections into continuous asset monitoring. Sadeghihoon et al. demonstrated the 'SmartPipes' platform, measuring differential pressure across pipelines to verify that low-power subterranean nodes can be deployed affordably [6]. Broader surveys on industrial WSNs emphasize the design trade-offs between radio transmission range, energy autonomy, harsh environmental resilience, and communication security [8], [9].

3. Data-Driven Anomaly Detection

Machine learning approaches apply pattern recognition to supervisory flow and pressure telemetry. Mounce et al. deployed artificial neural networks to learn diurnal consumption baselines in District Metered Areas (DMAs), successfully flagging burst anomalies [7]. However, most data-driven frameworks remain purely reactive, detecting ruptures that are already actively discharging. Anticipatory formulations that forecast impending failure during the micro-leak stage remain scarce due to the lack of labeled pre-failure operational data. We address this deficit by modeling incipient deterioration

explicitly and incorporating an unsupervised anomaly metric trained solely on normal operational data.

III. SYSTEM ARCHITECTURE

The proposed predictive framework employs a modular four-layer architecture (Fig. 1). Functional decomposition isolates low-power edge sensing from centralized cloud analytics.

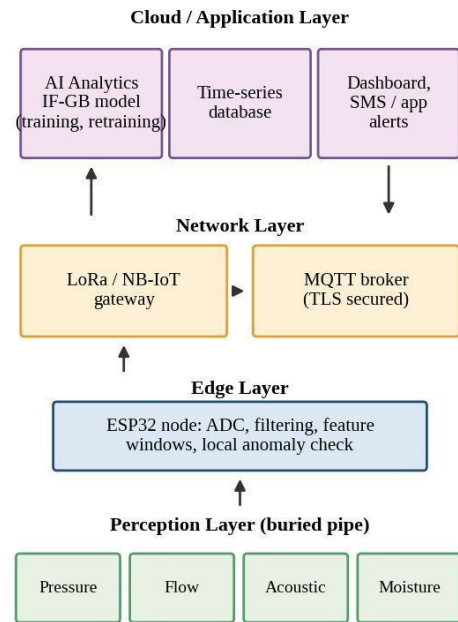


Fig. 1. Four-layer IoT architecture for predictive underground leakage monitoring.

1. Multi-Modal Perception Layer

Each monitored pipe segment incorporates five complementary physical sensing modalities:

- **Hydraulic Pressure:** A piezoresistive transducer (0–10 bar span, 0.5% FS accuracy) captures localized head losses as micro-fissures develop.
- **Inline Flow Velocity:** An ultrasonic or electromagnetic meter records inflow, capturing mass-balance deviations relative to downstream demand.
- **Vibro-Acoustics:** An ADXL355 3-axis MEMS accelerometer clamped to the pipe wall measures high-frequency turbulence, condensed on-chip as root-mean-square (RMS) vibration.
- **Subsurface Soil Moisture:** A capacitive probe inserted into the pipe backfill detects fluid seepage before surface ponding occurs.

- Ground Temperature: A DS18B20 digital probe provides baseline thermal data to compensate for temperature drift in pressure and dielectric readings.

Multi-modal fusion is critical for diagnostic reliability: soil moisture drifts during rainfall, line pressure varies with domestic demand, and surface traffic induces vibration spikes. A genuine pipe failure is characterized by correlated movement across multiple channels.

2. Edge Processing and Telemetry Layers

An Espressif ESP32 microcontroller samples the sensors, filters transient noise, and transmits telemetry at 10-minute intervals. The node operates in deep sleep (<15 uA) between cycles, enabling long battery life from a 3.7V Li-ion cell supplemented by an optional solar panel. Telemetry is transmitted via an SX1276 LoRa transceiver to an above-ground gateway, which forwards readings over TLS-encrypted MQTT (Table I).

