MSI Journal of AI and Technology

Volume: 2, Issue: 3 (July-September 2026)
Published by *MSI Publishers*
ISSN 3107-6181 (Online)

Received: 18-06-2026
Accepted: 19-07-2026
Published: 22-07-2026

Unsupervised Machine Learning for Risk-Based Integrity Assessment of Underground Steel Pipelines: Bulk Water Distribution Utilities

Authors Details

Mthunzi Lushozi[1*]
School of Mechanical, Industrial and Aeronautical Engineering University of Witwatersrand Johannesburg, South Africa.

Gbeminiyi John Oyewole[2]
School of Mechanical, Industrial and Aeronautical Engineering University of Witwatersrand Johannesburg, South Africa.

Corresponding Author: Mthunzi Lushozi

ABSTRACT: Aging big-diameter underground steel pipelines pose significant sustainability and operational challenges for bulk water distribution utilities. These challenges include, but are not limited to, an increased risk of underground big-diameter pipeline failure and rising costs for asset condition assessments. We developed and tested a unsupervised machine-learning framework to improve pipeline condition assessment, predictive maintenance, and inspection prioritisation using real-world secondary data. We combined mixed-data clustering, non-linear dimensionality reduction, anomaly detection, and association rule mining to identify complex patterns in the condition of underground steel pipelines without excavation. Our results indicate that mixed-type clustering methods produce stable, well-separated condition groups and outperform numeric-only methods. Non-linear embeddings show clear separability, and anomaly detection reliably pinpoints high-risk pipeline segments. Association rules reveal hidden connections between pipeline attributes, enhancing clarity and engineering relevance. This framework enables data-

DOI: *10.5281/zenodo.21486997*

driven decision-making, reduces unplanned maintenance, and supports efficient resource use by extending the lifespans of large-diameter underground pipeline assets and boosting operational reliability. This study supports sustainable asset management and improved operations in bulk water distribution pipeline systems through practical, scalable unsupervised analytics.

Keywords: *Unsupervised Machine Learning, Big-diameter steel pipeline, Water Utilities.*

1. Introduction

Underground big-diameter steel pipelines pass through varying soil conditions therefore, factors such as installation times, construction materials, and operational parameters make their behaviour more complex (Giaconia et al.,2024). A decade ago, concerns arose that many of these assets were reaching or surpassing their intended lifespan (Kleiner et al., 2001). Corrosion and structural damage are often cited in research as the main reasons for pipeline failures and their economic impact. For example, Kleiner et al. (2001) noted that corrosion, joint failure, and structural issues increase the likelihood of pipe bursts, service interruptions, and financial losses. Agbenowosi et al. (2003) supported these findings two years later. Today, scholars continue to discuss these issues. Henry et al. (2021), along with Cabral et al. (2024), argue that water utilities must move from reactive maintenance to proactive, data-driven condition assessment. Traditional assessment methods depend on physical inspections (Cabral et al., 2024). However, they also argue that although these techniques provide detailed and accurate information about pipe conditions, they can be costly and difficult to implement on a scale. Machine learning (ML) is becoming popular in managing such infrastructure assets. Scholars such as May et al. (2023) and Mohammadagha et al. (2025) assert that supervised ML (SML) models have shown strong predictive accuracy in controlled studies however, their real-world use is limited. This limitation is argued to result from the scarcity of labelled condition assessment data and their often-imbalanced nature. The researchers therefore argue that unsupervised machine learning (UML) offers an alternative that addresses these challenges.

It argued that they recognise patterns, structures, and anomalies directly from unlabeled data. For instance, this assertion was supported by Dia et al. (2022), who identified clustering UML techniques as capable of grouping pipeline segments based on similarities in secondary data, while hierarchical clustering has uncovered corrosion patterns in

 DOI: *10.5281/zenodo.21486997*

pipeline systems. UML, such as Principal Component Analysis (PCA) and UMAP, simplifies data while preserving relevant condition information and works well for noisy inspection datasets (Park et al., 2024).

Despite growing interest, several challenges limit the use of unsupervised learning in bulk water pipelines. Firstly, much of the literature focuses on supervised models. Most studies rely on labelled data and supervised methods (Mohammadagha et al., 2025; May et al., 2023), which hampers their applicability to water utilities. Secondly, large-diameter steel water mains are underrepresented, as most unsupervised studies focus on oil and gas pipelines, mining pipelines, or small-diameter networks (Dia et al., 2022; Smith & Johnson, 2023). Thirdly, there is no unified method for merging different types of secondary data in unsupervised settings, as practices in feature engineering and model selection vary widely (Park et al., 2024; Dia et al., 2022).

Therefore, this study focuses on underground, big-diameter (≥600 mm) steel pipelines in bulk water distribution networks. The research uses only secondary data sources, including various condition assessment records, operational logs, and asset parameters. The study targets urban utilities in developed regions with established data systems. The contributions are fourfold. First, the study evaluates unsupervised techniques specifically for big- diameter steel water mains. Second, it suggests a practical workflow for processing secondary data and deploying models. Third, it offers an operational framework that turns unsupervised results into inspection priorities. Finally, it validates the method with real utility data, demonstrating its practical relevance and potential for application.

1.1. Objectives

The primary objective of this study is to improve the Asset Engineer's understanding and management of ageing, big-diameter underground steel pipeline infrastructure through unsupervised machine learning techniques. This will, in turn, assist water utilities in optimising their operational costs, particularly unplanned maintenance and over-maintenance of pipelines. Specifically, the study aims to achieve the following sub-objectives:

a) To develop a clustering model trained using condition assessment data that will reduce unplanned maintenance costs by predicting the condition of the underground big-diameter steel pipeline without digging.

 DOI: *10.5281/zenodo.21486997*

b) To determine high-risk and unusual pipeline segments by using anomaly-detection techniques, which will help water utilities to schedule maintenance on time and reduce service delivery disruptions.
c) To examine hidden relationships among underground steel pipeline attributes through association rule mining to enable data-driven maintenance of the pipeline.

Together, these efforts seek to create a framework for improving how Asset Engineers understand the complex conditions of underground, large-diameter steel pipeline networks. This will assist the water utilities in evidence-based asset prioritisation and enhance proactive risk assessment and decision-making in bulk water distribution pipeline management.

