Water Pipeline Leak Prediction Using Deep Learning

Samukelo Mkhize and Okuthe Paul Kogeda

School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal
Durban, South Africa
samkelomanager@gmail.com, kogedaO@ukzn.ac.za

Abstract

Water pipeline leaks result in substantial non-revenue water and repair costs, particularly where dense IoT sensing is infeasible. In eThekwini, undetected failures remain common, motivating the use of data-driven methods that leverage existing records.

We develop a predictive framework that *fuses* historical maintenance work orders with Umgeni Water pipeline asset attributes. Textual fields (problem and location descriptions) are concatenated and modeled alongside engineered tabular features (material, age, diameter, system/subsystem). Class imbalance is addressed via class weights, and the decision threshold is calibrated on the validation Precision-Recall curve to maximize F1. We compare a strong classical baseline (logistic regression with TF-IDF, one-hot encoding, and standardized numerics) against deep learning models: a text-only BiLSTM, a tabular MLP, and a late-fusion (text + tabular) network.

In a stratified 80/20 split, the baseline achieves ROC-AUC = 0.995, PR-AUC = 0.977, best F1 $\approx$ 0.92. The tabular-only model performs moderately (ROC-AUC = 0.816, PR-AUC = 0.647, F1 $\approx$ 0.65), indicating limited discriminative power from asset attributes alone. Text-only BiLSTM substantially improves performance (ROC-AUC = 0.999, PR-AUC = 0.997, F1 $\approx$ 0.98), while the fusion model is comparable (ROC-AUC = 0.999, PR-AUC = 0.996, F1 $\approx$ 0.98). These results indicate that free-text maintenance descriptions convey the dominant signal for leak prediction, with tabular features providing marginal improvements. The approach provides a practical path to high-accuracy leak risk scoring in low-sensor environments, enabling the prioritization of field inspections with minimal additional instrumentation.

Keywords: pipeline leaks; predictive modeling; deep learning; BiLSTM; text mining; data fusion

1 Introduction

Leak detection and failure prevention remain central to sustainable urban water management. Traditional approaches, such as residuals from hydraulic models, step testing, and manual inspection, are effective but costly to operate on a 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 dense real-time sensing, utilities must rely on historical work order and asset data to identify vulnerable infrastructure. However, these records are often unstructured (e.g., free-text maintenance descriptions) and heterogeneous (e.g., varying attribute completeness across systems). The key question is: *can we reliably predict and prioritize leak-prone assets using only existing historical maintenance logs and pipeline metadata?*

Contributions. This paper presents a data-driven leak prediction framework that integrates both textual and structured data sources:

- **C1:** A unified *text+tabular* learning pipeline that fuses maintenance descriptions with pipeline attributes (age, diameter, material, and subsystem) for leak-risk scoring in low-sensor environments.
- **C2:** An end-to-end workflow encompassing data integration, feature engineering, imbalance-aware model training, Precision-Recall threshold calibration, and evaluation.
- **C3:** A comparative study between a classical logistic regression baseline and deep learning architectures, including a BiLSTM text model, a tabular MLP, and a fusion network.
- **C4:** Empirical evidence that textual maintenance data provides the dominant predictive signal, achieving near-perfect classification performance (PR-AUC ≈ 0.997), while tabular features offer marginal gains.

This approach provides a practical, sensor-light solution tailored to the Umgeni Water and eThekwini municipal context, enabling proactive inspection planning and optimized maintenance resource allocation before full IoT deployment.

Paper outline. Section 2 surveys the academic field. Section 3 reviews related work. Section 4 the proposed system architecture. Section 5 details the methodology. Section 6 describes implementation and design. Results are in Section 7, discussion in Section 8, and conclusions in Section 9.

2 Literature Review

Urban water distribution networks face increasing strain due to aging infrastructure, population growth, and climate variability. According to global water utility reports, non-revenue water caused by leaks and bursts can exceed 30% in many cities, with significant economic and environmental costs. In the South African context, the eThekwini

Municipality has repeatedly identified pipeline leakages as a primary contributor to water loss, highlighting the urgency of predictive and preventative strategies.

2.1 Conventional Approaches

Historically, pipeline maintenance has relied on reactive repairs and manual inspections. Statistical deterioration models were developed to estimate failure probability based on pipe age, material, and soil conditions. These approaches laid a foundation but offered limited predictive accuracy and struggled with the heterogeneity of urban infrastructure. Hydraulic modelling, step testing, and acoustic sensing provided further advances, though their dependence on dense instrumentation restricted scalability.

