UDK/*UDC*: 004.85:628.1(1-21)
Izvirni znanstveni članek – *Original scientific paper*
DOI: 10.15292/acta.hydro.2025.05

Prejeto/*Received:* 29.06.2025
Sprejeto/*Accepted:* 20.07.2025
Objavljeno na spletu/*Published online:* 18.08.2025

PREDICTING WATER DISTRIBUTION PIPE FAILURES USING MACHINE LEARNING AND CROSS-INFRASTRUCTURE DATA

NAPOVED OKVAR VODOVODNIH CEVI S STROJNIM UČENJEM IN PODATKI O SOSEDNJI INFRASTRUKTURI

Daniel Kozelj[1], David Abert Fernández[2]

[1]Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova cesta 2, 1000 Ljubljana, Slovenia

[2]LEQUIA, Institute of the Environment, University of Girona, Maria Aurèlia Capmany 69, E-17003, Girona, Catalonia, Spain

Abstract

Water pipeline failures in urban networks are a significant source of non-revenue water, service disruptions, and high maintenance costs. This study develops a machine learning model to predict pipeline failure probabilities and inform risk-based maintenance strategies. Trained on real-world assets and geospatial data from 2010 to 2025, the model incorporates standard pipe attributes – such as material, age, diameter, network type, and maintenance history – alongside spatially derived indicators of the surrounding infrastructure. Notably, it quantifies the predictive impact of adjacent infrastructure systems, including electricity grids, gas pipelines, district heating, sewage systems, and roads, utilizing spatial buffering and overlay techniques. Several of these cross-utility features, particularly road category, electricity voltage, and sewer type, showed meaningful predictive importance, reflecting their indirect but consistent influence on the risk of pipe failure. The ML model, built with the XGBoost algorithm and validated through stratified K-fold cross-validation, achieved high performance (ROC AUC: 0.9102, recall: 0.7750, accuracy: 0.8750). Despite lower precision due to class imbalance, the F1 score (0.2261) and LogLoss (0.2500) confirm its reliability. This study introduces a novel, spatially enriched approach to failure prediction, advancing urban infrastructure management through context-aware, data-driven insights.

Keywords: Water distribution systems, Pipe failure prediction, Machine learning, XGBoost, Spatial analysis, condition assessment.

Izvleček

Okvare vodovodnih cevi v urbanih omrežjih so pomemben vzrok komercialnih izgub zaradi neobračunane vode, motenj v oskrbi in visokih stroškov vzdrževanja. Ta študija razvija model strojnega učenja za napoved verjetnosti okvar cevovodov in podporo strategijam vzdrževanja, temelječim na tveganju. Model, izurjen na

[1] Stik / *Correspondence*: daniel.kozelj@fgg.uni-lj.si

podatkih o infrastrukturi in geoprostorskih podatkih iz obdobja 2010–2025, vključuje standardne lastnosti cevi – kot so material, starost, premer, vrsta omrežja in zgodovina vzdrževanja – ter prostorsko izpeljane kazalnike infrastrukture v neposredni soseščini. Posebej pomembno je, da model kvantificira napovedno vrednost sosednjih infrastrukturnih sistemov, vključno z električnim omrežjem, plinovodi, daljinskim ogrevanjem, kanalizacijo in cestnim omrežjem, z uporabo prostorske analize soseščine in tehnik prekrivanja. Več teh medinfrastrukturnih značilnosti, zlasti kategorija ceste, napetost električnega omrežja in tip kanalizacije, je pokazalo pomemben napovedni vpliv, kar odraža njihovo posredno, a dosledno povezavo z verjetnostjo okvare cevi. Model strojnega učenja, zasnovan z algoritmom XGBoost in validiran s slojevitim navzkrižnim preverjanjem (K-fold), je dosegel visoko zmogljivost (ROC AUC: 0,9102; priklic: 0,7750; natančnost: 0,8750). Kljub nižji preciznosti zaradi neravnovesja razredov rezultat F1 (0,2261) in LogLoss (0,2500) potrjujeta njegovo zanesljivost. Raziskava predstavlja nov, s prostorskimi podatki obogaten pristop k napovedovanju okvar in prispeva k naprednemu, na podatkih temelječem upravljanju urbane infrastrukture.

Ključne besede: vodovodni sistemi, napovedovanje okvar cevovodov, strojno učenje, XGBoost, prostorska analiza, ocenjevanje stanja.

1. Introduction

Water supply systems (WSS) are vital components of public infrastructure, providing clean drinking water to residents, a key factor for human well-being and societal prosperity. These systems, which include wells, reservoirs, pumps, and pipelines, are becoming increasingly challenging to manage due to, among other things, the need to maintain supplies amid changing urban settlements, infrastructure investment, and increased regulatory pressure (Ganjidoost et al., 2022). In addition to water availability and quality, the deterioration of WSS pipelines is a growing concern. Environmental conditions, material ageing, and operational stress contribute to this deterioration, leading to water losses, increased maintenance costs, and reduced service quality (Misiunas, 2006). The challenges faced by water utilities, particularly those related to incomplete or outdated infrastructure data, further complicate proactive asset management (Phan et al., 2019; Ganjidoost et al., 2022).

