

Article

Hierarchical Corrosion Assessment of Water Pipelines in a Hydroelectric Plant Through Statistical Analysis and Clustering Methods

Cleber Gustavo Dias ^{1,2,3,*} , Fabio Henrique Pereira ^{1,2,3,*} , Carlos Alberto Murad ¹, Autharis da Silva Peixoto ⁴, Fernando Hiroyuki Hamaji ⁴, Gilberto Francisco Martha de Souza ¹ , Ivan Eduardo Chabu ¹ , Idalina Vieira Aoki ¹  and Silvio Ikuyo Nabeta ¹ 

¹ Polytechnic School, Universidade de São Paulo, São Paulo 05508-220, São Paulo, Brazil; gfmsouza@usp.br (G.F.M.d.S.); snabeta@usp.br (S.I.N.)

² Acurai—Pesquisa, Desenvolvimento e Inovação Ltda, Marília 17500-030, São Paulo, Brazil

³ Informatics and Knowledge Management Graduate Program, Universidade Nove de Julho—UNINOVE, São Paulo 01525-000, São Paulo, Brazil

⁴ Axia Energia S.A., Rodovia BR-364, Porto Velho 78905-220, Rondônia, Brazil

* Correspondence: diascg2019@gmail.com (C.G.D.); fabio.pea@gmail.com (F.H.P.)

Abstract

Corrosion in carbon steel piping systems is a critical issue in hydroelectric power plants, as progressive wall thickness loss may compromise integrity and operational reliability. This study proposes a data-driven framework combining descriptive statistical analysis and distinct clustering methods to map and rank corrosion conditions in water pipeline systems of a hydroelectric power plant in Brazil. A total of 4916 pipe segments from 20 generators (generating units 1 and 2) were evaluated and the feature matrix included wall loss descriptors, measured thickness statistics, minimum allowable thickness, wall integrity indicators, pipe geometry, segment type, and system information. Descriptive analyses revealed heterogeneous corrosion patterns across generating units, segment types, diameters, and inspection points, with localized severe wall loss conditions in specific segments. A principal components analysis was applied to reduce the original feature space while preserving approximately 80% of the cumulative variance, as suggested by the literature. The best clustering solution was obtained using a weighted consensus model based on the Calinski–Harabasz index, resulting in five degradation/integrity profiles that support segment-level corrosion ranking and inspection prioritization. The proposed framework provides a structured basis for integrating inspection data, statistical descriptors, and integrity indicators into maintenance and decision support processes for hydroelectric power plant piping systems.

Keywords: corrosion assessment; hydroelectric plant; carbon steel pipes; clustering techniques



Academic Editor: Alberto Benato

Received: 23 June 2026

Revised: 16 July 2026

Accepted: 21 July 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Pipelines are among the most important assets in civil infrastructure because of their role in transporting materials such as oil, gas, and water [1]. In hydroelectric plants, pipes are widely used in water cooling, draining and filling, and raw water systems.

In ref. [2], for example, the authors evaluated a leakage occurred on the water cooling distribution pipe in a 175 MW hydroelectric power plant. In this work, the pipe, made

of low-carbon steel and meeting the standard of ASTM A595 (ASTM A595/A595M-22: Standard Specification for Steel Tubes) grade A, is used to convey cooling water from the penstock to the heat exchanger.

Nondestructive and metallographic analyses showed that the pipe leakage resulted from localized severe erosion and uniform corrosion on the inner wall, mainly caused by flow turbulence associated with the pipe elbow, for example, which led to progressive wall thinning.

Previous studies, such as those reported by [3,4], investigated systems with abrupt changes in flow direction and demonstrated that mass transfer is directly proportional to the corrosion rate and strongly dependent on the local flow geometry.

Other studies have also investigated the causes of pitting corrosion in stainless steel gate slots of a hydropower station [5]. In addition, other equipment and their associated pipes are commonly present in these systems and may also be affected by corrosion, including heat exchangers, pumps, filters, and other components [6].

In this sense, assessing pipeline integrity is a highly relevant task, particularly when considering the different corrosion patterns and their spatial distribution along the pipes. Among the main steps in pipeline integrity evaluation are defect detection and identification, which enable data collection from pipes through field inspections conducted by technical personnel, in-line inspection, and/or non-destructive evaluation (NDE) [7].

Recent studies have shown that corrosion management should not rely solely on isolated defect measurements, since degradation is typically heterogeneous and influenced by multiple environmental, operational, and pipe-related factors [8].

Other research has addressed corrosion assessment in pipelines not only through defect sizing, but also through the definition of severity levels, probability-of-failure [1,9,10], and risk-based inspection thresholds, reinforcing the importance of stratifying corrosion conditions before selecting mitigation or maintenance actions [11].

The work published by Ismail et al. [12] proposed a pipeline corrosion workflow that combines probabilistic reliability analysis with machine-learning classification based on in-line inspection (ILI) defect data. The core idea is to move beyond manual screening of large ILI datasets by automatically classifying corrosion defects and supporting corrosion-growth or remaining-life analysis.

In ref. [13], the authors highlighted that the manual processing of in-line inspection data is time consuming, labor intensive, and prone to errors, particularly when large datasets and multiple inspection runs must be matched. In ref. [14], for example, the authors proposed a framework for the automated matching of corrosion features obtained from ILI results. This highlights that, before modeling corrosion growth, risk, or mitigation strategies, a critical step lies in the treatment and organization of inspection data, especially when the objective is to track the spatial evolution of defects.

Another recent study, published by Hussain et al. [15], addressed an important gap between ILI data collection and data quality assessment. In that work, different machine learning models were tested for outlier detection in ILI data, including random forest regression, linear regression, and nearest-neighbors-based methods. As pointed out by Hussain et al. [15], data quality in ILI collection should be assessed by considering not only missing values and inconsistent measurements, but also outlier values.

It is worth noting that, in recent years, several machine learning-based methodologies have been proposed not only for corrosion growth evaluation, but also for supporting the overall integrity and condition assessment of steel structures and pipeline systems [16–18].

Therefore, mapping, ranking, and categorizing corrosion conditions before selecting mitigation actions is an important step to support inspection prioritization, risk-based maintenance, and more rational integrity management decisions. Based on this premise, the present study proposes a combined data-driven approach based on descriptive sta-

tistical analysis, principal component analysis (PCA), and consensus clustering to map and rank the corrosion condition of pipeline segments in a hydroelectric power plant. The proposed methodology was applied to more than 4000 pipe segments, integrating wall-loss descriptors, measured thickness statistics, minimum allowable thickness, wall integrity indicators, pipe geometry, segment type, and system information.

In addition, the best clustering configuration was obtained using a weighted consensus model, leading to the identification of five degradation/integrity profiles. The resulting cluster hierarchy demonstrated the potential of the proposed approach to support corrosion ranking and inspection prioritization. The next section provides a detailed description of the statistical and clustering methods used in this study.

2. Statistical Data Treatment and Clustering Methods

This section presents the statistical approach adopted in this study, as well as the main theoretical foundations of the clustering methods used to achieve the most appropriate categorization or ranking of the different corrosion levels observed in the examined water pipelines.

The mathematical concepts employed during the data preparation stage are presented below, considering the measurements collected for each generating unit and each pipeline segment type. This section also includes the tasks related to the exploratory descriptive analysis of the corrosion data for each segment and inspection point.

2.1. Statistical Feature Engineering and Descriptive Analysis

After the restructuring data stage, each row of the dataset represented a specific pipeline segment identified by its generating group (GG), generating unit (Ug), water system (System), pipe tag (Tag), segment label (Segment), diameter (Diam), nominal thickness (Nom_Thick), and wall pipe loss (Loss), as shown in Table 1.

Table 1. Example of measured data for selected pipeline segments and inspection points.

GG	Ug	System	Tag	Segment	Point	Diam.	Nom_Thick	Meas_Thick	Min_Thick	Delta_Thick	Loss
GG1	UG1	AB	14"-AB-0101-A1	HSP1	1	14"	9.52	8.26	3.60548	1.26	13.2353
GG1	UG1	AB	14"-AB-0101-A1	HSP1	2	14"	9.52	7.64	3.60548	1.88	19.7479
GG1	UG1	AB	14"-AB-0101-A1	HSP1	3	14"	9.52	6.31	3.60548	3.21	33.7185
GG1	UG1	AB	14"-AB-0101-A1	HSP1	4	14"	9.52	7.66	3.60548	1.86	19.5378
GG1	UG1	AB	14"-AB-0101-A1	CURVE-1	1	14"	9.52	6.74	3.60548	2.78	29.2017

The measurements collected at the different inspection points of each segment were then used to compute segment-level statistical descriptors. This representation made it possible to summarize the corrosion and thickness conditions of each segment while preserving the spatial variability observed among the inspection points.

In this research, we consider x_{ij} as a value of a given variable measured at inspection point j of pipeline segment i , where $j = 1, 2, \dots, n_i$, and n_i is the number of valid inspection points available for segment i .

For each segment, the arithmetic mean of the pipeline thickness loss, expressed as a percentage, was calculated and this metric provides a global estimate of the average condition of the segment.

As highlighted in previous studies [19,20], the analysis of spatial corrosion patterns, or more broadly the spatial distribution of corrosion, is highly relevant for integrity assessment. Therefore, in the present study, the range between the maximum and minimum corrosion values measured along each pipeline segment was adopted as a simple descriptor of spatial variability, as shown in Equation (1).

$$R_i = \max_{1 \leq j \leq n_i} (x_{ij}) - \min_{1 \leq j \leq n_i} (x_{ij}), \quad (1)$$

where R_i measures the spread between the maximum and minimum observed values.