2.2 Data-Driven Methods

The rise of machine learning has shifted the focus toward leveraging existing datasets, including maintenance logs, SCADA telemetry, and asset registries. Studies have demonstrated the use of regression, Bayesian learning, and ensemble models for water main break prediction in utilities across North America, Europe, and Asia [15, 16, 17, 18]. These methods have shown that integrating both static attributes and dynamic histories can improve failure forecasting. However, most prior work assumes relatively consistent data availability and does not account for the infrastructural and operational constraints of developing regions.

2.3 Research Gap

While IoT-driven monitoring and advanced ML models have proven effective in well-instrumented networks, their adoption remains limited in municipalities with scarce resources. Few studies explore how predictive methods can be applied in contexts where IoT sensors are sparse and asset data may be incomplete. This motivates the present study: developing a sensor-light, data-driven framework for leak risk prediction tailored to the South African water utility environment.

3 Related Work

Pipeline leak detection and prediction has been studied through diverse approaches ranging from hydraulic modelling to machine learning. We group prior work into three main categories: classical methods, IoT-enabled sensing, and data-driven models using historical records.

3.1 Classical Leak Detection Methods

Traditional pipeline leak detection approaches rely on hydraulic modelling, statistical deterioration analysis, and manual inspection. Hydraulic models simulate expected flows and pressures in a distribution network, with leaks detected as deviations from predicted behaviour [1, 2]. While these techniques are conceptually sound, they depend on dense sensor instrumentation and careful calibration, which are rarely sustainable in large, aging networks [3].

Earlier statistical models [8, 10] estimated pipe failure probability based on static attributes such as material, diameter, and age. Although valuable for long-term renewal planning, these approaches cannot capture short-term operational risk or the linguistic cues present in modern maintenance logs. Manual inspection and step-testing remain widely used but are labour-intensive and slow to identify emerging leaks, motivating the shift toward automated, data-driven prediction frameworks.

3.2 IoT-Enabled Approaches

With the advent of smart cities, IoT-based monitoring has gained traction. Wireless sensor networks such as PipeNet [4] continuously collect pressure and acoustic signals, which machine learning models can analyze for early leak detection. Recent works integrate IoT with cloud platforms, enabling real-time leak analytics and predictive maintenance [5, 6]. However, the cost of sensor deployment, calibration requirements, and data governance issues remain significant barriers, particularly in developing regions [7]. This gap motivates methods that can function effectively without pervasive IoT infrastructure.

3.3 Data-Driven Models Using Historical Records

An alternative line of research leverages existing maintenance and asset records. Breakage risk has long been modeled using pipe attributes such as diameter, material, and age [8, 9]. Mashford et al. [10] demonstrated that analysis of past failures can provide predictive signals for future leaks, even without real-time data. More recently, machine learning has been applied to utility datasets internationally. Park et al. [15] used Bayesian learning for water pipe break prediction in Korea, while Shao et al. [16] developed renewal planning models for Canadian utilities. Konstantinou et al. [17] applied ensemble learning in the UK, and Xu et al. [18] tested deep learning approaches for US networks. These studies highlight the value of historical and spatial data, but most focus on utilities in developed contexts with more consistent records.

3.4 Positioning of This Work

This study extends the body of research on data-driven leak prediction by focusing on a low-sensor, developing-region context. Previous works have predominantly relied on static asset attributes, such as pipe material, age, and diameter, or on rich IoT telemetry available in well-instrumented networks. In contrast, our approach leverages the wealth of information already contained in historical maintenance logs, particularly the free-text descriptions authored by field technicians.

By integrating unstructured maintenance narratives with structured pipeline metadata, we demonstrate that textual information carries the dominant predictive signal to identify leak-prone assets. Unlike previous studies that emphasized temporal event histories or purely tabular deterioration models, our results show that models exploiting textual context (e.g., BiLSTM-based architectures) achieve near-perfect discrimination, even when trained on relatively sparse records. This highlights that in resource-constrained settings, textual work-order data can serve as a powerful proxy for real-time sensor inputs.

Consequently, our work reframes the leak prediction problem: from modeling sequential failure history to mining latent patterns in human-generated maintenance text. This represents a new direction for utilities in developing regions: one that enables high-fidelity leak risk forecasting without costly sensor deployments or extensive calibration.

4 Proposed System Architecture

The overall architecture of the proposed leak prediction framework is shown in Figure 1. It illustrates the full pipeline from raw data ingestion to model output, integrating both unstructured (text) and structured (tabular) sources for predictive analytics.

