Towards Predictive Modeling of Time to Sewer Pipe Failure: A Preliminary Exploratory Study Combining Statistical Analysis and AI Techniques

Jingchao Yang[1], Tarek Zayed[1], Dramani Arimiyaw[1], Mohamed Nashat[1, 3], Xianyang Liu[2, *], Abdelazim Ibrahim[1,4]

1 Department of Building and Real Estate (BRE), Faculty of Construction and Environment (FCE), The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
2 Department of Civil and Architecture Engineering, Hainan University, Haikou 570228, China
3 Department of Public Works Engineering, Faculty of Engineering, Mansoura University, Mansoura, 35516, Egypt
4 Department of Civil Engineering, Benha Faculty of Engineering, Benha University, Benha 13518. Egypt

ABSTRACT: Urban sewer infrastructure is under increasing strain from population growth, network expansion, and aging. Consequently, pipe failures endanger public health and environmental safety. Traditional infrastructure management primarily relies on reactive maintenance. However, recent advances in predictive modeling have enabled proactive approaches to condition assessment. Most models assess current pipe conditions instead of predicting failure time, creating a significant gap in proactive maintenance planning. To address this gap, this preliminary study investigates sewer failure time prediction using statistical analysis and machine learning. This study uses sewer data from Hong Kong's Drainage Services Department, supplemented with parameters from open-source databases. Statistical analysis revealed distinct bimodal failure patterns, with peaks occurring at 30-40 years and 57-60 years for concrete pipes, and at 30-40 years and 52-60 years for vitrified clay pipes. Regional analysis further identified failure patterns across Hong Kong's major districts: Hong Kong Island showed peaks at 55-60 years, the Islands district displayed multiple dispersed peaks, the New Territories exhibited a concentrated distribution peaking at 30-40 years, while Kowloon demonstrated a gradual increase peaking at 55-60 years. The study also examined and compared two advanced machine learning models, namely Random Forest and XGBoost, for failure time prediction. Key challenges influencing prediction accuracy were identified, including complex failure mechanisms, feature engineering constraints, and issues with historical data. This study provides a foundational framework for sewer failure time prediction and highlights key methodological challenges requiring resolution to improve prediction accuracy. These insights advance the understanding of infrastructure deterioration patterns and offer a roadmap for developing advanced predictive models, ultimately leading to more efficient, proactive maintenance strategies for urban sewer systems.

Keywords: Failure Analysis; Artificial Intelligence; Remaining Service Life; Reliability Analysis

1. INTRODUCTION

With the dual pressures of rapid urbanization and aging infrastructure, the risk of failure in urban drainage pipelines is becoming increasingly prominent (Daulat et al. 2024). When pipelines rupture, collapse, or leak, it can not only have severe impacts on urban life and the environment but may also incur enormous economic and social costs (Salihu et al., 2022). However, traditional pipeline maintenance still relies largely on routine or post-incident inspections, limiting opportunities for proactive intervention and precise management. Moreover, focusing solely on inspection-based grading provides insufficient insight into the deterioration dynamics of pipelines at various stages of their service life (Fenner, 2000). Consequently, how to scientifically predict when pipelines will fail and develop more forward-looking operation and maintenance strategies has become a critical issue in urban infrastructure management.

In light of these challenges in recent years, two major methodological trends: statistical analysis and machine learning have emerged to address failure prediction and remaining service life estimation. Notably, statistical models can capture overall failure trends and properties (Lawless 2011), while machine learning algorithms enable more flexible modeling of complex, high-dimensional data (L’heureux et al. 2017). By leveraging both methods in tandem, this study exploits the strengths of each approach: while statistical techniques offer insights into broader distribution patterns and trends, the high-dimensional and nonlinear modeling capacity of machine learning can deliver more nuanced predictive power. Building on previous research that has established these two separate lines of investigation, this study revisits both perspectives and uses actual measured data provided by Drainage Services Department (DSD) in Hong Kong for case validation and performance evaluation.

Specifically, in this study, both statistical analysis and machine learning approaches are employed to systematically investigate pipeline failure. On the statistical side, actual pipeline failure data are utilized to conduct exploratory analysis of failure-age distributions for typical pipe materials such as concrete and vitrified clay. By applying nonparametric methods like Kernel Density Estimation (KDE), potential bimodal failure patterns and critical risk periods are thoroughly identified, providing a direct reference for more precise preventive maintenance strategies. At the same time, Random Forest (RF) and XGBoost are employed to handle multi-source heterogeneous data for feature extraction and model training, thereby quantitatively assessing their suitability in predicting pipeline failure time. Through comparing training and testing results, the study delves into issues including overfitting, data noise, and limitations in feature engineering.