Two segments with similar mean loss values may still exhibit substantially different corrosion profiles depending on the variability observed among their inspection points.

In addition, σ_i and σ_i^2 quantify the dispersion of the measurements around the mean value for segment i . Higher values indicate greater variability among inspection points, which may be associated with more heterogeneous deterioration patterns.

It should be noted that, since the number of valid inspection points was not necessarily the same for all pipe segments, the statistical descriptors were computed using only the available measurements of each segment. This aspect is particularly important because it explains the possible occurrence of undefined higher-order statistics, such as kurtosis and skewness, in segments with a reduced number of valid points. Thus, kurtosis and skewness were computed in order to characterize the shape of the distribution of the inspection values.

These statistical descriptors were computed for the main variables considered in the study, namely, wall-loss values, measured thickness values, and minimum-allowable thickness values.

Consequently, each pipeline segment was represented by a compact set of features capable of describing not only the average condition of the segment, but also the degree of spatial heterogeneity and the overall shape of the measurement distribution.

In addition to the statistical descriptors, integrity-related features were derived from the measured thickness and minimum-allowable thickness values at the common inspection points. We defined $t_{ij}^{(m)}$ as the measured thickness and $t_{ij}^{(\min)}$ as the minimum-allowable thickness at inspection point j of segment i . According to [21], the assessment of in-service piping systems should consider inspection and minimum thickness requirements.

Thus, the local thickness ratio was defined as shown in Equation (2):

$$\rho_{ij} = \frac{t_{ij}^{(m)}}{t_{ij}^{(\min)}}, \quad (2)$$

which expresses the relative safety level between the measured wall thickness and the minimum acceptable thickness.

From this local ratio, two aggregated features were computed. The mean thickness ratio was defined as Equation (3) and the minimum thickness ratio was defined as Equation (4).

$$\bar{\rho}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \rho_{ij}, \quad (3)$$

$$\rho_i^{\min} = \min_{1 \leq j \leq n_i} (\rho_{ij}). \quad (4)$$

The first metric provides an overall integrity indicator for the segment, whereas the second emphasizes the most critical local condition.

Similarly, the local thickness margin was defined as Equation (5):

$$m_{ij} = t_{ij}^{(m)} - t_{ij}^{(\min)}, \quad (5)$$

from which the average thickness margin and the minimum thickness margin were obtained as Equations (6) and (7), respectively, as follows:

$$\bar{m}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} m_{ij}, \quad (6)$$

$$m_i^{\min} = \min_{1 \leq j \leq n_i} (m_{ij}), \quad (7)$$

These variables directly quantify how far the measured wall thickness is from the minimum allowable limit, both on average and at the most critical inspection point of the segment.

In addition to the corrosion and thickness measurements, contextual variables such as segment type, nominal diameter, and local pressure measurements were preserved in the structured dataset in order to support the interpretation of the statistical and clustering results. Thus, the feature engineering stage produced a representation that combines descriptive statistical information, structural integrity indicators, and contextual attributes of the pipeline segments.

After the feature engineering stage, outlier detection was performed using a multivariate interquartile range (IQR) criterion applied to the loss mean and measured thickness mean. It should be noted that the segments flagged as outliers should not be interpreted as failed components a priori; however, they represent atypical conditions that may indicate increased structural vulnerability and therefore deserve priority in inspection and integrity assessment. In Section 4.3, the outlier cases are presented and discussed.

Therefore, the aforementioned procedure enabled the removal of segments with unusually extreme average corrosion or thickness values before the clustering stage. In the next subsection, we present the main concepts behind the clustering methods used and tested in this research.

2.2. Clustering Methods and Criticality Hierarchy

As discussed by Wang et al. [22], clustering techniques are suitable for classifying pipe segments because segments assigned to the same cluster share similar characteristics in the feature space.

In this work, two groups of clustering algorithms were employed. The first comprises classical methods used both as standalone baselines and as generators of the base partitions required by the consensus framework. The second group consists of the consensus-based methods themselves.

2.2.1. Classical Methods

In this research, the classical clustering methods, such as K-means and agglomerative hierarchical clustering, were also evaluated. The K-means algorithm and its main concepts can be found in the seminal work of MacQueen [23], whereas hierarchical clustering schemes were systematized in the classical work of Johnson [24].

2.2.2. Consensus-Based Methods

As mentioned in the previous section, this study also investigated the use of consensus-based clustering methods. For this purpose, the consensus clustering framework described by Unlu et al. [25] was adopted. In this framework, a pool of c base partitions, $\mathcal{P} = \{P_1, \dots, P_c\}$, is generated from the PCA-transformed dataset $\mathcal{X} = \{x_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^m$, using different clustering algorithms.

Each partition P_l is encoded as a binary membership matrix $\mathbf{H}_l \in \{0, 1\}^{n \times k}$, where $H_l[i, j] = 1$ if segment i belongs to cluster j .

In the standard scheme, all partitions contribute equally, yielding the consensus similarity matrix:

$$\mathbf{S} = \frac{1}{c} \mathbf{H} \mathbf{H}^\top, \quad \mathbf{H} = [\mathbf{H}_1 \mid \dots \mid \mathbf{H}_c], \quad (8)$$

where $S[i, j] \in [0, 1]$ is the fraction of partitions that co-assign segments i and j . In the weighted schemes, each partition is scaled by a quality weight w_l , giving the weighted hypergraph matrix $\mathbf{H}^* = [w_1 \mathbf{H}_1 \mid \cdots \mid w_c \mathbf{H}_c]$ and the weighted similarity matrix (Equation (9)):

$$\hat{\mathbf{S}} = \mathbf{H}^* \mathbf{H}^{*\top}. \quad (9)$$

The normalized weight $w_l = \tilde{w}_l / \sum_{l'} \tilde{w}_{l'}$ is derived from an internal validation index ξ : $\tilde{w}_l = \xi(P_l)$ for indices where higher values are better (Silhouette, Calinski–Harabasz), and $\tilde{w}_l = 1 / (1 + \xi(P_l))$ for the Davies–Bouldin index, where lower values are better. These metrics, commonly referred to as clustering validity indices, are widely used in the literature to assess clustering quality across different application domains [26,27].

Spectral clustering was then applied to $\hat{\mathbf{S}}$ to obtain the final partition. The optimal number of clusters k^* was determined by a composite score that aggregates the three indices via z-score standardization:

$$\phi_k = z_{\text{SH},k} - z_{\text{DB},k} + z_{\text{CH},k}, \quad k^* = \arg \max_k \phi_k, \quad (10)$$

where $z_{\text{SH},k}$, $z_{\text{DB},k}$, and $z_{\text{CH},k}$ denote the z-scores of the Silhouette, Davies–Bouldin, and Calinski–Harabasz indices evaluated for a given value of k , respectively. In the case of tied or similar composite-score values, the consensus index proposed by Vinh et al. [28] was used as a secondary criterion:

$$\text{CI}(P_k) = \frac{1}{\binom{c}{2}} \sum_{l < l'} \text{ARI}(P_l, P_{l'}), \quad (11)$$

which measures the mean pairwise agreement among the base partitions generated for the same candidate value of k . Higher CI values indicate more stable clustering structures. The cluster assignments obtained from this procedure were then used in the subsequent criticality assessment, in which each cluster was ranked according to corrosion severity indicators, as described in more detail in the next section.

3. Methodology for SAE-Based Corrosion Ranking

This section presents the main steps involved not only in pipeline corrosion inspection, but also in the treatment of the measured data and in the corrosion ranking procedure, which is based on a combination of statistical analysis and clustering methods, as shown in Figure 1.

As can be seen, the proposed methodology was structured as a sequential pipeline comprising three major parts: (i) pipe thickness measurement and data collection, (ii) data preparation and statistical analysis, and (iii) clustering, configuration selection, and criticality assessment.

3.1. Thickness Loss Inspection Approach and Field Measurement

This section presents the inspection approach applied to the water pipelines of the Santo Antonio Hydroelectric (SAE) Plant, Brazil, including the raw water (AB), cooling (AR), draining/filling (EE), and drainage systems (DR).

In the present study, wall thickness loss was assessed in the pipelines under investigation, which were manufactured from ASTM A53 carbon steel. Thickness measurements were performed using a B-scan ultrasonic thickness measurement device, model Epoch XT/650/GE DM5E, manufactured by Olympus/GE (Hamburg, Germany). A 12.7 mm diameter transducer was employed during the inspections and the equipment was calibrated to evaluate wall thicknesses up to 20 mm.

The inspections were carried out sequentially across generating units 1 to 5 of the SAE plant, as shown in Figure 2. At this stage of the study, only the wall thickness loss data associated with generating units 1 and 2 were considered, since these were the only datasets available for analysis at the time.

Additionally, the pipe sections considered in this study comprised horizontal straight pipe (HSP) and vertical straight pipe (VSP) segments, as well as fittings/accessories such as tees (TEE), bends/elbows (ELB), reducers (RED), and end caps (CAP), as shown in Figure 3, for pipelines with nominal diameters of 6", 8", 10", 12", 14", and 18".