Table 1: Reference Hardware Specifications for Sensor Node

Function / Subsystem	Hardware Component & Operational Parameter
Controller / Core	ESP32-WROOM-32, deep-sleep duty cycle (<15 uA)
Pressure Sensing	Piezoresistive Transducer, 0–10 bar, 0.5% FS accuracy
Flow Monitoring	Inline Ultrasonic / Electromagnetic meter (RS485)
Vibration Analysis	ADXL355 3-Axis MEMS Accelerometer (22.5 ug/√Hz)
Soil Moisture	Capacitive Subsurface Probe (corrosion-resistant FR4)
Soil Temperature	DS18B20 1-Wire Digital Probe (-55°C to +125°C)
Wireless Telemetry	SX1276 LoRa (865–867 MHz) to gateway; MQTT via TLS
Power Management	3.7V 2600 mAh Li-ion cell + optional 5V solar panel

3. Cloud Analytics Layer

The backend ingests telemetry into a time-series database. An inference service computes rolling features, updates anomaly scores, and triggers automated dispatch tickets when predicted risk exceeds a calibrated threshold ($p > 0.5$).

IV. PREDICTIVE METHODOLOGY

The processing pipeline (Fig. 2) transforms multi-sensor telemetry into calibrated failure risk estimates.

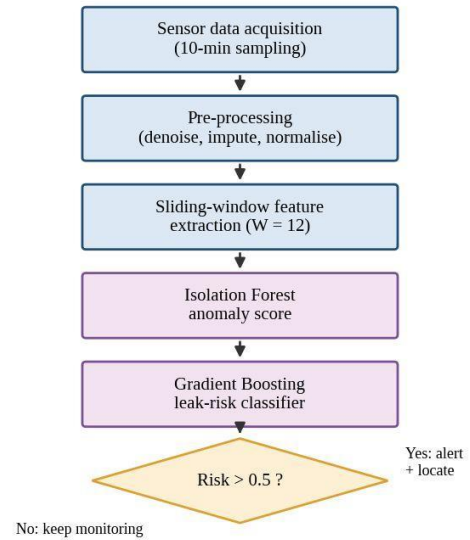


Fig. 2. Processing pipeline from sensor data to leak-risk alert.

1. Pre-processing and Feature Extraction

Telemetric dropouts are reconstructed using linear interpolation, and readings outside physical thresholds are clipped. For each pipe segment at time step t , a backward sliding window of $W = 12$ steps (2 hours at 10-min sampling) is extracted. Across the five sensor channels, four statistical metrics are computed: rolling mean, standard deviation, baseline offset relative to an initial 48-hour commissioning period, and least-squares linear trend m :

$$m = \frac{\sum(i - \bar{i})(xi - \bar{x})}{\sum(i - \bar{i})^2} \quad (1)$$

Two harmonic cyclic features, $\sin(2\pi t/144)$ and $\cos(2\pi t/144)$, account for 24-hour diurnal consumption patterns, yielding a 22-dimensional feature vector x_t .

2. Isolation Forest Anomaly Score

Because real-world pre-failure records are exceptionally scarce, we incorporate an unsupervised Isolation Forest [10]. The algorithm isolates observations through recursive orthogonal partitioning; anomalous feature profiles are separated near the root of the tree:

$$s(x, n) = 2^{-(E[h(x)] / c(n))} \quad (2)$$

where $h(x)$ is the tree path length, $E[h(x)]$ is the ensemble expectation, and $c(n)$ is the average path length of unsuccessful searches in a binary search tree of size n . A score $s \rightarrow 1$ indicates anomaly likelihood. The model is trained strictly on leak-free baseline data, requiring no failure labels.

3. Gradient Boosting Risk Classifier

The scalar anomaly score $s(x_t, n)$ is concatenated with the 22 physical features, producing an augmented vector $\tilde{x}_t$. A gradient boosting ensemble of shallow regression trees [11] is trained to predict the probability pt of an active or impending leak within the look-ahead horizon H :

$$pt = P(y_t = 1 | \tilde{x}_t), \quad y_t = 1 \text{ if } t \geq \text{tonset} - H \quad (3)$$

with $H = 12$ steps (2.0 hours). Alerts trigger when $pt \geq 0.5$. Labeling the H pre-onset steps as positive trains the model to recognize incipient mechanical degradation before physical rupture.