Accordingly, the main objectives of this research can be summarized as follows:

1.Employ statistical methods to analyze the failure-age distribution characteristics of different pipeline materials (e.g., concrete and vitrified clay), revealing potential bimodal failure patterns and high-risk intervals;

2.Using ensemble methods such as Random Forest (RF) and XGBoost as the core techniques, construct a pipeline remaining service life prediction model based on multidimensional features, and compare its performance on both training and test sets;

3.Investigate the sources of prediction bias and challenges, and propose directions for improvements in data quality, feature engineering, and model architecture.

2. METHODOLOGY

This study employs a systematic research methodology (Figure1) that begins with data acquisition from the Hong Kong Drainage Services Department and integration of open-source data, followed by failure data filtering, data cleaning, and processing, culminating in GIS database integration. The analytical approach follows two parallel tracks: a statistical analysis track incorporating failure age distribution analysis, and an artificial intelligence track utilizing Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) models for in-depth modeling and performance analysis. By combining the results from both statistical and AI analyses, the study provides a comprehensive evaluation of findings, leading to thorough discussions and identification of future research directions and practical recommendations, thus forming a complete research cycle.

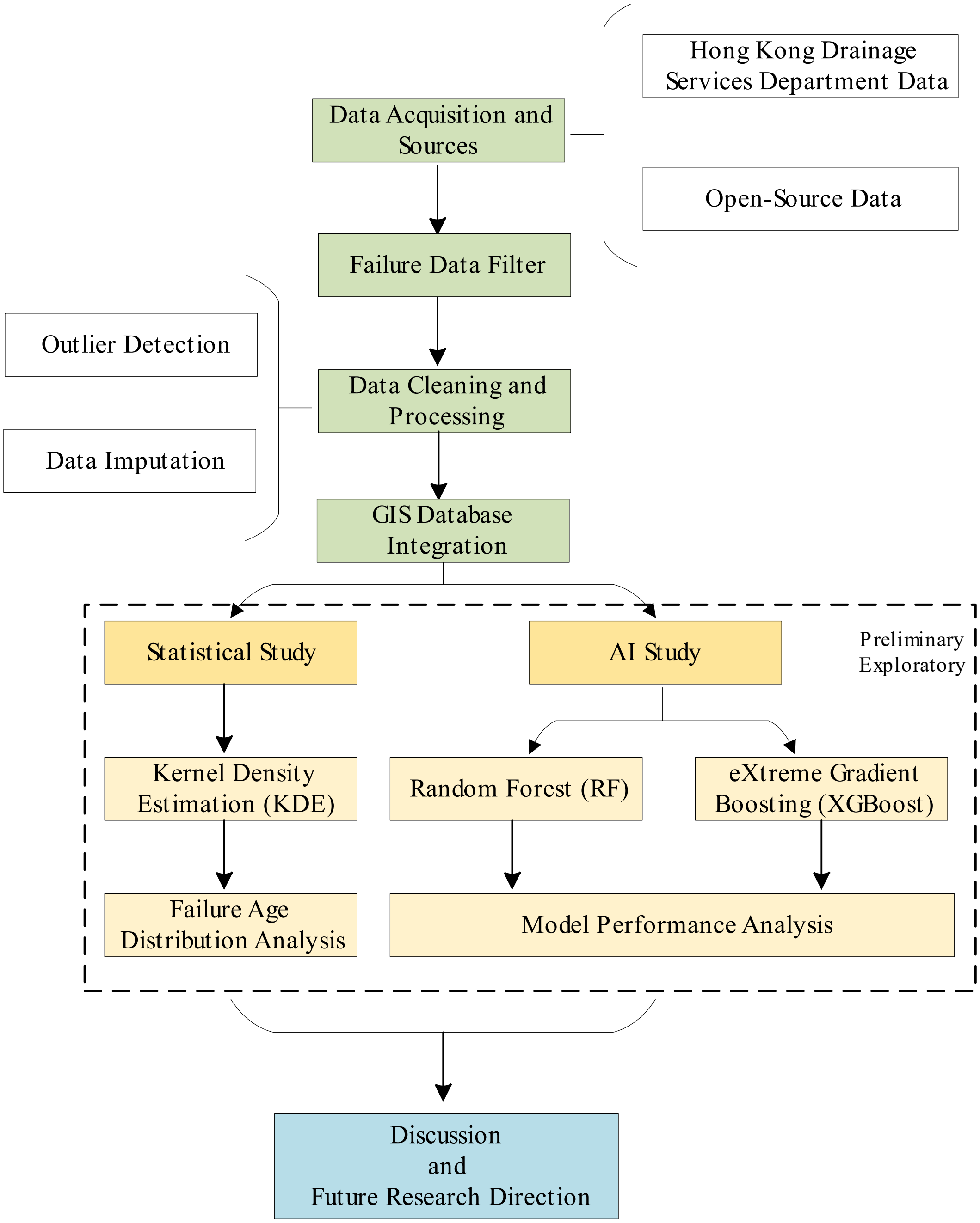

Hong Kong Drainage Services Department Data
Data Acquisition and Sources
Open-Source Data
Failure Data Filter
Outlier Detection
Data Cleaning and Processing
Data Imputation
GIS Database Integration
Statistical Study
AI Study
Preliminary Exploratory
Kernel Density Estimation (KDE)
Random Forest (RF)
eXtreme Gradient Boosting (XGBoost)
Failure Age Distribution Analysis
Model Performance Analysis
Discussion and Future Research Direction

Figure 1: Research flowchart

2.1 Data Collection

The primary dataset detailing internal drainage network conditions was obtained from the Hong Kong Drainage Services Department (DSD). This dataset includes extensive pipeline infrastructure data in GIS format. It covers spatial parameters (district location, layout configuration, connectivity), physical characteristics (diameter, material composition), temporal attributes (installation date, operational age), and CCTV inspection documentation from 2007 to 2021.

To account for environmental influences on pipeline deterioration, the dataset was augmented with external environmental parameters obtained from multiple governmental agencies. Traffic loading data, specifically Annual Average Daily Traffic (AADT), was procured from the Transport Department. Meteorological parameters, including diurnal temperature variations, relative humidity, and precipitation measurements, were extracted from the Hong Kong Observatory's historical meteorological database. Land use classification data were derived from statutory planning documentation and land utilization records maintained by the Planning Department and Lands Department. Geotechnical parameters were obtained from the Geotechnical Engineering Office's geological and ground investigation repository.