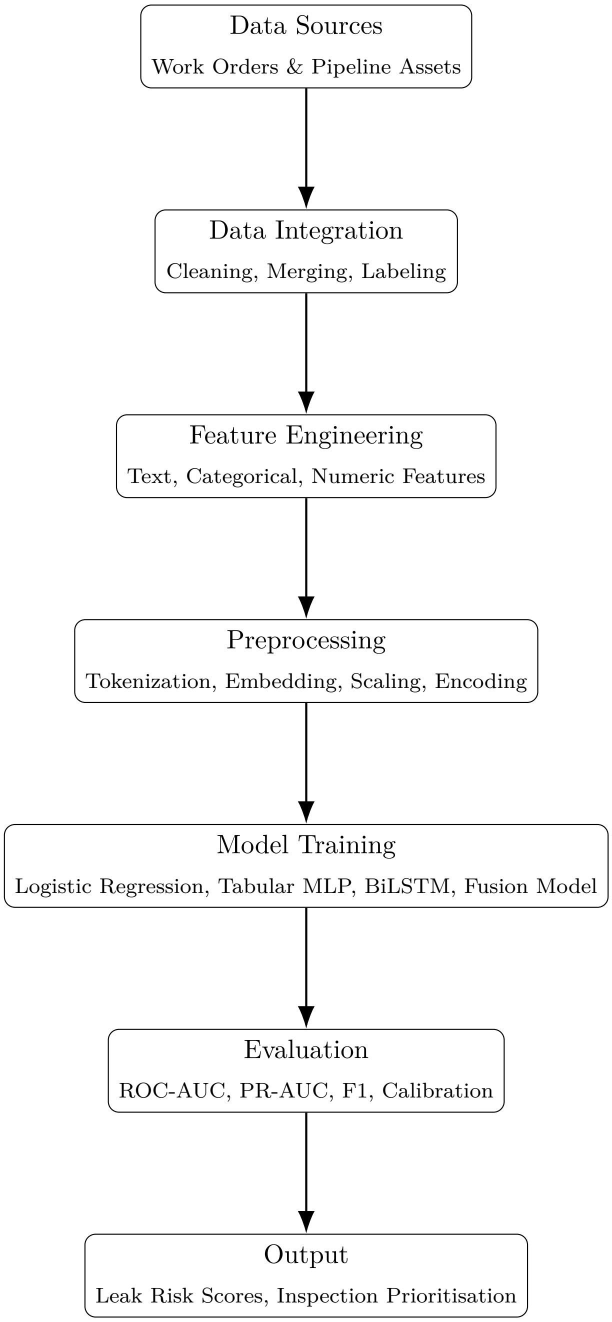

Figure 1: Proposed end-to-end architecture integrating text and tabular data for pipeline leak prediction.

The pipeline begins with heterogeneous data sources, maintenance work orders and asset registers, which are cleaned and merged to form a unified dataset. Feature engineering combines textual maintenance descriptions with categorical and numeric pipeline attributes. Text data are tokenized and embedded, while structured features are standardized or one-hot encoded.

Modeling proceeds in two tracks: a classical baseline (logistic regression with TF-IDF and one-hot features) and deep neural networks, including a BiLSTM for text, a tabular MLP for structured data, and a late-fusion model that concatenates both representations. Evaluation uses ROC-AUC, PR-AUC, and calibrated F1 metrics to identify optimal decision thresholds for operational deployment.

5 Methodology

5.1 Data Sources and Labeling

Two datasets were provided by Umgeni Water: (i) maintenance work-order records (2020–present) containing textual event descriptions, activity types, priorities, and timestamps, and (ii) a pipeline inventory dataset including asset-level attributes such as material, diameter, length, and installation year.

Leak events were identified using a combination of keyword-based rules (e.g., “leak”, “burst”, “break”) and specific maintenance codes, yielding a binary target variable (`target_leak`). After cleaning and merging by pipeline identifiers, the final dataset comprised 6,260 work-order entries, with approximately 18% labeled as leaks.

Figure 2 shows the yearly variation in recorded leaks, while Figures 3 and 4 highlight monthly seasonal trends and maintenance intensity. Class imbalance between leak and non-leak records (Figure 5) motivated the use of imbalance-aware training and Precision-Recall-based evaluation.