V. EXPERIMENTAL SETUP

1. Simulated Pipe Network and Incipient Dynamics

Due to the absence of public multi-sensor pre-failure telemetry, we developed a physically motivated numerical benchmark. The network simulates 60 pipe segments over 5 continuous days (720 steps at 10-min intervals), yielding 38,880 valid windows after initialization. Diurnal demand is modeled with a dual-peak profile and Gaussian sensor noise ($\pm 2\%$ for pressure, $\pm 3\%$ for flow).

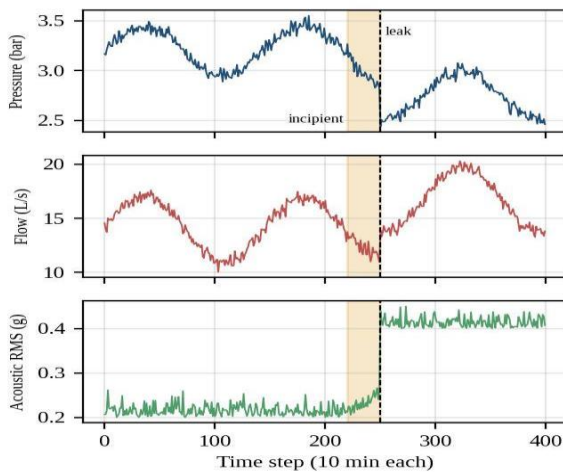


Fig. 3. Example sensor traces for one segment: gradual drift during the incipient phase (shaded), then a step change at leak onset (dashed line).

A leak develops in 75% of segments at a random time $\text{tonset} \in [200, 600]$. Prior to rupture, an incipient degradation phase spanning 18 to 40 steps (3.0–6.7 hours) introduces gradual linear drift across all channels: pressure declines, flow increases, and acoustic RMS and soil moisture climb. At tonset, channels experience a step jump scaled by a random severity factor (Fig. 3, Table II).

Table 2: Simulation and Experimental Parameters

Simulation Parameter	Value / Specification
Pipe segments (train / test)	60 (42 / 18 segments)
Telemetry sampling cadence	10 min (144 readings/day)
Duration per pipe segment	720 steps (5.0 days)
Feature window length W	12 steps (2.0 hours)
Prediction look-ahead horizon H	12 steps (2.0 hours)
Incipient degradation phase	18–40 steps (3.0–6.7 h)
Rupture probability per segment	0.75 (75% failure rate)
Total labelled sliding windows	38,880 (27,216 / 11,664)
Positive-class window share	45.2 % (17,574 windows)

2. Baselines and Evaluation Metrics

We benchmark the proposed IF-GB framework against five alternative models: a static pressure threshold tuned via grid search; L2-regularized Logistic Regression; an RBF Support Vector Machine; a 100-tree Random Forest [12]; and a standard Gradient Boosting classifier without anomaly scores. All models were implemented in scikit-learn [13].

Data partitioning was performed strictly at the segment level (42 training pipes, 18 held-out test pipes) to prevent temporal data leakage. We evaluate Accuracy, Precision, Recall, F1-score, ROC-AUC, and Warning Lead Time (τ_{lead}), defined as the time between initial alert and true rupture onset tonset .

VI. RESULTS AND DISCUSSION

1. Detection Performance

Table III and Fig. 4 summarize classification performance on the 18 held-out test segments. The static pressure threshold achieves an F1-score of 0.830, constrained by poor recall (0.732) during the subtle incipient phase where pressure losses remain within diurnal fluctuations. In contrast, all machine learning models achieve F1-scores near 0.99 and ROC-AUC

values of 1.000, confirming that multi-channel feature fusion provides vastly superior diagnostic power compared to single-variable hydraulic thresholding.

