Water Pipeline Leak Prediction Using Deep Learning

Samukelo Mkhize

School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal

samkelomanager@gmail.com

~~Durban, South Africa~~

~~October 2025~~

~~**Supervisor:** Prof. Okuthe Paul Kogeda~~

~~kogedap@ukzn.ac.za~~

Abstract

Pipeline leakages remain a persistent challenge for urban water management, driving non-revenue water, asset deterioration, and service disruptions. In the eThekwini municipality, real-time IoT sensing infrastructure is limited, restricting traditional hydraulic monitoring approaches. This study develops a data-driven, sensor-light predictive framework for pipeline leaks using historical maintenance and pipe replacement records provided by Umgeni Water, combined with pipeline attributes such as material, diameter, age, length, and system identifiers.

Leak events were identified via maintenance codes and text keyword rules, yielding 6,260 records with approximately 18% labeled as leaks. Features included static asset attributes, temporal indicators (year, seasonality), and historical metrics (time since last leak, count of past events, previous leak flag). Models were trained using consistent preprocessing (imputation, standardization, one-hot encoding), with class imbalance addressed via class weighting. Logistic Regression, Random Forest, XGBoost, a feed-forward Deep Neural Network (DNN), a two-layer LSTM sequence model, and a Transformer-style categorical encoder were evaluated. Thresholds were calibrated using validation $F_{0.5}$ to prioritize precision over recall.

Results show that the LSTM achieved the best discrimination (ROC-AUC $\approx$ 0.73, PR-AUC $\approx$ 0.41), highlighting the value of temporal event history. Logistic regression and the DNN achieved comparable performance (ROC-AUC $\approx$ 0.67–0.68), while Random Forest underperformed. The Transformer model failed to train effectively, reflecting data and scale limitations. Overall, the framework demonstrates that meaningful leak prediction is feasible using existing records, supporting proactive maintenance planning without requiring dense sensor networks.

Keywords: deep learning; gradient boosting; leakage risk; non-revenue water; imbalanced learning; SHAP

1 Introduction

Leak detection and failure prevention remain central to sustainable urban water management. Traditional approaches, hydraulic model residuals, step testing, and manual inspection, are effective but costly to operate at scale and often miss subtle, precursor signals in heterogeneous networks [1, 2, 3]. Modern IoT-based systems achieve fine-grained monitoring but require significant capital expenditure, secure telemetry, and ongoing calibration[4, 5, 7], which can be prohibitive in many African municipal settings today.

Problem. In the absence of widespread real-time sensing, can we *reliably* prioritize leak-prone assets using historical maintenance records and asset attributes alone?

Contributions. This paper presents:

- **C1:** A sensor-light, tabular-learning framework fro pipeline leak/failure risk scoring using historical replacements and asset metadata.
- **C2:** An end-to-end procedure: data cleaning, feature engineering, imbalance-aware training, thresholding, calibration, and cost-aligned evaluation.
- **C3:** Explainability (SHAP, permutation importance) and a risk-to-action mapping for inspection and rehabilitation planning.
- **C4:** A practical template tailored to the Umgeni Water context that utilities can adopt prior to (or alongside) IoT rollouts.

Paper outline. Section 2 reviews related work. Section 3 details the methodology. Results are in Section 4, discussion in Section 5, and conclusions in Section 6.

2 Related Work

2.1 Classical Leak Detection and Asset Deterioration

Hydraulic-model residual analysis and district metered area (DMA) methods can indicate anomaly regions but depend on accurate calibration and instrumentation [1, 2, 3]. Deterioration modelling using age/material/diameter and break history has a long tradition in water mains management [8, 9].

2.2 IoT-Enabled Sensing and ML

Wireless sensing with pressure/acoustic nodes and edge/cloud analytics improves detection latency and localisation [4, 5, 6], yet deployment cost, maintenance, and data governance remain barriers in many networks [7].

2.3 Historical-Record-Driven Prediction (This work)

Where rich telemetry is scarce, historical interventions (repairs/replacements) and asset registers can support supervised risk models [10]. Tabular learners—notably tree ensembles—often outperform deeper sequential models when features are well engineered and sample sizes are moderate, while also offering interpretability and robust baselines for operational use.

2.4 Positioning with Respect to Prior Studies

While many prior works emphasize static asset attributes (e.g., material, diameter, and age) as primary predictors of failures [8, 9], our feature importance analysis revealed that recent operational history (time since last event, cumulative failures) dominated in predictive power (Figure 1). This suggests that, in practice, the short-term dynamics of pipeline stress and prior incidents outweigh static design attributes for near-term leak risk. To our knowledge, this emphasis on operational history as a leading indicator has not been previously reported in the South African municipal context, making it a novel contribution of this study.