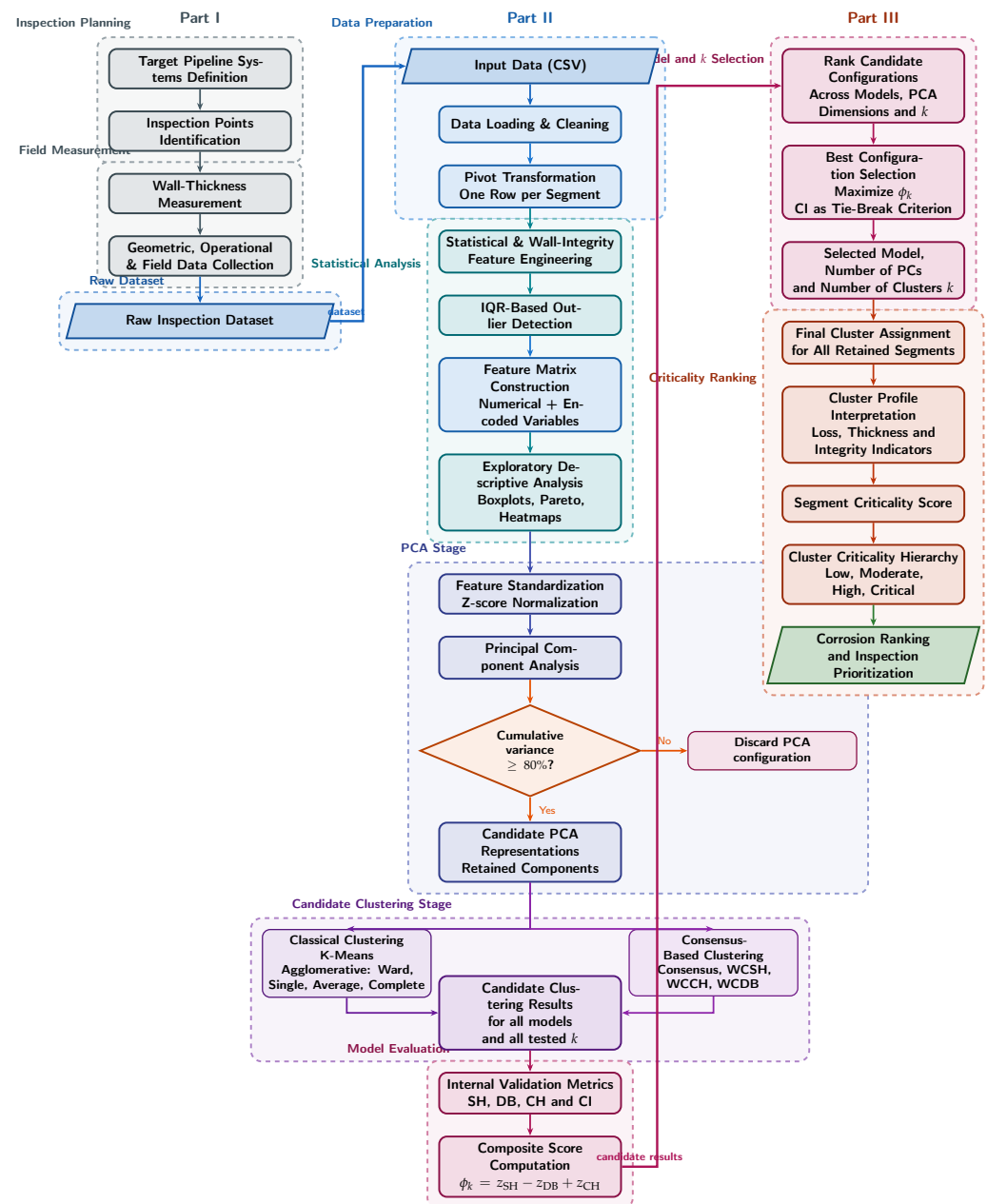


Figure 1. Main steps of the present methodology.

Furthermore, as shown in Figure 3, different inspection points were defined for each type of pipeline segment during the measurement procedure, allowing the evaluation of thickness loss at multiple locations along each geometry.

After the measurement phase for generating units 1 and 2 had been completed (total of 20 generators), the collected data were compiled, organized, and processed according to the methodology proposed and described in the following section.

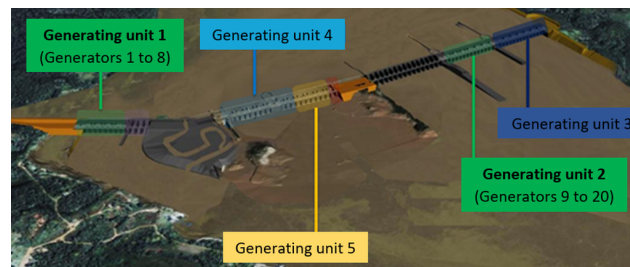


Figure 2. Distribution of the generating units for SAE Plant (adapted from [29]).

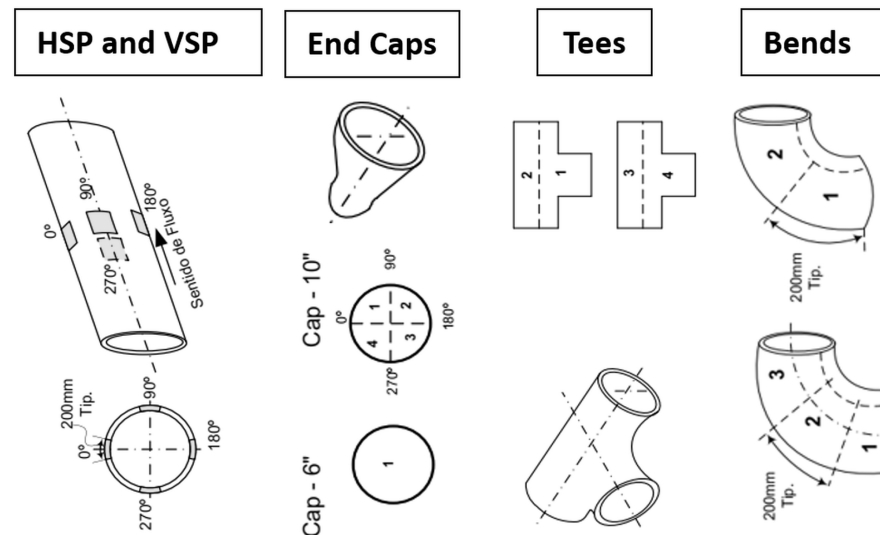


Figure 3. Pipe segments used for thickness loss inspection using B-Scan (adapted from [29]).

3.2. Data Preparation and Descriptive Statistical Analysis

At this stage, the data measured in the field were loaded and cleaned for the post-processing step. Each pipe segment was divided into one row in our dataset. As described in Section 2, statistical feature engineering was carried out in order to extract some metrics from the inspection points available for each segment, such as average condition, range between maximum and minimum corrosion values, standard deviation, variance, kurtosis, and skewness.

More specifically, numerical features were computed to support the initial assessment of pipe corrosion. The mean wall thickness loss calculated for each segment was adopted as the primary criterion for evaluating wall loss severity. The exploratory descriptive analysis was mainly based on selected corrosion-related variables, including mean wall thickness loss, loss variability, pipe diameter, segment type, and generating unit identifiers.

Boxplots, Pareto charts, and localized heatmaps were then used to examine the spatial distribution of corrosion, identify the most critical pipe segments, and support the assessment of corrosion severity at both the segment and generating-unit levels. Prior to this analysis, outliers were removed using an IQR-based criterion, as illustrated in Figure 1.

In addition, it should be noted that outlier removal was performed before PCA using the criterion applied to the mean thickness loss and, when available, the mean measured wall thickness. A segment was removed when at least one of these variables fell outside the interval defined by $Q1 - 2IQR$ and $Q3 + 2IQR$. Outlier screening in this study was based on an interquartile-range fence rule derived from Tukey's approach [30], with a multiplier of 2.0 adopted as a conservative threshold to reduce the influence of highly atypical segments on the clustering stage without excluding an excessive number of observations [31]. In Section 4.3, the outlier segments are discussed in the context of the present study and its objectives.

3.3. Clustering Approach and Criticality Assessment

This section presents the main steps involved in segment clustering and pipe criticality assessment. As shown in Figure 1, PCA was first applied to reduce the dimensionality of the feature space before performing the clustering analysis. However, before applying PCA, the feature matrix was constructed by combining numerical descriptors and encoded categorical variables, as shown in Table 2.

Table 2. Input features used in the PCA and clustering analyses.

Feature Group	Variables
Geometrical	DIAMETER_NUM
Wall-loss	LOSS_MEAN, LOSS_RANGE, LOSS_STD, LOSS_VAR, LOSS_KURTOSIS, LOSS_SKEWNESS
Measured thickness	MEAS_THICKNESS_MEAN, MEAS_THICKNESS_RANGE, MEAS_THICKNESS_STD, MEAS_THICKNESS_VAR, MEAS_THICKNESS_KURTOSIS, MEAS_THICKNESS_SKEWNESS
Minimum allowable thickness	MIN_THICKNESS_MEAN, MIN_THICKNESS_RANGE, MIN_THICKNESS_STD, MIN_THICKNESS_VAR, MIN_THICKNESS_KURTOSIS, MIN_THICKNESS_SKEWNESS
Nominal thickness	NOMINAL_THICKNESS
Wall-integrity	RATIO_MEAS_MIN_MEAN, RATIO_MEAS_MIN_MIN, THICKNESS_MARGIN_MEAN, THICKNESS_MARGIN_MIN, THICKNESS_MARGIN
Segment type encoding	segment_CURVE, segment_HSP, segment_RED, segment_TEE, segment_VSP
System encoding	system_AB, system_AR, system_DR, system_EE

The numerical descriptors included pipe diameter, wall loss statistics, measured-thickness statistics, minimum-allowable thickness statistics, nominal thickness, and wall integrity indicators. In addition, segment type and system information were converted into binary variables using one-hot encoding. As a result, the final input matrix used for PCA comprised 34 variables, including 25 numerical features and 9 encoded categorical features.

As described by Jolliffe [32], in many applications of PCA, the primary objective is to represent a large set of original features by a much smaller number of principal components (PCs), while retaining most of the information present in the data. Although there is no universally accepted minimum threshold for cumulative explained variance in PCA, practical cutoffs in the range of 70% to 90% are often used for retaining principal components, depending on the objective of the analysis [32].

