Proceedings of International Structural Engineering and Construction, 12(1), 2025
Integrating Technology for Advanced Practices in Structural Engineering and Construction
Edited by Vimonsatit, V., Askarinejad, H., Pour, M., Singh, A., and Yazdani, S.

www.doi.org/10.14455/ISEC.2025.12(1).ENV-03

MACHINE LEARNING APPROACH TO ROOT INTRUSION PREDICTION IN URBAN SEWERS USING CCTV AND ENVIRONMENTAL FEATURES

DRAMANI ARIMIYAW, TAREK ZAYED, MOHAMED NASHAT, JINGCHAO YANG, and EWALD KUORIBO

Dept of Building and Real Estate), The Hong Kong Polytechnic Univ, Hung Hom, Hong Kong

Sewer blockages from root intrusion pose significant economic and environmental challenges for water utilities worldwide. While machine learning (ML) offers promising solutions for infrastructure management, its application to this specific failure mode remains largely unexplored. This study develops a comprehensive ML framework for predicting root intrusion risk in sewer systems by integrating physical pipe attributes, environmental features, and demographic data from Hong Kong. Three classification algorithms; Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) were systematically evaluated and compared. Results showed that LR and RF achieved statistically equivalent and superior performance with an AUC-ROC of 0.933, while SVM performed marginally lower at 0.914. The comparable performance of the simpler linear model (LR) with the complex ensemble method (RF) indicates that the predictive relationships are predominantly linear in nature. Feature importance analysis revealed that geographic (District) and demographic (Total Population) contextual factors were more influential predictors than specific pipe characteristics. These findings provide water utilities with a highly interpretable and effective tool for proactive asset management, enabling targeted inspections, optimized maintenance scheduling, and improved resource allocation.

Keywords: Feature analysis, Preventive maintenance, Logistic regression, Support vector machine, Random forest.

1 INTRODUCTION

Sewer system plays a pivotal role in urban development by collecting and conveying sewage and excess stormwater (Post *et al.* 2015). However, blockages in sewers and drains represent a significant operational challenge for water utilities (Marlow *et al.* 2011). These blockages often cause sewer overflows and flooding, leading to substantial environmental pollution and public health risks (Kelly *et al.* 2024). A variety of factors contribute to sewer blockages, including physical sewer characteristics, population density, tree density, and precipitation (Kargar and Joksimovic 2024). Tree root intrusion through sewer joints is responsible for up to 50% of all sewer pipe blockages (Östberg *et al.* 2012). Timely detection of root-intruded pipes is therefore crucial for efficient network management and environmental risk prevention. Currently, closed-circuit television (CCTV) inspection is the predominant technique due to its availability, ease of use, and relatively low cost (Faris *et al.* 2024). However, the extensive nature of sewer networks

and limited utility budgets restrict periodic inspections to only 5-10% of the total network (Salihu *et al.* 2023).

Recent advances in machine learning offer promising solutions for infrastructure risk prediction; however, their application to wastewater pipeline management remains limited. Existing research primarily focuses on predicting structural failure (Castiblanco Ballesteros *et al.* 2022) or assessing corrosion (Wang *et al.* 2023), paying minimal attention to root intrusion despite its disproportionate economic impact. This gap is critical because root intrusion involves unique interactions between pipe materials and root systems, which depend on complex environmental factors beyond traditional structural parameters.

To address this gap, a comprehensive machine learning framework was developed that assesses root intrusion risk by integrating multi-source urban data, including physical pipe attributes (e.g., age, material, diameter), environmental context (e.g., land use, proximity to roads), and demographic factors. This study employed a strategic model selection to ensure both interpretability and performance. Two highly interpretable, global models (Logistic Regression, Support Vector Machines) were compared with a Random Forest classifier, an ensemble method previously demonstrated in sewer condition literature to achieve comparable or superior efficiency to gradient boosting methods, while often being less computationally intensive and easier to tune.

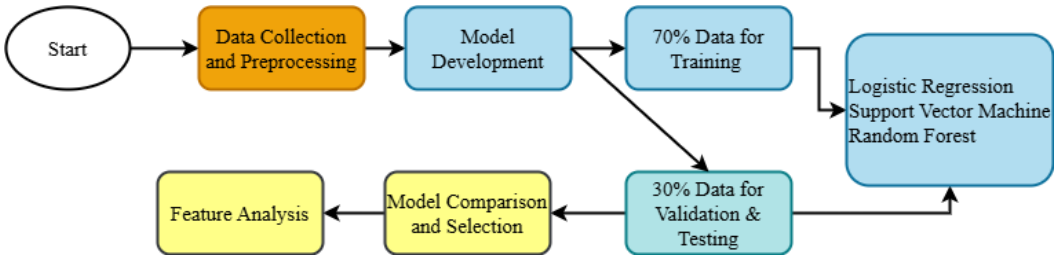

Figure 1. Research methodology.