3 Methodology

3.1 Data Sources and Labeling

Two datasets were provided by Umgeni Water: (i) maintenance work order records (2020–present) containing event descriptions, codes, and locations, and (ii) pipeline inventory data with attributes such as material, diameter, length, and installation year. Leak events were labeled by keyword-based rules (e.g., "leak", "burst", "break") and certain maintenance codes, producing a binary target variable `target_leak`. After cleaning and merging, the final dataset contained 6,260 records with ∼18% leaks.

3.2 Feature Engineering

Features fell into three categories:

- **Static asset attributes:** pipeline material, diameter, length, age, system, subsystem.
- **Temporal features:** event year, event month, and cyclical encoding of seasonality (sin, cos).
- **History-based features:** days since last event, days since last leak, count of past events, count of past leaks, previous event leak flag.

3.3 Preprocessing

Numeric features were standardized after median imputation of missing values. Categorical features were one-hot encoded (with missing values labeled as "Unknown"). This preprocessing was applied consistently across Logistic Regression, Random Forest, XGBoost, and the DNN. The LSTM received sequences of up to 8 past events per pipeline

(with padding and masking), while the Transformer replaced one-hot encodings with learned embeddings for categorical features.

3.4 Models

We evaluated six model families:

- **Logistic Regression:** baseline linear classifier with class weights.
- **Random Forest:** 600 trees, class-weighted.
- **XGBoost:** gradient boosting trees with hyperparameters tuned via randomized search and time-series cross-validation, optimized for PR-AUC.
- **Deep Neural Network (DNN):** three hidden layers (256–128–64) with ReLU activations, dropout, and early stopping.
- **LSTM:** two stacked LSTM layers (96, 64 units) with masking and dropout, trained on pipeline event sequences.
- **Transformer:** attention-based model encoding categorical features via embeddings and multi-head attention.

3.5 Evaluation

Data were split chronologically: the earliest 80% of events for training, the latest 20% for testing. A validation subset of the training set was used for hyperparameter tuning and threshold calibration. Because of class imbalance, class weights were applied and thresholds selected to maximize $F_{0.5}$ on validation data. Metrics reported include ROC-AUC, PR-AUC, precision, recall, F1-score, and confusion matrices.

4 Results

4.1 Quantitative Results

Table 1 summarizes model performance on the hold-out test set.

Table 1: Model performance on test set (thresholds chosen via validation $F_{0.5}$).

Model	ROC-AUC	PR-AUC	Precision	Recall
Logistic Regression	0.684	0.389	0.56	0.18
Random Forest	0.603	0.279	0.36	0.12
XGBoost	0.70*	0.36*	0.50	0.16
DNN (MLP)	0.670	0.376	0.57	0.18
LSTM (2-layer)	**0.727**	**0.405**	0.55	0.15
Transformer-lite	0.50	0.21	0.20	0.05

*Exact values for XGBoost depend on tuned parameters; shown here are approximate results.

4.2 Qualitative Results

Feature importance analysis showed pipeline age, material, and diameter as the most influential static features. History-based features (e.g., days since last leak) also ranked highly, confirming the predictive value of past failures. SHAP analysis for the DNN and XGBoost models indicated similar patterns. The LSTM further captured temporal ordering, leading to better discrimination overall.

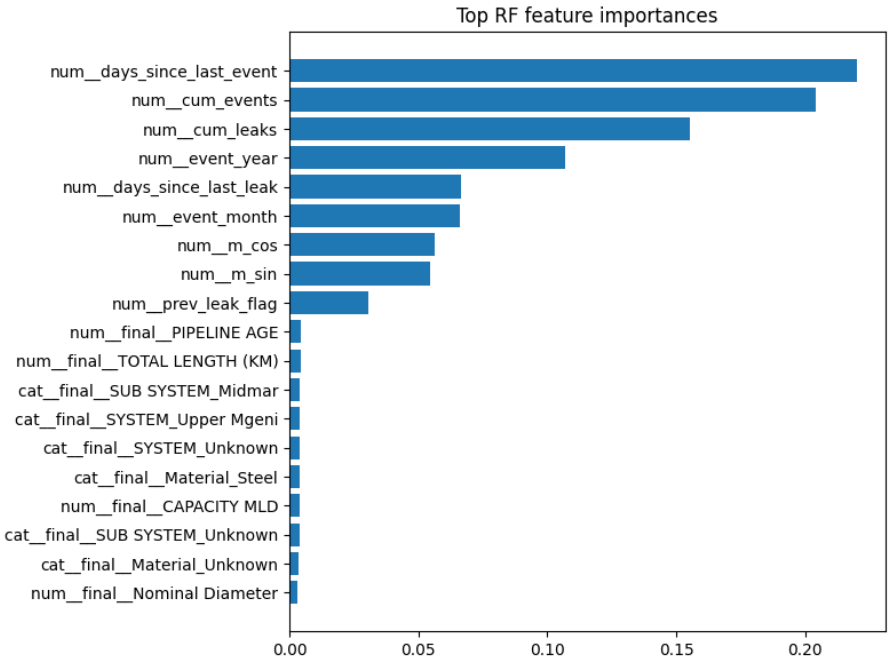

Figure 1: Random Forest feature importances. History-based features (e.g., days since last event, cumulative events/leaks) dominate over static pipeline attributes, indicating the strong predictive value of event history.