Since these variables have different physical meanings and numerical scales, the complete feature matrix was standardized before PCA using z-score normalization. This step ensured that variables with larger numerical magnitudes, such as thickness-related measurements, did not dominate the computation of the principal components. PCA was then applied to the standardized matrix, and the number of retained components was selected according to the cumulative explained variance criterion. In this study, only PCA configurations retaining at least 80% of the cumulative variance were considered for the subsequent clustering stage.

The resulting principal components were used as the input space for the clustering algorithms. Therefore, clustering was not performed directly on the original variables, but on a reduced and orthogonal representation of the corrosion-related information. This

procedure allowed the clustering stage to consider the combined effect of wall loss severity, spatial variability, pipe geometry, wall integrity condition, segment type, and hydraulic system, while reducing redundancy among correlated features.

It should be noted that, although PCA was applied to a feature matrix combining continuous and one-hot encoded binary variables, we acknowledge that this represents a pragmatic simplification rather than the statistically optimal treatment of mixed-type data. This choice was guided by the exploratory nature of the clustering step, for which linear PCA provided an interpretable and computationally efficient reduction. Given the presence of categorical variables, more specialized techniques such as Factor Analysis of Mixed Data (FAMD) or Multiple Factor Analysis (MFA), which jointly accommodate continuous and categorical variance structures, represent a valuable direction for future refinement of this pipeline.

4. Results and Discussion

This section presents the results obtained from both the descriptive statistical analysis and the clustering methods, with the aim of assessing the corrosion condition and criticality of each water pipeline system in the SAE plant. In total, 4916 pipe segments associated with the 20 generators were evaluated. Prior to the descriptive statistical analysis, PCA, and clustering, an IQR-based outlier detection procedure was applied, resulting in the exclusion of only 21 segments, which represents approximately 0.43% of the initial dataset. The remaining 4895 segments were retained for the descriptive and multivariate analyses, as shown in the following subsections.

4.1. Statistical Descriptive Results

As described, the first analysis was carried out using descriptive statistical metrics. Figure 4 shows the distribution of mean wall thickness loss for each generating unit (UG). As a first pipeline severity evaluation, we defined three corrosion degradation zones. The background bands indicate the severity zones: Zone 1, corresponding to losses below 15%, Zone 2, corresponding to losses between 15% and 30%, and Zone 3, corresponding to losses above 30%. Overall, most generators present median values within Zone 1, indicating that the central tendency of wall loss remains below the moderate severity threshold for most UGs.

However, several generators show wide interquartile ranges, long upper whiskers, and isolated high loss observations extending into Zones 2 and 3. This behavior suggests that, although the average degradation level is generally moderate to low, localized critical conditions may exist in specific pipe segments or inspection points. In particular, the presence of outliers above 30% indicates localized severe wall loss, which was further investigated in the subsequent criticality and clustering analyses (Section 4.2).

As can be seen, the boxplot also reveals differences in the dispersion of wall loss values among generators. Some UGs exhibit more compact distributions, suggesting more homogeneous corrosion behavior, whereas others show greater variability, indicating a more heterogeneous spatial distribution of wall loss, such as the case of UG8, for example. Therefore, this analysis supports the identification of generating units with higher corrosion variability and helps guide the prioritization of pipe segments for detailed integrity assessment.

It should be noted that the data associated with the group labeled UG1–UG8, highlighted in Figure 4, refer to a specific pipeline that serves all eight generating units simultaneously. In particular, this pipeline corresponds to the total downstream draft-tube dewatering line, which is used for the complete draining of the pit. Therefore, unlike the other boxplots, which represent measurements associated with individual generating units, the UG1–UG8 group represents a shared hydraulic structure encompassing the eight units.

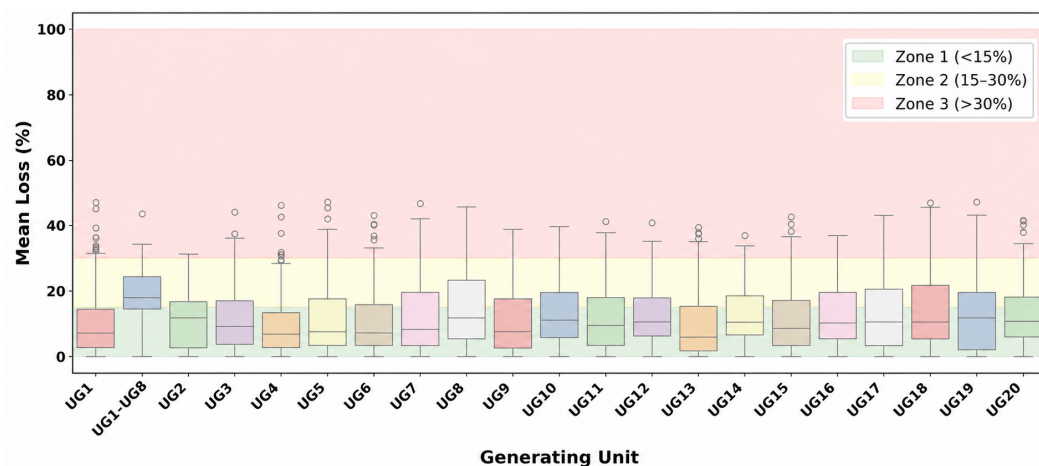


Figure 4. Wear severity by generating unit (UG).

Table 3 complements the boxplot analysis by quantifying the distribution of mean wall thickness loss for each generating unit. In general, most UGs presented mean loss values between approximately 9% and 14%, while the median values were generally lower than the corresponding means. This behavior indicates right-skewed distributions, in which a limited number of segments with higher wall loss increase the average loss values.

Table 3. Descriptive statistics of mean wall-thickness loss by generating group (GG) and generating unit (UG).

GG	UG	Mean	Median	Std	Q1	Q3
GG1	UG1	10.009	7.060	9.682	2.663	14.387
GG1	UG2	11.020	11.812	8.308	2.670	16.712
GG1	UG3	11.063	9.137	9.044	3.642	16.923
GG1	UG4	9.286	6.754	8.908	2.813	13.300
GG1	UG5	11.090	7.551	9.999	3.344	17.522
GG1	UG6	10.590	7.173	9.727	3.328	15.797
GG1	UG7	11.946	8.206	10.469	3.344	19.652
GG1	UG8	14.811	11.736	11.215	5.440	23.207
GG2	UG9	10.738	7.565	9.906	2.555	17.599
GG2	UG10	13.238	11.045	9.504	5.798	19.478
GG2	UG11	12.036	9.420	10.266	3.474	17.974
GG2	UG12	12.207	10.491	8.538	6.226	17.826
GG2	UG13	9.467	5.825	9.636	1.746	15.307
GG2	UG14	12.568	10.417	8.390	6.587	18.480
GG2	UG15	11.250	8.567	9.954	3.303	17.096
GG2	UG16	12.729	10.194	9.050	5.407	19.531
GG2	UG17	12.679	10.595	10.478	3.344	20.460
GG2	UG18	13.740	10.513	10.623	5.313	21.656
GG2	UG19	12.706	11.736	11.100	2.050	19.560
GG2	UG20	12.613	10.761	9.075	5.996	18.096

Among the generators, UG8 showed the highest mean loss in GG1, with a mean value of 14.811% and an upper quartile of 23.207%. In GG2, the highest mean values were observed for UG10, UG16, UG17, UG18, and UG19, all with mean losses above 12.6%.

The standard deviation values, mostly between 8% and 11%, also reveal considerable variability among pipe segments within the same generating unit. Therefore, the corrosion condition cannot be fully characterized by the mean loss alone. The combination of mean, median, quartiles, and maximum values suggests that some UGs present generally moderate degradation, while others contain localized critical segments that require further

investigation through the subsequent clustering-based criticality analysis, as shown in Section 2.

In addition, Figure 5 shows the distribution of mean wall-thickness loss by segment type and pipe diameter. The boxplot by segment type shows that corrosion severity is not uniformly distributed across the evaluated pipeline geometries. Among the analyzed categories, CURVE segments exhibited the highest median mean loss, indicating that this geometry tends to concentrate more severe degradation levels than the others.

In addition, CURVE, HSP, and VSP segments showed relatively large interquartile ranges and several upper outliers, suggesting that these segment types are not only more affected on average, but also more heterogeneous in terms of corrosion behavior. By contrast, RED and TEE segments generally presented lower median losses, although isolated severe cases were still observed.

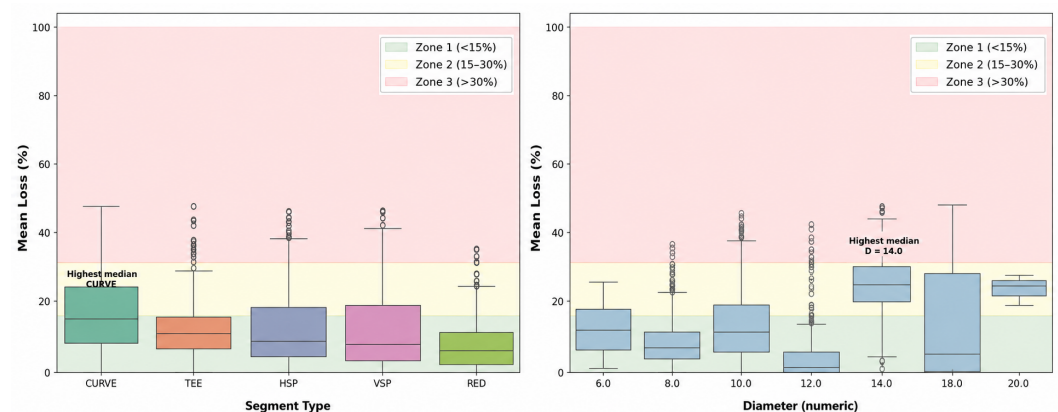


Figure 5. Mean loss distribution by segment type and diameter.