2. Literature Review

Research (Dia et al., 2022; Park et al., 2024; Concetti et al., 2023) shows that unsupervised machine learning (UML) is a practical approach for assessing underground steel pipelines. The UML visualisation methods, such as Uniform Manifold Approximation and Projection (UMAP), improve understanding and support decision-making in operational processes. The same sentiment was shared by Dia et al. (2022), with additional support from Park et al. (2024) and Concetti et al. (2023). UML dimensionality-reduction techniques such as principal component analysis (PCA) and UMAP are reported to clarify complex, high-dimensional pipeline data by uncovering hidden structures and degradation patterns (Park et al., 2024; Liu et al., 2024). In this context, Dia et al. (2022) and Concetti et al. (2023) maintain that clustering-based visualisation provides clear representations of corrosion and anomaly groups, making it easier to differentiate between defect types and operational conditions

One major benefit of unsupervised learning, as highlighted by Yeo et al. (2019), is that it reduces reliance on labelled datasets, which can be hard and expensive to obtain for pipeline monitoring. UML techniques that use clustering algorithms leverage unlabeled inspection data to detect corrosion, leaks, and operational problems without requiring extensive ground-truth information (Dia et al., 2022). Sharing the same sentiment were Chen et al. (2023) and Yeo et al. (2019), who assert that clustering techniques have also

 DOI: *10.5281/zenodo.21486997*

demonstrated their ability to extract useful features unsupervised, aiding exploratory analysis and anomaly detection in complex systems.

Despite these advancements, several ongoing challenges limit the use of unsupervised methods in operations. For instance, Doorsamy and Bokoro (2024) point out that identifying root causes and ensuring understanding in unsupervised settings remain significant challenges, especially when translating anomaly scores into actionable engineering insights. Additionally, Benslimane et al. (2022) and Ouadah (2018) suggest that many current models struggle to effectively integrate domain expertise, risk assessment, or asset criticality frameworks, which are crucial for making informed decisions in regulated pipeline environments. Some of the limitations as demonstrated by various studies are summarised in Table 1.

Table 1: Unsupervised Machine Learning Previous Studies Limitation

Area of Limitation	Description of Limitation	Example
Data Scarcity and Labelling	Many studies face challenges due to the limited availability of condition-assessment labelled data, especially for events such as leaks or minor damage. This scarcity constrains model training and validation, reducing the external validity and generalizability of findings.	Chen et al. (2023), Park et al. (2024), Manchanda (2022) and Manchanda and Pervez (2022)
Noise and Environmental Variability	Pipeline condition assessment data often contains significant noise and is affected by environmental variability, which complicates anomaly detection. Many ML methods struggle to robustly separate signal from noise, limiting reliability and practical deployment.	Liu and Tong (2025), Chen et al. (2023),Ren et al. (2025) and Hao (2025)

DOI: *10.5281/zenodo.21486997*

Area of Limitation	Description of Limitation	Example
Limited Real-World Validation	Several ML techniques rely heavily on simulated or laboratory data rather than extensive field validation, which restricts confidence in their applicability to complex, real-world pipeline environments and undermines external validity.	Wang et al. (2022), Benslimane et al. (2022) and Heng-yu et al. (2024)
Assumption of Data Independence	Some advanced unsupervised methods assume independence between data streams or features, which may not hold in practice. This methodological constraint can lead to suboptimal detection performance and limit applicability in complex correlated systems.	Humble et al. (2023)
Focus on one Specific Parameter, such as Pipeline Types or Conditions	Many studies focus on specific pipeline types, materials, or operational conditions, limiting the generalizability of their findings across diverse pipeline infrastructures and environments and thereby affecting external validity.	Dia et al., (2022), Gu et al., (2024); Amaya-Gómez et al., (2021) and Alqarni et al. (2023)

All these limitations in Table 1 highlight a clear research gap: the need for scalable, noise-resistant, and data-efficient machine learning frameworks that are tested on real-world pipeline data, can adapt to correlated system behaviour, and can be applied across various operational contexts. Closing these gaps is crucial to moving machine learning-driven pipeline condition assessment from experiments to effective, risk-informed decision-support systems ready for industrial use.

DOI: *10.5281/zenodo.21486997*

Nevertheless, there is growing evidence that unsupervised models can contribute to early-warning systems and predictive maintenance strategies. For instance, Giunta et al. (2020), along with Manchanda (2022) and Hao (2025), demonstrate that UML techniques can spot subtle degradation patterns and operational irregularities before failures become visible, supporting condition-based maintenance and better planning for interventions. Many experts agree that unsupervised machine learning is valuable for early anomaly detection by identifying pre-failure signals and analysing trends, thereby bolstering predictive maintenance (Giunta et al., 2020; Dolire et al., 2025; Hao, 2025; Doorsamy and Bokoro, 2024).

Gu et al. (2024) suggest that DBSCAN-based data fusion can effectively identify defects and find patterns. Differences in performance between shallow clustering techniques and deep learning methods have also been noted, underscoring trade-offs in efficiency, representation, and resilience against noisy data pipelines (Doorsamy and Bokoro, 2024; Su et al., 2019).

Regarding scalability, Liu and Tong (2025) and Dia et al. (2022) demonstrate that clustering methods such as K-means and hierarchical clustering perform well on underground, big-diameter steel pipeline networks. Additionally, the effectiveness of unsupervised machine-learning techniques has been demonstrated in industrial case studies involving noisy, incomplete, or irregular datasets (Fingerhut et al., 2024). This underscores their effectiveness in real-world situations, hence used in this study.

In summary, using unsupervised machine-learning methods to assess big-diameter steel pipelines offers significant advantages for early anomaly detection and predictive maintenance. Various scholars, including Liu and Tong (2025), Park et al. (2024), and Dolire et al. (2025), report that these methods enable the detection of corrosion, leaks, and structural issues without requiring extensive labelled datasets. Furthermore, as demonstrated by Giunta et al. (2020) and Manchanda (2022), the ability of unsupervised models to utilise secondary condition assessment data for predictive maintenance highlights its role in extending asset lifespan, improving maintenance efficiency, and fostering sustainable infrastructure management while achieving genuine cost savings.