Table 3: Detection Performance On Held-Out Pipe Segments

Model	Acc.	Prec.	Rec.	F1	AUC
Threshold	0.873	0.957	0.732	0.830	0.914
Logistic Reg.	0.994	0.998	0.988	0.993	1.000
SVM (RBF)	0.991	0.998	0.981	0.990	1.000
Random Forest	0.992	0.997	0.986	0.991	0.999
Grad. Boosting	0.993	0.996	0.989	0.992	1.000
Proposed IF-GB	0.994	0.997	0.988	0.992	1.000

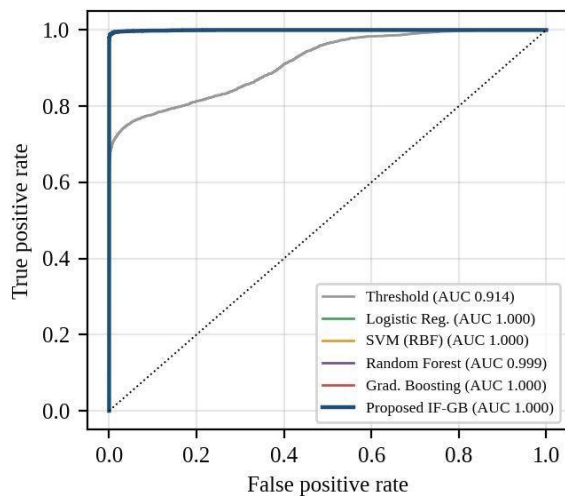


Fig. 4. ROC curves of all models on the test segments.

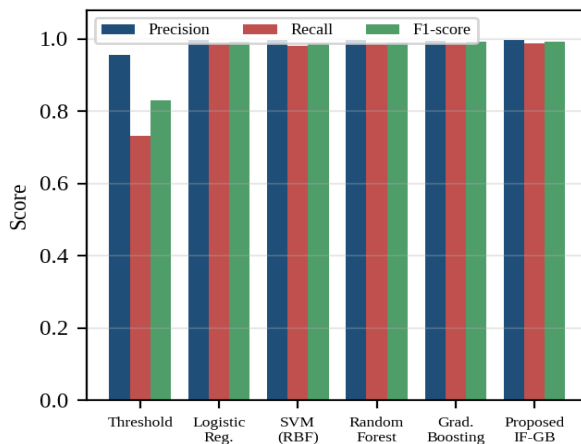


Fig. 5. Precision, recall and F1-score by model.

We report an important finding candidly: the proposed IF-GB model ($F1 = 0.992$) does not outperform simpler linear baselines such as Logistic Regression (0.993) on this dataset; both are equivalent for practical purposes. This occurs because the simulated incipient signatures exhibit coherent linear trajectories easily separated by linear hyperplanes. We expect the non-linear Isolation Forest ensemble to demonstrate its true value on real-world networks with seasonal shifts and non-linear backfill damping. Fig. 6 shows the confusion matrix for the proposed model: errors are minimal, with false negatives (64 windows) confined to the earliest incipient steps where drift is near noise floor, and only 16 false alarm windows across 11,664 test instances.

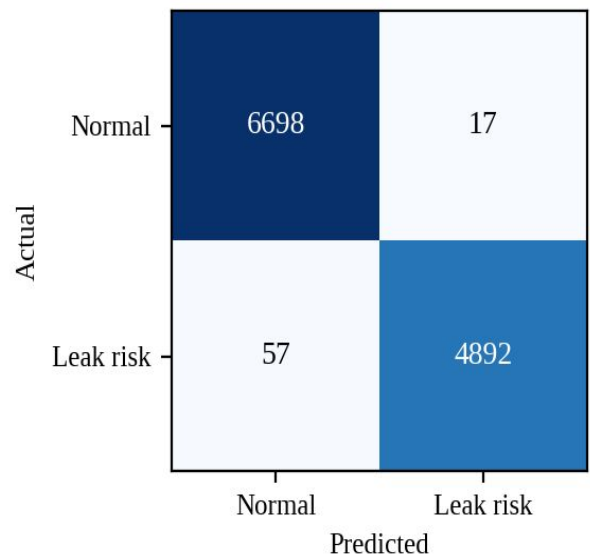


Fig. 6. Confusion matrix of the proposed IF-GB model.