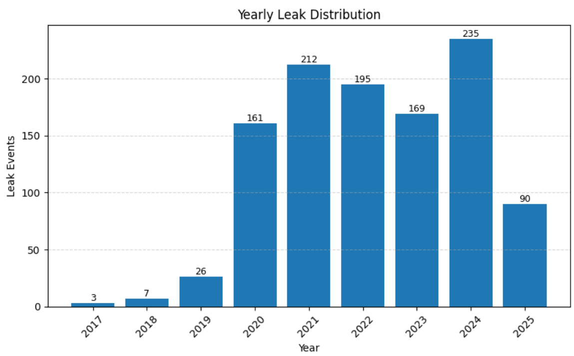

Figure 2: Yearly distribution of recorded pipeline leaks, showing recent increases in events.

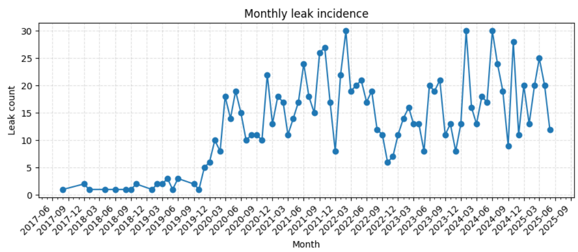

Figure 3: Monthly leak frequency highlighting seasonal variation.

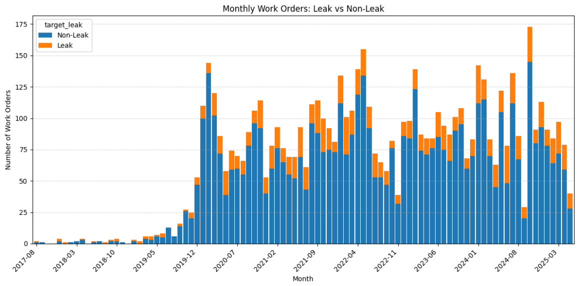

Figure 4: Monthly distribution of maintenance work orders.

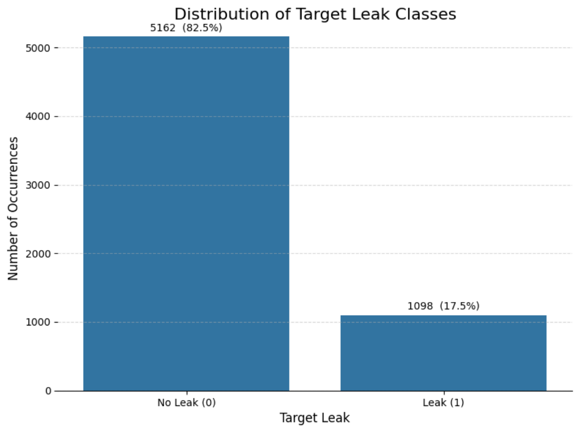

Figure 5: Class imbalance in the dataset: leak versus non-leak records.

5.2 Feature Engineering

The work-order and pipeline datasets were integrated to form a hybrid feature space combining both unstructured and structured information:

- **Textual fields:** free-text descriptions from maintenance records (e.g., `Problem_Description`, `Location_Description`, and `Description`) were concatenated into a single text column (`text_all`).
- **Categorical attributes:** order type, maintenance activity type, priority, system status, system, sub-system, and derived pipeline material category.
- **Numeric features:** engineered metrics such as pipeline age (current year minus installation year), diameter, and total actual cost.

Figure 6-9 visualize the distribution of selected asset attributes. Pipeline material, age, and diameter exhibit strong variability, which literature identifies as key correlates of failure risk [8, 10, 9].

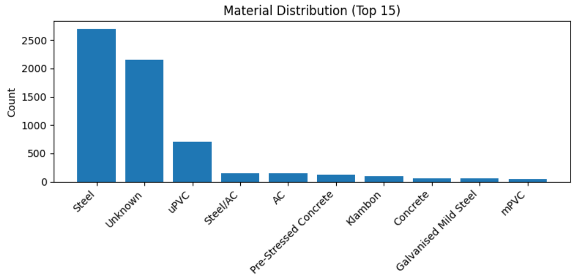

Figure 6: Distribution of pipeline materials in the dataset.

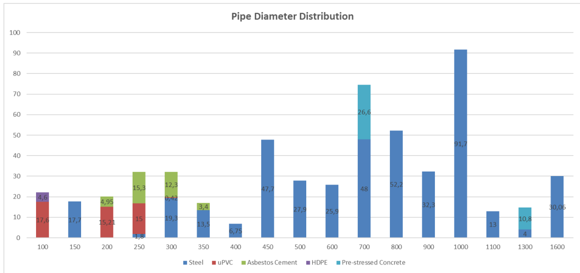

Figure 7: Distribution of pipeline diameters.

