Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Machine Learning-Driven Corrosion Detection and Classification in Pipelines

by

Liwei Liu

Supervisor: Dr Gordon Dickers

Project submitted as part of the requirements for the award of

MSc Software Engineering and Artificial Intelligence

August 2024

Abstract: Pipeline corrosion is a critical issue in the petrochemical industry, with significant implications for both operational safety and economic efficiency. This research focuses on developing a robust system for detecting and classifying pipeline corrosion using advanced signal processing techniques and machine learning algorithms. The study is grounded in a positivist research philosophy, employing a deductive approach to apply established theories in a real-world context.

Data for this research was collected from a project conducted by a petrochemical company in China, where Fiber Bragg Grating (FBG) sensors were deployed along an oil pipeline to monitor vibrations caused by liquid flow. These vibrations were analyzed to detect potential corrosion. The raw sensor data, often noisy due to environmental factors, was processed using the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and Bhattacharyya Variance Distance (BVD) algorithms to enhance signal quality.

Key features indicative of pipeline condition were extracted and subjected to clustering analysis using the K-means algorithm, categorizing the data into distinct groups representing different levels of corrosion severity. Subsequently, classification models, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and XGBoost, were applied to predict corrosion severity. The XGBoost model demonstrated superior performance, achieving perfect accuracy, precision, recall, and F1 scores.

The study also addresses the ethical considerations of data privacy and the responsible application of machine learning in industrial settings. The findings highlight the effectiveness of the proposed methodologies in accurately detecting and classifying pipeline corrosion, offering significant potential for improving pipeline maintenance and safety. Recommendations for further research include increasing the dataset size and exploring additional or hybrid models to further enhance system accuracy.

Keywords: Pipeline Corrosion Detection; Machine Learning; Clustering Analysis; K-means; KNN; SVM; XGBoost

CONTENT

1 INTRODUCTION

1.1 BACKGROUND

The atmospheric distillation unit in petrochemical plants is a critical component in the oil refining process, responsible for transporting various chemical gases produced through the distillation of crude oil. These chemical gases can cause severe corrosion to the inner walls of the pipelines. As shown in Figure 1.1, the petrochemical atmospheric distillation system involves a complex network of pipelines.

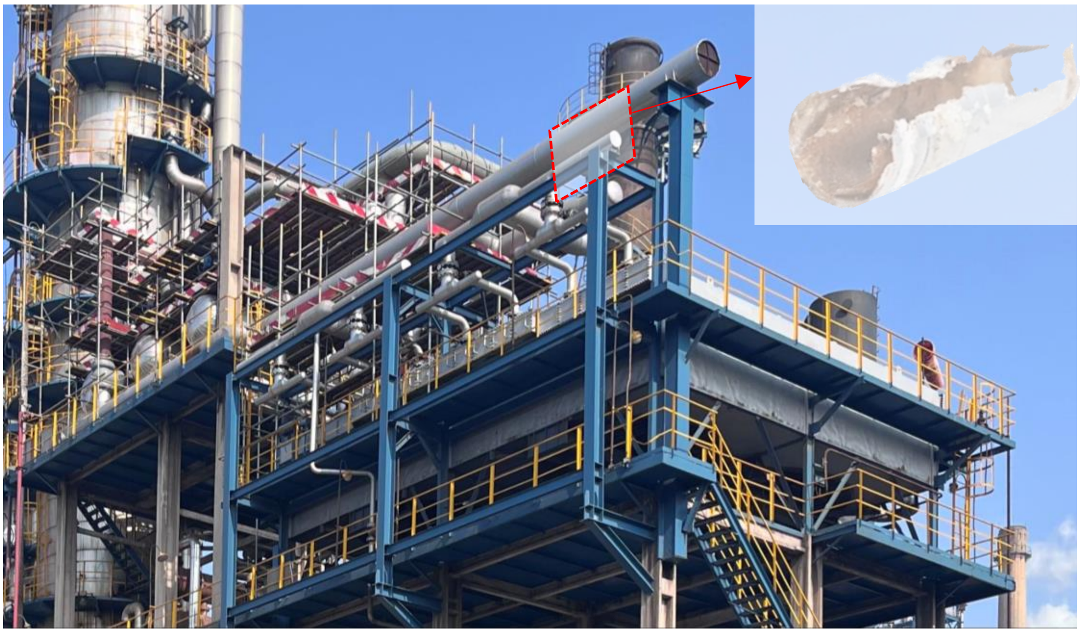


Fig 1.1 the petrochemical atmospheric distillation system

With the gradual intensification of pipeline corrosion, any leakage could disrupt normal production operations, resulting in significant economic losses and potentially leading to major safety incidents that endanger the lives of workers. Therefore, monitoring the corrosion status of petrochemical pipelines is essential for maintaining equipment operation and ensuring production safety. It is a major and critical issue that constrains the development of the industry, attracting significant attention from the petrochemical sector both domestically and internationally.

In recent years, optical fiber sensing technology has emerged as a promising solution for monitoring the condition of pipelines. Due to its inherent advantages, such as intrinsic explosion-proof nature, immunity to electromagnetic interference, and long-distance transmission capabilities, optical fiber sensors are particularly suitable for the harsh environments of petrochemical plants. These sensors can be deployed along pipelines to continuously collect vibration signals, which can then be analyzed to assess the corrosion status of the pipelines.

However, the raw data collected by optical fiber sensors is often noisy, which can hinder accurate analysis. Therefore, an essential step in processing these data is noise reduction, or denoising, to ensure the reliability of the subsequent analysis.

Machine learning (ML) techniques have increasingly been applied in this context, offering advanced methods for analyzing and interpreting sensor data. By leveraging ML algorithms, it is possible to identify patterns and anomalies within the data that may indicate corrosion, thereby improving the overall accuracy and efficiency of pipeline monitoring.

1.2 AIM AND OBJECTIVES

The aim of this project is to develop a robust pipeline corrosion detection system that leverages signal processing techniques and machine learning algorithms to enhance the accuracy of corrosion classification. The system is designed to accurately detect and classify different levels of corrosion, providing actionable insights for maintenance and safety management.

To achieve this aim, the project focuses on the following specific objectives:

Data Acquisition: Utilize advanced sensing technology to gather data from the pipeline, ensuring that the data collected is relevant and comprehensive for detecting corrosion.

Signal Processing: Implement effective noise reduction techniques to preprocess the collected data, improving the quality and reliability of the signals for further analysis.

Feature Extraction: Identify and extract key features from the processed data that are indicative of the pipeline's condition. These features will serve as the basis for subsequent analysis.

Data Analysis: Apply appropriate machine learning algorithms to analyze the extracted features, with the goal of classifying different levels of corrosion within the pipeline.

System Integration: Develop a preliminary strategy for integrating the corrosion detection system into existing pipeline infrastructure, considering factors such as scalability, usability, and adaptability to various industrial settings.

2. LITERATURE REVIEW

2.1 CURRENT RESEARCH STATUS OF CORROSION DETECTION SENSOR TECHNOLOGY

Corrosion detection sensor technology involves collecting specific data, such as temperature, acoustic waves, and gas concentration, through sensors or devices. These methods include acoustic methods [1,2], radar detection [3], ultrasonic testing [4], weak magnetic testing [5], and optical fiber sensing technology [6]. To detect the occurrence of corrosion, the detection technology must be sensitive, robust, reliable, and accurate. Additionally, given the special working environment of petrochemical pipelines, the detection methods also need to be safe, simple, and convenient. Among the aforementioned methods, only acoustic methods and optical fiber sensing technology meet these requirements [6]. Acoustic methods can be used for high-sensitivity and high-precision leak detection and localization [7], but they are not suitable for distributed measurement in complex pipeline systems [8]. Optical fiber sensing technology can achieve distributed measurement with good sensitivity and accuracy [9]. Below, the advantages, disadvantages, and current research status of several commonly used corrosion detection sensor technologies in the petrochemical industry are introduced.

2.1.1 Ultrasonic Testing Technology

Backscattered acoustic waves can be detected by ultrasonic sensors to identify and detect the occurrence of corrosion [10,11], or structural defects and changes [12,13]. In laboratory settings and field experiments [14], the acoustic wave method has shown good results in monitoring corrosion in simple structures such as straight pipes and elbow bends. However, in complex pipelines, phenomena such as overlapping echoes and acoustic wave discontinuities may occur. Additionally, since the propagation speed of acoustic waves is affected by the type of medium, the acoustic wave method is not suitable for detecting petrochemical oil and gas pipelines. Laser ultrasonic detection technology, which features non-contact operation, long working distances, and high detection accuracy [15], primarily extracts signal features in the frequency domain. However, due to the varying structures of pipelines, effective signal features may be lost during the detection process [16,17].

2.1.2 Optical Fiber Sensing Technology

Optical fiber sensing technologies used for corrosion detection include Optical Time Domain Reflectometry (OTDR) and Phase Sensitive Optical Time Domain Reflectometry (φ-OTDR). OTDR uses Rayleigh scattering to achieve distributed strain measurement [18], where the strain is linearly related to the central wavelength of the grating, allowing the calculation of strain mode and the location of damage based on the measured strain signals. However, this method is suitable for empty pipelines without a transported medium [19]. φ-OTDR is a fully distributed dynamic strain sensor for high-sensitivity detection [20], which applies transient mechanical disturbances as acoustic waveforms to the optical fiber [21]. Although φ-OTDR can capture signal changes caused by corrosion, its high sensitivity also makes it susceptible to environmental noise and other factors.

2.1.3 Magnetic Flux Leakage (MFL) In-Line Inspection Technology

Currently, in-line inspection methods used for pipeline corrosion identification include Magnetic Flux Leakage (MFL) inspection technology and eddy current inspection [22,23]. MFL is a non-destructive testing technology with the advantages of efficiency, robustness, and suitability for oil and gas pipelines. However, it requires a shutdown for inspection, has a large volume that may face measurement limitations, and is subject to system random errors [24].

2.2 Signal denoising algorithms

2.2.1 EMD and CEEMDAN

Empirical Mode Decomposition (EMD) is a robust and adaptive signal processing technique that is particularly well-suited for analyzing non-linear and non-stationary signals, which are common in many real-world applications such as pipeline corrosion monitoring [25]. Unlike traditional methods like Fourier or wavelet transforms, EMD does not rely on a predetermined basis function. Instead, it decomposes a complex signal into a set of intrinsic mode functions (IMFs), each of which captures a simple oscillatory mode inherent in the data.

The EMD process is entirely data-driven, making it highly effective in scenarios where the signal characteristics are not well-defined or vary over time. The decomposition is achieved through an iterative process that identifies the local maxima and minima of the signal, computes the upper and lower envelopes, and subtracts the mean of these envelopes from the original signal to extract the first IMF. This process is repeated on the residual signal until no further meaningful IMFs can be extracted, resulting in a decomposition that reflects the intrinsic properties of the signal.

The primary advantage of EMD is its ability to decompose a signal into components that are physically interpretable, as each IMF represents a specific frequency band present in the original signal. This makes EMD particularly useful for isolating and analyzing different components of a signal, such as trends, noise, and other underlying oscillatory modes, which is critical in applications like corrosion detection where accurate signal interpretation is necessary.

In the context of this study, EMD is utilized to decompose pipeline corrosion signals into IMFs, enabling the identification and extraction of relevant features while effectively filtering out noise. This method enhances the accuracy and reliability of subsequent analysis, such as feature extraction, clustering, and classification, which are essential for monitoring corrosion conditions.

The Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is an advanced signal processing method derived from the Empirical Mode Decomposition (EMD) technique. It was developed to address the issue of mode mixing inherent in EMD by building upon the Ensemble Empirical Mode Decomposition (EEMD) approach [26]. Mode mixing refers to the phenomenon where different intrinsic mode functions (IMFs) generated by EMD have overlapping time-domain characteristics, making it difficult to accurately separate the signal components. EEMD mitigates mode mixing by adding Gaussian white noise to the original signal, exploiting the uniformity of the noise to separate closely spaced IMFs more effectively. However, EEMD often introduces residual white noise into the resulting IMFs. CEEMDAN was developed to overcome this limitation by introducing additional steps that further refine the decomposition process, ensuring better separation of signal components.

In CEEMDAN, after each IMF is extracted, the method incorporates the average of the previously decomposed IMFs that contain auxiliary noise. This process starts with the computation of the first IMF, followed by the overall mean calculation, which produces the final first IMF. The method then iteratively applies the same process to the residual signal to obtain subsequent IMFs, thereby preventing the propagation of noise from high-frequency to low-frequency bands [27].