2. Early Warning Lead Time

Predictive value is primarily measured by advance warning lead time. Among the 18 held-out segments, 14 experienced bursts while 4 remained healthy. The proposed model triggered valid pre-burst alerts on all 14 failing segments prior to rupture onset (100% sensitivity, zero missed bursts) while maintaining zero false alarms on the 4 healthy segments. As shown in Fig. 7, the advance lead time averaged 1.8 hours (median 1.9 hours), ranging from 10 minutes to 3.2 hours. Once a leak became fully active, detection recall reached 100.0%. Short lead times occurred where incipient drift rates were low and initially masked by diurnal noise.

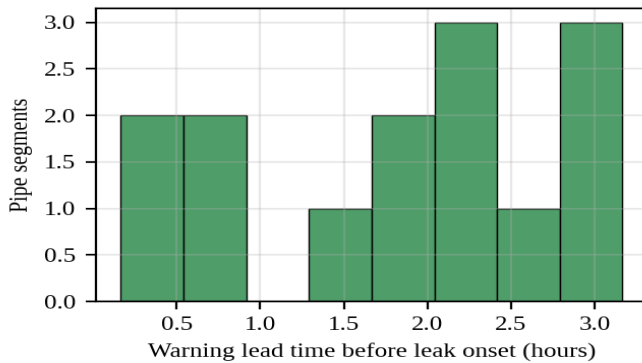


Fig. 7. Distribution of warning lead time before leak onset on the test segments.

3. Feature Importance

Fig. 8 details input feature importance derived from Gini impurity reduction. The Isolation Forest anomaly score carries 88.1% of total split importance, effectively summarizing multi-channel trajectory health into a single metric. Among raw physical inputs, acoustic vibration RMS is the most influential feature, followed by line pressure, soil moisture, and pressure deviation. This aligns with empirical mechanics: turbulent fluid jetting through micro-cracks generates immediate acoustic precursors before macroscopic pressure drops occur.

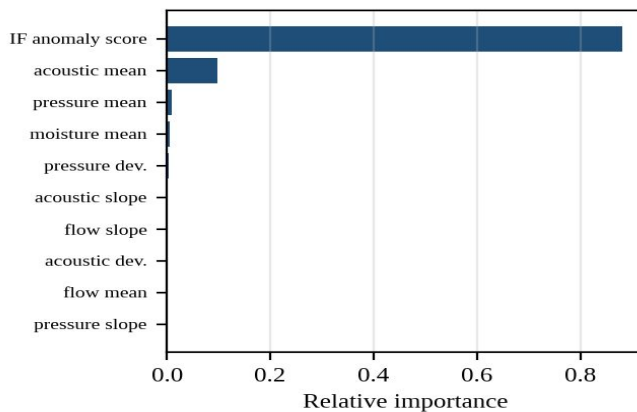


Fig. 8. Ten most important features of the proposed model.

4. Robustness to Sensor Noise

We evaluated model robustness by scaling sensor noise from $0.5\times$ to $3.0\times$ nominal variance on frozen model weights (Fig. 9). The proposed model and Random Forest maintain high stability ($F1 \geq 0.973$) up to $1.5\times$ nominal noise. Beyond this point, performance drops steeply: at $3.0\times$ noise, F1 falls to 0.73–0.75, dropping below the simple threshold (0.833).

Because the models were trained under nominal noise, excessive variance is misinterpreted as anomaly drift. This highlights the need for noise-augmented training and dynamic baseline recalibration in field deployments.