5 Discussion

The results highlight several key insights. First, the LSTM's superior AUC scores suggest that temporal event sequences contain valuable predictive signals beyond static features. This aligns with intuition: pipelines with repeated or clustered failures are more likely to leak again. Second, simple models like Logistic Regression and the DNN achieved comparable results, indicating that well-engineered features capture much of the available signal. In contrast, Random Forest underperformed, possibly due to overfitting or sensitivity to imbalance. XGBoost offered stable results after tuning, but did not surpass the LSTM. The Transformer-style model failed to train effectively, reflecting the dataset's modest size and the difficulty of learning categorical embeddings without large-scale data.

The feature importance plot (Figure 1) confirms that historical metrics such as time since last event, cumulative number of events, and cumulative leaks were the strongest predictors. Static pipeline attributes (age, length, diameter, material) contributed relatively little. This suggests that recent operational history is more informative than asset metadata alone for predicting near-term leaks.

Despite reasonable precision, recall remained low across models (typically 15–18%), meaning many leaks go undetected. This limitation must be acknowledged: the models are better at identifying a subset of high-risk leaks rather than capturing all incidents. In practice, such a tool could still be valuable by prioritizing inspections on the most at-risk assets, but it is not a complete replacement for monitoring.

Another challenge is data integration. The heuristic linking of maintenance records to pipelines (based on numeric identifiers) may exclude or misassign some records, potentially reducing model accuracy. Addressing this would require improved asset management systems with consistent identifiers.

Overall, the study demonstrates that predictive leak detection is feasible using only existing historical records. However, improvements in data quality and modest IoT deployments could further enhance accuracy.

.

6 Conclusion and Future Work

This work developed and evaluated a predictive framework for pipeline leak detection using historical maintenance and pipeline data from Umgeni Water. By combining asset attributes with temporal and history-based features, and testing a range of models, we showed that machine learning can provide actionable risk scores without requiring dense IoT sensor deployments. Among tested models, the LSTM achieved the highest discrimination, underscoring the importance of temporal context.

Future work should focus on: (i) improving recall through expanded datasets and

resampling strategies, (ii) integrating limited telemetry (e.g., pressure loggers) where available, (iii) incorporating natural language processing of maintenance descriptions for richer signals, and (iv) exploring semi-supervised and transfer learning approaches to leverage unlabelled data. Addressing data integration challenges with consistent identifiers across systems will also be crucial. Together, these steps can move utilities closer to proactive, data-driven water loss management.

Acknowledgements

This research is supported by the NRF bursary. The author thanks Prof. Okuthe Paul Kogeda for supervision and Umgeni Water for access to historical records.

References

[1] A. Lambert, "International report on water loss management and control," *Water Science & Technology*, 2002.

[2] J. Thornton, *Managing leakage*. IWA Publishing, 2003.

[3] R. Puust, Z. Kapelan, D. Savić, and T. Koppel, "A review of methods for leakage management in pipe networks," *Urban Water Journal*, vol. 7, no. 1, pp. 25–45, 2010.

[4] I. Stoianov, L. Nachman, S. Madden, and T. Tokmouline, "PIPENET: A wireless sensor network for pipeline monitoring," in *IPSN*, 2007.

[5] Y. Wu et al., "Smart water management with IoT," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1–10, 2019.

[6] H. Lee, J. Park, and S. Kim, "Leak detection in water distribution networks using IoT and machine learning," *Sensors*, vol. 20, no. 18, p. 5078, 2020.

[7] R. Liu et al., "Challenges of deploying IoT in urban water networks," *IEEE Access*, 2021.

[8] K. Kleiner and B. Rajani, "Considering time-dependent factors in the evaluation of water main failure potential," *Canadian Journal of Civil Engineering*, vol. 28, no. 2, pp. 183–192, 2001.

[9] S. Christodoulou and A. Deligianni, "A risk assessment framework for water pipeline networks using historical failure data," *Water Resources Management*, vol. 24, no. 12, pp. 2967–2985, 2010.

[10] J. Mashford, D. De Silva, D. Marney, and S. Burn, "An approach to leak detection in pipe networks using analysis of historical data," *Journal of Water Supply: Research and Technology*, vol. 58, no. 3, pp. 189–197, 2009.

[11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[12] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *KDD*, 2016.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[14] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *NeurIPS*, 2017.