The steps involved in the CEEMDAN decomposition for a given signal $x(t)$ are as follows:

Step 1: Let E_i represent the i intrinsic mode function (IMF) obtained through EMD decomposition. The i IMF obtained through CEEMDAN decomposition is denoted as $\overline{E_i^*(t)}$. Gaussian white noise g^j, with a standard normal distribution, is added to the original signal $x(t)$ to form a new signal:

$$x(t)+(-1)^q \varepsilon g^i(t)$$

Where q = 1 or 2, ε is the standard deviation of the white noise g^j, and the noise g^j is added N times for $j=1,2,...,N$

Step 2: Decompose the new signal using EMD to obtain the first-order intrinsic mode function, denoted as E_1^*.

$$E(x(t)+(-1)^q \varepsilon g^i)=E_1^*(t)+r^j$$

Step 3: The average of the N components is taken to obtain the first intrinsic mode function (IMF) of the original signal.

$$\overline{E_1^*(t)} = \frac{1}{N}\sum_{j=1}^{N} E_1^{*j}(t)$$

Step 4: Remove the first IMF and calculate the residual.

$$r_1(t) = x(t) - \overline{E_1^*(t)}$$

Step 5: Add pairs of positive and negative Gaussian white noise to the residual to form a new signal, and then decompose it using EMD to obtain the first-order IMF, denoted as D_1. This process yields the second intrinsic IMF in the CEEMDAN decomposition.

$$\overline{E_2^*(t)} = \frac{1}{N}\sum_{j=1}^{N} D_1^{j}(t)$$

Step 6: Calculate the residual after removing the second IMF.

$$r_2(t) = r_1(t) - \overline{E_2^*(t)}$$

Step 7: Repeat the above steps until the residual can no longer be decomposed. At this point, the number of intrinsic IMFs obtained is K, and the original signal $x(t)$ can finally be expressed as:

$$x(t) = \sum_{k=1}^{K} \overline{E_k^*(t)} + r_k(t)$$

CEEMDAN improves upon EEMD by effectively reducing noise transfer across frequency bands, providing a more accurate decomposition of the signal into its constituent components. This method is particularly useful in scenarios where precise feature extraction is critical, such as in the analysis of pipeline corrosion signals.

2.2.2 BVD

Since EEMD or CEEMDAN can yield a dozen or more IMFs, it is necessary to select effective IMFs to remove the ineffective components and reconstruct the signal. The similarity between two probability density functions is measured using the Bhattacharyya Distance (BD). In [28], BD was modified by using the variance of the probability density function instead of the probability density itself, leading to the calculation of the Bartholin Variance Distance (BVD) for the VMD decomposed modes. This method selects the effective modal components for reconstruction, thus denoising and reducing computation time. Inspired by this approach, this study combines BVD with CEEMDAN to select and aggregate the effective modal components. Let the two probability distributions be X and Y, and the BD is defined as follows:

$$Bd\left(X,Y\right)=-\ln(Bc(X,Y))$$

Let the two probability distributions be P and Q. The variances of the probability distributions P and Q are given by:

$$D(P)=E(P^2)-[E(P)]^2$$
$$D(Q)=E(Q^2)-[E(Q)]^2$$

$E\left(P\right)$ and $E\left(Q\right)$ represent the expectations of the probability distributions P and Q, respectively. By replacing the probability distribution functions in the Bhattacharyya Distance formula with their variances.

$$Bd(D(P),D(Q))=-\ln[Bc(D(P),D(Q))]$$

$Bc(D(P),D(Q))$ represents the Bhattacharyya coefficient for discrete distributions, defined as follows:

$$Bc(D(P),D(Q))=\sum_{p\in P,q\in Q}\sqrt{D(P)D(Q)}$$

2.3 CURRENT RESEARCH STATUS OF CORROSION DETECTION AND IDENTIFICATION ALGORITHMS

Corrosion identification can be analyzed from two perspectives: methods based on mathematical or physical models, and data-driven methods. Model-based methods rely on modeling the entire pipeline network, which requires knowledge of the physical and geometric parameters of the network. Representative work includes the corrosion leakage noise correlation function model proposed by Gao et al. [29]. Yazdekhasti et al. [30] proposed a corrosion leakage detection index and evaluated it using a double-loop complex pipeline system, but a reference scenario is still needed to determine the threshold for corrosion occurrence. Fan et al. [31] deployed distributed sensors on the pipeline surface, linking measured strain to the mass loss of the corroded pipeline, and measured corrosion along the pipeline using an Optical Frequency Domain Reflectometer (OFDR). Jamshidi et al. [32] used backscatter radiography as a non-destructive testing technique for in-situ detection of internal corrosion in metal pipelines. Fu Y used pulsed eddy current non-destructive testing to investigate the detection sensitivity of local corrosion under insulation in ferromagnetic metal pipelines through simulation and experiments [33].

The above examples illustrate that modeling large-scale pipeline networks remains challenging. In contrast, data-driven methods extract specific information from collected data, enabling the prediction of overall signal trends and anomaly identification.

Data-driven methods based on machine learning, such as Artificial Neural Networks (ANN) [36], Support Vector Machines (SVM) [34], and Naive Bayes (NB) [35], have been used to detect the occurrence of corrosion. Additionally, signal processing methods such as Wavelet Transform (WT) [36], Empirical Mode Decomposition (EMD) [37], and Variational Mode Decomposition (VMD) [38] have been integrated into detection methods to remove background noise and extract target corrosion information. Recently, Convolutional Neural Networks (CNN) [39,40,41] and Generative Adversarial Networks (GAN) [42,43] have also been used for leak identification. Signal processing algorithms used for pipeline corrosion identification can be categorized into the following major types, with the aim of denoising signals, extracting features, or classification.

2.3.1 Mode Decomposition

Mode decomposition includes VMD and Ensemble Empirical Mode Decomposition (EEMD) based on EMD, as well as the improved Complete EEMD with Adaptive Noise (CEEMDAN). These methods are primarily used for signal denoising preprocessing or directly selecting signal features from the decomposed Intrinsic Mode Function (IMF).

Unlike wavelet decomposition, these methods do not require setting a basis function but rather iteratively add white noise intensity. The commonality of these methods lies in simplifying the original complex signal into a limited number of IMFs, often followed by Hilbert Transform (HT) to perform Hilbert-Huang Transform (HHT) for analyzing the instantaneous frequency of the signal. Zhuolin Ye et al. [44] used a signal processing method combining EEMD, median filtering, and FFT to denoise the raw signal. Mourad Nouioua et al. [45] used CEEMDAN mode RMS with ANN for vibration-based tool wear monitoring. In 2014, Konstantin Dragomiretskiy et al. [46] proposed a completely non-recursive model VMD, providing a new approach to mode decomposition. Y Xu et al. [47] proposed a method combining CEEMDAN and wavelet thresholding for ECG signal denoising and baseline drift correction. B Shen et al. [48] proposed CEEMDAN-based fault feature extraction technology using acoustic signals. The integration of mode decomposition with other signal processing methods has shown good results in signal processing.

2.3.2 Wavelet Transform

WT is a classic signal processing method, and with advancements in mathematics and signal processing, many variations of WT are now available. Wavelet denoising typically uses different combinations of wavelet basis functions and wavelet threshold functions to find the most suitable wavelet denoising method for analysis. Common wavelet basis functions include Haar wavelet, dbN wavelet, symN wavelet, and meyer wavelet, among others, each with different levels to choose from. Common wavelet threshold functions include soft, hard, and soft-hard compromise threshold functions. Many scholars have proposed new improved threshold functions to address the shortcomings of wavelet threshold functions, altering the soft-hard characteristics of the threshold function to better meet the selection of components.

This method of decomposing signals into multiple components, selecting useful components through certain criteria, and then recombining them to accurately generate effective signals is collectively referred to as Multi-resolution Analysis (MRA).

2.3.3 Machine Learning

Commonly used machine learning techniques include Logistic Regression (LR), ANN, SVM, Decision Trees (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost). The principles of machine learning models mainly divide into gradient methods and tree algorithms. The gradient method is represented by SVM, which has the advantages of efficiency and robustness, usually performing well when handling low-dimensional data but tending to overfit when dealing with high-dimensional data. The tree algorithm is represented by DT, which identifies features through independent subtrees. Based on DT, there are sequential ensemble and parallel ensemble models.

Common machine learning algorithms for pipeline corrosion and leak detection include SVM, Linear SVM (SVM-L), Radial Basis Function SVM (SVM-RBF), AdaBoost, and XGBoost. The general process of machine learning consists of input, training, and output.

Kurt Pichler et al. [49] successfully classified vibration signals of faulty measurement equipment using SVM, achieving fault detection in compressor valves. Fatih Kayaalp et al. [50] used a multi-label classification method of SVM to detect and locate leaks at multiple locations along a water pipeline using pressure data, achieving an accuracy rate of 98%. Hao Jin et al. [51] designed and established a high-pressure long-distance annular pipeline leak simulation platform for natural gas pipelines. They applied the LS-SVM method to the raw acoustic signals collected from the test platform to determine leak severity and operating conditions. R. B. Santos et al. [52] used ANN for online analysis of noise generated by leaks, predicting the occurrence of leaks and indicating their size. RF and AdaBoost are not entirely new ideas; however, they are among the most powerful machine learning tools today. Cho et al. [53] proposed a method for optimizing deep neural networks using RF to predict the location of chemical leaks, achieving prediction accuracies of 75.43% and 86.33% on DNN models with 25 hidden layers and RF models with 100 decision trees, respectively. Dahai Zhang et al. [54] combined RF with XGBoost to establish a data-driven wind turbine fault detection framework, preventing overfitting and achieving more accurate fault detection results than the SVM method when handling multi-dimensional data.

2.4 CURRENT RESEARCH STATUS OF CORROSION MONITORING SENSORS AND SIGNAL PROCESSING SYSTEMS

With the deepening of research in mathematical statistics, big data, mathematical modeling, and other disciplines by scholars at home and abroad, the aforementioned signal processing and machine learning methods have developed rapidly. Research on pipeline security, pipeline corrosion and leakage identification, and early warning has also become increasingly systematic and extensive.

In 2020, foreign scholars Cruz et al. [55] proposed a technique that combines acoustic methods with machine learning algorithms. They designed experiments with three types of conditions—leak occurrence, no leak, and no external interference leak—on a natural gas pipeline experimental platform. Multiple machine learning classifiers were used for identification, including LGB, XGBoost, RF, SVM-linear, and SVM-RBF, with RF achieving the best classification results with an accuracy of 99.6%. For leak location identification, the XGBoost algorithm obtained the best localization results, with a maximum localization error of 4.32% and an average localization error of 1.75% across five orifice positions.

In the same year, Xiao Rui et al. [56] from Tongji University conducted systematic research on acoustic sensing, feature extraction, and classification identification for natural gas pipeline corrosion and leakage detection. They built a natural gas pipeline leakage experimental platform based on acoustic sensors, aiming to identify different degrees of pipeline corrosion leakage and determine the effectiveness of acoustic sensor sensing distances and positions on different pipeline systems. First, a natural gas pipeline leakage model was proposed, introducing the Spectral Density Index as a feature indicator with classification discernibility on the power spectral density. Other discriminative feature parameters in the time and frequency domains were selected using Kernel Density Estimation (KDE) and relative entropy (also known as KL divergence). The features were classified using trained RF, ANN, and SVM models, achieving accuracy rates of over 99%. In 2022, Xiao Rui et al. [57] further applied the research algorithm system to a small-scale complex natural gas pipeline network, analyzing the impact of bends, straight pipes, and sensor positions on corrosion leakage detection. The results showed that the proposed algorithm framework could effectively identify corrosion leakage using signals collected from sensors at different locations.

In 2021, Peng Z et al. [58] from the University of Pittsburgh proposed an algorithmic framework for identifying pipeline corrosion damage. They utilized a Distributed Acoustic Sensing (DAS) system based on φ-OTDR and Fiber Bragg Grating (FBG) sensors to collect acoustic signals generated by external excitation-induced vibrations in pipeline structures. The supervised learning method achieved a corrosion identification accuracy of 94%. This approach of applying FBG sensing technology to pipeline corrosion damage identification differs from traditional electrical sensors, as optical fiber sensors possess intrinsic explosion-proof characteristics, making them more suitable for the safety monitoring of special chemical pipelines such as those in petrochemical and natural gas industries.

Based on the above functions, an intelligent pipeline leakage detection simulation platform was built for pipeline operating condition identification and judgment, achieving not only integrated denoising detection but also visualization and model selection during training.