2 METHODOLOGY

This study employed a structured four-step approach to predict sewer root intrusion: (1) data collection and preprocessing, (2) model development, (3) model evaluation, and (4) feature importance analysis.

2.1 Data Collection and Preprocessing

The dataset integrated CCTV inspection reports (2007–2021) from Hong Kong's Drainage Services Department with GIS pipeline attributes (district, installation date, material, diameter, spatial layout), population data from the Census and Statistics Department, and land use information from the Planning Department. The final dataset comprised 1,511 pipe records (Table 1). Missing values were imputed using median for numerical features and mode for categorical features. Numerical features were standardized (zero mean, unit variance), and categorical features were one-hot encoded.

Table 1. Descriptive statistics of sample input data.

	Total Population	Distance	Pipe Age	Nominal Diameter	Computed Length
Count	1511	1511	1511	1511	1511
Mean	17324.81	3606.47	34.23	408.80	19.07

Std	3092.98	1665.79	15.61	323.12	16.33
Min	6014.0	156.16	1.0	100.0	0.34
Max	25343.0	10002.12	116.0	2550.0	150.68

2.2 Model Development and Evaluation

Three classifiers were selected: Logistic Regression (LR) for interpretability and linear relationship modeling, Support Vector Classifier (SVC) for handling non-linear decision boundaries, and Random Forest (RF) for ensemble robustness (Loganathan *et al.* 2024, Woldesellasse and Tesfamariam 2024). The dataset was split 70%-15%-15% for training, validation, and testing. Class imbalance was addressed using class_weight='balanced'. Key hyperparameters included LR (max_iter=1000), SVC (probability=True), and RF (n_estimators=200, max_depth=10). Performance was evaluated using confusion matrices, quantitative metrics (Accuracy, Precision, Recall, F1-score), and ROC-AUC analysis.

2.3 Feature Importance Analysis

The top-performing model underwent feature importance analysis using the Gini importance metric, ranking features by their contribution to reducing node impurity across all trees, thereby enhancing interpretability and guiding future optimization.

3 RESULTS

3.1 Comparative Model Performance

The performance comparison on the independent test set, summarized in Table 2, reveals that LR and RF achieved statistically equivalent and superior performance across key metrics, including Accuracy (~0.869) and F1-score (~0.865). The SVC, in contrast, demonstrated lower predictive capability on this dataset. The confusion matrices presented in Figure 2 provide granular insight into this performance. The matrices for both LR and RF show a balanced and high count of true positives and true negatives, confirming their robust and similar classification ability. The SVC matrix, however, reveals a higher misclassification rate, particularly a greater number of false negatives (20), which directly explains its lower recall and F1-score as seen in Table 2. This hierarchy of model performance is further validated by the Receiver Operating Characteristic (ROC) curves in Figure 3. The ROC curves for LR and RF are nearly superimposed, both achieving a high AUC of approximately 0.933, which indicates excellent and identical discriminative power between the two classes. The SVC showed only a marginally lower level of performance with an AUC-ROC of 0.914, indicating that all three models possess strong discriminative ability for this task.

This suggests that the underlying relationships between the selected infrastructure and environmental predictors and root intrusion risk may be predominantly linear in nature. The strong performance of the SVC (AUC-ROC = 0.914) further supports this, as it effectively identified a stable separation hyperplane in the feature space. The high performance across all models confirms that the engineered feature set comprehensively captures the primary mechanisms driving root intrusion.

Table 2. Test performance of classification models.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.868	0.842	0.889	0.865

Model	Accuracy	Precision	Recall	F1 Score
Support Vector Classifier	0.846	0.823	0.861	0.842
Random Forest	0.868	0.842	0.889	0.865

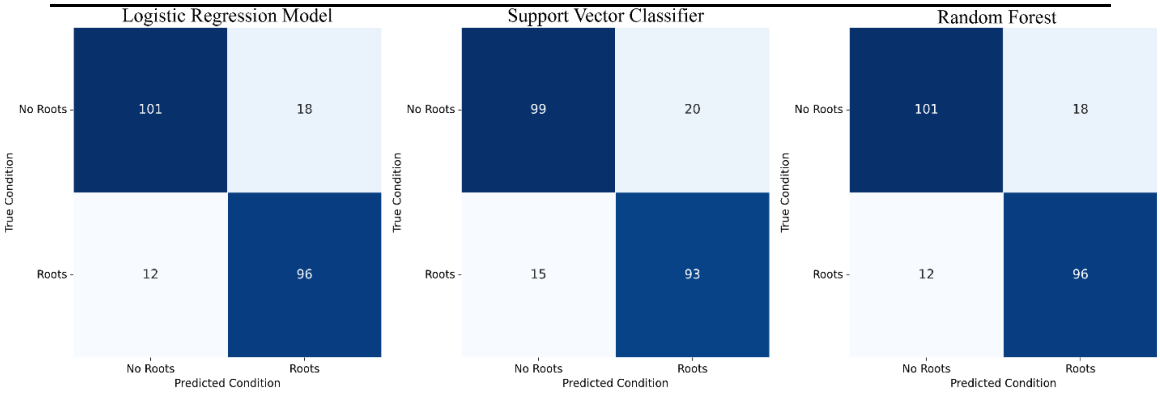

Figure 2. Confusion matrices for models.