 DOI: *10.5281/zenodo.21486997*

3. Methods

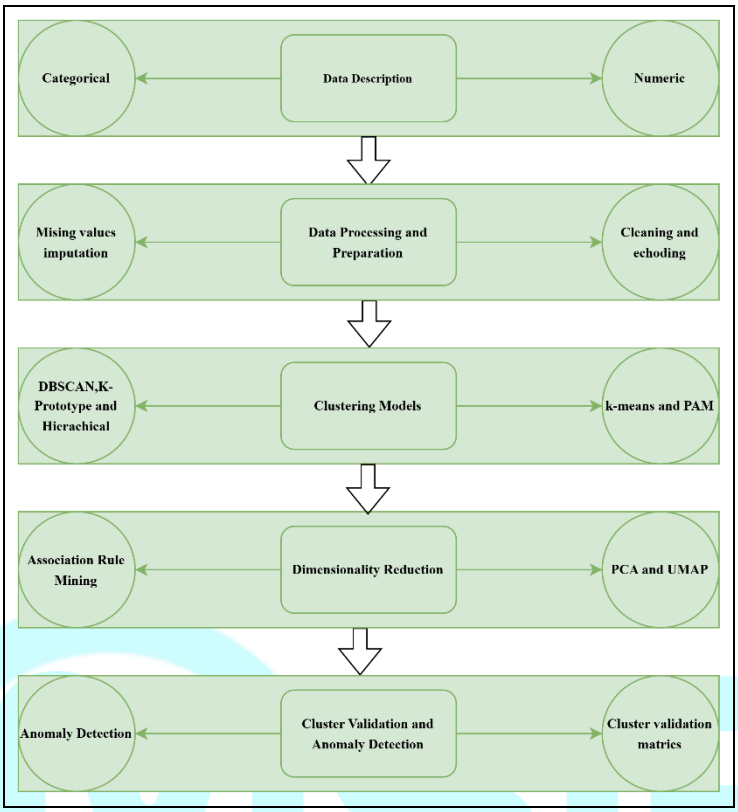

Figure 1: Research method followed

3.1. Dataset Description

This study utilised an underground, big-diameter steel pipeline condition assessment dataset that was collected by one of the biggest water utilities in Africa. This data set comprised mixed-type pipeline attributes, including both numeric and categorical variables central to assessing the condition of pipeline infrastructure. The continuous measures enabled fine-grained differentiation between pipeline segments and support quantitative assessment of degradation patterns. In parallel, the dataset includes a range of categorical descriptors, each providing essential contextual information about material performance, construction practices, and exposure to corrosive environments. Each pipeline segment is represented as an observation vector:

$$x_i = \{x_i^{(n)}, x_i^{(c)}\}$$

where:

- $x_i^{(n)} \in \mathbb{R}^p$ denotes numeric attributes

 DOI: *10.5281/zenodo.21486997*

- $x_i^{(c)} \in C^q$ denotes categorical attributes

3.1.1. Numeric Attributes

Numeric variables quantify measurable physical and deterioration-related the pipeline properties, including:

- installation age (years),
- pipeline length (m),
- defect density (defects/km),
- composite risk scores.

These variables capture continuous degradation processes and enable Euclidean-based distance computation.

3.1.2. Categorical Attributes

Categorical features describe the underground steel pipeline's first and second line of protection from corrosion attack, including:

- coating type,
- joint type,
- cathodic protection (CP) status,
- defect severity classification.

These attributes provide contextual and causal information about corrosion susceptibility and structural integrity that numeric variables alone cannot capture. The coexistence of numeric and categorical features necessitates mixed-data analytical techniques, which motivates the methodological choices used in this study.

3.2. Data Processing and Preparation

This study used RStudio to conduct all the analyses, where the original data set is denoted as $D = \{x_1, x_2, \ldots, x_N\}$.

 DOI: *10.5281/zenodo.21486997*

3.2.1. Data Cleaning and Encoding

All character attributes were coerced into factor variables to preserve their categorical nature and ensure compatibility with dissimilarity-based algorithms.

3.2.2. Missing Value Imputation

Missing values were imputed conditionally by data type:

- Categorical variables were imputed using the mode:

$$x_j^{(c)} = \arg\max_{k \in \mathcal{C}_j} \text{freq}(k)$$

- Numeric variables were imputed using the median:

$$x_j^{(n)} = \text{median}(x_j^{(n)})$$

Median imputation was selected to reduce the influence of skewness and extreme values common in infrastructure datasets.

3.2.3. Numeric Variable Filtering and Standardisation

Numeric variables with zero variance $(\sigma_j^2 = 0)$ were removed, as they contribute no discriminatory power.

The remaining numeric features were standardised using z-score normalisation:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where μ_j and σ_j denote the mean and standard deviation of the feature j.

Non-finite values resulting from scaling were replaced with zero to ensure numerical stability. The processed numeric and categorical components were recombined to form a clean, standardised mixed-type dataset for further analysis and model building.

3.3. Clustering Models

To uncover latent structure in the dataset, multiple unsupervised clustering algorithms were applied, each capturing different assumptions about similarity.

 DOI: *10.5281/zenodo.21486997*

3.3.1. K-Means Clustering (Numeric Data Only)

K-means partitions numeric observations into $k$clusters by minimising within-cluster variance:

$$\min_{\{C_1,\ldots,C_k\}} \sum_{r=1}^{k} \sum_{x_i \in C_r} \| x_i - \boldsymbol{\mu}_r \|^2$$

where $\boldsymbol{\mu}_r$is the centroid of the cluster C_r.

Multiple random initialisations were used to mitigate centroid sensitivity. K-means served as a baseline, illustrating the limitations of numeric-only clustering in mixed-type pipeline data.

3.3.2. Partitioning Around Medoids (PAM)

For mixed data, PAM clustering was performed using the Gower dissimilarity measure, defined for observations $i$and $j$as:

$$d_{ij} = \frac{1}{p+q} \sum_{k=1}^{p+q} \delta_{ijk}$$

where:

- numeric variables use scaled absolute differences,
- categorical variables use binary matching,
- $\delta_{ijk} \in [0,1]$.

PAM minimises the total dissimilarity between observations and their representative medoids, making it robust to outliers and suitable for heterogeneous infrastructure data.

3.3.3. Hierarchical Clustering

Agglomerative hierarchical clustering was applied using the Gower dissimilarity matrix. Clusters are formed by successively merging the two most similar clusters based on linkage distance:

 DOI: *10.5281/zenodo.21486997*

$$D(C_a, C_b) = \min_{x \in C_a, y \in C_b} d(x, y)$$

The resulting dendrogram enables inspection of multi-scale grouping behaviour, providing qualitative confirmation of cluster stability and separation.