2.2 Kernel Density Estimation (KDE)

KDE is a non-parametric method pioneered by Rosenblatt (Rosenblatt 1956) and Parzen (Parzen 1962) for estimating the probability density function (PDF) of a continuous random variable. Unlike parametric approaches, KDE does not presuppose a specific distributional form. Instead, kernels are placed over the data points and summed to generate a smooth density curve (Chen, 2017).

Mathematically, the KDE for a dataset $\{x_1, x_2, \ldots, x_n\}$ can be expressed as:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

where n enotes the total number of data points, h is the bandwidth controlling the smoothness of the density estimate, K is the chosen kernel function (e.g., Gaussian), and x represents the target point at which the density is being estimated.

The selection of the bandwidth h is crucial for balancing the bias-variance trade-off—larger bandwidths produce smoother estimates (potentially with higher bias), while smaller bandwidths capture more detailed structure (but may increase variance). In practice, automated methods such as cross-validation or rule-of-thumb estimators can be utilized to determine an appropriate bandwidth value.

2.3 Random Forest (RF)

Random Forest (RF), introduced by Breiman (Breiman 2001), is an ensemble learning method that constructs multiple decision trees for classification or regression tasks, subsequently aggregating their predictions. By randomly sampling both the training data (i.e., bootstrap sampling) and subsets of features at each split, RF mitigates the tendency of individual trees to overfit and thus reduces overall variance. This approach often leads to higher predictive accuracy compared to a single decision tree.
Formally, let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be the training dataset, where $\mathbf{x}_i$ represents the feature vector and y_i the response (e.g., pipeline condition or failure time). RF draws T bootstrap samples from the original dataset and trains an individual decision tree $h_t(\mathbf{x})$ on each subsample. The final prediction $\hat{y}$ for a new input $\mathbf{x}$ is derived by aggregating the individual tree predictions. For regression tasks, this is typically the average:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(\mathbf{x}) \tag{2}$$

in classification scenarios, it is usually the majority vote:

$$\hat{y} = \mathrm{mode}\big(h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_T(\mathbf{x})\big) \tag{3}$$

Additionally, RF features out-of-bag (OOB) error estimation, allowing performance assessment without a separate validation set (Adelabu et al. 2015). While gradient-boosted tree methods may demonstrate advantages in certain domains (e.g., noisier data, highly sparse features), RF’s versatility in handling heterogeneous data types—together with its robust ensemble structure—makes it particularly suitable for pipeline failure prediction. Its ability to reduce overfitting while still capturing relevant nonlinearities and interactions proves advantageous when modeling complex factors underlying infrastructure deterioration.