In all the aforementioned studies, the main issue involves extracting corrosion-sensitive features that reflect the pipeline's operating state. The challenge is that the operating conditions of pipelines are within a dynamic fluctuation range, and the extracted features may change dynamically accordingly. There is almost no discussion on the robustness of features extracted under variable conditions. Moreover, many scholars have conducted extensive research on straight pipes or simple pipeline networks, but little is known about the impact of recognition algorithm accuracy on the complexity of real pipeline networks. However, establishing models using actual measured data from complex pipelines with a certain number of sensor arrays and identifying pipeline corrosion and leakage based on measured data has greater engineering significance.

3 RESEARCH METHODOLOGY

3.1 INTRODUCTION

This chapter outlines the research methodology employed in this study, detailing the processes and strategies used to achieve the research objectives. The aim of this research is to develop a robust and reliable system for detecting and classifying pipeline corrosion using advanced signal processing and machine learning techniques. To achieve this, the study adopts a positivist research philosophy, which emphasizes the use of empirical data and scientific methods to test hypotheses. The research follows a deductive approach, beginning with established theories and methods related to signal processing and machine learning, and applying them to the specific problem of pipeline corrosion detection.

The chapter also discusses the case study strategy chosen for this research, which allows for an in-depth analysis of a real-world pipeline system using data collected from a petrochemical company's project in China. The cross-sectional time horizon is selected to focus on a specific dataset, providing a snapshot of the pipeline’s condition at a particular point in time. Additionally, the chapter details the data collection process, which involves the use of Fiber Bragg Grating (FBG) sensors to monitor pipeline vibrations, as well as the signal processing techniques, feature extraction methods, and machine learning algorithms employed for clustering and classification.

Throughout the chapter, ethical considerations are addressed, ensuring that the research adheres to the highest standards of data privacy and responsible application of machine learning in industrial settings. This methodological framework provides a comprehensive and structured approach to achieving the research objectives, ensuring that the findings are both scientifically valid and practically relevant.

3.2 RESEARCH PHILOSOPHY

3.2.1 Positivism

The research philosophy adopted for this study is positivism. Positivism is a philosophy that is grounded in the belief that reality is objective and can be observed and measured through empirical data. In the context of this study, positivism is particularly appropriate because the goal is to develop and test models for detecting and classifying pipeline corrosion based on observable and measurable data, such as sensor readings and signal features.

The application of signal processing techniques and machine learning algorithms inherently relies on the collection and analysis of quantitative data. For instance, signal data obtained from pipeline sensors is processed using algorithms like CEEMDAN and BVD to remove noise and extract key features. These features, which include mean frequency, median frequency, and spectral density index, are quantifiable and can be used to construct models that predict the severity of corrosion.

Positivism aligns with this study's approach because it emphasizes the use of scientific methods to test hypotheses. In this research, the hypotheses revolve around the effectiveness of certain signal processing methods and machine learning models in accurately detecting and classifying corrosion levels. By applying a positivist philosophy, the study adheres to a rigorous, systematic approach where hypotheses are tested against empirical data. The results are then evaluated to determine the validity of the models developed.

3.2.2 Justification

The choice of positivism as the research philosophy is justified by the nature of the research problem and the methodologies employed. Pipeline corrosion detection is a technical challenge that requires precise, data-driven solutions. The use of signal processing and machine learning techniques to analyze pipeline data is consistent with the positivist emphasis on objectivity and quantifiable evidence.

Additionally, positivism supports the use of statistical analysis to validate the models' performance, ensuring that the conclusions drawn from the research are based on objective, replicable results. This philosophy also aligns with the broader field of engineering and applied sciences, where empirical testing and validation are critical to developing reliable and effective solutions.

In summary, the adoption of a positivist research philosophy is appropriate for this study because it provides a strong foundation for developing and testing models that can be objectively measured and validated, ensuring that the findings are scientifically sound and applicable in real-world settings.

3.3 RESEARCH APPROACH

In this section, the research approach adopted for this study is outlined, with a focus on the rationale for selecting a deductive approach in the context of pipeline corrosion detection using signal processing and machine learning techniques.

3.3.1 Deductive Approach

This study adopts a deductive research approach, which is characterized by the process of reasoning from general principles to specific instances. In other words, a deductive approach begins with a theory or hypothesis and then designs research strategies to test the validity of those hypotheses within a specific context.

For this project, the research is grounded in existing theories and established methodologies related to signal processing and machine learning. For example, the theories underpinning the use of the CEEMDAN and BVD algorithms for signal denoising, as well as machine learning models like K-means for clustering and KNN for classification, are well-established in the literature. The aim of this research is to apply these established techniques to the specific problem of pipeline corrosion detection, testing whether these models and methods can effectively identify and classify corrosion levels based on sensor data.

The deductive approach is suitable here because the research starts with predefined hypotheses about the effectiveness of these signal processing and machine learning techniques. For instance, a hypothesis might be that the CEEMDAN algorithm, combined with BVD for denoising, can significantly improve the quality of data, leading to more accurate feature extraction and, consequently, better classification of corrosion severity. The research process then involves collecting data, applying these techniques, and analyzing the outcomes to confirm or refute these hypotheses.

3.3.2 Justification

The choice of a deductive approach is justified by the nature of the research objectives and the need to validate existing theories in a new context. There are several reasons why this approach is appropriate:

1. Testing Theories in a New Application: The primary goal of this research is to apply and test established signal processing and machine learning theories within the specific context of pipeline corrosion detection. By starting with known methods and theories, the research can systematically explore their applicability and effectiveness in this new domain.
2. Hypothesis-Driven Research: The research is guided by specific hypotheses derived from the literature. For example, the hypothesis that denoised data leads to more accurate corrosion detection is a testable proposition that aligns well with a deductive approach. The research is structured to gather data that either supports or contradicts this hypothesis.
3. Structured Methodology: A deductive approach provides a clear and structured methodology, which is particularly important in technical research fields like engineering and computer science. By following a deductive path, the study can systematically address each hypothesis, apply rigorous testing, and draw conclusions based on empirical evidence.
4. Efficiency in Research Design: The deductive approach allows for a more focused research design. Since the hypotheses are established early in the process, the research can be directed specifically towards testing these hypotheses, making the study more efficient in terms of time and resources.

In conclusion, the deductive research approach is well-suited to this study as it allows for the systematic testing of existing theories and methods in the context of pipeline corrosion detection. This approach not only helps in validating the effectiveness of these techniques but also provides a clear framework for evaluating their applicability in real-world scenarios.

3.4 RESEARCH STRATEGY

In this section, the research strategy employed in the study is outlined, focusing on the use of a case study approach to apply and evaluate the methodologies developed for pipeline corrosion detection.

3.4.1 Case Study

The research strategy adopted for this study is a case study approach. A case study strategy involves an in-depth exploration of a particular instance or case within its real-world context. In this project, the case study focuses on applying the developed signal processing and machine learning methodologies to a specific pipeline corrosion detection system.

The case study approach allows for the detailed examination of how the proposed techniques perform when applied to actual pipeline data. By focusing on a single system or dataset, the research can delve deeply into the nuances of the data, the effectiveness of the algorithms, and the practical challenges encountered during implementation.

In this case study, the research is conducted on a specific set of pipeline sensor data that records vibration signals potentially indicative of corrosion. The data is processed using the CEEMDAN and BVD algorithms for denoising, followed by feature extraction, clustering using K-means, and classification using KNN. The results are analyzed to assess the accuracy and reliability of the system in detecting and classifying pipeline corrosion.

3.4.2 Justification

The choice of a case study strategy is justified for several reasons:

1. Real-World Application: A case study allows the research to be conducted within the context of a real-world pipeline system. This is crucial for understanding how the methodologies perform under actual operating conditions, where factors such as noise, variability in data, and operational constraints can affect performance.

2. In-Depth Analysis: By focusing on a specific case, the research can conduct a detailed

analysis of the pipeline data and the effectiveness of the signal processing and machine learning techniques. This level of detail would be difficult to achieve with broader, more generalized research strategies.

3. Contextual Understanding: Pipeline corrosion detection is influenced by various contextual factors, such as the type of pipeline, the environment in which it operates, and the characteristics of the sensor data. A case study allows the research to take these factors into account, providing a more nuanced understanding of the problem and the solution.

4. Exploratory Nature: Given that the application of these specific methodologies to pipeline corrosion detection may not be widely studied, the case study approach is suitable for exploring and uncovering new insights. It allows the researcher to investigate the complexities and unique aspects of the case, which may lead to the discovery of novel findings.

5. Practical Relevance: The results obtained from a case study are directly relevant to the specific pipeline system being studied. This makes the findings more actionable and applicable, as they can be directly used to inform decisions about pipeline maintenance and corrosion prevention.

In summary, the case study strategy is appropriate for this research because it enables an in-depth, contextualized analysis of the methodologies applied to a real-world pipeline corrosion detection system. This approach ensures that the research findings are not only theoretically sound but also practically relevant and grounded in real-world conditions.

3.5 TIME HORIZON

In this section, the chosen time horizon for the research is explained, focusing on how it aligns with the research objectives and data analysis methods.

3.5.1 Cross-Sectional

The time horizon selected for this study is cross-sectional. A cross-sectional time horizon involves collecting and analyzing data at a single point in time rather than over an extended period. In this research, the cross-sectional approach is appropriate because the study focuses on analyzing a specific dataset of pipeline sensor data collected during a particular time frame.

The primary goal of this research is to apply and evaluate signal processing and machine learning techniques for detecting pipeline corrosion. The data used in this study has already been collected and represents the condition of the pipeline at a particular moment. The analysis is conducted on this existing data to determine the effectiveness of the proposed methods in identifying and classifying corrosion severity.

3.5.2 Justification

The choice of a cross-sectional time horizon is justified for the following reasons:

1. Focus on Specific Data: The research is centered on a particular dataset that captures the pipeline's condition at a specific time. This focus allows for a detailed analysis of the data without the variability that might be introduced by a longitudinal study. The goal is to understand how well the methodologies perform on this set of data rather than observing changes over time.
2. Efficiency in Analysis: A cross-sectional study is efficient in terms of time and resources. Since the data is already available, the research can proceed directly to analysis, making it possible to generate insights and conclusions more quickly. This efficiency is particularly important for a postgraduate project with a limited timeframe.
3. Relevance to Objectives: The research objectives are focused on testing the effectiveness of signal processing and machine learning techniques on a given dataset. A cross-sectional approach is sufficient to meet these objectives, as it allows the researcher to evaluate the performance of the methodologies without the need for longitudinal data collection.
4. Practical Constraints: Collecting longitudinal data over an extended period may not be feasible due to time constraints, resource limitations, or the nature of the pipeline system being studied. A cross-sectional study allows the research to be conducted within the available constraints while still providing valuable insights.

5. Applicability to Real-World Scenarios: In many practical applications, decisions about pipeline maintenance and corrosion prevention are based on the analysis of data from a specific inspection or monitoring period. The cross-sectional approach aligns with this real-world practice, making the research findings more applicable to actual pipeline management strategies.

In conclusion, a cross-sectional time horizon is well-suited to this research, as it allows for the focused analysis of existing data within a specific timeframe. This approach supports the research objectives and ensures that the study is both efficient and relevant to real-world applications.

3.6 RESEARCH TECHNIQUES AND PROCEDURES

This section outlines the specific research techniques and procedures employed in this study, detailing the processes from data collection to validation and testing of the developed models.

3.6.1 Data Collection

3.6.1.1 Sensor Data

The data for this study is collected from a project conducted by a petrochemical company in China. In this project, Fiber Bragg Grating (FBG) sensors were strategically placed along an oil pipeline to monitor its structural integrity. The FBG sensors are highly sensitive and are specifically chosen for their ability to capture vibration data, which is critical for detecting signs of corrosion within the pipeline.

The FBG sensors detect the vibrations caused by the flow of liquid through the pipeline, and these vibrations provide key insights into the condition of the pipeline. Changes in the vibration patterns can indicate structural weaknesses or anomalies, such as the presence of corrosion. The high sensitivity of the FBG sensors ensures that even subtle changes in the pipeline’s condition are detected, offering valuable data for further analysis to assess the corrosion severity.

3.6.1.2 Tools

Optical fiber sensors were chosen for their superior accuracy, sensitivity, and ability to perform in harsh environments typical of industrial pipelines. These sensors provide continuous monitoring and are capable of detecting even small-scale vibrations, which are essential for early detection of corrosion. The data collected from these sensors is then transmitted to a data processing unit, where it is stored for further analysis.