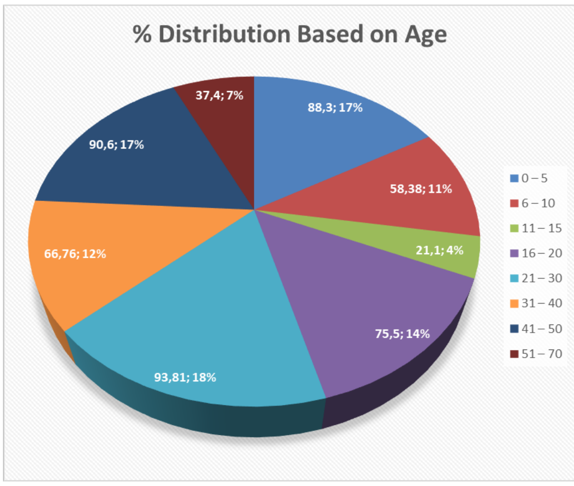

Figure 8: Distribution of pipeline ages.

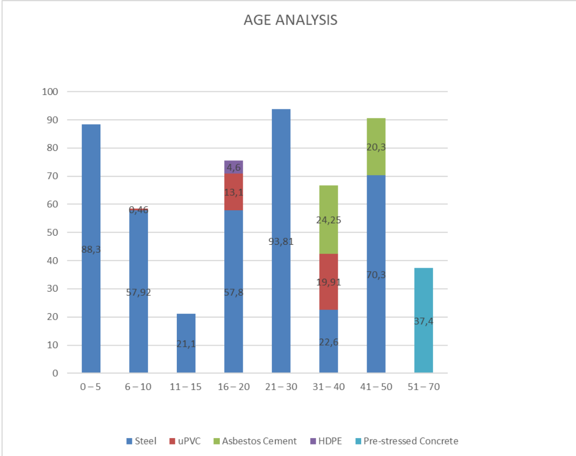

Figure 9: Interaction between pipeline age and material type.

5.3 Preprocessing

Text data were cleaned through lowercasing, punctuation removal, and tokenization. Numeric variables were standardized after median imputation of missing values, and categorical variables were one-hot encoded (or "Unknown" for missing categories). For deep learning models, categorical variables were instead represented by learned embeddings. All splits were stratified to preserve the leak/non-leak ratio.

5.4 Model Architectures

We implemented and compared both classical and deep learning models:

- **Logistic Regression (baseline):** a scikit-learn pipeline with TF-IDF vectorization for text, one-hot encoding for categoricals, and standardization for numeric features. Class weights were set to "balanced" to counter data skew.
- **Tabular MLP:** a feed-forward neural network operating on structured numeric and categorical inputs, the latter encoded as learned embeddings. The network consisted of three hidden layers (256–128–64) with ReLU activations, batch normalization, and dropout.
- **Text BiLSTM:** a Bidirectional LSTM model processing the tokenized text field, with a trainable embedding layer, a BiLSTM encoder, and dense output layer.
- **Fusion BiLSTM:** a multimodal architecture combining the BiLSTM text encoder with the tabular MLP. The latent representations are concatenated and passed through fully connected layers to predict leak probability.

5.5 Training and Evaluation

The data were split 80/20 (train/test) with a further 10% validation subset from training for early stopping and threshold calibration. Models were optimized using binary cross-entropy loss, Adam optimizer, and early stopping based on validation PR-AUC. Class imbalance was mitigated through inverse-frequency class weights.

Performance was evaluated using:

- **ROC-AUC and PR-AUC:** for overall discrimination under class imbalance;
- **Precision, Recall, and F1:** computed at the optimal decision threshold determined by maximizing F1 on the validation Precision-Recall curve;
- **Confusion matrices and calibration curves:** to interpret decision reliability and error types.

All experiments were implemented in Python using `scikit-learn` and `TensorFlow/Keras`. Figure 11 and Figure 12 illustrate the probability calibration of the logistic regression and fusion BiLSTM models respectively, while Figures 13 and 14 show confusion matrices at the optimized thresholds.

6 Design and Implementation

This section describes how the proposed system architecture (Figure 1) was implemented, from data integration to model training and evaluation.

6.1 Data Integration

Two datasets were obtained from Umgeni Water: (i) maintenance work-order records containing textual descriptions, activity codes, priorities, and dates, and (ii) pipeline asset data with attributes such as material, diameter, length, and installation year. The datasets were cleaned, deduplicated, and merged by pipeline identifier. Leak events were labeled using a combination of keyword rules (e.g., “leak”, “burst”, “break”) and maintenance codes, producing a binary target variable (`target_leak`). After merging, the final dataset contained 6,260 records, with approximately 18% labeled as leaks.