2.4 eXtreme Gradient Boosting (XGBoost)

XGBoost, introduced by Chen and Guestrin (Chen and Guestrin 2016), is an advanced implementation of gradient boosting designed for high efficiency, scalability, and predictive accuracy. Similar to other gradient boosting frameworks, XGBoost iteratively trains new decision trees to correct the prediction residuals of previously built models (Elavarasan and Vincent 2020, Wade and Glynn 2020). However, it distinguishes itself through several key innovations, including a second-order (Taylor expansion) approximation of the loss function for more precise optimization, as well as configurable regularization terms to mitigate overfitting (Kavzoglu and Teke 2022).

Formally, let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be the training dataset, where $\mathbf{x}_i$ represents the feature vector and y_i the corresponding target. In the t-th boosting iteration, XGBoost learns a new tree $f_t(\mathbf{x})$ to minimize an objective $\mathcal{L}$ that incorporates both the loss function and a regularization term $\Omega(f_t)$:

$$\mathcal{L} \approx \sum_{i=1}^{N} \left(g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i [f_t(\mathbf{x}_i)]^2 \right) + \Omega(f_t)$$

where g_i and h_i are the first- and second-order derivatives (gradients) of the loss function evaluated at $\mathbf{x}_i$. By employing this second-order approximation and integrating it with a suitable regularization term, XGBoost achieves a compelling balance between model complexity and predictive accuracy.

To further enhance computational efficiency, XGBoost provides both exact and approximate split-finding strategies, adapting well to datasets of varying size and sparsity. It also supports custom loss functions, enabling flexible application in diverse domains (Liu et al. 2024). Together, these features—precise gradient optimization, robust regularization, and scalable split-finding—make XGBoost particularly advantageous for complex tasks such as infrastructure failure prediction, where both performance and interpretability are essential (Nalluri et al. 2020).

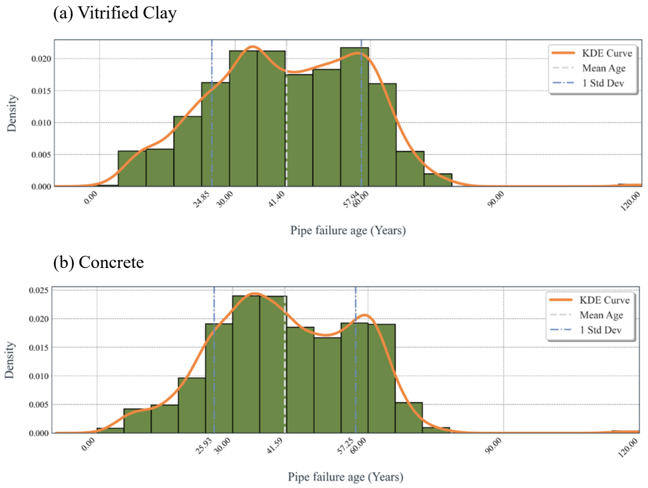
3. RESULT ANALYSIS

This study analyzes the lifespan of pipeline systems in two main parts. First, it employs statistical analysis to identify the lifetime distribution characteristics of pipeline systems under various materials. Second, it develops and validates prediction models to ensure accurate lifespan forecasting.

3.1 Failure Age Distribution Analysis

Figures 2 & 3 integrate normalized histograms and Kernel Density Estimation (KDE) to depict the lifetime distribution characteristics of pipeline systems from two analytical perspectives: material types and geographical regions. The green bars represent the normalized frequencies (frequencies divided by class width) of failure ages, the orange curve illustrates the smoothed probability density trend, the blue dashed lines indicate the boundaries of one standard deviation interval (mean ± standard deviation), and the white dashed line indicates the mean failure age.

Figure 2 demonstrates the failure patterns across different material types. For vitrified clay pipelines show an average failure age of approximately 41.4 years, with a standard deviation interval spanning 24.85 to 57.94 years. These pipes also display bimodal characteristics with peaks in the 30-40 year and 52-60 year intervals, respectively. Similarly, for concrete pipelines, the average failure age is 41.59 years, with a standard deviation interval spanning 25.93 to 57.25 years. The distribution exhibits two distinct failure concentration periods, with a primary peak in the 30-40 year range and a smaller secondary peak in the 57-60 year range.