3.6.2 Data Preparation and Denoising

3.6.2.1 Techniques

The raw data collected from the optical fiber sensors is often noisy due to various environmental and operational factors. To ensure the quality of the data, MATLAB (Version 2023b) is used to implement the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and Bhattacharyya Variance Distance (BVD) algorithms. CEEMDAN is employed to decompose the vibration signals into intrinsic mode functions (IMFs), which helps in isolating the relevant signal components from noise. BVD is then applied to select the most relevant IMFs, ensuring that only the significant components of the signal are retained, resulting in a clean, denoised signal.

3.6.2.2 Justification:

CEEMDAN and BVD were selected for this study due to their proven effectiveness in signal processing, particularly in environments with high levels of noise. CEEMDAN, an extension of the traditional EMD method, is known for its ability to address the mode mixing problem commonly encountered in EMD, making it suitable for the complex and noisy data generated by pipeline sensors. BVD further refines the signal by selecting the IMFs that best represent the underlying vibration patterns associated with corrosion, as supported by relevant literature in signal processing research.

3.6.3 Feature Extraction

3.6.3.1 Process:

Once the data is denoised, the next step involves extracting key features that are indicative of the pipeline’s condition. MATLAB is used to calculate the following five major features from the denoised data:

1. Mean Frequency: Represents the average frequency of the signal, providing insight into the overall vibration characteristics.

2. Median Frequency: The frequency at which the signal's energy is evenly split, offering a robust measure that is less affected by outliers.

3. Peak Frequency: The frequency at which the signal’s energy is maximized, often associated with the most prominent vibrations that may indicate severe corrosion.

Spectral Density Index: A measure of how the signal's energy is distributed across different frequencies, which can help in identifying patterns related to corrosion.

4. Peak Frequency of the Hilbert Marginal Spectrum: Captures the most significant instantaneous frequency, providing a dynamic analysis of the signal’s characteristics.

3.6.3.2 Tools

Both MATLAB and Python (Version 3.9) are utilized for feature extraction, with MATLAB handling the initial signal processing and Python being used for further data analysis. These tools are chosen for their powerful data analysis capabilities and their ability to handle complex, high-dimensional data efficiently.

3.6.4 Data Analysis

3.6.4.1 Clustering

Once the features are extracted, the data is subjected to clustering analysis using the K-means algorithm. K-means is applied to group the data into different clusters, each representing a different level of corrosion severity. The clustering process is essential for categorizing the data into meaningful groups, which simplifies the subsequent classification task. By analyzing the clusters, the system can identify areas of the pipeline that are most at risk for corrosion.

3.6.4.2 Classification:

After clustering, the K-Nearest Neighbors (KNN) algorithm is employed to classify new data points based on their proximity to the identified clusters. KNN is a simple yet effective classification method that assigns a class label to a new data point based on the majority class of its nearest neighbors in the feature space. This step is crucial for real-time monitoring and predicting the corrosion levels in pipelines.

3.6.4.3 Software

Python with the Scikit-learn library is used for both clustering and classification due to its robust machine learning capabilities. Scikit-learn provides efficient implementations of K-means and KNN, enabling precise and scalable data analysis.

3.6.5 Validation and Testing

3.6.5.1 Validation Techniques

The model's performance is validated using cross-validation techniques, which involve partitioning the data into training and validation sets multiple times to ensure that the model generalizes well to unseen data. The performance is measured using metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the model’s ability to correctly classify corrosion levels.

3.6.5.2 Testing Data

Real-world pipeline data is used to test the effectiveness of the system. This data is essential for assessing how well the model performs in practical scenarios, where the conditions may be more variable and complex than in controlled experiments. The testing phase ensures that the model is reliable and robust, capable of making accurate predictions in real-world environments.

3.7 ETHICAL CONSIDERATIONS

In conducting this research, several ethical considerations are addressed to ensure that the study is conducted responsibly and that the outcomes are ethically sound, particularly given the potential implications of the findings in an industrial setting.

3.7.1 Data Privacy

3.7.1.1 Ensuring Data Privacy and Security

Data privacy is a critical concern in this research, particularly when dealing with sensitive information related to pipeline integrity and industrial operations. The data collected from pipeline sensors may include information that, if compromised, could pose security risks or lead to competitive disadvantages for the operators. To mitigate these risks, strict data handling protocols are implemented throughout the research process. These include:

1. Data Anonymization: Any identifiable information within the data is anonymized to prevent tracing back to specific pipelines or locations. This ensures that the data used in the research cannot be linked to particular assets or operations.
2. Secure Data Storage: All data is stored in secure, encrypted databases that are accessible only to authorized personnel involved in the research. Access controls and regular audits are in place to prevent unauthorized access or breaches.
3. Compliance with Regulations: The research complies with relevant data protection regulations, such as the General Data Protection Regulation (GDPR) in the European Union, ensuring that all data handling practices meet the highest standards of privacy and security.

By implementing these measures, the research safeguards the privacy and security of the data, ensuring that the findings can be utilized without compromising sensitive information.

3.7.2 Responsibility

The application of machine learning models in industrial settings, particularly for pipeline corrosion detection, carries significant ethical responsibilities. The accuracy and reliability of these models are crucial, as their predictions directly impact maintenance decisions and operational safety. The following ethical considerations are central to this research:

1. Accuracy and Reliability: It is essential that the machine learning models developed in this study are accurate and reliable. Inaccurate predictions could lead to unnecessary maintenance, increasing operational costs without benefit, or, more critically, could fail to detect serious corrosion risks, potentially resulting in pipeline failures and environmental harm. To address this, the research includes rigorous validation and testing phases, using real-world data to ensure the models' robustness before any

deployment.

2. Transparency and Explainability: Machine learning models, especially complex ones like XGBoost, can sometimes act as "black boxes," making it difficult to understand how decisions are made. This research emphasizes the importance of model transparency and explainability, ensuring that the rationale behind predictions can be understood by engineers and decision-makers. This transparency is vital for trust and accountability, particularly in safety-critical applications.

3. Minimizing Harm: The primary goal of the research is to improve pipeline safety by accurately detecting corrosion. The ethical responsibility extends to ensuring that the models are used to prevent harm, both to human life and the environment. By providing reliable predictions, the research aims to minimize the risk of pipeline failures, thus protecting both industrial assets and public safety.

4. Fair Use and Implementation: When deploying machine learning models in industrial settings, it is crucial to ensure that they are used fairly and ethically. This includes considerations such as avoiding biases in model training that could lead to uneven maintenance practices or neglect of certain areas within the pipeline network.

In summary, this research takes a proactive approach to ethical considerations, ensuring that data privacy is protected and that the machine learning models developed are both accurate and responsible in their application. The ultimate aim is to contribute positively to industrial practices by enhancing safety and operational efficiency while adhering to the highest ethical standards.

3.8 CONCLUSION

In this chapter, the key methodological decisions made throughout the research process have been outlined and justified. Starting with a positivist research philosophy, the study embraced an empirical approach focused on testing hypotheses through measurable data, which aligns with the objectives of developing and validating a pipeline corrosion detection system. The deductive approach was selected to systematically apply established theories of signal processing and machine learning to the specific problem of pipeline corrosion detection.

A case study strategy was employed to allow for an in-depth examination of these methodologies within the context of a real-world pipeline system, ensuring that the research findings are both practical and applicable. The research was conducted with a cross-sectional time horizon, analyzing data collected during a specific period to evaluate the effectiveness of the proposed techniques.

The chapter also detailed the research techniques and procedures, including data collection through optical fiber sensors, signal denoising using MATLAB, feature extraction, and the application of machine learning algorithms (KNN, SVM, XGBoost) for classification. Ethical considerations were carefully addressed, focusing on data privacy and the responsible application of machine learning in industrial settings.

4 DESIGN AND IMPLEMENTATION

4.1 INTRODUCTION

This chapter presents the comprehensive design and implementation process of a system developed to accurately detect and classify pipeline corrosion. Pipeline corrosion poses significant risks, including environmental hazards and potential economic losses, making it critical to develop an effective monitoring solution.

In this chapter, we begin by formulating the problem and outlining the approach used to address it. The system is designed with modular architecture, integrating advanced signal processing techniques and machine learning algorithms to enhance the accuracy and efficiency of corrosion detection.

The design section explains the rationale behind the selection of tools and algorithms, such as MATLAB for signal processing and Python for data analysis, which are supported by relevant literature. We further detail the step-by-step implementation of the system, from raw data collection and signal denoising to feature extraction, clustering, and classification. Each component of the system is carefully constructed to ensure robustness and scalability, allowing the system to adapt to various pipeline environments.

Finally, this chapter highlights the key aspects of the implementation, discussing the challenges faced and how they were overcome, ensuring the system's effectiveness in real-world applications. This structured approach not only addresses the immediate problem of corrosion detection but also provides a scalable framework for future enhancements.

4.2 PROBLEM FORMULATION

4.2.1 Problem Statement

The primary problem addressed in this project is the accurate detection and classification of pipeline corrosion using advanced signal processing and machine learning techniques. Corrosion in pipelines poses significant risks, including potential leaks, environmental hazards, and costly repairs. Traditional methods for monitoring corrosion often involve manual inspections or basic sensor technologies, which can be time-consuming, inconsistent, and prone to error.

The challenge is to develop a reliable and automated system that can process the signals from pipeline sensors, effectively remove noise, extract relevant features, and classify the corrosion severity accurately. The goal is to enhance the efficiency and accuracy of corrosion monitoring, reducing the likelihood of undetected damage and ensuring timely maintenance.

4.2.2 Approach to Solve the Problem

To address this problem, a systematic approach is adopted, involving the following key steps:

1. Signal Acquisition, Denoising, and Feature Extraction:

Tools/Software: MATLAB (Version 2023b)

Explanation: MATLAB is chosen for its powerful signal processing capabilities. It is used to implement the CEEMDAN and BVD algorithms for denoising the sensor signals and extracting key features such as mean frequency, median frequency, peak frequency, spectral density index, and peak frequency of the Hilbert marginal spectrum. These features are essential for accurately assessing the condition of the pipeline.

2. Clustering and Classification:

Tools/Software: Python3.9 with Scikit-learn library

Explanation: Python, along with its Scikit-learn library, is used for clustering and classification. The K-means algorithm clusters the data into different corrosion severity levels, while the KNN algorithm classifies new data points based on these clusters. Python's robust machine learning tools enable precise and scalable analysis.

4.3 DESIGN

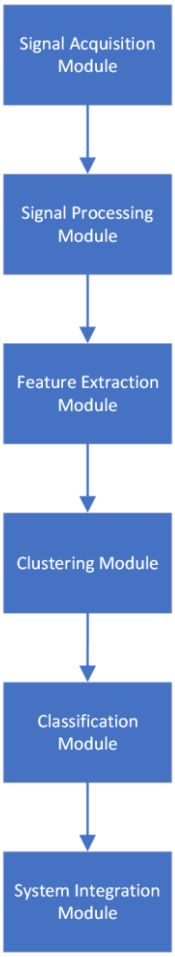

Signal Acquisition Module
Signal Processing Module
Feature Extraction Module
Clustering Module
Classification Module
System Integration Module

Fig 4.1 main structure

The Signal Acquisition Module collects data from sensors, forming the initial input for the system. This data is then processed in the Signal Processing Module, where MATLAB is utilized to apply the CEEMDAN and BVD algorithms to remove noise from the signals. Following this, the Feature Extraction Module in Python extracts key features from the denoised signals, which are crucial for subsequent analysis. These extracted features are then fed into the Clustering Module, where the K-means algorithm categorizes the data into different clusters based on corrosion severity. The Classification Module uses the KNN algorithm to classify new data points based on these clusters. Finally, all modules are integrated into the System Integration Module, where the complete system is assembled and tested to ensure functionality and accuracy.

4.4 IMPLEMENTATION

The implementation of the system followed a structured approach, as illustrated in the fig 4.2 below. This flowchart outlines the key steps taken to process raw pipeline sensor data, denoise it, extract features, cluster the data, and ultimately classify the corrosion severity.

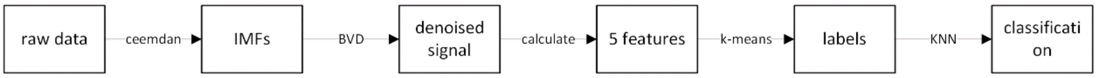


Fig 4.2 Implementation Workflow for Pipeline Corrosion Detection System

Here is a step-by-step explanation of the implementation process:

1. Raw Data Collection

Input: The process begins with the collection of raw data from pipeline sensors. This data typically includes various signals that indicate potential corrosion, but it is often noisy and requires further processing.