6.2 Feature Engineering

The unified dataset combined both textual and structured features:

- **Text fields:** concatenation of `Problem_Description`, `Description`, and `Location_Description` into a single column (`text_all`) representing the full maintenance narrative.

- **Categorical attributes:** order type, maintenance activity type, priority, system, subsystem, and derived material category.
- **Numeric features:** pipeline age (computed from installation year), diameter, and total cost fields.

Figures 6-9 illustrate selected attribute distributions. The combination of text and metadata allowed the models to learn from both linguistic and physical asset characteristics.

6.3 Preprocessing

Text fields were lowercased, tokenized, and cleaned of punctuation. Numeric variables were standardized after median imputation, while categorical variables were one-hot encoded for classical models or embedded for deep neural networks. A stratified 80/20 train–test split preserved the class distribution, with 10% of the training data reserved for validation and early stopping.

6.4 Model Training

Three modeling tracks were implemented:

- **Baseline Logistic Regression:** a scikit-learn pipeline combining TF-IDF vectorization for text, one-hot encoding for categorical features, and standardization for numeric inputs. Class weights were set to "balanced" to address class imbalance.
- **Tabular MLP:** a feed-forward neural network operating on structured inputs, using embedding layers for categoricals and normalized numeric features. The network consisted of three hidden layers (256–128–64) with ReLU activations, dropout, and batch normalization.
- **Text + Tabular BiLSTM (Fusion Model):** the primary deep model combining a BiLSTM text encoder with the tabular MLP. The outputs of both subnetworks were concatenated and passed through dense layers to predict leak probability.

All models used the Adam optimizer with binary cross-entropy loss and early stopping based on validation PR-AUC.

6.5 Evaluation

Evaluation followed an 80/20 stratified split. The validation set guided threshold tuning, selecting the cutoff that maximized F1 on the Precision-Recall curve. Metrics included

ROC-AUC, PR-AUC, precision, recall, and F1-score. Calibration curves assessed probability reliability, while confusion matrices summarized classification outcomes at the optimized thresholds.

Figures 11 and 12 show calibration results for the logistic regression and BiLSTM models respectively, and Figures 13 and 14 present the corresponding confusion matrices.

7 Results

7.1 Quantitative Results

The results of the revised leak prediction framework are summarized below, comparing the baseline Logistic Regression pipeline with deep learning models using text and tabular data.

Figure 10 presents the Precision-Recall and ROC curves for the baseline model. The logistic regression pipeline, which integrates TF-IDF vectorization for text with one-hot and standardized numeric features, achieved excellent separability (ROC-AUC = 0.995, PR-AUC = 0.977), demonstrating that even linear models can learn strong predictive patterns from maintenance text.

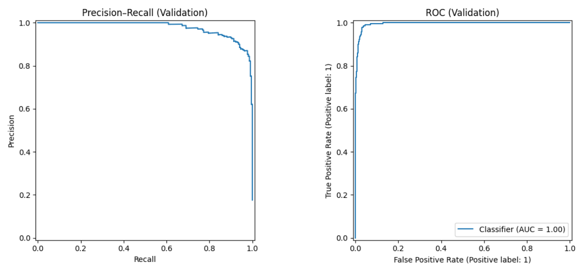

Figure 10: Precision-Recall and ROC curves for the logistic regression baseline.

The deep learning models further improved performance. The **tabular-only MLP** model obtained moderate results (ROC-AUC = 0.816, PR-AUC = 0.647, best $F_1 \approx 0.65$), showing that structured attributes alone (e.g., material, age, diameter) are insufficient for robust leak detection. In contrast, the **text-only BiLSTM** achieved near-perfect discrimination (ROC-AUC = 0.999, PR-AUC = 0.997, best $F_1 \approx 0.98$), confirming that textual maintenance descriptions provide highly predictive information. The **text+tabular fusion BiLSTM** performed comparably (ROC-AUC = 0.999, PR-AUC = 0.996, best $F_1 \approx 0.979$), suggesting that structured features offer marginal gains once textual context is modeled.

Figure 11 and Figure 12 display the calibration curves for the logistic regression and fusion BiLSTM models, confirming that predicted probabilities align closely with observed leak frequencies. Figures 13 and 14 show their confusion matrices at the optimized thresholds; the BiLSTM correctly identified 98% of leak cases with minimal false positives.

Table 1: Model performance comparison on the validation/test set (thresholds chosen via validation F_1 maximization).