One of the most important technical indicators to follow in WSS performance is non-revenue water (hereafter referred to as NRW), and it is one of the most significant challenges for water utility operators. NRW is the difference between the abstracted amount of water and the amount of water sold. NRW is composed of three primary components: real losses, apparent losses, and unbilled consumption. The proportion of NRW varies considerably from country to country, depending on the state of the water supply networks, and ranges from 8% to 57% in Europe (European Commission, 2025). In Slovenia, data on NRW have been available since 2010, following the introduction of the national reporting system, the Information System of Public Environmental Protection Services (IJSVO) (MOPE, 2025). At the national level, NRW is increasing and accounts for 29.4%, according to the latest data from 2023 (SURS, 2025). While water losses are often viewed through an economic lens – primarily due to the costs associated with treatment and pumping – they also pose significant challenges for water system management and public health. Addressing water losses effectively requires a comprehensive strategy developed by the system operator and approved by the owner to find an appropriate balance between repair, rehabilitation, and replacement of deteriorated pipelines.

In recent decades, utilities and researchers have developed various models to assess pipeline conditions and predict failure using indirect indicators. These tools are crucial for prioritizing rehabilitation efforts when resources are limited and complete pipe replacement is not economically feasible (Large et al., 2015; Le Gat et al., 2023). However, predicting pipe deterioration is not solely reliant on one feature. The presence of older pipes that still function adequately complicates the use of age as the sole predictor, highlighting the need for multifactorial assessment approaches. Condition assessment methods and risk-based planning are

increasingly integrating GIS-based spatial modelling and data analysis. These tools are instrumental in understanding deterioration patterns through visualizing the spatial distribution of deterioration factors and optimizing investment strategies by identifying the most critical areas for rehabilitation (Mackey et al., 2014; Ganjidoost et al., 2022).

Various models have been developed to assist in pipeline renewal decisions, each with distinct data needs, methodologies, and optimization goals (Liu et al., 2012; Bakhtawar et al., 2025). A key challenge is to determine which pipe segments should be prioritized for rehabilitation. According to Bakhtawar et al. (2025), pipe condition assessment approaches typically fall into two categories: physical or statistical. Physical models rely on structural and environmental inputs, while statistical models leverage historical data to predict failures. These include deterministic models, which are based on fixed input values from laboratory tests or standards, and probabilistic models, which estimate the probability of failure or remaining useful life (Rajani & Kleiner, 2001; Rezaei et al., 2015; Bakhtawar et al., 2025). Unlike deterministic models, probabilistic methods account for uncertainty, offering a range of likely outcomes rather than a single estimate. Due to the limited and costly data requirements, statistical or empirical models are often used as an alternative. These rely on historical default data to predict future trends (Kleiner & Rajani, 2001) and can be deterministic or probabilistic. Deterministic approaches group pipes by similar characteristics and model failure patterns based on age and history, while probabilistic models estimate the lifetime or probability of failure (Misiunas, 2006).

Over the last few decades, machine learning (ML) approaches have emerged as efficient tools, offering improved prediction power and adaptability in handling complex, nonlinear interdependencies between variables. Recent advancements in ML have significantly improved the prediction of pipeline failures and leak detection (Latifi et al., 2024). Standard statistical models, though sometimes helpful, suffer from data insufficiency and a lack of adaptability as they struggle with complex, nonlinear interdependencies between variables (Cabral et al., 2023; Mohammadagha et al., 2025).

In comparison, ML algorithms such as gradient boosting (e.g. XGBoost) have demonstrated significant superiority in terms of failure prediction, particularly with imbalanced datasets. In a comparative study, Asadi (2024) found that XGBoost achieved a recall rate of 0.795 – substantially higher than the 0.683 recorded by logistic regression – highlighting its effectiveness in identifying vulnerable pipeline segments. Warad et al. (2024) reported that their ensemble learning model, optimized via Bayesian tuning, achieved an exceptionally low root mean squared error (RMSE) of 0.0023, indicating high prediction accuracy.

Despite their potential, the development and widespread application of ML models are still limited, primarily due to data availability, the variability of degradation mechanisms, the need for user-specific calibration, and a lack of data integration (Jafar et al., 2010; Le Gat et al., 2023). While some of the limitations are inherent to the model's application, there are still untapped possibilities in integrating data to enrich condition assessment and strategic planning for improved and proactive asset management. We have observed that most of the published works have not considered the possible effects of adjacent utility infrastructure in urban environments. This therefore raises the question of whether integrating nearby infrastructure systems influences the probability of pipe failure. Such a multidimensional approach could prove beneficial in improving the prediction accuracy of pipe failure algorithms rather than only using WSS-specific attributes. These improvements could then be used in another effective and widespread strategy for reducing water losses, i.e. the creation of permanent or temporary District Metered Areas (DMAs). DMAs enable the targeted monitoring of pressure and flow in isolated sections of water distribution networks (WDN) and enable the control of water losses within these zones. The probability of pipe failure can play a vital role in shaping DMA design. These probabilities can be used as input variables to steer and influence the size and shape of WDN segmentation, which can

lead to an optimized structure and cost-effectiveness of DMAs. As Kozelj et al. (2017) emphasize, spectral graph partitioning integrated with hydraulic data can optimize network segmentation into hydraulically functional DMAs, facilitating quicker detection of anomalies. Zevnik et al. (2019) further demonstrate that incorporating failure probability and topological, hydraulic, and cost criteria into the design process yields more efficient and objective DMA configurations. Since water utilities are responsible for maintaining the public water infrastructure and thus must ensure a timely, safe, and sufficient supply of drinking water, there is and will be a growing need for models that assess the likelihood of pipeline failure.