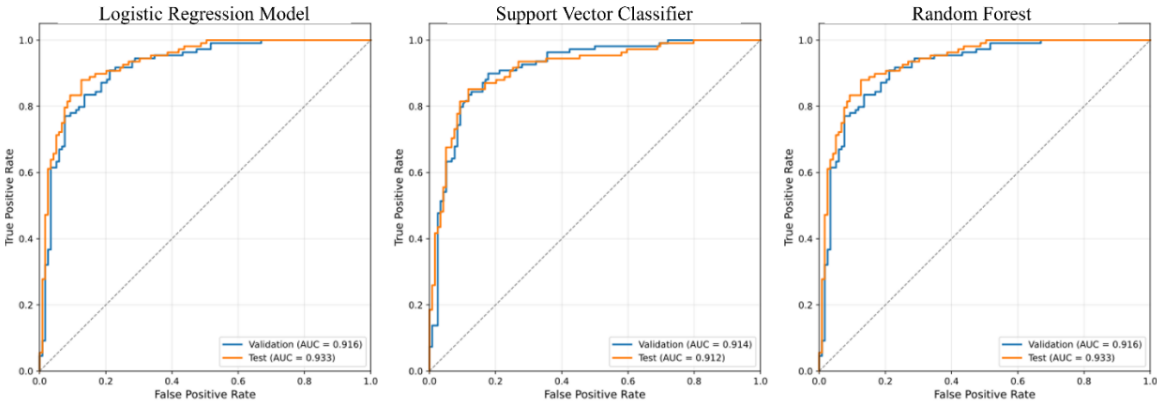

Figure 3. ROC analysis curves.

3.2 Feature Analysis

The RF model, which demonstrated equivalent performance to Logistic Regression, was subjected to feature importance analysis using the Gini importance metric to identify the most influential predictors of root intrusion. The results revealed a distinct hierarchy of predictive features, with District emerging as the dominant factor, followed by Total Population, Computed Length, Distance, and Land Use (Figure 4). The prominence of District as the primary predictor suggests significant spatial heterogeneity in root intrusion risk across Hong Kong's urban landscape. This likely reflects complex interactions between unobserved district-specific factors, including variations in soil composition, prevalent tree species with differential root aggressiveness, historical maintenance practices, and infrastructure age profiles. The high importance of Total Population aligns with established understanding of hydraulic loading and wastewater composition effects on pipe condition, while Length and Distance parameters capture fundamental physical vulnerabilities related to pipe exposure and proximity to root sources.

Notably, the feature importance ranking reveals that broader geographic and demographic contextual factors outweighed specific pipe attributes in predictive power for this dataset. This

finding has direct implications for asset management strategy, suggesting that utilities should prioritize inspection resources according to district-level risk profiles and population density metrics, potentially before conducting more granular, pipe-specific assessments. The strong performance of these contextual features further explains why the Random Forest model, capable of capturing these complex non-linear interactions, achieved equivalent performance to the simpler Logistic Regression approach.

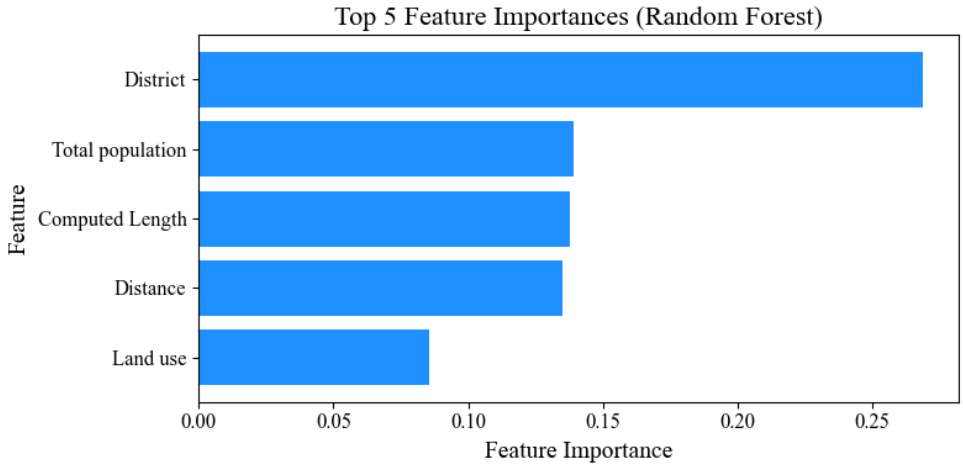

Figure 4. Feature importance plot.