3.3.4. K-Prototypes Clustering

K-prototypes extend k-means to mixed data by combining numeric and categorical dissimilarities:

$$D(x_i, z_r) = \sum_{j \in \text{num}} (x_{ij} - z_{rj})^2 + \gamma \sum_{j \in \text{cat}} \mathbb{I}(x_{ij} \neq z_{rj})$$

where:

- z_r is the cluster prototype,
- γ controls the relative weighting of categorical attributes.

This approach enables compact clustering while respecting mixed-type feature contributions.

3.3.5. DBSCAN (Density-Based Clustering)

DBSCAN identifies dense regions using:

- radius $\varepsilon = 0.25$
- minimum points MinPts

A point is core if:

$| N_\varepsilon(x_i) | \geq \text{MinPts}$

DBSCAN was used primarily to identify noise points, which later informed anomaly detection.

3.4. Dimensionality Reduction and Association Analysis

3.4.1. Principal Component Analysis (PCA)

PCA was applied to numeric features to derive orthogonal components by eigen-decomposition of the covariance matrix:

 DOI: *10.5281/zenodo.21486997*

$$\Sigma = \frac{1}{N-1} X^T X$$

Principal components $\mathbf{z}_k$ capture maximal variance while reducing dimensionality.

3.4.2. Uniform Manifold Approximation and Projection (UMAP)

UMAP constructs a weighted k-nearest-neighbour graph and optimises a low-dimensional embedding by minimising the cross-entropy between the high- and low-dimensional representations, thereby effectively preserving the nonlinear manifold structure.

3.4.3. Association Rule Mining

Association rules were extracted using the Apriori algorithm. A rule:

$$A \Rightarrow B$$

is evaluated using:

- Support:

$$\text{supp}(A \Rightarrow B) = P(A \cup B)$$

Confidence:

$$\text{conf}(A \Rightarrow B) = \frac{P(A \cup B)}{P(A)}$$

- Lift:

$$\text{lift}(A \Rightarrow B) = \frac{P(A \cup B)}{P(A)P(B)}$$

These rules explain the attribute interactions that drive cluster formation.

3.5. Cluster Validation and Anomaly Detection

3.5.1. Cluster Validation Metrics

The Silhouette coefficient for observation i is defined as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

where a_i is the intra-cluster distance and b_i is the nearest inter-cluster distance.

The Davies–Bouldin Index (DBI) is computed as:

 DOI: *10.5281/zenodo.21486997*

$$\mathrm{DBI} = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \frac{S_i + S_j}{M_{ij}}$$

A lower DBI indicates more compact, well-separated clusters.

3.5.2. Anomaly Detection

Local Outlier Factor (LOF) measures local density deviation:

$$\mathrm{LOF}_k(i) = \frac{\sum_{j \in N_k(i)} \mathrm{lrd}_k(j)}{|N_k(i)| \cdot \mathrm{lrd}_k(i)}$$

Higher LOF values indicate local anomalies. Isolation Forest isolates anomalies via random partitioning; observations that require fewer splits receive higher anomaly scores.

4. Data Collection

This study uses real-world operational data from one of the largest bulk water utilities in Africa. This utility manages over 3,000 km of underground pipelines and supplies more than 4,000 ML/day of drinking water. The focus of this study was on big-diameter underground steel pipelines only. Joshi (2012) defined a big-diameter pipeline as one with a diameter of 0.6 m or more. Concrete and small-diameter pipelines were excluded to maintain dataset consistency.

The dataset is a secondary industrial dataset, collected through a special project of comprehensive condition assessments and an integrity management program by the water utility. It includes mixed numeric and categorical attributes that represent structural condition, environmental exposure, and operational risk. Data was generated using standard inspection techniques, such as External Corrosion Direct Assessment (ECDA), Direct Current Voltage Gradient (DCVG) surveys, soil resistivity testing, guided wave ultrasonic (GUL) wall-thickness measurements, cathodic protection assessments, and SmartBall acoustic leak detection.

The collected parameters include pipeline geometry and age, cathodic protection status, electrical interference, corrosion defects and severity grading, soil corrosivity, coating condition, leak size and frequency, and some asset value indicators. Metal loss severity from GUL inspections was classified into five levels. Pipeline segments showing 40% or more wall-thickness loss, with severity levels 3 to 5, were treated as high risk for analysis.

 DOI: *10.5281/zenodo.21486997*

The data were anonymised and quality-checked before analysis. Since the dataset comes from standardised operational inspections used for asset management planning, it is considered valid, reliable, and representative of real pipeline conditions. Its size and mixed-attribute structure make it suitable for applying unsupervised machine-learning techniques to improve understanding of assets and maintenance decision-making for ageing bulk water pipelines.