This study aims to address this need by utilising indirect assessment methods, such as predictive modelling, to identify the primary factors that influence pipeline deterioration. By integrating historical failure data, detailed pipe characteristics, and publicly available cadastral data on other urban infrastructure, we could observe influences that can improve the effectiveness of ML applications for the proactive management of water network risks. The goal is to develop a model that can evaluate the condition of pipelines in a specified area and pinpoint the most critical sections that require interventions (repair, rehabilitation, or replacement). These findings will not only support a more targeted and cost-effective pipeline renewal strategy but also have the potential to significantly improve water supply management based on asset management principles.

2. Methodology

2.1 Ljubljana water supply system

The presented methodology was developed and evaluated using data from the Ljubljana WSS (Ljubljana, Slovenia). The Ljubljana WSS supplies drinking water to more than 355,000 registered residents in the city and its surroundings. In 2024, the revenue water amounted to 20,212,381 cubic meters, while a total of 29,168,618 cubic meters was abstracted from water sources. To ensure the provision of drinking water, the system consumed approximately 11.7 million kilowatt-hours (kWh) of electricity during the year (VOKAS, 2025). The annual real water losses within the system were estimated at 22.8%; NRW represented 28.2% of the total water volume for the year 2024. The yearly occurrence of pipe failures from 2010 to 2024 amounts to 2,281 records, ranging from a minimum of 64 to a maximum of 229, with an average total of 164 annual pipe failures.

Our research considered the cadastral base of the Ljubljana WSS, which comprised 52,605 records of pipe sections, totaling 1,190 kilometers of the distribution network. Pipe sections consisted of different materials such as asbestos cement (AC; 3.1%), reinforced glass fiber plastics (GRP; 0.5%), steel (JE; 2.2%), cast iron (LZ; 21.2%), ductile iron (NL; 43.0%), polyethylene (PE; 16.0%), polyvinyl chloride (PVC; 13.7%), and some unknown (NZ; 0.3%). The pipe diameters of the Ljubljana WSS consist of pipe sections: $d \leq 50$ mm (1.1%), $50 < d \leq 80$ mm (9.6%), $80 < d \leq 100$ mm (19.0%), $100 < d \leq 200$ mm (47.9%), and $d > 200$ mm (22.6%). From this pipe diameter distribution, we obtain a classification of primary (28%) and secondary (72%) network types.

2.2 XGBoost modelling

To develop a predictive ML model of pipe failure probabilities, we focused our research towards the well-established XGBoost algorithm (Chen & Guestrin, 2016), which can accurately predict water loss resulting from pipe failures in a city's water supply systems (Asadi, 2024; Warad et al., 2024). To predict pipeline failure risk in a real-world WSS, two primary datasets were integrated. The first originated from the national utility cadastre managed by GURS (2025) and contained detailed spatial and attribute data for 52,605 individual pipe segments across the municipal water network. This dataset included intrinsic pipe characteristics such as material, diameter, function, and year of installation, which are often identified as influential for pipe failures (Karadirek et al., 2024; Bakhtawar et al., 2025). The second dataset, provided by the municipal utility operator VOKA Snaga Ljubljana, comprised a register of 2,281 documented pipe bursts, including the age of the pipe at the time of failure. These historical failure events were spatially

mapped to their corresponding pipe segments, allowing for the construction of a binary target variable indicating the presence (True) or absence (False) of leaks for each segment.

To enhance predictive capability, contextual spatial features were incorporated by integrating data on nearby urban infrastructures, including roads, electricity grids, sewerage systems, telecommunications cables, district heating networks, natural gas pipelines, and rail lines – sourced from the national cadastre. Proximity measures were derived by spatially intersecting pipe segments with buffered zones around these infrastructure types, thus enriching the dataset with localized environmental and operational attributes that potentially influence pipe degradation and failure risk.

The pipeline failure prediction was framed as a binary classification task. The model's target variable indicated whether a given pipe segment had experienced a documented leak based on historical records. The XGBoost algorithm (Chen & Guestrin, 2016), a gradient-boosting framework utilising ensembles of decision trees, was selected due to its superior performance on structured tabular data and its native support for categorical variables (Chen et al., 2023), which eliminated the need for manual encoding and streamlined preprocessing. Since leak events represent a small fraction of all pipe segments, the *scale_pos_weight* parameter in XGBoost was empirically set to 50 to address class imbalance. This adjustment increased the influence of rare leak instances during training, prioritizing sensitivity to leaks over false positives. Such a trade-off is acceptable in infrastructure risk modelling, where missing actual leaks can have significant consequences.

Hyperparameter optimization was conducted using Optuna (Akiba et al., 2019), which efficiently explored parameter spaces to identify optimal values for learning rate, tree depth, and regularization terms, thereby enhancing both predictive accuracy and generalizability.

A stratified five-fold cross-validation procedure was employed to validate the model, ensuring that each fold contained a representative proportion of leak and non-leak samples. In each of the five iterations, 80% of the data was used for training and 20% for validation. This approach generated out-of-fold predictions for all pipe segments, allowing comprehensive risk estimates without reusing training data or introducing bias. The balanced distribution of leak cases across folds mitigated sampling bias and improved model robustness. The procedure is schematically depicted in Figure 1.