A complementary evaluation is provided by the boxplot by diameter, which indicates that the 14-inch pipelines displayed the highest median mean loss among the evaluated nominal diameters. This result suggests that, in the present dataset, this diameter class may be associated with less favorable operating or geometric conditions from the corrosion standpoint.

Nevertheless, other diameters, such as 10 and 18 inches, also exhibited substantial dispersion and several upper outliers, indicating that severe degradation is not restricted to a single diameter category. In our study, these two figures suggest that both segment geometry and pipe diameter influence the observed corrosion patterns, with geometry appearing as a particularly relevant factor in explaining the concentration and variability of wall-thickness loss.

As shown in Figure 6, the Pareto graph highlights that HSP, CURVE, and VSP are the main contributors to the accumulated wall thickness loss, with HSP and CURVE alone accounting for most of the total contribution (almost 80% of the total mean loss). The segments TEE and RED presented comparatively smaller contributions. This result suggests that corrosion degradation is concentrated in specific segment types, supporting the prioritization of HSP, CURVE, and VSP segments in the subsequent criticality assessment. In Figure 7, for example, a corrosion case is illustrated in a HSP segment.

It should be noted that, the study reported by [11], for example, also defined three severity levels for pipeline corrosion. However, as discussed in the next subsection, the clustering based criticality analysis proposed in the present study allowed the identification of different degradation profiles among pipe segments, resulting in a data-driven classification that was not restricted to three predefined severity levels.

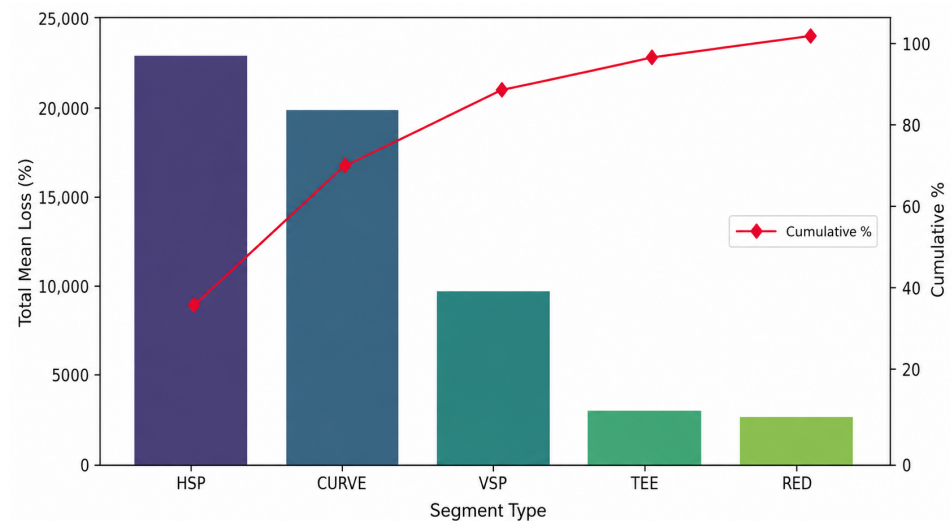


Figure 6. Pareto chart: Loss contribution by segment type.



Figure 7. Corrosion case found in the HSP segment.

4.2. Clustering Results and Corrosion Ranking

As described in Section 2, before applying the clustering methods, in the present study, for each candidate value of k , several base partitions were generated from the PCA-transformed data. As shown in Figure 8, the cumulative explained variance increased progressively with the number of retained principal components (varying from 3 to 12). However, because the first configuration exceeding 80% cumulative explained variance was obtained with nine principal components, only PCA configurations with nine or more components were considered in the clustering stage.

After applying PCA, different clustering configurations were evaluated for both classical and consensus clustering methods. The best configuration was then selected based on the combined score, as described in Section 2 and defined in Equation (10). As can be seen in Figure 9 and Table 4, the WCCH model achieved the highest combined score, which indicates that the weighted consensus clustering approach based on the Calinski–Harabasz (WCCH) index provided the best clustering configuration among the evaluated methods.

This composite score was applied uniformly across all candidate models, using nine principal components, both classical and consensus-based, and all values of k , varying from 3 to 8, so that the final configuration (model and number of clusters) was jointly determined by the highest ϕ_k , as shown in Table 4. Thus, in our study, the best case was obtained using nine principal components, five clusters ($k = 5$) and the WCCH method.

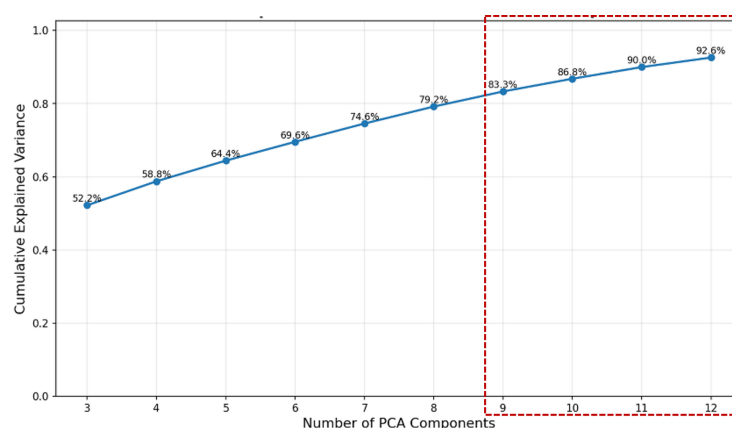


Figure 8. Cumulative explained variance vs. number of PCA components.

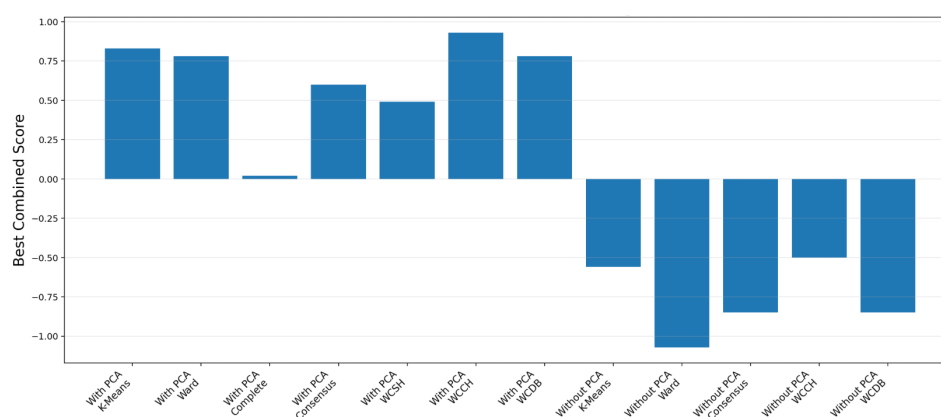


Figure 9. Best combined clustering score by model.

In comparison, the PCA-based models generally achieved better results than the corresponding solutions obtained without dimensionality reduction. Among them, WCCH presented the highest combined score, followed by K-Means and Ward, while WCDB, Consensus, and WCH also showed positive and competitive performance. In contrast, most models without PCA yielded negative combined scores, indicating a poorer balance among cohesion, separation, and global partition quality in the original feature space.

Table 4. Combined scores for clustering approaches with and without PCA.

Approach	Model	Combined Score
With PCA	K-Means	0.8379
With PCA	Ward	0.7835
With PCA	Complete	0.0243
With PCA	Consensus	0.5994
With PCA	WCH	0.4873
With PCA	WCCH	0.9309
With PCA	WCDB	0.7779
Without PCA	K-Means	−0.5588
Without PCA	Ward	−1.0663
Without PCA	Consensus	−0.8533
Without PCA	WCCH	−0.4960
Without PCA	WCDB	−0.8533

Table 5 summarizes the main parameter settings adopted in the dimensionality reduction and clustering stages, including the tested PCA configurations, the cluster range, the

evaluated methods, and the final selected solution. This table is intended to improve the clarity and reproducibility of the proposed methodology.

Figure 10 shows the projection of the final clustering solution onto the first two principal components. Although the best WCCH model was obtained using nine principal components, the two-dimensional projection already reveals a good organization of the clusters in the PCA space.

Table 5. Main parameter settings adopted in the PCA and clustering stages.

Parameter/Setting	Value	Description
PCA candidate range	3–12	Tested number of retained principal components.
Minimum cumulative explained variance	80%	Threshold adopted to retain PCA configurations for clustering.
PCA configurations used in clustering	9–12	Only configurations exceeding 80% cumulative explained variance were considered.
Selected PCA configuration	9	First configuration exceeding the 80% threshold; used in the best final solution (Figure 8).
Cluster number range (k)	3–8	Candidate number of clusters tested for all clustering models.
Classical clustering methods	5 methods	K-Means, Ward, Average, Complete, and Single.
Consensus-based methods	4 methods	Consensus, WCSH, WCCH, and WCDB.
Best clustering method	WCCH	Weighted consensus clustering based on the Calinski–Harabasz index.
Best number of clusters	$k = 5$	Number of clusters associated with the highest combined score.
Model selection criterion	ϕ_k	Combined score defined in Equation (10), used to rank candidate solutions.