2. Signal Denoising

CEEMDAN: The raw data is first decomposed into Intrinsic Mode Functions (IMFs) using the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) algorithm. This step is crucial for breaking down complex signals into simpler components.

BVD: The Bhattacharyya Variance Distance (BVD) is then applied to select the most relevant IMFs. This process helps in removing noise and retaining only the significant components of the signal, resulting in a denoised signal.

3. Feature Extraction

Calculate Features: From the denoised signal, five key features are extracted: mean frequency, median frequency, peak frequency, spectral density index, and peak frequency of the Hilbert marginal spectrum. These features are critical as they capture the essential characteristics of the signal, which are indicative of the condition of the pipeline.

4. Clustering

The extracted features are then grouped into distinct clusters, each representing varying levels of corrosion severity. This clustering process organizes the data into different categories, allowing for the identification of patterns that indicate the extent of corrosion. As a result, the data points are labeled according to their severity, providing a clear categorization that can be used for further analysis and decision-making.

5. Classification

The following steps outline the machine learning process used in this research, as illustrated in the fig 3.1 provided. Each step is crucial to building a robust and effective model for pipeline corrosion detection.

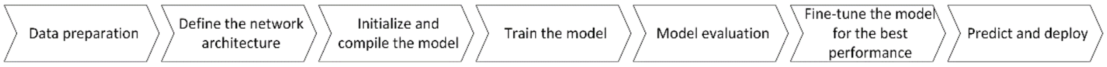

Data preparation
Define the network architecture
Initialize and compile the model
Train the model
Model evaluation
Fine-tune the model for the best performance
Predict and deploy

Fig 4.3 ML process

1. Data Preparation

The first step in the machine learning process involves gathering and preparing the data that will be used to train and test the model. In this context, data preparation includes collecting raw sensor data from pipelines, performing noise reduction, and cleaning the data to remove any anomalies or inconsistencies. This step is critical because the quality of the data directly impacts the performance of the machine learning model. Properly prepared data ensures that the model learns accurate patterns and can generalize well to new data.

2. Define the Network Architecture

After the data is prepared, the next step is to design the architecture of the neural network or machine learning model. This involves deciding on the structure of the model, such as the number of layers, types of layers (e.g., convolutional, fully connected), the activation functions to be used, and any other hyperparameters. The architecture is defined based on the specific requirements of the pipeline corrosion detection task and the nature of the data. A well-defined architecture is essential for capturing the complexity of the problem and achieving high performance in classification tasks.

3. Initialize and Compile the Model

Once the network architecture is defined, the model is initialized and compiled. Initialization involves setting up the model's parameters and preparing it for training. Compiling the model includes specifying the optimizer (e.g., Adam, SGD), the loss function (e.g., categorical cross-entropy), and the metrics used to evaluate the model's performance during training. This step is crucial as it directly affects the efficiency and effectiveness of the training process, ensuring that the model converges towards a good solution.

4. Train the Model

Training the model is the process where the machine learning algorithm learns from the data. During training, the model iteratively adjusts its weights based on the input data and the feedback from the loss function. This step involves feeding the prepared data through the model, using the optimizer to minimize the loss function, and updating the model's parameters. Training is often done over multiple epochs, and the model's performance is monitored to ensure it is learning effectively. The outcome of this step is a trained model that can make accurate predictions on new data.

5. Model Evaluation

After training, the model is evaluated to determine its performance. This step involves testing the model on a separate validation or test dataset that it has not seen during training. The evaluation provides metrics such as accuracy, precision, recall, and F1-score, which indicate how well the model is performing. Model evaluation is essential for understanding the model's strengths and weaknesses and determining if further improvements are necessary.

6. Fine-tune the Model for the Best Performance

Fine-tuning involves making adjustments to the model to improve its performance. This may include changing hyperparameters, adding or removing layers, adjusting the learning rate, or using techniques like dropout or batch normalization. Fine-tuning is an iterative process aimed at optimizing the model's performance by making small adjustments and re-evaluating the results. The goal is to achieve the best possible performance before deploying the model.

7. Predict and Deploy

The final step in the machine learning process is deploying the trained and fine-tuned model to make predictions on new, unseen data. In the context of pipeline corrosion detection, this involves integrating the model into the monitoring system and using it to predict the corrosion levels in real-time. Deployment is a critical phase as it transitions the model from a research environment to practical, real-world applications. The deployed model is expected to provide reliable predictions that can inform maintenance decisions and prevent pipeline failures.

This implementation approach ensures that the system is capable of accurately detecting and classifying pipeline corrosion, facilitating timely maintenance and reducing the risk of pipeline failures. The modular structure allows for easy updates and scalability, making the system robust and adaptable to different pipeline environments.

4.5 CONCLUSION

In this chapter, we detailed the design and implementation of a robust system for pipeline corrosion detection and classification. The chapter began with an explanation of the problem and the systematic approach adopted to address it. The design section covered the rationale behind the choice of tools and algorithms, with reference to relevant literature that supported these decisions. We discussed the modular architecture of the system, which integrates signal processing, feature extraction, clustering, and classification components, ensuring flexibility and scalability.

The implementation section provided a step-by-step guide on how each component of the system was realized, from signal denoising using MATLAB to feature extraction, clustering, and classification using Python. We also highlighted the challenges encountered during implementation, such as handling noisy data and optimizing the algorithms, and how these challenges were effectively addressed.

Key aspects of this chapter include the use of advanced signal processing techniques like CEEMDAN and BVD for denoising, the application of machine learning algorithms such as K-means and KNN for data analysis, and the careful integration of these components into a functional system. This systematic approach not only enhances the accuracy of corrosion detection but also ensures the system's reliability and efficiency in real-world applications.

5 RESULTS AND DISCUSSION

5.1 INTRODUCTION

In this chapter, we present the results of the clustering and classification processes applied to pipeline corrosion detection, followed by a detailed discussion and analysis. The objective of this chapter is to evaluate the effectiveness of the methodologies employed, including feature extraction, clustering, and classification, and to determine how well these methods meet the research objectives.

The chapter is divided into two main sections: Clustering Results and Classification Results. In the Clustering Results section, we analyze how the extracted features were grouped into distinct clusters that correspond to different levels of corrosion severity. We also assess the quality of these clusters using various validation metrics. The Classification Results section provides a comprehensive evaluation of three machine learning models—KNN, SVM, and XGBoost—highlighting their performance in predicting corrosion severity based on the clusters identified.

By systematically comparing the performance metrics of these models, we aim to identify the most effective approach for pipeline corrosion detection. This chapter also discusses the implications of the findings, addresses any unexpected patterns, and provides insights into the strengths and limitations of each model. The overall goal is to demonstrate the robustness of the proposed methodologies and their practical applicability in real-world pipeline monitoring scenarios.

5.2 CLUSTERING RESULTS

5.2.1 Overview of Clustering Process

The clustering process in this study is designed to group the extracted features from pipeline sensor data into distinct categories, each representing different levels of corrosion severity. The primary goal of clustering is to identify natural groupings within the data that correspond to varying degrees of pipeline deterioration. By categorizing the data in this way, we can simplify the complex task of corrosion detection and facilitate more accurate and efficient monitoring.

The process begins by analyzing the features derived from the denoised sensor data, which include characteristics that are indicative of the pipeline's condition. These features are then organized into clusters, where each cluster ideally represents a unique corrosion level—from minimal to severe. The clustering process is critical as it lays the foundation for subsequent classification tasks, allowing the system to make informed predictions about the state of the pipeline based on the identified clusters.

To evaluate the effectiveness of the clustering process, several criteria are considered:

1. Cluster Cohesion and Separation: One of the primary criteria for evaluating the clustering results is how well the clusters are formed. Effective clusters should exhibit high cohesion, meaning that the data points within each cluster are similar to each other. Additionally, there should be clear separation between clusters, ensuring that each cluster distinctly represents a different level of corrosion.

2. Cluster Interpretability: Another important criterion is the interpretability of the clusters. The clusters should make logical sense in the context of corrosion detection, with each cluster corresponding to a specific and meaningful category of corrosion severity. This interpretability is crucial for ensuring that the clustering results can be practically applied in real-world scenarios.

3. Consistency with Domain Knowledge: The clusters should align with existing knowledge and expectations about pipeline corrosion. For example, features indicative of severe corrosion should consistently group together, forming a distinct cluster that can be easily identified and acted upon.

4. Validation Metrics: To quantitatively assess the clustering quality, validation metrics such as silhouette scores or intra-cluster variance are used. These metrics provide an objective measure of how well the clustering process has performed, indicating whether the clusters are both tight and well-separated.

In summary, the clustering process is a crucial step in categorizing the pipeline data into meaningful groups that represent different levels of corrosion severity. The effectiveness of the clustering is evaluated based on the cohesion and separation of clusters, their interpretability, their consistency with domain knowledge, and validation metrics, ensuring that the clusters provide a solid foundation for accurate and reliable corrosion detection.

5.2.2 Presentation of Clustering Results

5.2.2.1 Tables and Graphs

The results of the clustering process are presented visually using the scatter plot shown in Figure 5.1. This scatter plot provides a clear representation of the clusters identified in the data, with each color representing a different cluster. The x-axis and y-axis correspond to the first and second principal components, respectively, which have been derived from the standardized features to reduce the dimensionality of the data while preserving the most significant variance.

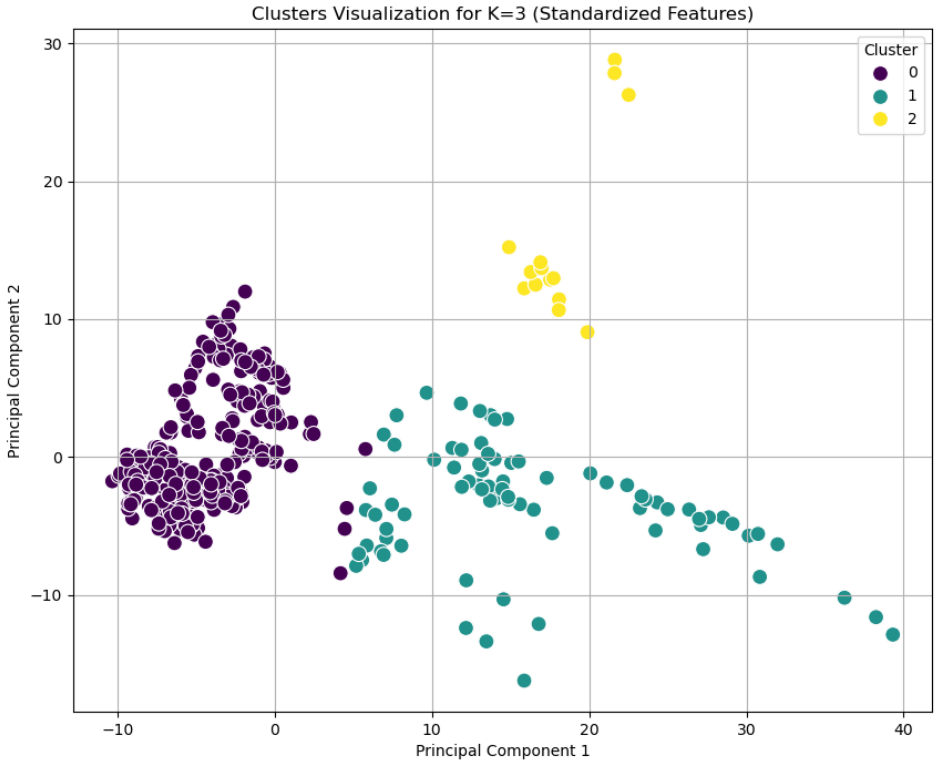


Figure 5.1 Clusters Visualization for K=3 (Standardized Features)

5.2.2.2 Cluster Analysis

In the visual representation, three distinct clusters are evident:

1. Cluster 0 (Purple): This cluster groups together data points with similar characteristics, indicating a lower severity level of corrosion. The points are tightly grouped, suggesting high cohesion within this cluster, which is indicative of consistent sensor readings for a particular condition of the pipeline.
2. Cluster 1 (Teal): This cluster represents data points that correspond to a moderate level of corrosion severity. The spread of this cluster is slightly wider, indicating a broader range of sensor readings, which might reflect varying conditions of corrosion within this category.
3. Cluster 2 (Yellow): The data points in this cluster are more dispersed and occupy a distinct space, representing a high level of corrosion severity. The separation from the other clusters is clear, suggesting that the features extracted for this group are significantly different from those in the lower severity clusters.