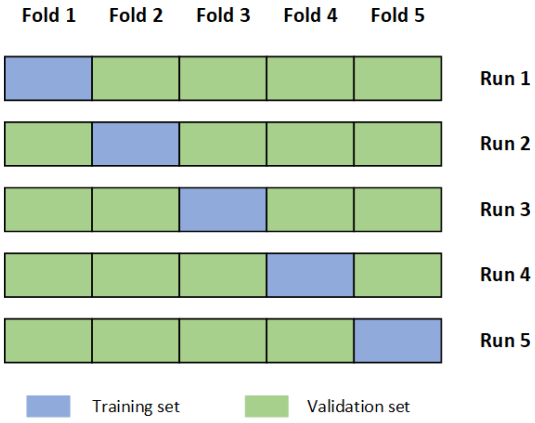

Figure 1: *5-fold stratified cross-validation of dataset (authors' visualization).*

Slika 1: *5-kratna slojevita navzkrižna validacija podatkovnega nabora (vizualizacija avtorjev).*

Feature importance analysis was conducted using XGBoost's built-in metrics: gain, relative frequency, and cover. These metrics quantify each variable's contribution to the model based on tree splits. Specifically, gain measures the improvement in split quality contributed by a feature across all tree splits, relative frequency (also known as weight) indicates how often a feature is used in tree splits, and cover represents the number of samples affected by those splits. To standardize comparisons, we expressed frequency as a proportion – each feature's count divided by the total number of tree splits. This analysis highlighted the most influential intrinsic pipe attributes and contextual spatial features, offering valuable insights for asset management and pipeline condition assessment.

This integrative, data-driven approach combining historical failure data, detailed pipe characteristics, and urban infrastructure context demonstrates the effective application of machine learning for the

proactive management of water network risk. By using XGBoost, the ML model predicts not only leak likelihoods but also estimates the relative contribution of each input feature by leveraging the model's inherent ability to compute metrics on feature importance.

2.3 Evaluation metrics

The model was evaluated using standard classification metrics implemented via the scikit-learn library (Pedregosa et al., 2011). Accuracy quantifies the ratio of correct predictions – both positive and negative – to the total number of predictions:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1),$$

where: TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives.

Precision measures how many of the instances predicted as positive were positive, indicating the model's effectiveness in minimizing false alarms:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2).$$

Recall evaluates the proportion of true positive cases correctly identified by the model, serving as an indicator of detection sensitivity:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3).$$

The F1 score balances precision and recall by calculating their harmonic mean, which is especially useful when dealing with class imbalance:

$$\text{F1 Score} = \frac{2TP}{2TP+FP+FN} \quad (4).$$

ROC AUC (Receiver Operating Characteristic – Area Under the Curve) quantifies how well the model distinguishes between classes across all classification thresholds by analyzing the trade-off between true positive and false positive rates. It is a metric used to evaluate the performance of binary classification models, especially for imbalanced datasets.

Logarithmic loss (Logloss) evaluates how well predicted probabilities align with actual labels, heavily penalizing overconfident misclassifications:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \quad (5)$$

where: y_i: Actual label (0 or 1), p_i: Predicted probability for class 1, N: Number of samples.

3 Results and discussion

For each pipeline segment in the dataset, the developed XGBoost model generated a continuous probability score (0 to 1) for that segment, reflecting the estimated likelihood of a failure event. These probability levels were directly derived from validation subsets used during five-fold stratified cross-validation, which implies that all predictions were generated on validation data not included in the training, ensuring the independence of the model's assessments. This cross-validation strategy reduced the risk of overfitting and allowed for a more reliable evaluation of the model's predictive ability across the entire network.

The probability of output results was integrated within a GIS framework (QGIS.org, 2025), where the probabilities could be visualized and ranked into four ordered risk classes, enhancing interpretability and decision-making at the operational level (Figure 2). The classification thresholds were: (0.0–0.2) for segments of low risk; (0.2–0.4) for medium-low risk segments; (0.4–0.6) for segments of medium risk; (0.6–0.8) for medium-high risk segments; and (0.8–1.0) for segments of high risk or virtual certainty of default. This ordered classification enabled a clear presentation of pipeline segments at risk, categorized by their respective levels of risk. This risk-based classification allows utility operators to target high-risk segments for inspection or preventive maintenance.