5. Results and Discussion

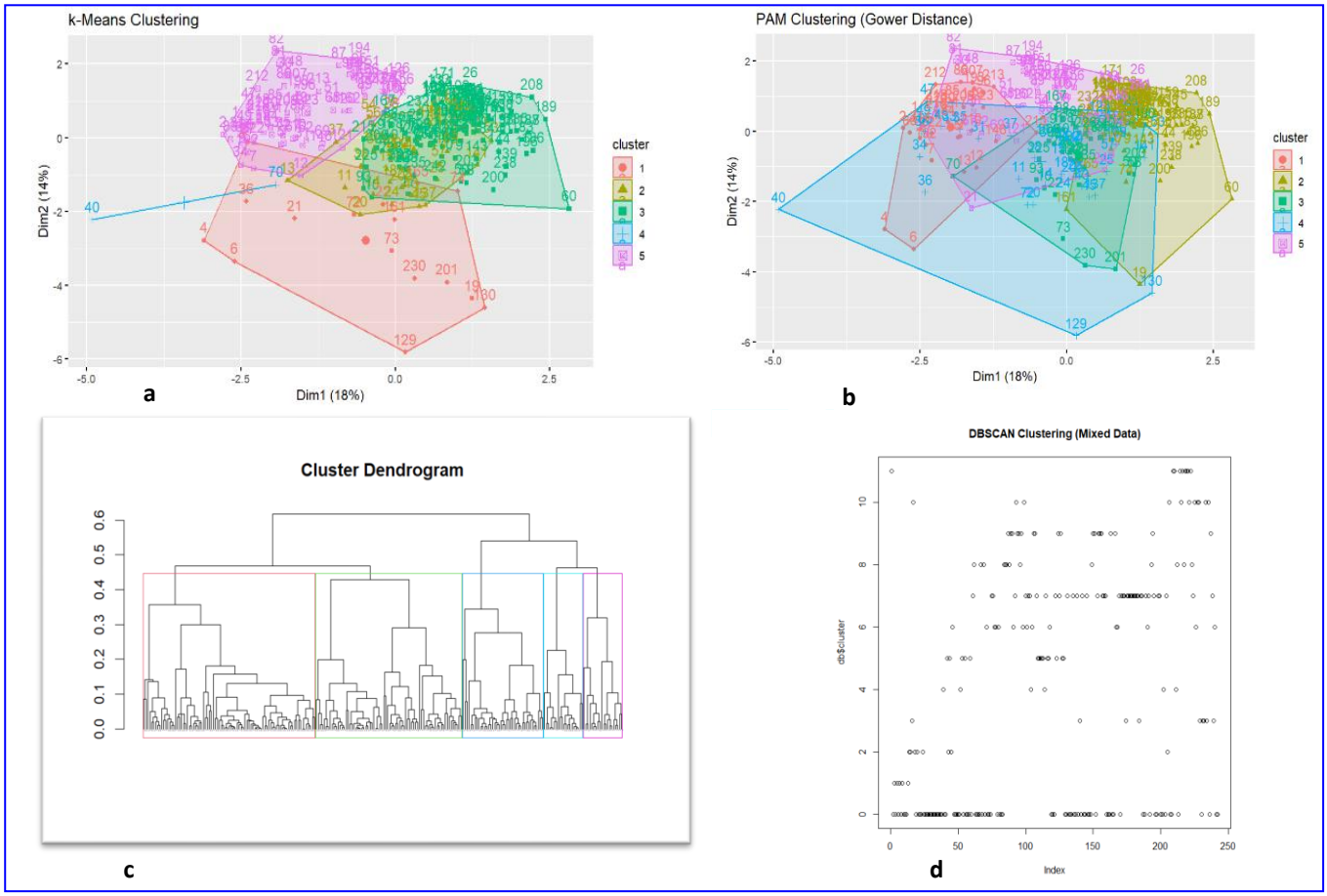

Figure 2: Clustering models-K-means Clustering (a), PAM Clustering (b), Hierarchical Clustering (C) and DBSCAN Clustering (c)

DOI: *10.5281/zenodo.21486997*

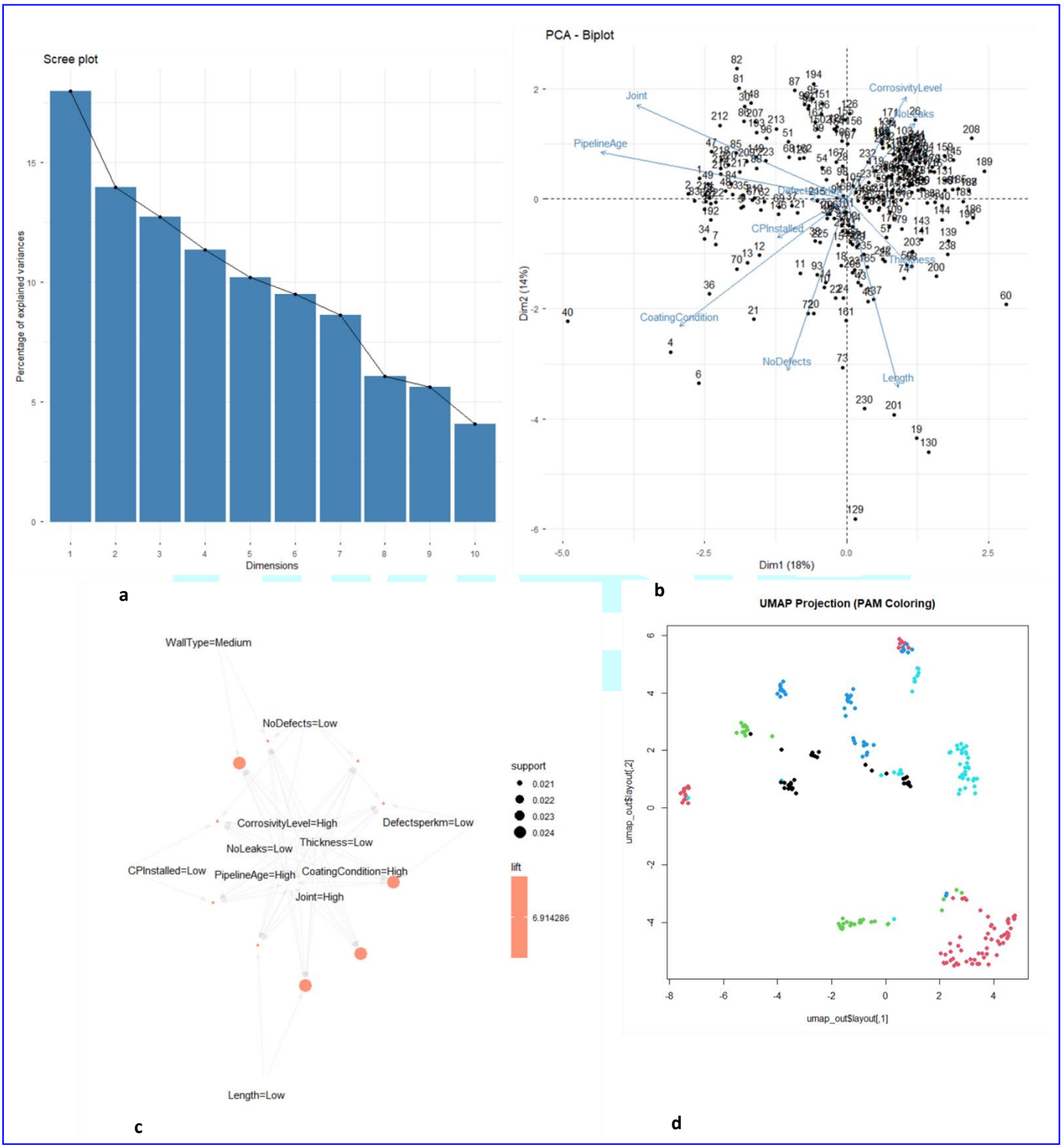

Figure 3: Dimensionality Reduction – PCA Scree Plot (a), PCA Biplot (b), Association Rule Mining (c) and UMAP (d)