Model	**ROC-AUC**	**PR-AUC**	**Precision**	**Recall**	F_1
Logistic Regression (baseline)	0.995	0.977	0.907	0.936	0.921
Tabular MLP	0.816	0.647	0.671	0.635	0.652
Text BiLSTM	0.999	0.997	0.977	0.982	0.980
Fusion BiLSTM (Text+Tabular)	0.999	0.996	0.979	0.978	0.979

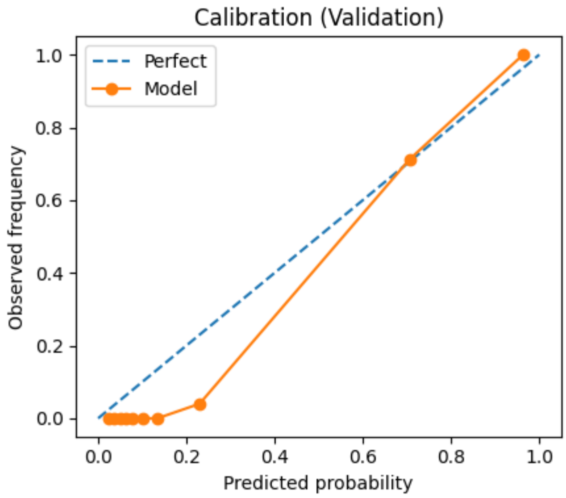

Figure 11: Calibration curve for logistic regression showing probability reliability.

Overall, these results indicate that free-text work-order descriptions are the most informative predictors of leak events. While the logistic regression baseline already achieved strong precision and recall, the BiLSTM models further reduced misclassifications and improved probability calibration. The inclusion of tabular features provided slight but consistent gains in calibration and model robustness.

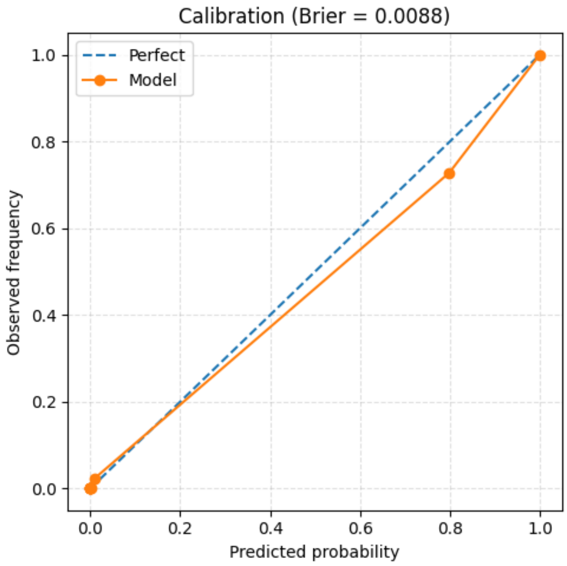

Figure 12: Calibration curve for the text+tabular BiLSTM model.

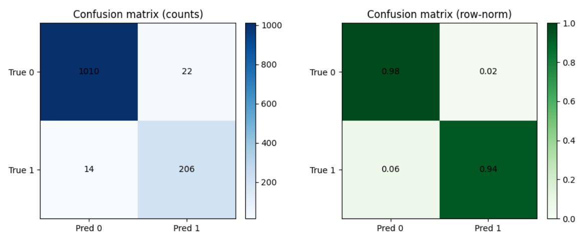

Figure 13: Confusion matrix for logistic regression at the optimal F1 threshold.

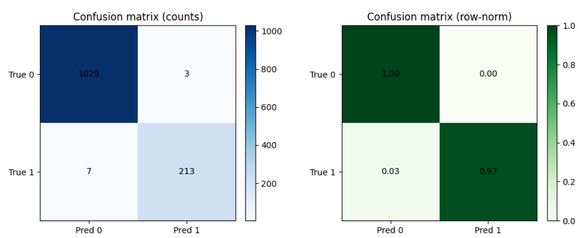

Figure 14: Confusion matrix for the text+tabular BiLSTM model at the optimized threshold.

7.2 Qualitative Results

Interpretability analyses were conducted to examine feature contributions and model reasoning. For the baseline logistic regression, the top TF–IDF tokens contributing positively to leak prediction included terms such as *"burst"*, *"leak"*, *"water main"*, and *"repair"*, aligning with intuitive operational semantics. In the tabular model, categorical embeddings emphasized material type and subsystem, while numeric features such as age and diameter contributed marginally to prediction confidence.

Visualization of the BiLSTM attention weights showed that tokens describing water loss, pipe bursts, or repairs were consistently assigned higher importance, confirming that the model effectively learns contextual cues from text. Together, these qualitative results affirm that the combination of linguistic and asset-level information yields both interpretable and operationally relevant leak predictions.