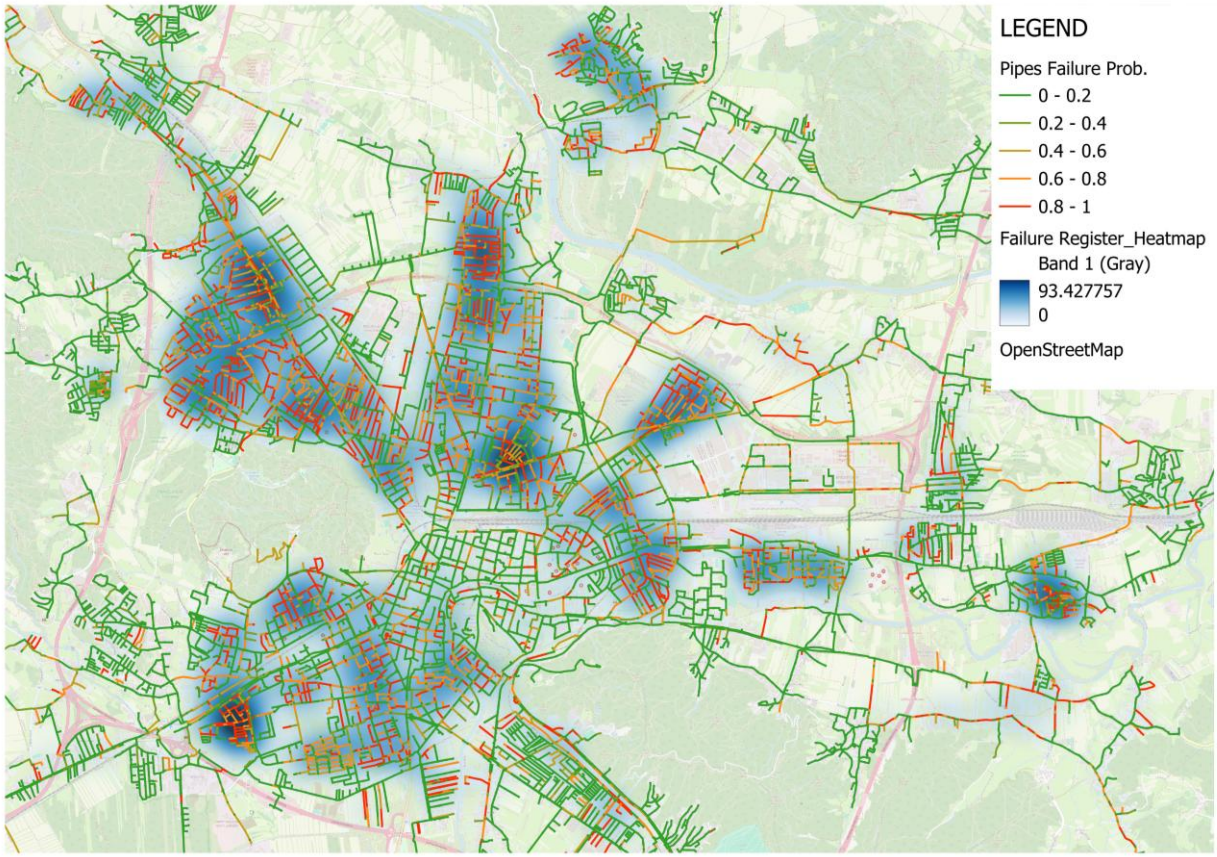

Figure 2: *Pipe failure probabilities (leakage) and heatmap the spatial distribution of historical failures (authors' analysis; Data source: GURS, 2025; VOKAS, 2025).*

Slika 2*: Verjetnost okvar cevovodov (puščanje) in toplotna karta prostorske porazdelitve zgodovinskih okvar (analiza avtorjev; vir podatkov: GURS, 2025; VOKAS, 2025).*

Table 1 presents the model's explainability through three standard feature importance metrics derived from the ensemble structure: gain, relative frequency, and cover. Gain quantifies the total improvement in the model's objective function attributable to splits involving each feature, serving as a direct measure of predictive contribution. Relative frequency captured how often each feature was used to split nodes across all trees, normalized by the total of 101,222 splits, enabling standardized comparison across variables. Cover reflected the cumulative number of training instances affected by those splits, offering insight into the feature's impact across the dataset. Together, these metrics provided a comprehensive view of how each feature contributed to the model's learning and decision-making processes.

As anticipated, pipe-related features – most notably *WS_Material* – contributed the most to gain (~103.8), indicating a strong impact on model decision paths (Table 1). Despite a moderate relative frequency (~5.7%), the significant gain here indicates that pipe material is a key differentiator between high-risk and low-risk segments within the model's structure.

The *WS_Year_Install*, or year of pipe installation (i.e. the age of the pipe), exhibited the greatest relative frequency (~24.8%), demonstrating its widespread application across tree splits in the ensemble. Like its lower value compared to pipe material, its gain (~52.7) also substantiates its significant predictive contribution. It arises from the model's dependency on the age of the pipes, as a high age value is always correlative with processes of deterioration and the probability of breakdown associated with ageing infrastructure.

WS_Diameter, as the variable for pipe diameter, also exhibited a moderate impact, with a relative frequency of ~13.8% and a gain of ~22.1. The

diameter would presumably contribute to the risk of failure through its association with inner pressure dynamics and mechanical load. These findings align with results reported by Karadirek et al. (2024), who also identified material type and pipe age as dominant predictors of failure in urban water networks.

Table 1: *Feature importance metrics for XGBoost model (authors' analysis).*

Preglednica 1: *Vpliv vhodnih spremenljivk v modelu XGBoost (analiza avtorjev).*

Feature[3]	Gain	Relative Frequency	Cover
WS_Material	103.79	5.72%	1281.97
WS_Year_Install	52.7	24.79%	588.25
WS_Diameter	22.12	13.85%	728.26
EE_Voltage	19.4	6.45%	780.93
WS_Type_Network	17.48	2.89%	735.49
TC_Type_Cable	16.09	9.87%	630.84
WW_Type_Sewer	15.8	7.83%	657.12
RO_Road_Categ	13.44	9.98%	635.12
DH_Diameter	12.69	5.42%	681.1
NG_Diameter	11.86	13.03%	560.67

The XGBoost model's performance evaluation revealed good overall performance, particularly in terms of discrimination capability. With an ROC AUC of 0.9102, the model demonstrated excellent classification ability. While precision varied – showing an overall rate of 0.8750 but dropping to 0.1324 in the minority class – recall remained strong at 0.7750, confirming the model's effectiveness in identifying most failure cases. The F1 score of 0.2261 illustrates the inherent trade-off in class-imbalanced settings. As the objective function, the LogLoss value of 0.2500 indicates good calibration of predicted probabilities, ensuring that the model is suitable assessing pipe condition.

[3] WS – Water Supply; EE – Electric Energy; TC – telecommunications; WW – sewerage; RO – Roads; DH – District Heating; NG – Natural Gas

Additionally, to the geospatial representation of the probabilistic results generated by the XGBoost algorithm (Figure 2), we can observe the 52,605 pipeline segments in a statistical analysis by observing the distribution of pipe-specific features, such as material and diameter, in pipe failure probability classes. Figure 3 illustrates the pipe failure probability of pipes depending on the material of the pipe and the age.