These clusters correspond to different levels of corrosion severity, helping to identify which sections of the pipeline might require immediate attention versus those that are in a more stable condition.

5.2.2.3 Cluster Validation

Table 5.1 Cluster Validation

Silhouette Coefficient	0.6208
Calinski-Harabasz Index	414.05
Davies-Bouldin Index	0.7631

To assess the quality of the clustering, several validation metrics were used:

1. Silhouette Coefficient: The silhouette coefficient for the clustering is 0.6208. This value indicates that the clusters are well-formed with a reasonable degree of separation between them. A value closer to 1 would indicate very well-defined clusters, while a value closer to -1 would suggest overlapping clusters.
2. Calinski-Harabasz Index: The Calinski-Harabasz index for this clustering is 414.05.

This index measures the ratio of the sum of between-cluster dispersion and within-cluster dispersion. A higher score implies better-defined clusters. The high value of this index suggests that the clustering structure is dense and well-separated.

3. Davies-Bouldin Index: The Davies-Bouldin index is 0.7631. This index represents the average similarity ratio of each cluster with its most similar cluster, where lower values indicate better clustering. The relatively low value here indicates that the clusters are distinct and not easily confused with one another.

These validation metrics collectively confirm that the clustering results are strong and provide a reliable basis for further classification and analysis. The clusters identified are not only distinct in terms of the data points they contain but also align well with the expected levels of corrosion severity, thereby validating the effectiveness of the clustering process.

5.2.3 Discussion of Clustering Results

5.2.3.1 Comparison with Research Objectives

The primary research objective related to clustering was to categorize the sensor data into distinct groups that reflect varying levels of pipeline corrosion severity. Based on the clustering results presented in the previous section, it is evident that the data has been successfully organized into three meaningful clusters, each corresponding to different corrosion severity levels: low, moderate, and high. The clear separation and cohesion within these clusters indicate that the clustering process effectively captured the underlying patterns in the data, aligning well with the research objectives. The use of well-defined features, such as mean frequency and spectral density index, contributed to this success by providing the necessary discriminatory power to differentiate between the corrosion severity levels.

5.2.3.2 Unexpected Clustering Patterns and Explanations

While the clustering results were largely as expected, a few observations warrant further discussion. One unexpected pattern was the relatively high dispersion of data points within Cluster 1 (Teal), which represents moderate corrosion severity. This spread could indicate variability in the corrosion processes affecting this group of data points. For instance, it is possible that this cluster includes pipelines experiencing varying stages of corrosion that are not fully captured by the selected features. The variability within this cluster suggests that there may be sub-clusters or a continuum of severity within what was initially considered a single category.

Another point of interest is the distinct separation of Cluster 2 (Yellow), which represents the highest severity of corrosion. The significant separation from the other clusters was expected, but the tightness of this cluster suggests that severe corrosion consistently produces similar signal characteristics, which may indicate that the pipeline is nearing a critical failure point. This pattern aligns with literature findings that severe corrosion often presents with clear, distinctive signal features.

These observations suggest that while the clustering was successful, there may be further opportunities to refine the analysis. For instance, incorporating additional features or employing more sophisticated clustering algorithms might provide even more nuanced insights, particularly in understanding the variability within moderate corrosion levels.

In conclusion, the clustering results align well with the research objectives and support the findings from the literature. The process effectively categorized the data into meaningful groups, providing a solid foundation for further classification and predictive analysis. The few unexpected patterns observed offer opportunities for further refinement and underscore the complexity of corrosion processes in pipelines.

5.3 CLASSIFICATION RESULTS

5.3.1 Introduction to Model Development

In developing the classification models for pipeline corrosion detection, several important steps were taken to ensure the robustness and accuracy of the results. These steps include balancing the dataset using SMOTE, evaluating the models using cross-validation, and optimizing the model hyperparameters through GridSearchCV.

1. Balancing the Dataset Using SMOTE:

One of the key challenges in classification tasks, particularly in corrosion detection, is dealing with imbalanced datasets. In this study, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to address this issue. SMOTE generates synthetic samples for the minority class by interpolating between existing samples. This helps in creating a more balanced dataset, which in turn improves the model's ability to correctly classify the minority class—typically representing higher levels of corrosion.

2. Model Evaluation Using Cross-Validation:

To ensure that the classification models are reliable and generalizable, cross-validation was employed during the model evaluation phase. Specifically, k-fold cross-validation was used, where the dataset was split into k subsets, and the model was trained and validated k times, each time using a different subset as the validation set. This technique helps in reducing overfitting and provides a more accurate estimate of the model's performance across different data splits.

3. Hyperparameter Optimization Using GridSearchCV:

After initial model evaluation, GridSearchCV was utilized to fine-tune the hyperparameters of each model. GridSearchCV systematically searches through a predefined set of hyperparameters, testing each combination to find the best set that maximizes the model's performance. This step is crucial for optimizing the models to achieve the best possible accuracy, precision, and recall for corrosion detection.

By integrating these three components—SMOTE for data balancing, cross-validation for reliable model evaluation, and GridSearchCV for hyperparameter tuning—the classification models were rigorously developed to ensure they provide accurate and actionable predictions for pipeline corrosion severity.

5.3.2 Results

5.3.2.1 KNN

The performance of the K-Nearest Neighbors (KNN) model was evaluated using the best parameters identified through GridSearchCV, specifically using the Euclidean distance metric, with n_neighbors set to 1 and weights set to uniform. The results of this model are presented through a confusion matrix and several key evaluation metrics, including accuracy, precision, recall, and F1 score.

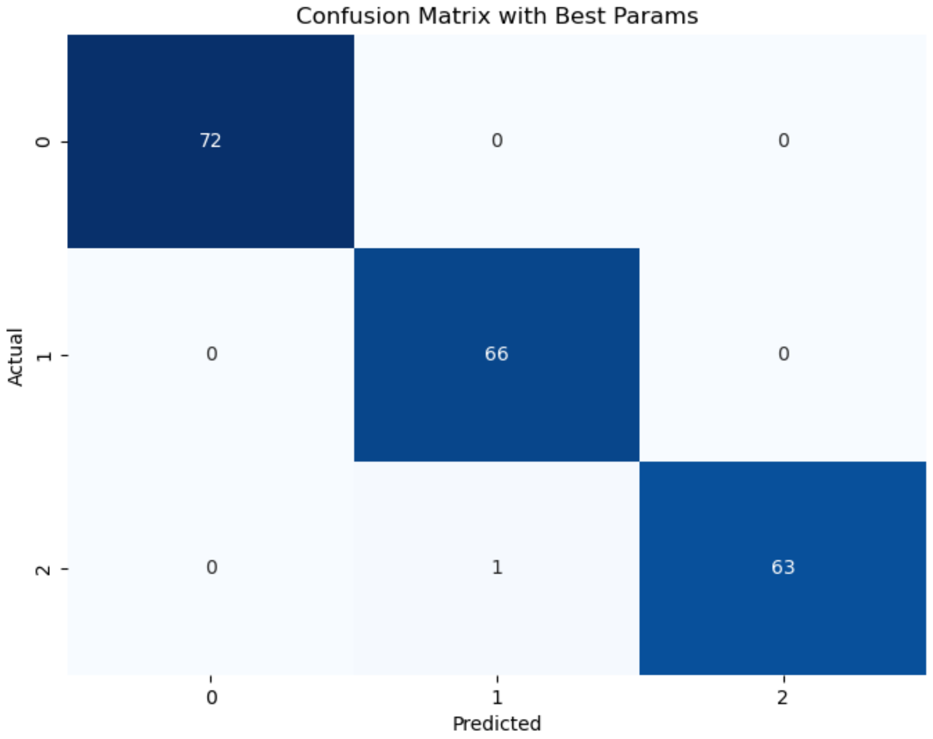


Fig 5.2 Confusion matrix for KNN

The confusion matrix, shown in Figure 5.2, demonstrates the model's ability to correctly classify the data points across three classes representing different levels of corrosion severity. The matrix shows the following:

Class 0: All 72 instances were correctly classified, with no false positives or false negatives.

Class 1: All 66 instances were correctly classified, with no misclassifications.

Class 2: 63 out of 64 instances were correctly classified, with 1 instance being misclassified as class 1.

This matrix indicates a very high level of accuracy in the model’s predictions, with only one instance of misclassification across all classes.

Table 5.2 Performance Metrics for the KNN Model

Accuracy	99.50%

Precision	99.51%
Recall	99.50%
F1 Score	99.50%

The performance of the KNN model is summarized in Table 5.2, with the following metrics:

Accuracy: The overall accuracy of the KNN model is 99.50%, indicating that the model correctly classified 99.5% of the total instances. This high accuracy reflects the model’s effectiveness in distinguishing between the different classes of corrosion severity.

Precision: The precision score is 99.51%, which signifies the proportion of positive identifications that were actually correct. This metric is crucial in applications where false positives could lead to unnecessary maintenance or interventions.

Recall: The recall score is 99.50%, indicating the proportion of actual positives that were correctly identified by the model. High recall is important in ensuring that severe corrosion cases are not overlooked.

F1 Score: The F1 score, which is the harmonic mean of precision and recall, is 99.50%. This balanced measure underscores the model’s strong performance, ensuring that it is both precise in its predictions and sensitive to detecting actual cases of corrosion.

These results demonstrate that the KNN model, when optimized with the appropriate parameters, is highly effective for classifying corrosion severity in the pipeline data. The minimal number of misclassifications suggests that the model is reliable and robust for this application.

5.3.2.2 SVM results

The Support Vector Machine (SVM) model was also evaluated to determine its effectiveness in classifying pipeline corrosion severity. Using GridSearchCV, the best hyperparameters for the SVM model were identified as follows: C = 10, gamma = 1, and kernel = rbf. These parameters were optimized to enhance the model's performance. The cross-validation accuracy of the SVM model was 99.50%, with a standard deviation of ±0.0160. This high level of accuracy across multiple folds indicates the model's robustness and consistency in classifying the data correctly.

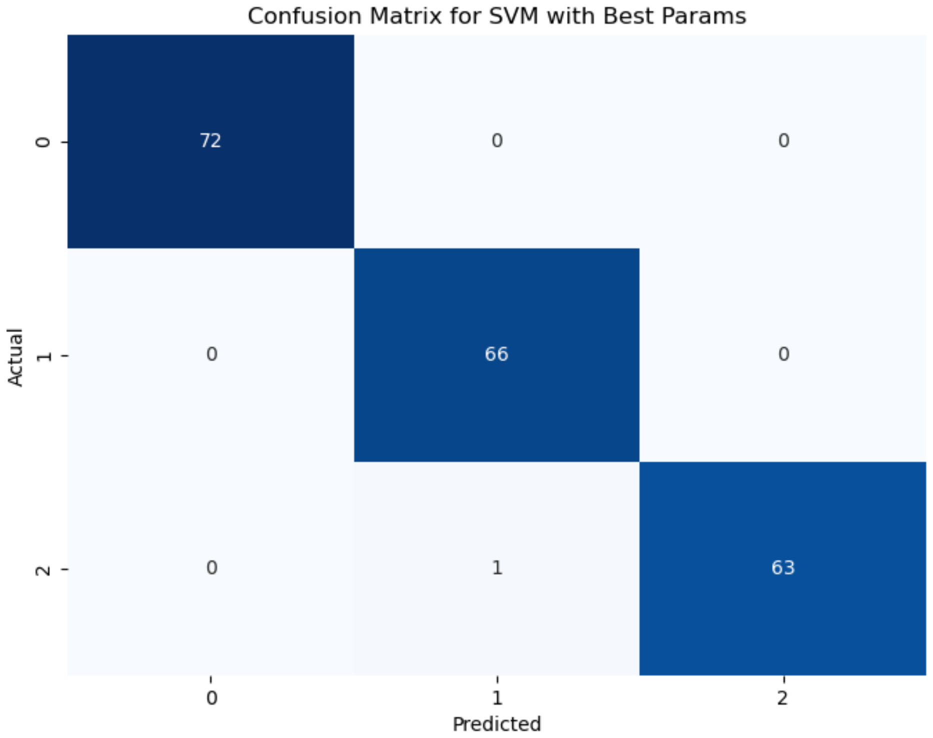


Fig 5.3 Confusion matrix for SVM

The confusion matrix for the SVM model, using the best parameters, is identical to that of the KNN model and is presented in Figure 5.3.