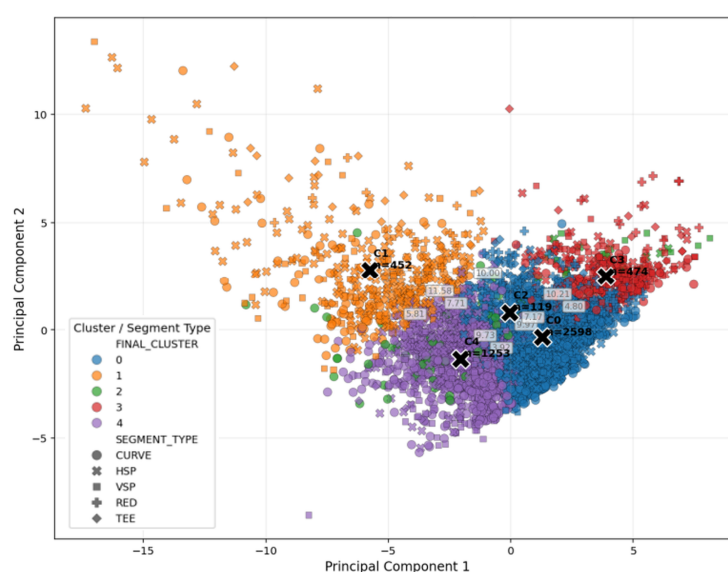


Figure 10. Final cluster distribution in PCA space for best configuration (WCCH/PCA = 9/ k = 5).

Clusters C1 and C3, for example, occupy more distinct regions, while C0, C2, and C4 are located in intermediate areas with partial overlap. This behavior indicates that relevant differences among the degradation/integrity profiles can already be observed in the first two principal components, even though the final clustering was performed in the nine-dimensional PCA space. The centroid positions also support the separation among the identified groups. Cluster C1 is mainly distributed toward negative values of the first principal component, whereas C3 is concentrated toward positive values of both principal components. Cluster C0 represents the largest group (with 2498 pipe segments) and occupies a central-to-lower region of the PCA plane, while C2 and C4 appear as intermediate profiles. Therefore, the scatter plot provides visual evidence that the proposed PCA-based weighted consensus clustering approach produced distinguishable corrosion/integrity groups.

An additional insight provided by the PCA scatter plot is that the internal dispersion of the clusters is not uniform. In particular, Cluster C1 exhibits a much broader spread, especially along the negative direction of the first principal component, which suggests a more heterogeneous degradation/integrity profile. By contrast, Cluster C3 appears more compact and concentrated in the positive region of both principal components, indicating a more cohesive and well-defined group in the reduced feature space. This difference in compactness reinforces that the clustering solution captured not only distinct group locations, but also different levels of internal variability among the identified corrosion/integrity profiles.

Another important observation is that the visual separation among the groups seems to be driven mainly by the first principal component, whereas the second component plays a complementary role in refining the distribution of the intermediate clusters. Moreover, the partial overlap observed among C0, C2, and C4 suggests the existence of gradual transitions between some degradation states rather than completely rigid boundaries. From an engineering perspective, this behavior is consistent with the continuous nature of corrosion processes, in which intermediate conditions may arise between more clearly separated profiles. Therefore, the scatter plot supports the interpretation of the final clustering solution as a hierarchy of distinguishable but physically connected corrosion or integrity states.

In Figure 11, the cluster profile analysis shows that the WCCH model separated the pipe segments into distinct degradation and integrity profiles. Cluster C1 exhibited the highest mean wall-thickness loss, together with the largest loss range and standard deviation, indicating severe and heterogeneous degradation. In a similar case, the Cluster C4 also showed high mean loss, although with lower variability than C1.

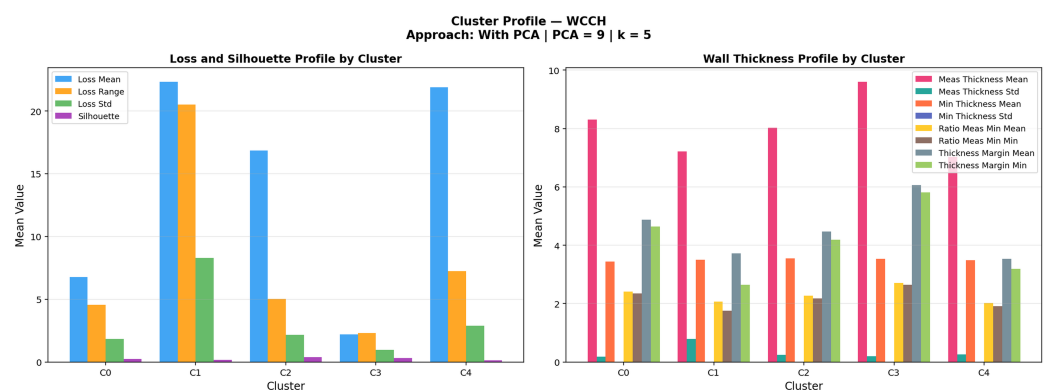


Figure 11. Loss, silhouette, and wall thickness profile by cluster (WCCH/PCA = 9/k = 5).

In contrast, C3 presented the lowest loss-related indicators and the highest wall-thickness margins, suggesting the most favorable integrity condition. Therefore, the cluster

profiles confirm that the proposed clustering approach captured not only differences in average wall loss, but also differences in spatial variability and remaining wall-thickness margin among the inspected segments.

Another type of analysis, as depicted in Figure 12, shows the mean criticality score for each cluster obtained from the final clustering solution. The results indicate a clear criticality hierarchy among the identified groups, with C1 presenting the highest mean score, followed by C4. Both clusters were classified as high-criticality groups, suggesting that they represent the most severe degradation profiles among the inspected pipe segments. Clusters C2 and C0 showed moderate criticality levels, whereas C3 presented the lowest mean score and was classified as a low-criticality group. These findings confirm that the present clustering-based approach was able to distinguish pipe segments according to different levels of corrosion severity and wall-integrity condition.

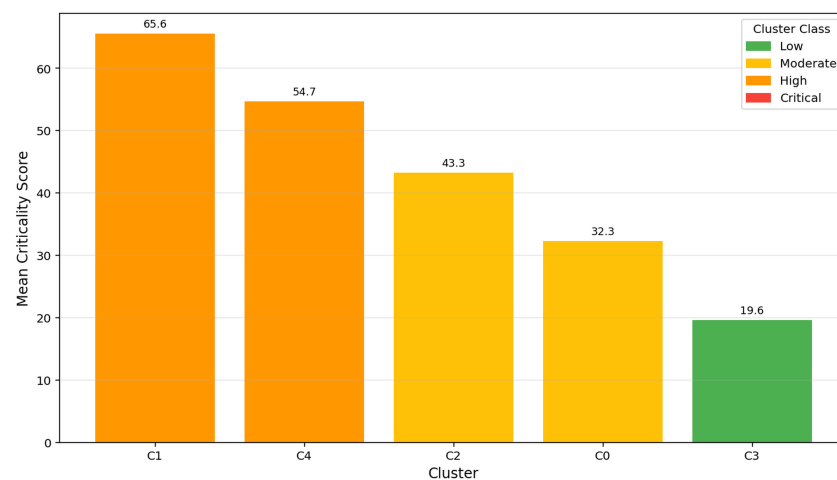


Figure 12. Mean criticality level by cluster.

In Figure 13, we present another kind of evaluation highlighting the distribution of segment criticality scores within each cluster. The results confirm the criticality hierarchy previously observed from the mean scores, with C1 and C4 classified as high-criticality clusters, C2 and C0 as moderate-criticality clusters, and C3 as a low-criticality cluster.

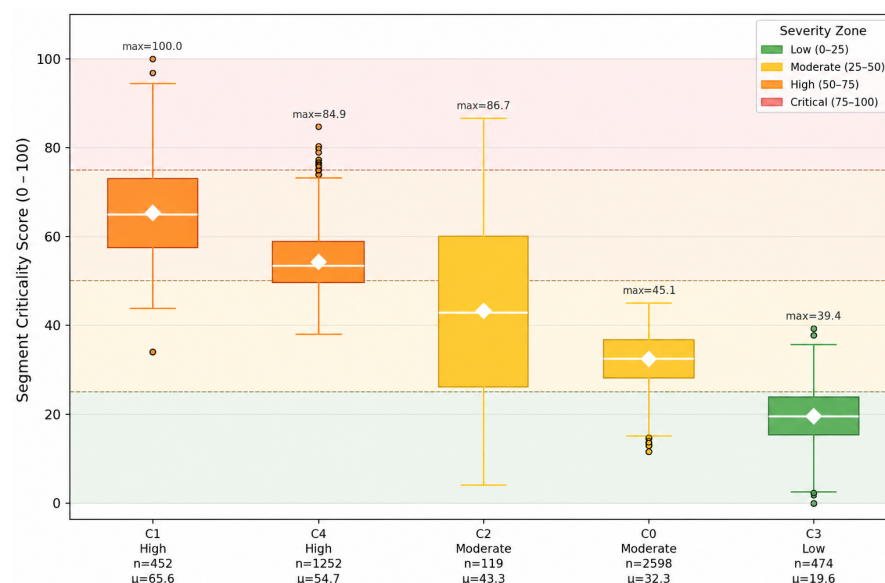


Figure 13. Distribution of segment criticality score by cluster.

Cluster C1 presented the highest mean score and included the most severe cases, with criticality values reaching 100. On the other hand, C3 showed the lowest mean score and the most favorable integrity condition. Although C2 was classified as moderate on average, it exhibited a wide dispersion and included some high-criticality segments, indicating a heterogeneous degradation profile. These findings show that the clustering-based approach not only ranked the clusters by average criticality but also revealed the variability of segment-level severity within each group.