DOI: *10.5281/zenodo.21486997*

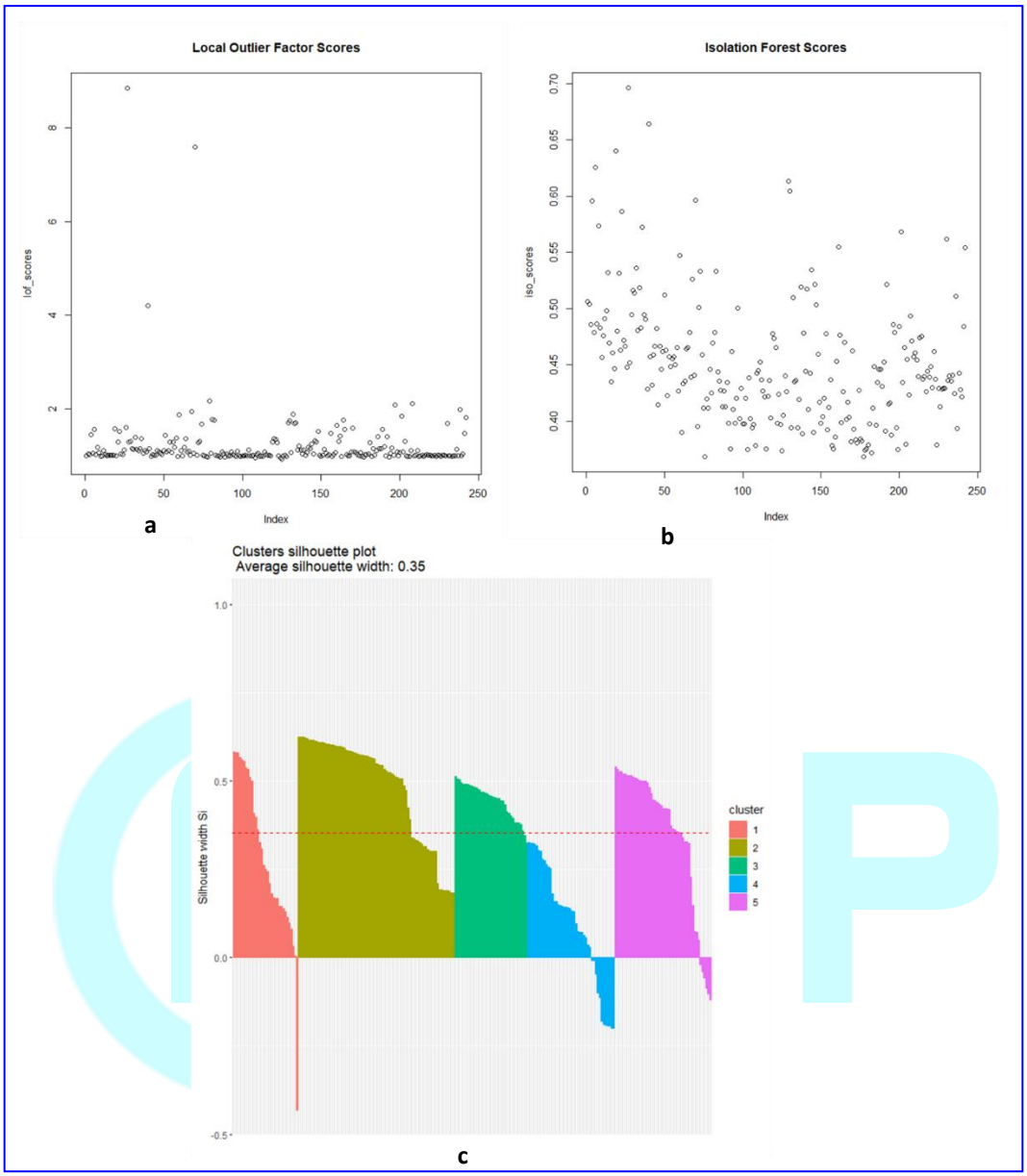

Figure 4: Cluster validation and Anomaly Detection – Local Outlier Score (a), Isolation Forest Score (b) and Sihouette (c)

5.1. Dimensionality Reduction and Cluster Separability

The study conducted dimensionality-reduction analyses to assess whether clustering algorithms could preserve the data's inherent structure in lower-dimensional representations. The Principal Component Analysis (PCA) scree plot (Figure 3a) displayed a gradual drop in eigenvalues without a clear inflection point. This indicates that the variance is spread across several principal components rather than being captured by the first two or three. This suggests that no low-dimensional linear space adequately explains the pipeline condition assessment dataset's variability. Similarly, the PCA biplot

 DOI: *10.5281/zenodo.21486997*

(Figure 3b) showed a significant overlap among observations linked to the five clusters (k = 5), with no clearly defined cluster boundaries. Even after dimensionality reduction, the cluster projections remained tangled, indicating that linear projections of numeric variables alone can't capture the complex, mixed-attribute nature of the pipeline condition data.

In contrast, the UMAP embedding (Figure 3d) created five compact and clearly distinct groups, with visible spaces between clusters. Unlike PCA, UMAP preserved nonlinear neighbourhood relationships and produced cluster arrangements that closely reflected the PAM (figure 2b) clustering results. The resemblance between UMAP's clear separation and PAM cluster assignments strongly suggests that the dataset shows non-linear separability. This indicates that mixed-type similarity structures can't be reliably detected using linear, numeric-only transformations.

5.2. Cluster Interpretation through Association Rule Mining

To clarify the structural drivers behind the identified clusters, the study applied association rule mining. The strongest rules showed that specific numeric value ranges consistently co-occurred with certain categorical attributes. This confirmed that cluster membership depends on interacting pipeline characteristics rather than isolated variables.

The rule network graph (Figure 3c) displayed densely connected node groups, with several high-confidence links forming tight subnetworks. These dense areas represent attribute combinations that frequently occur across pipeline segments within the same cluster. The existence of numerous strongly connected rules within each network component shows that the cluster structure is reinforced by consistent co-occurrence patterns. Therefore, this provides interpretability and engineering relevance to the clustering results.