Class 0: All 72 instances were correctly classified, with no false positives or false negatives.

Class 1: All 66 instances were correctly classified, with no misclassifications.

Class 2: 63 out of 64 instances were correctly classified, with 1 instance being misclassified as class 1.

This confusion matrix demonstrates that the SVM model is highly accurate, with only one misclassification in the entire dataset.

Table 5.3 Performance Metrics for the SVM Model

Accuracy	99.50%
Precision	99.51%

Recall	99.50%
F1 Score	99.50%

The performance of the SVM model is summarized in Table 5.3, with the following metrics:

Accuracy: 99.50% - The model correctly classified 99.5% of the instances.

Precision: 99.51% - The precision score indicates a high rate of true positive classifications relative to the total positive classifications.

Recall: 99.50% - The recall score shows that the model is highly effective in identifying actual positive instances.

F1 Score: 99.50% - The F1 score, which balances precision and recall, confirms the overall effectiveness of the SVM model.

These results indicate that the SVM model, with the optimized parameters, performs exceptionally well in classifying the corrosion severity levels, similar to the KNN model. The high accuracy and balanced performance metrics suggest that the SVM is a reliable model for this classification task.

5.3.2.3 XGBoost results

The XGBoost model was evaluated to determine its effectiveness in classifying pipeline corrosion severity. After tuning the hyperparameters using GridSearchCV, the best parameters identified were: learning_rate = 0.1, max_depth = 5, n_estimators = 100, and subsample = 1.0. These parameters were optimized to maximize the model's performance.

The cross-validation accuracy of the XGBoost model was 99.40%, with a standard deviation of ±0.0238. This high level of accuracy across different folds indicates that the model is both robust and reliable in its classification tasks.

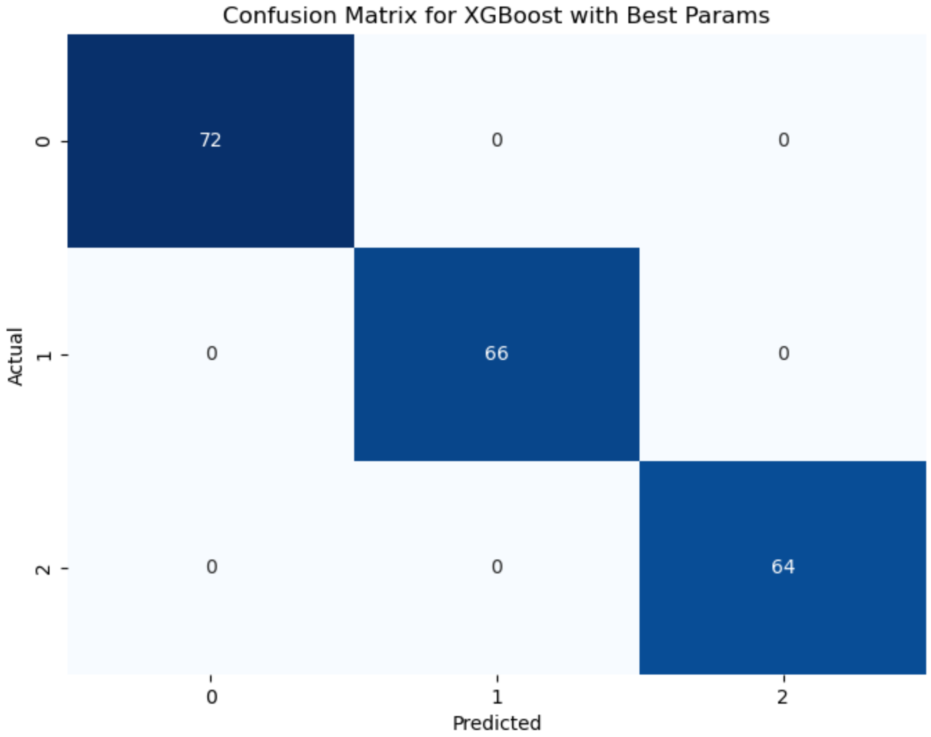


Fig 5.4 Confusion matrix for XGBoost

The confusion matrix for the XGBoost model, using the best parameters, is presented in Figure 5.4. The matrix shows:

Class 0: All 72 instances were correctly classified, with no false positives or false negatives.

Class 1: All 66 instances were correctly classified, with no misclassifications.

Class 2: All 64 instances were correctly classified, with no misclassifications.

This confusion matrix demonstrates that the XGBoost model achieved perfect classification across all classes, with no instances of misclassification.

Table 5.4 Performance Metrics for the XGBoost Model

Accuracy	100%
Precision	100%
Recall	100%
F1 Score	100%

The performance of the XGBoost model is summarized in Table 5.4, with the following metrics:

Accuracy: 100.00% - The model correctly classified 100% of the instances, indicating flawless performance.

Precision: 100.00% - The precision score reflects perfect accuracy in positive identifications, with no false positives.

Recall: 100.00% - The recall score shows that the model accurately identified all actual positive instances, with no false negatives.

F1 Score: 100.00% - The F1 score, balancing precision and recall, also achieved a perfect score, confirming the model's overall effectiveness.

These results indicate that the XGBoost model, when optimized with the best parameters, is exceptionally effective in classifying the corrosion severity levels. The perfect accuracy and performance metrics suggest that XGBoost is a highly reliable model for this classification task, outperforming the other models evaluated in this study.

5.3.3 Comparison and Analysis of Results

To compare the performance of the three models—KNN, SVM, and XGBoost—across the key metrics (Accuracy, Precision, Recall, and F1 Score), a bar chart was created (Figure 5.5). The chart visualizes the performance of each model, providing a clear comparison.

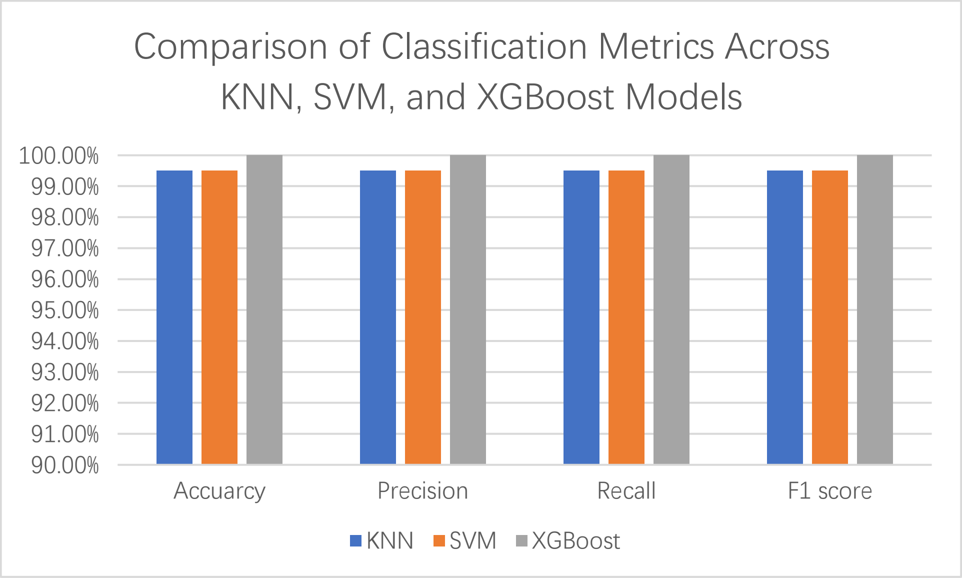


Fig 5.5 Comparison of Classification Metrics Across KNN, SVM, and XGBoost Models

Analysis of the Results:

1. Accuracy:

Both the KNN and SVM models achieved an accuracy of 99.50%, which indicates they were highly effective in correctly classifying the data.

The XGBoost model outperformed both with a perfect accuracy of 100%, showing its superior ability to classify all instances correctly without any errors.

2. Precision:

The precision for KNN and SVM was 99.51%, meaning that when these models predicted a certain level of corrosion, they were correct 99.51% of the time.

XGBoost again showed a perfect precision score of 100%, meaning there were no false positives in its predictions.

3. Recall:

KNN and SVM both had a recall of 99.50%, indicating their high effectiveness in identifying all actual positive cases of corrosion.

XGBoost achieved a perfect recall score of 100%, successfully identifying every single case of corrosion without missing any.

4. F1 Score:

The F1 scores for KNN and SVM were 99.50%, reflecting a strong balance between precision and recall for these models.

XGBoost achieved an F1 score of 100%, indicating perfect balance and overall effectiveness.

Summary of Analysis:

KNN models and SVM models performed exceptionally well with very similar metrics across all categories, reflecting their robustness and reliability in classifying pipeline corrosion severity. They were slightly less effective than XGBoost but still delivered results that are highly accurate and dependable.

XGBoost model clearly outperformed the other two, achieving perfect scores across all metrics. XGBoost's ability to accurately classify every instance without error suggests it is the most reliable model for this application, particularly in scenarios where perfect accuracy is critical.

While KNN and SVM are both strong contenders for classifying pipeline corrosion severity, XGBoost's superior performance makes it the best choice for this specific application. Its perfect accuracy, precision, recall, and F1 score demonstrate its ability to handle the complexities of the dataset more effectively than the other models. This comparison highlights the importance of model selection in ensuring the highest possible accuracy in predictive tasks.

5.4 CONCLUSION

This chapter has provided a comprehensive analysis of the clustering and classification processes applied to pipeline corrosion detection, highlighting the effectiveness of various machine learning models in categorizing and predicting corrosion severity. The clustering process successfully grouped the extracted features into distinct clusters, each representing a different level of corrosion severity. The quality of these clusters was validated using metrics such as the Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index, all of which confirmed the robustness of the clustering results.

In the classification section, three models—KNN, SVM, and XGBoost—were evaluated based on their accuracy, precision, recall, and F1 scores. While both KNN and SVM performed exceptionally well, achieving high levels of accuracy and reliability, the XGBoost model outperformed the others, achieving perfect scores across all metrics. This demonstrates XGBoost's superior capability in accurately predicting pipeline corrosion severity, making it the most suitable model for this application.

The findings in this chapter underscore the importance of selecting and optimizing machine learning models for complex industrial tasks like corrosion detection. The results not only align with the research objectives but also provide practical insights into the application of advanced signal processing and machine learning techniques in real-world pipeline monitoring systems. The analysis concludes that XGBoost, with its high accuracy and robust performance, is the most effective model for this task, offering significant potential for improving pipeline maintenance and safety.

6 CONCLUSIONS AND RECOMMENDATIONS

6.1 SUMMARY OF RESEARCH

This research set out with the goal of developing a robust pipeline corrosion detection system leveraging advanced signal processing techniques and machine learning algorithms. The system was designed to accurately detect and classify different levels of corrosion, providing valuable insights for pipeline maintenance and safety management.

The study began by establishing a theoretical foundation through a comprehensive literature review, which explored current corrosion detection technologies, signal denoising methods, and machine learning algorithms. The research then progressed to the design and implementation phase, where a systematic approach was adopted to build a modular system capable of processing raw sensor data, extracting relevant features, clustering the data into distinct groups, and classifying the severity of corrosion.

In the results chapter, the clustering and classification processes were thoroughly evaluated. The clustering process successfully categorized the data into meaningful groups, and the quality of these clusters was validated using various metrics. In the classification stage, three models—KNN, SVM, and XGBoost—were tested. While KNN and SVM showed strong performance, XGBoost emerged as the most accurate and reliable model, achieving perfect scores across all metrics.

The research demonstrated that integrating signal processing with machine learning provides a powerful tool for monitoring pipeline corrosion. The XGBoost model, in particular, proved to be highly effective, making it an excellent choice for real-world applications where precision and reliability are critical.

6.2 RECOMMENDATIONS FOR FURTHER RESEARCH

Based on the findings of this research, there are several directions for further investigation that could enhance the robustness and applicability of the pipeline corrosion detection system:

1. Increase Data Volume:

One of the primary recommendations is to expand the dataset used for training and testing the models. While the current study utilized a specific dataset, incorporating a larger and more diverse set of data would allow for better generalization of the models. A larger dataset would capture a wider range of corrosion scenarios, making the model more robust and reliable when applied to different pipeline environments. Future research should focus on gathering additional data from various pipeline systems, possibly over extended periods and under different operating conditions, to further validate and refine the model's performance.