Finally, the results shown in Table 6 reinforce the practical relevance of the proposed methodology. Rather than providing only a global corrosion indicator for the plant, the proposed approach assigns each pipe segment to a final cluster and associates it with an individual criticality score. Since the corresponding pipe TAGs are preserved throughout the analysis, maintenance professionals can directly identify the specific sections that require higher inspection priority, particularly those classified in high-criticality clusters and presenting criticality scores above 90. This is especially relevant considering the approximately 18 km of pipelines subject to inspection in the hydroelectric power plant. Therefore, the final ranking provides a direct link between the multivariate clustering results and actionable maintenance decisions, supporting a more objective and data-driven corrosion management strategy.

Table 6. Top-five most critical pipe segments in each final cluster.

GG	UG	System	Tag Pipe	Segment	Final Cluster	Criticality Score
GG2	UG11	EE	10" EE-1108A1	VSP1	3	32.7325
GG1	UG1-UG8	EE	14" EE-0807A1	VSP1	3	33.2682
GG1	UG3	EE	12" EE-0303A1	CURVE-1	3	33.7625
GG2	UG18	EE	12" EE-1803A1	CURVE-4	3	34.4085
GG2	UG17	EE	12" EE-1703A1	CURVE-4	3	34.5514
GG2	UG10	AR	10" AR-1007A1	HSP7	0	44.2044
GG2	UG20	AR	10" AR-2005A1	HSP5	0	44.2077
GG2	UG13	AR	18" AR-1301A1	HSP4	0	44.2407
GG1	UG6	AR	10" AR-0605A1	HSP5	0	44.3278
GG2	UG14	AR	10" AR-1407A1	HSP4	0	44.3812
GG2	UG13	AR	14" AR-1303A1	CURVE-1	2	76.1691
GG2	UG17	AR	14" AR-1701A1	CURVE-1	4	76.4239
GG1	UG8	DR	8" DR-0844A1	CURVE-6	2	76.8853
GG2	UG11	AR	14" AR-1101A1	CURVE-2	2	77.1359
GG1	UG6	AB	14" AB-0601A1	CURVE-4	2	78.3890
GG2	UG18	AR	10" AR-1805A1	CURVE-5	4	78.8182
GG1	UG7	AR	14" AR-0702A1	CURVE-1	2	79.1199
GG2	UG19	AR	18" AR-1901A1	TEE1	4	79.8405
GG1	UG1-UG8	EE	14" EE-0808A1	CURVE-2	4	80.4071
GG1	UG1	AR	14" AR-0103A1	CURVE-1	4	84.9335
GG1	UG8	AR	14" AR-0802A1	VSP2	1	93.8585
GG1	UG7	AR	10" AR-0705A1	CURVE-1	1	94.1364
GG1	UG8	AR	10" AR-0806A1	HSP1	1	94.4717
GG1	UG7	AB	14" AB-0701A1	VSP1	1	96.8575
GG1	UG5	AB	14" AB-0501A1	HSP9	1	100.0000

4.3. Discussion of Outlier Segments

In this section, an additional discussion of the outlier segments is provided. As mentioned earlier, after processing the raw data during the feature engineering stage, the outlier cases were separated, resulting in 21 pipeline segments. The analysis of these segments, flagged as outliers, provides additional engineering insight beyond the clustering results.

As previously described, a segment was removed whenever at least one of the variables, namely, mean thickness loss, or mean measured wall thickness, fell outside the interval defined by $Q1 - 2IQR$ and $Q3 + 2IQR$, as discussed in Section 3.2.

As shown in Figure 14, the outlier segments ranked by mean loss exhibited markedly elevated degradation levels, with several segments presenting average loss values close to or above 50%, indicating highly atypical corrosion conditions when compared with the remaining dataset. In particular, the most severe outliers were mainly associated with CURVE segments, suggesting that these geometries may be more susceptible to localized or intensified degradation under the operating conditions considered in this study. Nevertheless, some outlier segments did not present measurable thickness loss, which indicates that the outlier detection was based on a multivariate criterion rather than on mean loss alone.

Although the outlier ranking by mean loss highlights segments with atypically high degradation values, the comparison between measured and minimum-allowable thickness shows that many of these cases still remain above the structural limit. In some segments, the measured thickness is even higher than the nominal thickness, indicating that these outliers should not be interpreted a priori as imminent failure cases, but rather as atypical conditions requiring individual engineering assessment.

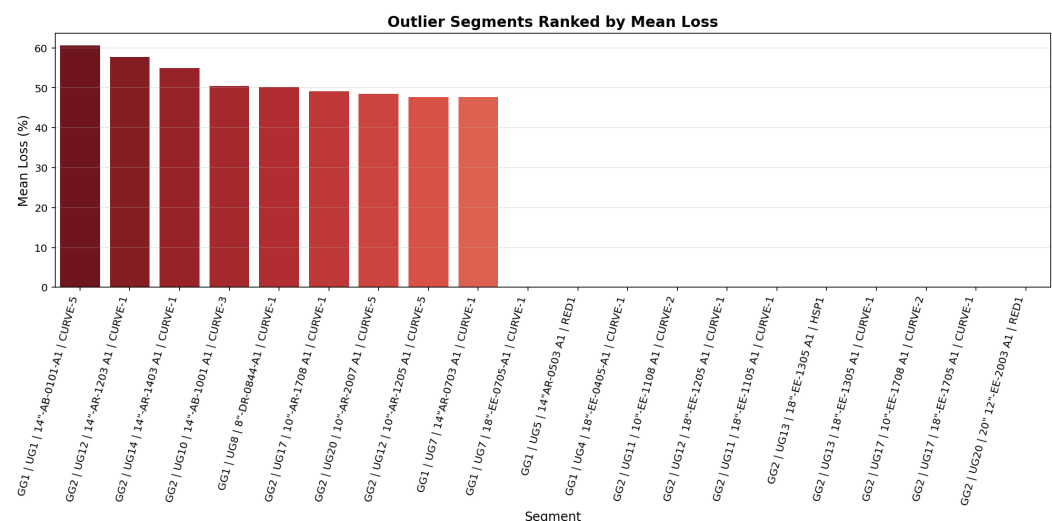


Figure 14. Outlier pipe segments split by mean loss.

Thus, while Figure 14 emphasizes the severity of the outlier cases in terms of mean loss, Figure 15 shows that their structural interpretation must be complemented by thickness-based indicators, since high loss values do not necessarily imply that the segment is close to the minimum-allowable condition.

In other words, Figure 15 shows that the outlier segments do not represent a uniform group in terms of structural condition. While some segments exhibit measured thickness values only slightly above the minimum-allowable thickness, suggesting reduced integrity margins, others remain well above the admissible limit and therefore cannot be interpreted as imminent failure cases. This result indicates that the statistical identification of outliers should not be directly equated with critical structural condition, but rather interpreted as evidence of atypical behavior requiring individual engineering assessment.

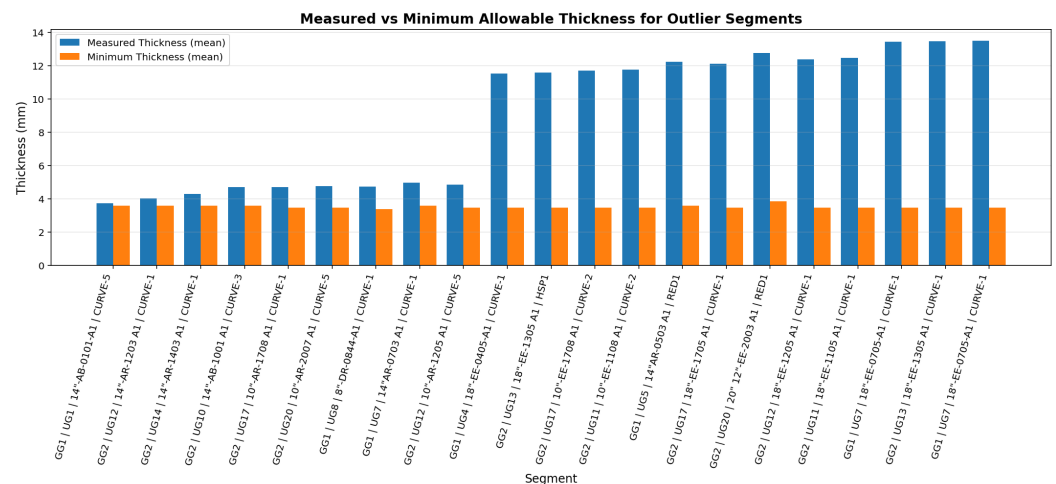


Figure 15. Measured and minimum-allowable thickness for outliers cases.

5. Conclusions

This study proposed a data-driven framework combining descriptive statistical analysis, PCA, and clustering methods to support corrosion mapping and criticality ranking of water pipeline systems in a hydroelectric power plant. A total of 4916 pipe segments associated with 20 generators (generating units 1 and 2) were initially evaluated.

The descriptive statistical analysis showed that wall thickness loss was not uniformly distributed across the plant. Differences were observed among generating units, segment types, pipe diameters, and inspection points. Although most segments presented median losses within the low severity range, localized severe wall loss conditions were identified in specific segments. Pareto analyses indicated that HSP, CURVE, and VSP segments were the main contributors to the accumulated wall-thickness loss, reinforcing the importance of considering segment geometry in inspection prioritization.

The PCA step reduced the dimensionality of the original feature matrix while preserving most of the information contained in the data. Nine principal components retained approximately 83.3% of the cumulative explained variance and were therefore used as the reduced feature space for clustering. This representation allowed the clustering stage to integrate wall loss descriptors, wall thickness statistics, wall integrity indicators, pipe geometry, segment type, and system information in a compact and less redundant form.