4 LIMITATIONS AND FUTURE RESEARCH

4.1 Limitations

This study is constrained by its geographical focus on a single metropolitan area (Hong Kong), which may limit the direct transferability of the model to regions with differing environmental conditions, utility practices, or pipe materials. Furthermore, the dataset represents a static snapshot, excluding temporal degradation dynamics and potential seasonal variations in root growth.

5 CONCLUSIONS

This study successfully developed and validated a machine learning framework for predicting root intrusion in sewer pipes. The key finding, that a highly interpretable Logistic Regression model matched the performance of a complex Random Forest, indicates that the underlying risk factors are predominantly linear and well-captured by the selected features. The feature importance analysis further revealed that geographic (District) and demographic (Total Population) contextual factors were more influential than specific pipe attributes. This provides water utilities with a practical, transparent, and effective tool for transitioning from reactive maintenance to a proactive, risk-based asset management strategy, enabling optimized inspection targeting and significant potential cost savings.

Acknowledgments

This work was supported by the Research Grants Council of the University Grants Committee [RGC-15209022] and the General Research Fund (GRF) [GRF-15202524] in Hong Kong.

References

Castiblanco Ballesteros, S., Cardenas Mercado, L., Valle Mendoza, J. E., Espitia Layton, S. P., Vanegas Granados, L. C., Caicedo, A., and Torres, A., *SVM-based Predictive Model for the Most Frequent Structural Failure in Bogota Sewer System*, International Journal of Critical Infrastructures, Inderscience Publishers, 18(4), 366-380, 2022.

Faris, N., Zayed, T., Aghdam, E., Fares, A., and Alshami, A., *Real-time Sanitary Sewer Blockage Detection System Using IoT*, Measurement, Elsevier, 226, 114146, February 2024.

Kargar, K., and Joksimovic, D., *Analysis of Sewer Blockage Causes Using Open Data*, Water Practice and Technology, IWA Publishing, 19(9), 3855-3866, September 2024.

Kelly, D. A., Garden, M., Sharif, K., Campbell, D., and Gormley, M., *A Novel Approach to Detecting Blockages in Sewers and Drains: The Reflected Wave Technique*, Buildings, MDPI, 14(10), 3138, October 2024.

Loganathan, K., Najafi, M., Kermanshachi, S., Maduri, P. K., and Pamidimukkala, A., *Inspection Prioritization of Gravity Sanitary Sewer Systems Using Supervised Machine Learning Algorithms*, Journal of Infrastructure Preservation and Resilience, Springer, 5(1), 9, July 2024.

Marlow, D. R., Boulaire, F., Beale, D. J., Grundy, C., and Moglia, M., *Sewer Performance Reporting: Factors That Influence Blockages*, Journal of Infrastructure Systems, ASCE, 17(1), 42-51, March 2011.

Östberg, J., Martinsson, M., Stål, Ö., and Fransson, A. M., *Risk of Root Intrusion by Tree and Shrub Species into Sewer Pipes in Swedish Urban Areas*, Urban Forestry and Urban Greening, Elsevier, 11(1), 65-71, 2012.

Post, J., Pothof, I., ten Veldhuis, M.-C., Langeveld, J., and Clemens, F., *Statistical Analysis of Lateral House Connection Failure Mechanisms*, Urban Water Journal, Taylor & Francis, 13(1), 69-80, 2015.

Salihu, C., Mohandes, S. R., Kineber, A. F., Hosseini, M. R., Elghaish, F., and Zayed, T., *A Deterioration Model for Sewer Pipes Using CCTV and Artificial Intelligence*, Buildings, MDPI, 13(4), 952, April 2023.

Wang, W., Xu, X., Cao, J., Zeng, M., and Zhang, W., *Deciphering and Predict Corrosion Effect, Influencing Factors and Microbial Mechanism of Sewer Concrete Corrosion Based on Extensive Data Analysis and Machine Learning*, Urban Water Journal, Taylor & Francis, 20(9), 1219-1230, 2023.

Woldesellasse, H., and Tesfamariam, S., *Data Augmentation Using Conditional Generative Adversarial Network (cGAN): Applications for Sewer Condition Classification and Testing Using Different Machine Learning Techniques*, Journal of Hydroinformatics, IWA Publishing, 26(7), 1471-1489, July 2024.