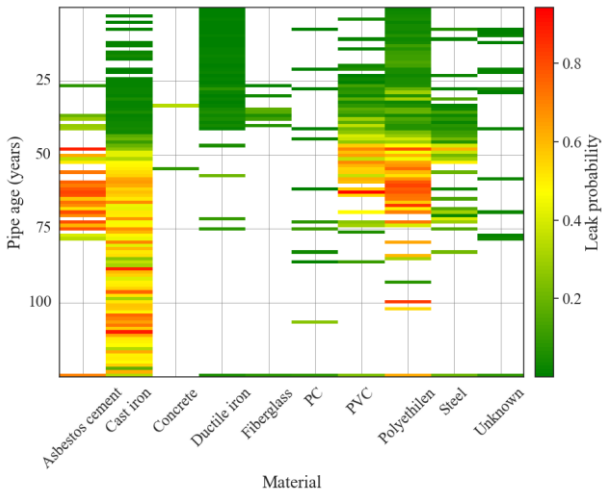

Figure 3: *Heat map of the pipe failure probabilities (leak) depending on the pipe material and age (authors' analysis).*

Slika 3: *Toplotna karta verjetnosti okvar cevi (puščanj) glede na material in starost cevi (analiza avtorjev).*

Additionally, Figure 4 provides categorization by material and pipe length, along with their corresponding failure probability classes. For example, older asbestos-cement (AC), cast-iron (LZ), polyethylene (PE), and polyvinyl chloride (PVC) pipes have a higher proportion of higher-risk probabilities. In contrast, newer materials, such as ductile iron (NL) and steel (JE) pipes, generally have lower risk classes, reflecting improved durability and resistance to failure.

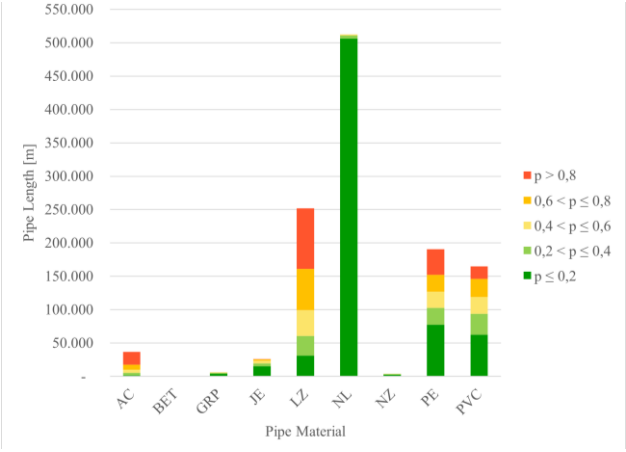

Figure 4: *Length of pipes classified by pipe material and failure probabilities (authors' analysis).*

Slika 4: *Dolžina cevi glede na material in verjetnost okvar cevi (analiza avtorjev).*

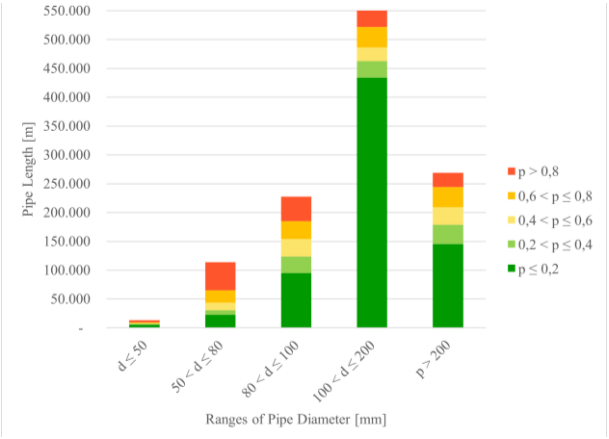

Figure 5: *Length of pipes classified by diameter of pipes and pipe failure probabilities (authors' analysis).*

Slika 5: *Dolžina cevi glede na premer in verjetnost okvar cevi (analiza avtorjev).*

Similarly, Figure 5 illustrates the lengths of pipes by pipe diameter ranges. In these diameter ranges, specifically the range $100 < d \leq 200$, ductile iron (NL) pipes have a large share, i.e. 68.1%. Additionally, we can observe that the smaller diameter ranges, $50 < d \leq 80$ and $80 < d \leq 100$, have a larger proportion of medium to high pipe failure risk.

Examining the feature importance results (Table 1) helps us infer the influence of nearby utility infrastructure. We will focus our observation on the gain metric, which measures how much a feature improves the model's accuracy when used in a decision split. *EE_Voltage* showed a moderate to high gain, suggesting that the presence or proximity of high-voltage electrical infrastructure may be correlated with an increased risk of pipe failure. This may reflect the complexity of construction or restricted access, where areas with underground electricity are more prone to delayed maintenance or suboptimal installation practices. *RO_Road_Categ* also exhibited substantial gain, indicating that road class – e.g. whether a pipe segment runs under arterial roads, local streets, or highways – has significant predictive value. Heavily trafficked roads can contribute to mechanical stress and vibration, which accelerates fatigue in pipes – a finding consistent with urban pipeline vulnerability studies. *WS_Type_Network*, although a WSS-native attribute, yielded less gain than *WS_Material* or *WS_Year_Install*, suggesting that network classification alone (e.g. distribution vs. trunk mains) may not be as discriminatory for failure likelihood. However, it still contributes to nuanced distinctions in pipe function and expected pressure regimes.