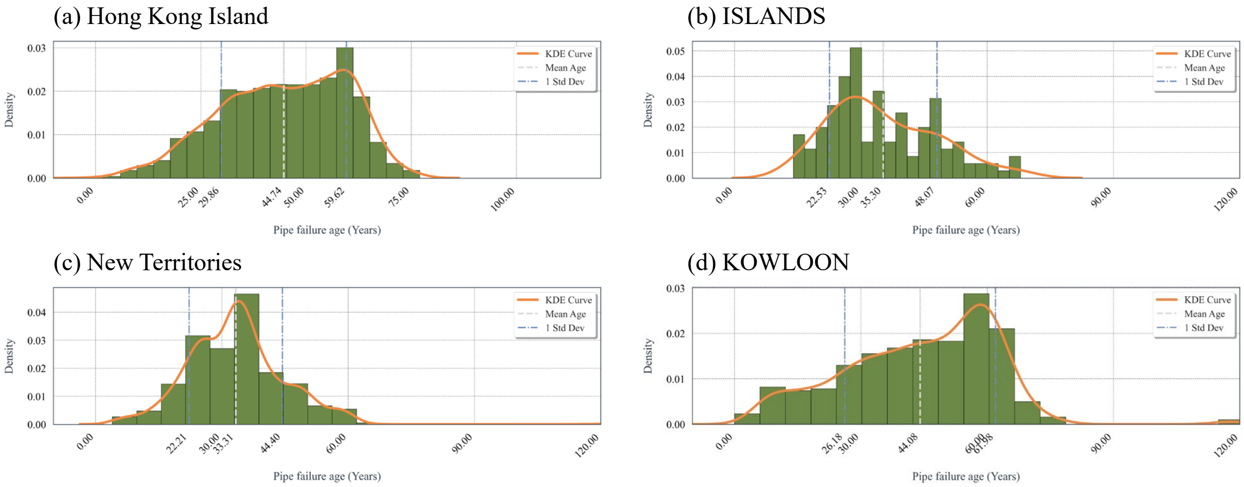


(a) Vitrified Clay

(b) Concrete

Figure 2. Density distribution of failure age by material types

Figure 3 presents the geographical distribution of pipeline failures across Hong Kong's major regions. Hong Kong Island shows a distinctive density pattern with peaks 55-60 years, while the Islands district displays a more dispersed distribution with multiple smaller peaks. The New Territories demonstrates a more concentrated distribution with a pronounced peak around 30-40 years. Kowloon presents a gradually increasing trend with a peak at approximately 55-60 years. These regional variations suggest that local factors such as geological conditions, infrastructure age, and operational demands may significantly influence pipeline lifespans.



(a) Hong Kong Island

(b) ISLANDS

(c) New Territories

(d) KOWLOON

Figure 3. Density distribution of failure age by geographical regions

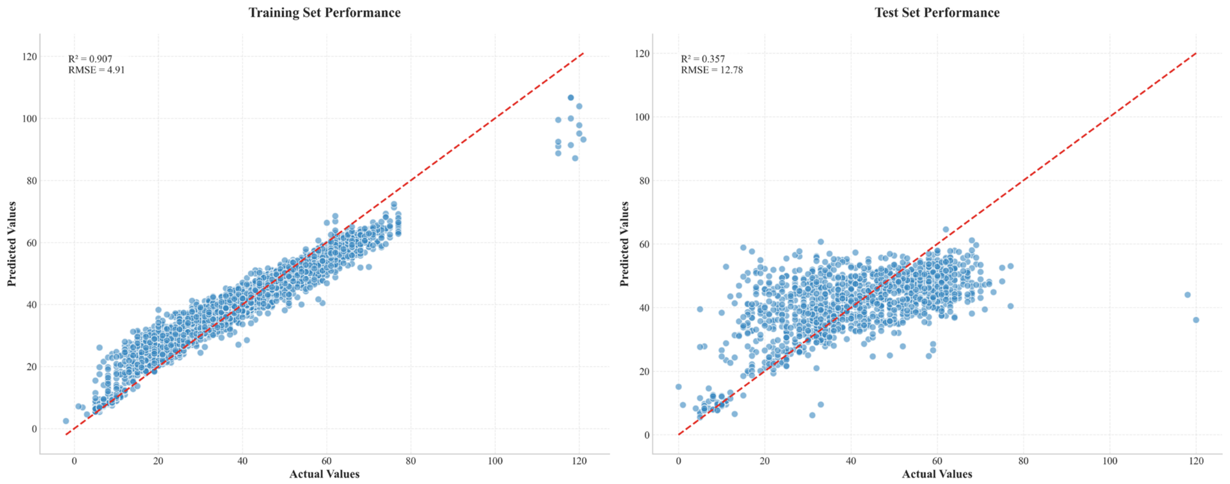
3.2 AI-Based Predictive Modeling Results Analysis