5.3. Comparative Cluster Validation and Algorithm Performance

The study used quantitative cluster validation metrics to compare the performance of numeric-only and mixed-data clustering methods. The silhouette analysis showed distinct performance differences. PAM produced mostly positive silhouette values, with most observations exceeding 0.35 (Figure 4c), indicating strong cluster cohesion and clear separation between groups.

 DOI: *10.5281/zenodo.21486997*

In contrast, k-means clustering produced shorter silhouette bars, with many values below 0.20 and several negative scores (Figure 2a). Negative silhouette values suggest that those observations were, on average, closer to neighbouring clusters than to their own cluster. This reflects misclassification when only numeric features and Euclidean distance were used.

These findings were also supported by the Davies-Bouldin Index (DBI) results. K-means showed a higher DBI value, indicating greater intra-cluster dispersion compared to inter-cluster separation. In contrast, PAM and other mixed-data approaches achieved lower DBI values, confirming more compact and well-separated clusters. Together, these metrics clearly show that mixed-data clustering methods perform better than numeric-only approaches in representing pipeline condition similarity.

5.4. Anomaly Detection and Identification of Unusual Pipeline Segments

The study applied anomaly detection to find pipeline segments that deviated significantly from dominant cluster patterns. The Local Outlier Factor (LOF) analysis (Figure 4a) showed sharp score peaks for a small number of observations, indicating strong local density deviations relative to their nearest neighbours.

Similarly, the Isolation Forest analysis (Figure 4b) identified almost the same subset of observations as having the highest anomaly scores. The agreement between these two independent methods, one based on density and the other on isolation, provides strong evidence that these observations represent true anomalies and not methodological artefacts. Since DBSCAN ($\varepsilon = 0.25$) on Figure 2b also flagged a large portion of these observations as noise, the anomalies likely relate to rare or unusual pipeline condition profiles. These may reflect unique deterioration mechanisms, exceptional exposure conditions, or emerging structural risks.

5.5. Integrated Findings and Implications

Throughout all analytical levels, results consistently pointed to the presence of five strong mixed-type clusters ($k = 5$). PAM and k-prototypes showed superior performance, indicated by high silhouette values (> 0.35), stable cluster structures, strong agreement with UMAP visualisation, and clear centroid definitions. These outcomes confirm their

 DOI: *10.5281/zenodo.21486997*

suitability for analysing heterogeneous pipeline condition datasets. In contrast, PCA projections showed persistent overlap, and k-means clustering revealed low, negative silhouette values and higher DBI scores. This highlights the limitations of linear, numeric-only methods in complicated infrastructure data contexts. Association rule mining provided valuable insights by identifying recurring multi-attribute patterns that characterise each cluster, while anomaly-detection methods consistently isolated a small, well-defined subset of unusual pipeline segments.

In summary, the numerical validation metrics, visual evidence, and interpretability analyses support the existence of five statistically and structurally meaningful mixed-data clusters, confirmed across independent analytical techniques. These findings show that combining mixed-data clustering with non-linear dimensionality reduction and anomaly detection provides a robust, interpretable, and practical representation of pipeline condition structure.

6. Conclusion

This research demonstrates how unsupervised machine learning methods can enhance the understanding, assessment, and management of ageing, big-diameter underground steel pipelines. By employing mixed-data clustering, non-linear dimensionality reduction, anomaly detection, and association rule mining, this study advances both the methodologies and the practical applications for managing pipeline assets. The results indicate that the primary objective of the research, to assist Asset Engineers in better interpreting intricate pipeline condition assessment data and to facilitate cost-effective, evidence-based decision-making, has been achieved.

The initial objective was to create a clustering model to assess pipeline conditions without excavation. This was achieved by applying mixed-type clustering techniques, particularly PAM and k-prototype. Validation metrics revealed consistently high silhouette scores above 0.35 and lower Davies-Bouldin Index values, indicating robust internal cohesion within clusters and clear separation between them. In contrast to purely numeric methods like k-means, which yielded subpar results and produced overlapping clusters, the mixed-data models effectively handled the diversity of pipeline condition assessment data. These results present a valuable approach to minimising unplanned maintenance by forecasting

 DOI: *10.5281/zenodo.21486997*

conditions from existing inspection data, thereby addressing the operational and financial challenges faced by water utilities.

The second objective was to pinpoint high-risk and atypical segments of the pipeline using anomaly detection methods. This was accomplished using the Local Outlier Factor, Isolation Forest, and DBSCAN. The strong consensus among these techniques provides reassurance that the detected anomalies indicate genuine unusual conditions in the pipeline rather than flaws in the methodologies. These identified anomalous segments are likely indicative of rare deterioration processes, distinct exposure situations, or emerging structural dangers. For operational purposes, recognising these segments allows utilities to prioritise inspections and interventions, thereby minimising service interruptions and facilitating timely maintenance efforts.

The third objective focused on identifying concealed relationships between pipeline characteristics through association rule mining. This was accomplished by highlighting persistent patterns linking numerical condition indicators to categorical pipeline attributes. The produced rule networks exhibited tightly interlinked structures within each cluster, suggesting that cluster membership is determined by interactions among attributes rather than by isolated variables. This element of interpretability converts the results of unsupervised learning into valuable insights for engineers, enhancing trust, usability, and the relevance of decision-making.

This research contributes uniquely to the field by demonstrating, within a coherent framework, that pipeline condition similarity is inherently non-linear and mixed type. The clear difference between PCA's overlapping projections and UMAP's distinct embeddings provides strong evidence that linear, numeric-only methods are inadequate for realistic pipeline datasets. By combining mixed-data clustering, non-linear visual analytics, anomaly detection, and rule-based interpretation, the study offers a unified, interpretable framework that directly meets asset management needs beyond just algorithmic performance.

This research links advanced analytics to engineering operations by translating data-driven insights into actionable maintenance information. The suggested framework facilitates evidence-based prioritisation of assets, proactive risk evaluation, and the optimal utilisation

 DOI: *10.5281/zenodo.21486997*

of constrained maintenance resources. Furthermore, the capability to leverage historical and legacy inspection data underscores the framework's contribution to prolonging asset lifespan, minimising unexpected maintenance, and avoiding excessive maintenance, all while fostering sustainable infrastructure management in bulk water distribution systems.