Feature *WW_Type_Sewer* demonstrated a relatively high frequency in XGBoost model splits, even if its gain was moderate. This suggests the consistent predictive value of sewer type presence (e.g. combined vs. separate systems), possibly due to shared trenching or construction timelines, which may impact the integrity of both networks. *DH_Diameter* and *NG_Diameter* appeared frequently in decision paths, reflecting that collocation with large-diameter district heating or gas lines may indirectly signal urban density or excavation difficulty, both of which are risk factors associated with legacy infrastructure and reduced intervention frequency. The use of these features across numerous tree splits suggests they offer complementary context, enriching the model's ability to account for environmental complexity.

Compared to standard WSS features like pipe material (*WS_Material*), installation year (*WS_Year_Install*), and diameter (*WS_Diameter*) – which exhibited the highest gain and frequency – the cross-utility contextual variables were

secondary but meaningful predictors. Traditional features dominate due to their direct linkage to physical deterioration mechanisms (e.g. corrosion, age-related degradation). However, contextual infrastructure features expand the model's scope, enabling it to capture risk factors related to environmental setting, construction history, and collocated utility impacts – dimensions not typically represented in legacy asset management models.

The coverage metric, which measures how much a feature affects the dataset, tracks the number of affected training patterns (segments) when that feature is used in a split. *WS_Type_Network* and *RO_Road_Category* showed broad coverage, indicating that these features contribute to splits that affect large portions of the pipeline network. These features are likely generalizable across various segment types and thus are valuable for system-wide risk classification. *EE_Voltage* and *NG_Diameter*, on the other hand, had lower cover but higher gain, suggesting that while they influence fewer segments, their predictive power is strong in specialised contexts, such as industrial zones or critical urban corridors.

The XGBoost algorithm has demonstrated overall performance in terms of pipe failure prediction and achieved a high recall rate, which proves its effectiveness in identifying vulnerable pipeline segments. The presented ML pipe failure model could therefore be applied in further WDN studies to optimize DMA design and their cost-effectiveness in terms of identifying water losses in WSS.

4 Conclusions

In this study, we showed that ML models can be highly effective in proactively managing large asset inventories, like in the case of WSS. Additionally, we incorporated a new perspective of quantified assessment of non-WSS features (i.e. adjacent infrastructure systems) in pipeline failure predictions. This multidimensional approach is rare in existing literature, where most models are confined to WSS-specific attributes and historical failure records alone (Kleiner & Rajani, 2001; Bakhtawar et al., 2025). For data-driven predictive models, integrating variables that capture the influence of nearby utility infrastructure – such as construction activity, operational stresses from nearby transportation infrastructure, and environmental impacts from adjacent utility operations – is shown to significantly enhance model accuracy in predicting pipeline failures. By using spatial buffering and overlay analysis to integrate water infrastructure with broader urban infrastructure datasets, this study highlights the multisystem interdependence of buried utility infrastructure. It aligns with actual urban planning and utility collocation constraints, offering a scalable framework for cross-infrastructure evaluations that incorporates risk perception from other utilities. The implications of these findings can significantly refine prioritization strategies, foster proactive asset management, and extend the reliability and service life of critical water infrastructure.

Acknowledgment

This paper was reviewed and corrected for grammar and style using Grammarly.com.

Author Contributions

Conceptualization: D.K.; Methodology: D.K. and D.A.F.; Software: D.A.F.; Validation: D.K. and D.A.F.; Formal analysis: D.K. and D.A.F.; Investigation: D.K. and D.A.F.; Resources: D.K.; Data curation: D.K.; Writing – original draft: D.K. and D.A.F.; Writing – review & editing: D.K.; Visualization: D.K. and D.A.F.; Supervision: D.K. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data that supports the findings of this study is available from GURS and JP VOKA SNAGA d.o.o. Restrictions apply to the availability of JP VOKA SNAGA d.o.o. data, which were used under license for the current research and are not publicly available. However, data are available from the corresponding author upon reasonable request and with permission from JP VOKA SNAGA d.o.o.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

This research received no external funding.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, United States, Pages 2623–2631. https://doi.org/10.1145/3292500.3330701.

Asadi, Y. (2024). Employing machine learning in water infrastructure management: predicting pipeline failures for improved maintenance and sustainable operations. *Industrial Artificial Intelligence* **2(8),** 1-13. https://doi.org/10.1007/s44244-024-00022-w.

Bakhtawar, B., Zayed, T., Elshaboury, N. (2025). Time-to-failure based deterioration factors of water networks: Systematic review and prioritization. *Reliability Engineering & System Safety*, **263**, 111246. https://doi.org/10.1016/j.ress.2025.111246.

Cabral, M., Gray, D., Brentan, B., Covas, D. (2024). Assessing Pipe Condition in Water Distribution Networks. *Water*, **16(10)**, 1318. https://doi.org/10.3390/w16101318.

Chen, T., Guestrin, C. (2016). "XGBoost: A scalable tree boosting system." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794. https://doi.org/10.1145/2939672.2939785.

Chen, T., et al. (2023). XGBoost documentation. XGBoost. https://xgboost.readthedocs.io/.

European Commission. (2025). European Water Resilience Strategy (COM(2025) 280 final). Brussels: European Commission. https://environment.ec.europa.eu/publications/european-water-resilience-strategy_en.

Ganjidoost, A., Haghighi, A., Klise, K. A. (2022). Pipe failure prediction in water distribution networks using machine learning models. *Journal of Water Resources Planning and Management*, **148(5)**, 04022017. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001557.

GURS. (2025). Data on public infrastructure = Zbirni kataster gospodarske javne infrastrukture. Republic of Slovenia, Surveying and Mapping Authority of the Republic of Slovenia. Slovenia. https://www.e-prostor.gov.si/.