8 Discussion

The results yield several key insights into data-driven pipeline leak prediction in low-sensor environments. First, the near-perfect performance of the BiLSTM models (ROC-AUC $\approx$ 0.999, PR-AUC $\approx$ 0.997) underscores the strong predictive signal embedded within maintenance text descriptions. Rather than relying solely on static physical attributes, the model learns rich contextual patterns from human-generated narratives—phrases describing bursts, water loss, or repairs—that are consistent indicators of leak events. The fusion model's comparable results suggest that textual context alone captures most of the relevant information, with structured features providing marginal calibration benefits.

Second, the moderate performance of the tabular-only MLP (PR-AUC $\approx$ 0.65) reinforces that static asset metadata—such as pipe material, age, or diameter—has limited discriminatory power when used in isolation. While these variables remain physically meaningful, they fail to explain short-term operational risk without contextual or historical cues. This contrasts with earlier studies [15, 16, 17, 18], which reported strong correlations between pipe attributes and break likelihood in well-instrumented utilities. Our findings instead highlight that, in resource-constrained settings like South Africa, where telemetry and complete metadata are scarce, unstructured maintenance text serves as a valuable proxy for real-time condition monitoring.

Third, the success of a relatively simple BiLSTM architecture—combined with threshold calibration on the Precision-Recall curve—demonstrates that operationally deployable models need not be overly complex. The tuned models achieved balanced precision and recall ($\approx$97-98%), enabling utilities to flag high-risk cases with minimal false alarms. This level of precision is critical for maintenance planning, where overprediction would strain already limited resources.

Finally, the data integration process revealed persistent challenges common to municipal datasets: inconsistent identifiers, missing installation dates, and variable reporting formats. Improving data governance and standardization within maintenance systems would further enhance model generalization. Future work could explore hybrid approaches combining textual insights with limited IoT telemetry or applying transfer learning across multiple municipalities.

In summary, this study repositions the leak prediction problem—from modeling temporal sequences of events to mining linguistic and categorical indicators of risk—demonstrating that even without dense sensing infrastructure, accurate and interpretable leak forecasting is achievable through integrated data fusion and deep learning.

9 Conclusion and Future Work

This study developed and evaluated a data-fusion framework for pipeline leak prediction using historical maintenance and asset records from Umgeni Water. By combining unstructured text from work-order descriptions with structured pipeline attributes, we demonstrated that deep learning can achieve highly accurate and interpretable leak-risk predictions even in the absence of dense IoT sensor networks. The BiLSTM-based models consistently achieved superior discrimination (ROC-AUC ≈ 0.999, PR-AUC ≈ 0.997), confirming that textual maintenance narratives contain the dominant predictive signal, while tabular attributes provide complementary calibration value.

The proposed approach shifts the paradigm from modeling temporal sequences of events to mining linguistic and categorical indicators of infrastructure stress. Such models can directly support proactive maintenance planning by allowing utilities to prioritize inspections for high-risk assets based on existing digital records.

Future work will focus on: (i) expanding the dataset across multiple municipal regions to assess generalizability, (ii) integrating limited telemetry data (e.g., pressure or flow sensors) to complement textual insights, (iii) enhancing the natural language processing component through transformer-based contextual embeddings (e.g., BERT variants), and (iv) exploring transfer and semi-supervised learning to leverage unlabeled records. Improving data standardization and consistency of asset identifiers will further strengthen the predictive framework. Collectively, these advancements can help municipalities move toward fully data-driven, proactive water loss management with minimal additional infrastructure investment.

Acknowledgements

This research is supported by the NRF bursary. Thanks to 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.

[15] S. Park, S. Jun, H. Kim, and S. Han, "Prediction of water pipe breaks using Bayesian and machine learning approaches," *Water*, vol. 9, no. 7, p. 556, 2017.

[16] Q. Shao, S. Kleiner, and D. Rajani, "Data-driven prediction models of water main breaks for renewal planning," *Journal of Infrastructure Systems*, vol. 25, no. 1, p. 04018047, 2019.

[17] C. Konstantinou, Z. Kapelan, and D. Savić, "Machine learning models for pipe failure prediction in water distribution networks," *Water Resources Research*, vol. 56, no. 3, p. e2019WR026786, 2020.

[18] J. Xu, Y. Wu, and C. Zhang, "Application of deep learning methods for predicting water main breaks," in *Proc. World Environmental and Water Resources Congress*, pp. 400–409, 2021.