3.2.1 Random Forest Model Performance Analysis

To estimate the lifespan of the pipeline system, a random forest model was developed, utilizing 80% of the dataset for training and the remaining 20% for testing. The model exhibited notable discrepancies between the training and testing phases (Figure 4):

During the training phase, the model exhibited strong fitting performance on the training data. The evaluation metrics indicated an R² value of 0.907 and a Root Mean Square Error (RMSE) of 4.91 years. In the scatter plot comparing predicted versus actual lifespans, data points were closely clustered around the 45-degree ideal prediction line (red dashed line), suggesting that the model effectively captured the distribution characteristics of the training data.

However, when the model was evaluated on the test data, its prediction accuracy declined significantly. The test set evaluation revealed an R² value of 0.357 and an RMSE of 12.78 years. The scatter plot for the test set demonstrated a marked increase in dispersion between predicted and actual lifespans, with numerous data points deviating significantly from the ideal prediction line. This pronounced difference in performance between the training and test sets suggests an overfitting issue in the model, implying that it has overfitted to the features of the training data, thereby compromising its generalization capability.


Training Set Performance

Test Set Performance

Figure 4. Random forest model performance

3.2.2 XGBoost Model Performance Analysis

Employing the same data partitioning scheme as the random forest, the XGBoost model exhibited consistent yet suboptimal performance in predicting pipeline lifespan. The evaluation metrics for the training set revealed an R² value of merely 0.579 and an RMSE of 10.42 years. The scatter plot for the training set suggested a weak correlation between predicted and actual lifespans, with data points exhibiting pronounced dispersion (Figure 5).

The performance on the test set was comparable to that on the training set, yielding an R² value of 0.350 and an RMSE of 12.84 years. While the consistency between test and training set performance suggests that the model avoided overfitting, its overall prediction accuracy fell short of the requirements for engineering applications.

Compared to the random forest model, XGBoost demonstrated superior capability in avoiding overfitting, attributed to its built-in regularization mechanism. However, both models exhibited prediction errors nearing 13 years on the test set (with RMSEs of 12.84 and 12.78 years, respectively), highlighting substantial shortcomings in the current modeling approach, necessitating enhancements in data quality, feature engineering, and model architecture.

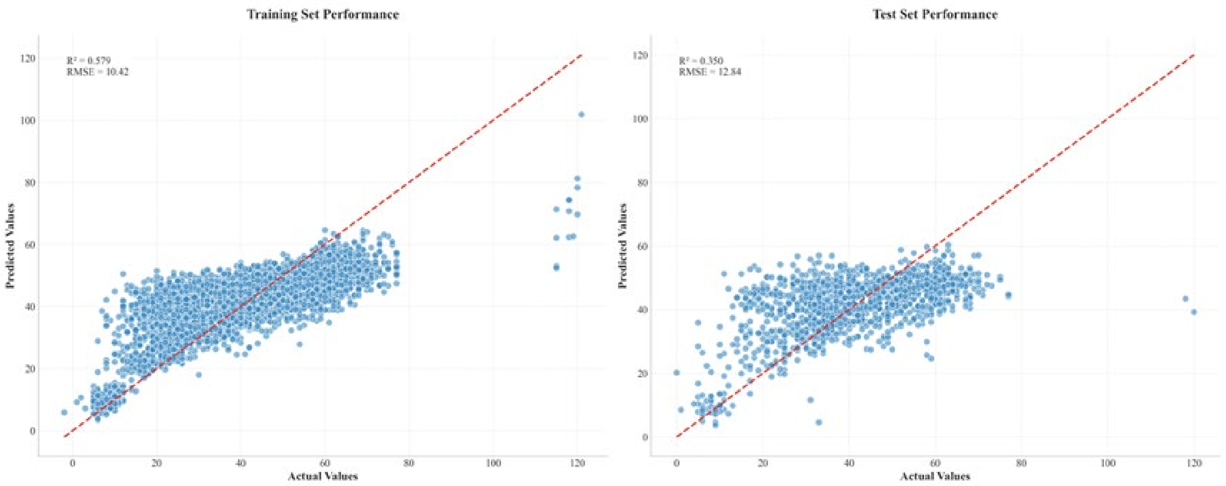


Figure 5. XGBoost Model Performance

3.2.3 Comparison of RF and XGBoost Result

Table 1 summarizes the primary performance metrics of RF and XGBoost. Overall, RF exhibits strong overfitting, as evidenced by a high R² of 0.907 and low RMSE of 4.91 years on the training set, but a significant drop to R² = 0.357 and RMSE = 12.78 years on the test set. By contrast, XGBoost's built-in regularization appears to mitigate overfitting, maintaining comparable (though still suboptimal) performance across training (R² = 0.579, RMSE = 10.42 years) and test sets (R² = 0.350, RMSE = 12.84 years). Both models require further refinement to achieve engineering-grade accuracy, notably through improved data preprocessing, feature engineering, and hyperparameter optimization to better capture the inherent complexities of pipeline deterioration.