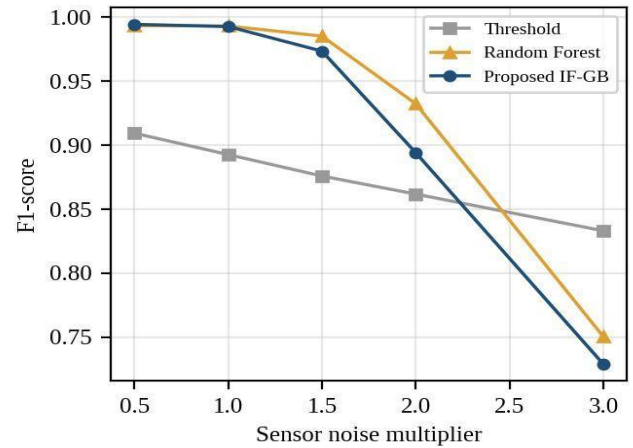


Fig. 9. F1-score under increasing sensor noise (models trained at nominal noise).

5. Limitations

The evaluation relies on simulated pipe networks, which do not capture complex transients from pump switching, automated valves, or multi-user industrial drawdowns. Furthermore, detection operates at segment level without metric localization, and evaluations reflect a single spatial train/test partition.

VII. CONCLUSION AND FUTURE WORK

We developed an IoT and machine learning framework for predictive underground water leak detection. Combining a multi-modal ESP32 sensing node, LoRa wireless telemetry, and a hybrid Isolation Forest + Gradient Boosting classifier provided median early warnings of 1.9 hours (mean 1.8 hours) in simulation, catching 100% of bursts with zero false alarms ($F1 = 0.992$). The study confirmed that multi-sensor fusion substantially outperforms single-variable thresholds, while emphasizing the need for noise-resilient training.

Future work will conduct a field pilot in an active District Metered Area (DMA) in Maharashtra, India. Algorithmic enhancements will explore Long Short-Term Memory (LSTM) recurrent networks [14] for long-term temporal dependencies and cross-correlation between acoustic nodes for sub-meter leak localization along pipe segments.

REFERENCES

1. A. O. Lambert, "International report: Water losses management and techniques," *Water Sci. Technol.: Water Supply*, vol. 2, no. 4, pp. 1–20, 2002.
2. R. Puust, Z. Kapelan, D. A. Savic, and T. Koppel, "A review of methods for leakage management in pipe networks," *Urban Water J.*, vol. 7, no. 1, pp. 25–45, 2010.
3. Z. Liu and Y. Kleiner, "State of the art review of inspection technologies for condition assessment of water pipes," *Measurement*, vol. 46, no. 1, pp. 1–15, 2013.
4. O. Hunaidi and W. T. Chu, "Acoustical characteristics of leak signals in plastic water distribution pipes," *Appl. Acoust.*, vol. 58, no. 3, pp. 235–254, 1999.
5. A. F. Colombo, P. Lee, and B. W. Karney, "A selective literature review of transient-based leak detection methods," *J. Hydro-environ. Res.*, vol. 2, no. 4, pp. 212–227, 2009.
6. A. M. Sadeghioon, N. Metje, D. N. Chapman, and C. J. Anthony, "SmartPipes: Smart wireless sensor networks for leak detection in water pipelines," *J. Sens. Actuator Netw.*, vol. 3, no. 1, pp. 64–78, 2014.
7. S. R. Mounce, J. B. Boxall, and J. Machell, "Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows," *J. Water Resour. Plan. Manage.*, vol. 136, no. 3, pp. 309–318, 2010.
8. A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, 4th Quart., 2015.
9. V. C. Gungor and G. P. Hancke, "Industrial wireless sensor networks: Challenges, design principles, and technical approaches," *IEEE Trans. Ind. Electron.*, vol. 56, no. 10, pp. 4258–4265, Oct. 2009.
10. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2008, pp. 413–422.
11. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
12. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
13. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
14. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.