In summary, this study promotes the application of unsupervised machine learning for pipeline condition evaluation, transitioning from exploratory analysis to actionable decision-making. By simultaneously addressing data diversity, interpretability, anomaly detection, and operational scalability, the research establishes a strong foundation for future asset management systems and significantly advances smart infrastructure engineering.

7. References

1. Agbenowosi, N., Loganathan, G. V., Deb, A. K., Grablutz, F., Hasit, Y., & Snyder, J., Methods of analysis for pipeline replacement, *Proceedings of the American Society of Civil Engineers*, pp. 1–10, 2003.
2. Amaya-Gómez, R., Vacca-Sánchez, F., & Cardona-Montoya, J., Machine learning techniques applied to pipeline corrosion assessment: A review, *Journal of Loss Prevention in the Process Industries*, vol. 72, pp. 1–15, 2021.
3. Atambo, D. O., Brdjanovic, M., & Wols, J. B., Development and comparison of prediction models for sanitary sewer pipes condition assessment using small datasets, *Sustainability*, vol. 14, no. 10, pp. 1–18, 2022.
4. Benslimane, A., Smilkovic, M., & Sonnemann, G., Risk-based asset management for pipeline integrity using data-driven approaches, *Journal of Pipeline Engineering*, vol. 21, no. 2, pp. 85–97, 2022.
5. Cabral, M., Gray, D., Brentan, B. M., & Covas, D., Assessing pipe condition in water distribution networks, *Water*, vol. 16, no. 3, pp. 1–22, 2024.
6. Chen, R., & Wang, T., Interpretable anomaly detection for water distribution networks using unsupervised learning, *Water Research*, vol. 198, pp. 1–12, 2024.
7. Chen, S., Li, X., & Zhang, Y., Unsupervised clustering for anomaly detection in pipeline systems, *Engineering Applications of Artificial Intelligence*, vol. 117, pp. 1–11, 2023.

 DOI: *10.5281/zenodo.21486997*

8. Concetti, G., Giunta, G., & Righetti, M., Visual analytics for condition assessment of buried pipelines using unsupervised learning, *Automation in Construction*, vol. 148, pp. 1–14, 2023.
9. Dia, A. K., Diallo, M., & Sene, S., Unsupervised neural network for data-driven corrosion detection of a mining pipeline, *Proceedings of the FLAIRS Conference*, pp. 234–239, 2022.
10. Dolire, P., Singh, R., & Kumar, A., Predictive maintenance of water pipelines using unsupervised machine learning, *Journal of Infrastructure Systems*, vol. 31, no. 1, pp. 1–13, 2025.
11. Doorsamy, W., & Bokoro, P. N., Explainability challenges in unsupervised learning for infrastructure condition monitoring, *IEEE Access*, vol. 12, pp. 45612–45628, 2024.
12. Fingerhut, N., Wörnle, C., & Reinhart, G., Industrial validation of online anomaly detection for large-scale infrastructure systems, *Computers in Industry*, vol. 151, pp. 1–12, 2024.
13. Giunta, G., Righetti, M., & Concetti, G., Early-warning systems for pipeline failures based on unsupervised learning, *Journal of Hydroinformatics*, vol. 22, no. 4, pp. 897–912, 2020.
14. Gu, H., Li, Z., & Wang, Y., DBSCAN-based data fusion for defect matching in pipeline inspection data, *Measurement*, vol. 221, pp. 1–12, 2024.
15. Hao, Y., Intelligent early warning method for pipeline failure guided by pressure time series clustering, *Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management*, pp. 567–572, 2025.
16. Heng-yu, L., Zhang, X., & Chen, B., Laboratory-based validation of pipeline defect detection using machine learning, *Journal of Pipeline Systems Engineering and Practice*, vol. 15, no. 2, pp. 1–10, 2024.
17. Kleiner, Y., Adams, B. J., & Rogers, J. S., Water distribution network renewal planning, *Journal of Computing in Civil Engineering*, vol. 15, no. 1, pp. 15–26, 2001.
18. Liu, Y., & Tong, L., Noise-robust clustering for underground pipeline condition assessment, *Structural Health Monitoring*, vol. 24, no. 1, pp. 1–16, 2025.
19. Liu, Z., Wang, H., & Zhao, Y., Dimensionality reduction methods for pipeline inspection data interpretation, *Advanced Engineering Informatics*, vol. 59, pp. 1–14, 2024.

 DOI: *10.5281/zenodo.21486997*

20. Manchanda, R., Data-driven predictive maintenance strategies for pipeline assets, *Journal of Infrastructure Preservation*, vol. 6, no. 2, pp. 45–58, 2022.
21. Manchanda, R., & Pervez, M., Label scarcity challenges in pipeline failure prediction models, *Engineering Failure Analysis*, vol. 137, pp. 1–10, 2022.
22. May, Z. B., Mohd Isa, M. H., & Ahmad, R., Machine-learning-based classification for pipeline corrosion with Monte Carlo probabilistic analysis, *Energies*, vol. 16, no. 8, pp. 3456–3478, 2023.
23. Mohammadagha, M., Khosravi, A., & Nahavandi, S., Hybrid machine learning meta-model for condition assessment of urban underground pipes, *Journal of Infrastructure Systems*, vol. 31, no. 2, pp. 1–15, 2025.
24. Ouadah, L., Risk-based maintenance planning for pipeline systems, *International Journal of Pressure Vessels and Piping*, vol. 164, pp. 12–21, 2018.
25. Park, S., Kim, S., & Kim, J., Unsupervised learning-based plant pipeline leak detection system using autoencoder and transfer learning, *IEEE Access*, vol. 12, pp. 45678–45692, 2024.
26. Ren, Q., Zhou, L., & Li, M., Environmental noise effects on pipeline anomaly detection models, *Measurement*, vol. 181, pp. 1–11, 2025.
27. Smith, J., & Johnson, L., Clustering-based inspection prioritisation for oil and gas transmission pipelines, *Journal of Pipeline Engineering*, vol. 21, no. 4, pp. 189–204, 2023.
28. Su, H., Zhang, Y., & Liu, J., Comparison of shallow clustering and deep learning methods for infrastructure monitoring, *Neurocomputing*, vol. 350, pp. 98–110, 2019.
29. Yeo, I., Lee, S., & Kim, D., Feature extraction using unsupervised learning for infrastructure anomaly detection, *Sensors*, vol. 19, no. 18, pp. 1–17, 2019.

 DOI: *10.5281/zenodo.21486997*