Table 1: Performance metrics of RF and XGBoost

Model	Dataset	R²	RMSE (years)
Random Forest (RF)	Training	0.907	4.91
	Test	0.357	12.78
XGBoost	Training	0.579	10.42
	Test	0.350	12.84

4. DISCUSSION

During the model development and validation phase, the predictive performance fell short of expectations, highlighting the challenges of accurately predicting sewer pipeline failure lifespans. A detailed analysis identified several key factors contributing to this predictive bias.

Firstly, the conventional definition of failure age may be flawed when viewing the pipeline as a single system; it is more appropriate to consider it as two distinct entities separated by major repairs. Each major repair serves as a clear demarcation point, effectively splitting the pipeline into two distinct entities.

Secondly, enhancing data quality and performing additional cleaning procedures is essential. The current dataset likely still includes outliers, all of which significantly impair predictive accuracy. Stricter data quality control mechanisms, such as systematic anomaly detection, effective missing value handling, and data consistency checks, are strongly recommended.

Thirdly, due to the complexity of the problem, starting with simplified models is advisable. Initial modeling can focus on data from unrepaired pipelines, thereby avoiding the effects of repairs and establishing a baseline model. Gradually increasing complexity allows for better understanding of influencing factors and guides future research directions.

Fourthly, feature engineering requires further optimization. Current model features may inadequately capture critical information about pipeline failures. For instance, time series feature complexity may be underrepresented, while existing variables may exhibit redundancy or high correlation. Feature selection using importance and correlation analysis is recommended, alongside introducing new features.

Fifthly, the strategy for splitting training and test sets warrants careful consideration. Due to the temporal correlation in pipeline data, random splitting may fail to accurately assess predictive performance. The varying performance patterns between RF and XGBoost models provide interesting insights for model selection. RF demonstrates strong pattern recognition capabilities (training $R^2 = 0.907$), while XGBoost shows more consistent performance across different data sets (training $R^2 = 0.579$, test $R^2 = 0.350$), likely due to its built-in regularization mechanisms. These distinctive characteristics suggest that future implementations could benefit from combining the strengths of different algorithms, alongside more sophisticated cross-validation strategies and advanced model optimization techniques.

Moreover, the interpretation of uncertainties in AI-based infrastructure lifespan prediction requires careful consideration. These uncertainties emerge from multiple sources: inherent variability in physical deterioration processes, data quality issues, and model-related uncertainties. While ensemble methods like RF and XGBoost offer certain mechanisms for handling variability in predictions, their current performance suggests that these models still struggle to fully capture and interpret the complex uncertainties in infrastructure lifespan prediction. To better address these uncertainties in practical applications, future implementations should consider incorporating probabilistic outputs rather than point estimates, enabling more informed maintenance scheduling decisions while acknowledging the inherent uncertainties in lifespan prediction.

Finally, this study predominantly focuses on analyzing and predicting data from failed pipelines, utilizing mainly failure time information from CCTV inspections. This focus somewhat restricts the comprehensiveness of studies and reduces the accuracy of predictions. Future research should incorporate both pipelines still operating normally (survival data) and multi-dimensional CCTV information including condition ratings evolution and defect patterns. This expanded scope would offer a more comprehensive perspective on the pipeline lifecycle, mitigate sample selection bias, and allow for a more accurate evaluation of pipeline deterioration mechanisms. Such integration facilitates the use of advanced analytical methods, which can further improve the predictive performance and practical utility of models.

Addressing these key issues is anticipated to significantly enhance the model's predictive performance, offering reliable decision support for sewer pipeline predictive maintenance. These recommendations also outline clear pathways for future research.