Among the evaluated clustering configurations, the best result was obtained using the WCCH weighted consensus model, with nine principal components and five clusters. This result indicates that weighting the base partitions according to the Calinski–Harabasz index improved the quality of the final consensus partition. The resulting cluster hierarchy identified two high-criticality groups, two moderate-criticality groups, and one low-criticality group. In particular, clusters C1 and C4 represented the most critical degradation/integrity profiles, whereas C3 corresponded to the most favorable condition.

Finally, the results demonstrated that the proposed PCA-based weighted consensus clustering approach can distinguish different corrosion and integrity profiles among pipe segments. The methodology provides a structured basis for corrosion ranking, inspection prioritization, and maintenance planning, supporting more rational integrity management decisions in hydroelectric power plant piping systems.

Future studies should evaluate the proposed approach using data from all five generating units, corresponding to a total of 50 generators.

Author Contributions: C.G.D. designed the study, developed the methodology, performed the experiments, analyzed the data and wrote the manuscript. F.H.P. contributed to the methodology evaluation, data interpretation and wrote de manuscript. C.A.M., G.F.M.d.S., I.E.C., I.V.A. and S.I.N. contributed data interpretation and revision of the manuscript. A.d.S.P. and F.H.H. contributed to manuscript revision. All authors contributed to the discussion of the results and approved the final version of the manuscript.

Funding: This research received funding from Axia Energia S.A., project number 2361.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from company Axia Energia S.A and are available with the permission of the company.

Acknowledgments: Fabio Henrique Pereira is supported by the National Council for Scientific and Technological Development/Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil under the grant #313369/2023-2. Idalina Vieira Aoki is also grateful for the support provided by the CNPq through grant 308564/2023-5.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Cleber Gustavo Dias and Fabio Henrique Pereira are employees of Acurai—Pesquisa, Desenvolvimento e Inovação Ltda; Autharis da Silva Peixoto and Fernando Hiroyuki Hamaji are employees of Axia Energia S.A. The authors declare that this study received funding from Axia Energia S.A. The funder had no role in the design of the study; in the collection, analysis, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Cui, B.; Wang, H. Pipeline corrosion prediction and uncertainty analysis with an ensemble Bayesian neural network approach. *Process Saf. Environ. Prot.* **2024**, *187*, 483–494. [[CrossRef](#)]
2. Nurbanasari, M.; Purwanto, T.S.; Piliang, R.S.; Kristyadi, T.; Saefudin, E.; Irwan, Y. Leakage on Water Cooling Distribution Pipe in a Hydroelectric Power Plant. *J. Fail. Anal. Prev.* **2019**, *19*, 412–417. [[CrossRef](#)]
3. Irshad, H.M.; Toor, I.U.; Badr, H.M.; Samad, M.A. Evaluating the Flow Accelerated Corrosion and Erosion–Corrosion Behavior of a Pipeline Grade Carbon Steel (AISI 1030) for Sustainable Operations. *Sustainability* **2022**, *14*, 4819. [[CrossRef](#)]
4. Wang, J.; Shirazi, S.A. A CFD based correlation for mass transfer coefficient in elbows. *Int. J. Heat Mass Transf.* **2001**, *44*, 1817–1822. [[CrossRef](#)]
5. Zhu, Y.; Dong, Y.; Teng, K.; Xu, Y.; Wang, H.; Zhang, Q. Analysis of the causes of pitting corrosion in the 304 stainless steel gate slot of a hydropower station. *Eng. Fail. Anal.* **2025**, *170*, 109268. [[CrossRef](#)]
6. An, J.; Huang, G.; Yao, Y.; Hu, Y.; Wu, J. Scaling and corrosion analysis of stainless steel heat exchange tubes in Jinsha River hydropower station. *Eng. Fail. Anal.* **2026**, *186*, 110533. [[CrossRef](#)]
7. Xie, M.; Tian, Z. A review on pipeline integrity management utilizing in-line inspection data. *Eng. Fail. Anal.* **2018**, *92*, 222–239. [[CrossRef](#)]
8. Hussein Farh, H.M.; Ben Seghier, M.E.A.; Taiwo, R.; Zayed, T. Analysis and ranking of corrosion causes for water pipelines: A critical review. *npj Clean Water* **2023**, *6*, 65. [[CrossRef](#)]
9. de Siqueira Motta, R.; Marques Ferreira, A.D.; Afonso, S.M. Probabilistic assessment of complex corrosion in pipelines considering River-Bottom Profile information. *Eng. Fail. Anal.* **2024**, *165*, 108801. [[CrossRef](#)]
10. Ruggeri, F.; Cagno, E.; Caron, F.; Mancini, M.; Pievatolo, A. Bayesian Assessment of Corrosion-Related Failures in Steel Pipelines. *Entropy* **2024**, *26*, 1111. [[CrossRef](#)] [[PubMed](#)]
11. Cui, B.; Wang, H. Analysis and prediction of pipeline corrosion defects based on data analytics of in-line inspection. *J. Infrastruct. Preserv. Resil.* **2023**, *4*, 14. [[CrossRef](#)]
12. Ismail, M.; May, Z.; Asirvadam, V.; Nayan, N. Machine-Learning-Based Classification for Pipeline Corrosion with Monte Carlo Probabilistic Analysis. *Energies* **2023**, *16*, 3589. [[CrossRef](#)]
13. Liu, H.; Liu, Z.; Taylor, B.; Dong, H. Matching pipeline In-line inspection data for corrosion characterization. *NDT E Int.* **2019**, *101*, 44–52. [[CrossRef](#)]

14. Dann, M.R.; Dann, C. Automated matching of pipeline corrosion features from in-line inspection data. *Reliab. Eng. Syst. Saf.* **2017**, *162*, 40–50. [\[CrossRef\]](#)
15. Hussain, M.; Zhang, T. Machine learning-based outlier detection for pipeline in-line inspection data. *Reliab. Eng. Syst. Saf.* **2025**, *254*, 110553. [\[CrossRef\]](#)
16. Coelho, L.B.; Zhang, D.; Van Ingelgem, Y.; Steckelmacher, D.; Nowé, A.; Terry, H. Reviewing Machine Learning of Corrosion Prediction in a Data-Oriented Perspective. *npj Mater. Degrad.* **2022**, *6*, 8. [\[CrossRef\]](#)
17. Huang, X.; Duan, Z.; Hao, S.; Hou, J.; Chen, W.; Cai, L. A Deep Learning Framework for Corrosion Assessment of Steel Structures Using Inception v3 Model. *Buildings* **2025**, *15*, 512. [\[CrossRef\]](#)
18. Wang, Q.; Wu, D.; Zhou, K.; Gong, H.; Jin, S.; Wang, J.; Chen, Z.; Ma, L.; Zhang, D. Quantification and Prediction of Carbon Steel Surface Corrosion Based on Machine Vision. *npj Mater. Degrad.* **2026**, *10*, 17. [\[CrossRef\]](#)
19. Amaya-Gómez, R.; Sánchez-Silva, M.; Muñoz, F.; Schoefs, F.; Bastidas-Arteaga, E. Spatial characterization and simulation of new defects in corroded pipeline based on In-Line Inspections. *Reliab. Eng. Syst. Saf.* **2024**, *241*, 109697. [\[CrossRef\]](#)
20. Dhungana, H.; Øvsthus, K.; Kampen, A.L. A measurement-driven corrosion progression model for maintenance planning in subsea pipelines. *Measurement* **2026**, *258*, 119263. [\[CrossRef\]](#)
21. *API Standard 570*; API 570: Piping Inspection Code—In-Service Inspection, Rating, Repair, and Alteration of Piping Systems. American Petroleum Institute: Washington, DC, USA, 2024.
22. Wang, H.; Yajima, A.; Liang, R.Y.; Castaneda, H. Bayesian Modeling of External Corrosion in Underground Pipelines Based on the Integration of Markov Chain Monte Carlo Techniques and Clustered Inspection Data. *Comput.-Aided Civ. Infrastruct. Eng.* **2015**, *30*, 300–316. [\[CrossRef\]](#)
23. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June–18 July 1967; Volume 1, pp. 281–297.
24. Johnson, S.C. Hierarchical Clustering Schemes. *Psychometrika* **1967**, *32*, 241–254. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Ünlü, R.; Xanthopoulos, P. Estimating the number of clusters in a dataset via consensus clustering. *Expert Syst. Appl.* **2019**, *125*, 33–39. [\[CrossRef\]](#)
26. Ekemeyong Awong, L.E.; Zielinska, T. Comparative Analysis of the Clustering Quality in Self-Organizing Maps for Human Posture Classification. *Sensors* **2023**, *23*, 7925. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Todeschini, R.; Ballabio, D.; Termopoli, V.; Consonni, V. Extended Multivariate Comparison of 68 Cluster Validity Indices. A Review. *Chemom. Intell. Lab. Syst.* **2024**, *251*, 105117. [\[CrossRef\]](#)
28. Vinh, N.X.; Epps, J. A Novel Approach for Automatic Number of Clusters Detection in Microarray Data Based on Consensus Clustering. In Proceedings of the 2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering, Taichung, Taiwan, China, 22–24 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 84–91. [\[CrossRef\]](#)
29. Almeida, V.; Mentz, M. *Relatório de Inspeção - Medição de Espessura (Casa de Força 1)*; Technical Inspection Report No. ME111277/25; Prepared for Santo Antonio Hydroelectric Plant; MKS Company: Porto Alegre, Brazil, 2025.
30. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley: Reading, MA, USA, 1977.
31. Schwertman, N.C.; Owens, M.A.; Adnan, R. A simple more general boxplot method for identifying outliers. *Comput. Stat. Data Anal.* **2004**, *47*, 165–174. [\[CrossRef\]](#)
32. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2002.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.