MOPE. (2025). Information System of Public Environmental Protection Services = Informacijski sistem za spremljanje gospodarskih javnih služb varstva okolja (IJSVO). Republic of Slovenia, Ministry of Natural Resources and Spatial Planning, Slovenia. (in Slovenian) https://www.ijsvo.si/.

Jafar, R., Shahrour, I., Juran, I. (2010). Application of artificial neural networks (ANN) to model the failure of urban water mains. *Mathematical and Computer Modelling*, **51(9–10)**, 1170–1180. https://doi.org/10.1016/j.mcm.2009.12.033.

Karadirek, I. E., Kaya-Basar, E., Akdeniz, T. (2024). A study on pipe failure analysis in water distribution systems using logistic regression. *Water Supply*, **24(1),** 176–186. https://doi.org/10.2166/ws.2023.335.

Kleiner, Y., Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: Statistical models. *Urban Water*, **3(3)**, 131–150. https://doi.org/10.1016/S1462-0758(01)00033-4.

Kozelj, D., Gorjup, M., Kramar Fijavž, M. (2017). Uporaba teorije grafov za zasnovo merilnih območij v vodovodnem omrežju = An application of spectral graph partition for designing district metered areas in water supply networks. *Acta hydrotechnica*, **30(53)**, 81–96. https://www.dlib.si/details/URN:NBN:SI:doc-AEEWAAH7.

Latifi, M., Zali, R.B., Javadi, A.A., Farmani, R. (2024). Efficacy of Tree-Based Models for Pipe Failure Prediction and Condition Assessment: A Comprehensive Review. *Journal of Water Resources Planning and Management*, **150(7)**, 03124001. https://doi.org/10.1061/JWRMD5.WRENG-6334.

Large, A., Le Gat, Y., Elachachi, S. M., Renaud, E., Breysse, D., Tomasian, M. (2015). Improved modelling of 'long-term' future performance of drinking water pipes. *Journal of Water Supply: Research and Technology—AQUA*, **64(4)**, 415–425. https://doi.org/10.2166/aqua.2015.115.

Le Gat, Y., Curt, C., Werey, C., Caillaud, K., Rulleau, B., Taillandier, F. (2025). Water infrastructure asset management: state of the art and emerging research themes. *Structure and Infrastructure Engineering*, **21(4)**, 539-562, https://doi.org/10.1080/15732479.2023.2222030.

Liu, Z., Kleiner, Y., Rajani, B., Wang, L., Condit, W. (2012). Condition assessment technologies for water transmission and distribution systems (EPA/600/R-12/017). U.S. Environmental Protection Agency. https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=241510&Lab=NRMRL.

Mackey, T., Cashman, A., Cumberbatch, R. (2014). Identification of factors contributing to the deterioration and losses in the water distribution system in Barbados (CERMES Technical Report No. 68, 73 pp.). The University of the West Indies, Centre for Resource Management and Environmental Studies. https://www.cavehill.uwi.edu/cermes/docs/technical_reports/mackey_et_al_2014_pipe_deterioration_and_water_los.aspx.

Misiunas, D., Vítkovský, J., Olsson, G., Lambert, M., Simpson, A. (2006). Failure monitoring in water distribution networks, *Water Science & Technology*, **53 (4-5)**, 503–511. https://doi.org/10.2166/wst.2006.154.

Mohammadagha, M., Najafi, M., Kaushal, V., Jibreen, A. (2025). Machine Learning Models for Reinforced Concrete Pipes Condition Prediction: The State-of-the-Art Using Artificial Neural Networks and Multiple Linear Regression in a Wisconsin Case Study. arXiv, cs.LG, 2502.00363. https://doi.org/10.48550/arXiv.2502.00363.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Phan, H. C., Dhara, A. S., Hu, G., Sadiq, R. (2019). Managing water main breaks in distribution networks—A risk-based decision making. *Reliability Engineering & System Safety*, **191**, 106581. https://doi.org/10.1016/j.ress.2019.106581.

QGIS.org, 2025. QGIS Geographic Information System. QGIS Association. http://www.qgis.org.

Rajani, B., Kleiner, Y. (2001). Comprehensive review of structural deterioration of water mains: Physically based models. *Urban Water*, **3(3)**, 151–164. https://doi.org/10.1016/S1462-0758(01)00032-2.

Rezaei, H., Ryan, B., Stoianov, I. (2015). Pipe failure analysis and impact of dynamic hydraulic conditions in water supply networks. *Procedia Engineering*, **119**, 253–262. https://doi.org/10.1016/j.proeng.2015.08.879.

SURS. (2025). Public Water Supply - Water supplied from public water supply (1000 m3). Statistical Office of the Republic of Slovenia. https://pxweb.stat.si/SiStat/en.

VOKAS. (2025). Annual report 2024. Javno podjetje VODOVOD KANALIZACIJA SNAGA d.o.o. (in Slovenian) https://www.vokasnaga.si/sites/www.jhl.si/files/dokumenti/letno_porocilo_2024.pdf.

Warad, A.A.M., Wassif, K. Darwish, N.R. (2024). An ensemble learning model for forecasting water-pipe leakage. *Sci Rep* **14**, 10683. https://doi.org/10.1038/s41598-024-60840-x.

Zevnik, J., Kramar Fijavž, M., Kozelj, D. (2019). Generalized normalized cut and spanning trees for water distribution network partitioning. *Journal of Water Resources Planning and Management*, **145(10)**, 1–12. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001100.