5. CONCLUSION

This study represents a pioneering effort in predicting specific failure times of sewer pipelines through the integration of statistical analysis and machine learning, revealing distinct lifetime distribution patterns for different pipe materials, with both concrete and vitrified clay pipes showing characteristic bimodal failure distributions around 30–40 years and 52–60 years of service life. These findings offer valuable guidance for maintenance scheduling and resource allocation, while the implementation of machine learning models, specifically Random Forest and XGBoost, demonstrates the technical feasibility of pipeline failure prediction despite current accuracy limitations. This underscores both the potential and challenges in applying machine learning to infrastructure maintenance prediction. Furthermore, the study identifies critical limitations in current data collection and processing methods, particularly in handling repair history and integrating environmental factors, thereby providing clear directions for improved data protocols and database design. Although the present prediction accuracy requires further enhancement for practical engineering use, this research lays important groundwork for developing more effective predictive maintenance strategies in urban sewer systems and makes a significant contribution to the evolving field of infrastructure asset management.

ACKNOWLEDGMENTS

This work was supported by the Research Grants Council (RGC)-General Research Fund (GRF) under grant number 15209022. The authors would also like to thank the Hong Kong Drainage Services Department (DSD) for the data support and the Hong Kong Utility Training Institute (UTI) for providing the Hong Kong Conduit Condition Evaluation Codes technical reference materials and for their valuable industry communication and guidance.

REFERENCES

Adelabu, S., Mutanga, O., and Adam, E. 2015. Testing the Reliability and Stability of the Internal Accuracy Assessment of Random Forest for Classifying Tree Defoliation Levels Using Different Validation Methods. Geocarto International, 30(7): 810-821.

Breiman, L. 2001. Random Forests. Machine Learning, 45: 5-32.

Chen, T. and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 785-794.

Chen, Y.C. 2017. A Tutorial on Kernel Density Estimation and Recent Advances. Biostatistics & Epidemiology, 1: 161-187.

Daulat, S., Rokstad, M.M., Klein-Paste, A., Langeveld, J., and Tscheikner-Gratl, F. 2024. Challenges of Integrated Multi-Infrastructure Asset Management: A Review of Pavement, Sewer, and Water Distribution Networks. Structure and Infrastructure Engineering, 20(4): 546-565.

Elavarasan, D. and Vincent, D.R. 2020. Reinforced XGBoost Machine Learning Model for Sustainable Intelligent Agrarian Applications. Journal of Intelligent & Fuzzy Systems, 39(5): 7605-7620.

Fenner, R.A. 2000. Approaches to Sewer Maintenance: A Review. Urban Water, 2: 343-356.

Kamyab, H., Khademi, T., Chelliapan, S., SaberiKamarposhti, M., Rezania, S., Yusuf, M., and Ahn, Y. 2023. The Latest Innovative Avenues for the Utilization of Artificial Intelligence and Big Data Analytics in Water Resource Management. Results in Engineering, 17: 101566.

Kavzoglu, T. and Teke, A. 2022. Advanced Hyperparameter Optimization for Improved Spatial Prediction of Shallow Landslides Using Extreme Gradient Boosting (XGBoost). Bulletin of Engineering Geology and the Environment, 81(5): 201.

L’heureux, A., Grolinger, K., Elyamany, H.F., and Capretz, M.A. 2017. Machine Learning with Big Data: Challenges and Approaches. IEEE Access, 5: 7776-7797.

Lawless, J.F. 2011. Statistical Models and Methods for Lifetime Data, John Wiley & Sons, Hoboken, NJ, USA.

Liu, Y., Kang, Y., Zou, T., Pu, Y., He, Y., Ye, X., and Yang, Q. 2024. Vertical Federated Learning: Concepts, Advances, and Challenges. IEEE Transactions on Knowledge and Data Engineering, 36(7): 3615-3634.

Nalluri, M., Pentela, M., and Eluri, N.R. 2020. A Scalable Tree Boosting System: XG Boost. Int. J. Res. Stud. Sci. Eng. Technol, 7(12): 36-51.

Parzen, E. 1962. On Estimation of a Probability Density Function and Mode. The Annals of Mathematical Statistics, 33: 1065-1076.

Rosenblatt, M. 1956. Remarks on Some Nonparametric Estimates of a Density Function. The Annals of Mathematical Statistics, 27: 832-837.

Salihu, C., Hussein, M., Mohandes, S.R., Zayed, T., 2022. Towards a comprehensive review of the deterioration factors and modeling for sewer pipelines: A hybrid of bibliometric, scientometric, and meta-analysis approach. Journal of Cleaner Production 351, 131460.

Wade, C. and Glynn, K. 2020. Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform Accessible Machine Learning and Extreme Gradient Boosting with Python, Packt Publishing Ltd, Birmingham, UK.