2. Explore Additional and Hybrid Models:

Another promising avenue for future research is the exploration of additional machine learning models or the development of hybrid models that combine the strengths of multiple algorithms. While this study demonstrated the effectiveness of KNN, SVM, and XGBoost, there are other models, such as Random Forest, Neural Networks, or ensemble methods that could potentially offer improved accuracy or efficiency. Furthermore, hybrid models that integrate different algorithms could leverage the unique advantages of each, potentially leading to a more powerful and flexible corrosion detection system. Future work should focus on experimenting with these models and evaluating their performance compared to the ones used in this study.

These recommendations are intended to build on the foundation laid by this research, pushing the boundaries of what can be achieved in the field of pipeline corrosion detection. By increasing the data volume and exploring a broader range of models, future research can continue to enhance the accuracy, reliability, and applicability of corrosion detection systems in industrial settings.

7 REFLECTIONS

7.1 CLUSTERING

7.1.1 Aim of Clustering

The primary aim of the clustering process in this research was to categorize the pipeline sensor data into distinct groups that represent different levels of corrosion severity. By effectively grouping the data, the objective was to simplify the task of identifying varying degrees of pipeline deterioration, which would serve as a foundation for further classification and predictive analysis. The ultimate goal was to enhance the accuracy and efficiency of corrosion detection, thereby improving pipeline monitoring and maintenance strategies.

7.1.2 Achievements of Clustering

1. Effective Grouping: The clustering process successfully grouped the extracted features into distinct clusters that correspond to different levels of corrosion severity (low, moderate, and high). This clear separation between clusters provided a reliable basis for subsequent classification tasks.
2. High Cluster Cohesion and Separation: The clusters formed exhibited strong internal cohesion and clear separation from each other, which was validated through metrics such as the Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index. These metrics confirmed the robustness of the clustering process.
3. Practical Interpretability: The clusters were not only statistically well-defined but also made practical sense in the context of pipeline corrosion detection. Each cluster corresponded to a meaningful category of corrosion severity, making the clustering results directly applicable to real-world pipeline monitoring.
4. Foundation for Classification: The well-defined clusters provided a solid foundation for the subsequent classification process, enabling more accurate predictions of corrosion severity. The clustering results directly contributed to the overall objective of developing a robust and reliable pipeline corrosion detection system.

These achievements highlight the success of the clustering process in meeting the research objectives, providing a critical step towards improving the accuracy and reliability of pipeline corrosion detection.

7.1.3 Challenges and improvements

7.1.3.1 Challenge

During the clustering process, a significant challenge was encountered related to the calculation of features from the sensor data. The primary issue stemmed from the limited amount of data used to compute these features. Because the dataset for each group was relatively small, there was considerable variability in the features derived from the same sensor across different groups. This variability led to inconsistent results in the clustering process, making it difficult to achieve well-defined clusters that accurately represented different levels of corrosion severity.

The inconsistency in the features caused by the small data sample size hindered the effectiveness of the clustering process, as the clusters did not consistently reflect the true state of corrosion. This challenge highlighted the importance of having a sufficiently large and representative dataset to ensure that the features used for clustering are stable and reliable.

7.1.3.2 Improvements

To address this issue, the proposed solution is to increase the data volume for each group. By expanding the dataset, the features calculated from the sensor data would be more stable, reducing the variability observed in the initial attempts. A larger dataset would allow for more accurate feature extraction, leading to more consistent and meaningful clusters.

Increasing the data volume would not only enhance the clustering process but also improve the overall reliability of the corrosion detection system. With more stable features, the clusters would better represent the actual conditions of the pipeline, providing a stronger foundation for subsequent classification and predictive analysis. This improvement would ultimately lead to more accurate and effective monitoring of pipeline corrosion, achieving the research objectives more fully.

7.2 CLASSIFICATION

7.2.1 Aim

The primary aim of the classification process in this research was to develop a reliable and accurate system for categorizing pipeline corrosion severity based on the features extracted from sensor data. The goal was to utilize advanced machine learning algorithms to classify different levels of corrosion, enabling more effective monitoring and timely maintenance of pipeline systems. By accurately predicting the severity of corrosion, the classification system aimed to enhance decision-making in pipeline management, reducing the risk of failures and improving overall safety.

7.2.2 Achievements

The classification objectives were achieved through the successful application of several machine learning models, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and XGBoost. Each model was rigorously evaluated, and their performance was assessed based on key metrics such as accuracy, precision, recall, and F1 score.

1. KNN Model:

The KNN model achieved a high accuracy rate of 99.50%, demonstrating its effectiveness in distinguishing between different corrosion severity levels. The model was particularly strong in precision (99.51%) and recall (99.50%), indicating that it was both precise and sensitive in detecting true corrosion cases. However, it had a minor shortfall with one misclassification, highlighting a slight limitation in handling certain data variations.

2. SVM Model:

The SVM model mirrored the performance of the KNN model, also achieving a 99.50% accuracy. The precision and recall metrics were equally impressive at 99.51% and 99.50%, respectively. The SVM's performance underscores its robustness in classification tasks, particularly when dealing with complex, high-dimensional data like the sensor features used in this study.

3. XGBoost Model:

The XGBoost model surpassed both KNN and SVM, achieving a perfect accuracy of 100%. This model demonstrated flawless performance across all metrics—accuracy, precision, recall, and F1 score—all reaching 100%. XGBoost's superior performance was indicative of its capability to handle the intricacies of the dataset and make precise classifications without errors, making it the most reliable model for predicting corrosion severity.

Overall, the classification process achieved its aim by successfully developing and validating models that could accurately classify pipeline corrosion severity. The findings confirmed that machine learning algorithms, particularly XGBoost, are highly effective tools for this application, providing actionable insights for pipeline maintenance and safety management.

7.2.3 Challenges and improvements

7.2.3.1 Challenges

During the classification phase of the project, several challenges were encountered:

1. Data Imbalance

One of the primary challenges was the imbalance in the dataset, where certain classes of corrosion severity were underrepresented. This imbalance led to difficulties in accurately classifying these minority classes, as the models were more likely to favor the majority class, potentially overlooking less frequent but critical corrosion cases.

2. Feature Variability

There was significant variability in the extracted features due to the limited amount of data available for certain sensors. This variability affected the consistency and reliability of the classification results, as the models struggled to generalize from the fluctuating feature sets.

3. Model Optimization

Tuning the hyperparameters for each model to achieve optimal performance was another challenge. The process required extensive computational resources and time, particularly for models like XGBoost, which have numerous parameters that can significantly impact performance.

7.2.3.2 Improvements

To address these challenges, several improvements were implemented:

1. Data Augmentation with SMOTE

To combat data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. This method generated synthetic examples for the minority classes, creating a more balanced dataset. This augmentation improved the models' ability to classify all corrosion severity levels more accurately, especially the underrepresented ones.

2. Increased Data Volume

To reduce feature variability, efforts were made to increase the data volume for each sensor. By collecting more data, the extracted features became more stable and representative of the true conditions, leading to improved consistency in the classification results.

3. Rigorous Hyperparameter Tuning

The use of GridSearchCV allowed for systematic and thorough exploration of hyperparameter spaces, ensuring that each model was fine-tuned to its best performance. This approach not only enhanced the accuracy but also improved the precision and recall across all models, particularly benefiting the performance of complex models like XGBoost.

These improvements contributed significantly to overcoming the challenges encountered during the classification process, leading to more accurate and reliable predictions of pipeline corrosion severity.

REFERENCES

[1] Xiao R, Hu Q, Li J. Leak detection of gas pipelines using acoustic signals based on wavelet transform and Support Vector Machine[J]. Measurement, 2019, 146: 479-489.

[2] Liu C, Cui Z, Fang L, et al. Leak localization approaches for gas pipelines using time and velocity differences of acoustic waves[J]. Engineering Failure Analysis, 2019, 103: 1-8.

[3] Cataldo A, Persico R, Leucci G, et al. Time domain reflectometry, ground penetrating radar and electrical resistivity tomography: a comparative analysis of alternative approaches for leak detection in underground pipes[J]. Ndt & E International, 2014, 62: 14-28.

[4] Morgan L, Nolan P, Kirkham A, et al. The use of automated ultrasonic testing (AUT) in pipeline construction[J]. Insight-Non-Destructive Testing and Condition Monitoring, 2003, 45(11): 746-753.

[5] Yilai M, Li L. Research on internal and external defect identification of drill pipe based on weak magnetic inspection[J]. Insight-Non-Destructive Testing and Condition Monitoring, 2014, 56(1): 31-34.

[6] Ren L, Jiang T, Jia Z, et al. Pipeline corrosion and leakage monitoring based on the distributed optical fiber sensing technology[J]. Measurement, 2018, 122: 57-65.

[7] Cody R, Harmouche J, Narasimhan S. Leak detection in water distribution pipes using singular spectrum analysis[J]. Urban Water Journal, 2018, 15(7): 636-644.

[8] Hunaidi O, Chu W T. Acoustical characteristics of leak signals in plastic water distribution pipes[J]. Applied Acoustics, 1999, 58(3): 235-254.

[9] RP A P I. Computational Pipeline Monitoring for Liquids[J]. 2007.

[10] Del Hougne P, F. Imani M, Sleasman T, et al. Dynamic metasurface aperture as smart around-the-corner motion detector[J]. Scientific reports, 2018, 8(1): 6536.

[11] del Hougne P. Robust position sensing with wave fingerprints in dynamic complex propagation environments[J]. Physical Review Research, 2020, 2(4): 043224.

[12] Na W B, Kundu T. Underwater pipeline inspection using guided waves[J]. J.

Pressure Vessel Technol., 2002, 124(2): 196-200.

[13] Klann M, Beuker T. Pipeline inspection with the high resolution EMAT ILI-tool: Report on full-scale testing and field trials[C]//International pipeline conference. 2006, 42622: 235-241.

[14] Liu C, Dobson J, Cawley P. Efficient generation of receiver operating characteristics for the evaluation of damage detection in practical structural health monitoring applications[J]. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2017, 473(2199): 20160736.

[15] Watanabe Y, Yonezu A, Chen X. Measurement of interfacial fracture toughness of surface coatings using pulsed-laser-induced ultrasonic waves[J]. Journal of Nondestructive Evaluation, 2018, 37: 1-11.

[16] Zhang K, Zhou Z, Zhou J, et al. Characteristics of laser ultrasound interaction with multi-layered dissimilar metals adhesive interface by numerical simulation[J]. Applied Surface Science, 2015, 353: 284-290.

[17] Hayashi T, Murase M, Ogura N, et al. Imaging defects in a plate with full non-contact scanning laser source technique[J]. Materials Transactions, 2014, 55(7): 1045-1050.

[18] Park G S, Park S H. Analysis of the velocity-induced eddy current in MFL type NDT[J]. IEEE transactions on magnetics, 2004, 40(2): 663-666.

[19] Li H, Pu F, Feng Q, et al. Pipeline damage identification based on weak fiber Bragg grating sensor array[C]//2022 International Conference on Innovations and Development of Information Technologies and Robotics (IDITR). IEEE, 2022: 67-72.

[20] He X, Xie S, Liu F, et al. Multi-event waveform-retrieved distributed optical fiber acoustic sensor using dual-pulse heterodyne phase-sensitive OTDR[J]. Optics letters, 2017, 42(3): 442-445.

[21] Bao X, Chen L. Recent progress in distributed fiber optic sensors[J]. sensors, 2012, 12(7): 8601-8639.

[22] Peng X, Zhang C, Anyaoha U, et al. Parameterizing magnetic flux leakage data for

pipeline corrosion defect retrieval[C]//2019 IEEE 28th International Symposium on Industrial Electronics (ISIE). IEEE, 2019: 2665-2670.

[23] Bao X, Chen L. Recent progress in distributed fiber optic sensors[J]. sensors, 2012, 12(7): 8601-8639.

[24] Huang L, Hong X, Yang Z, et al. CNN-LSTM network-based damage detection approach for copper pipeline using laser ultrasonic scanning[J]. Ultrasonics, 2022, 121: 106685.

[25] Das A B, Bhuiyan M I H. Discrimination of focal and non-focal EEG signals using entropy-based features in EEMD and CEEMDAN domains[C]//2016 9th International Conference on Electrical and Computer Engineering (ICECE). IEEE, 2016: 435-438.

[26] Gang L, Hongyan X, Guixian H. The adaptive hybrid algorithm for sea clutter denoising based on CEEMDAN[C]//2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI). IEEE, 2017: 471-476.

[27] Liu H, Mi X, Li Y. Comparison of two new intelligent wind speed forecasting approaches based on wavelet packet decomposition, complete ensemble empirical mode decomposition with adaptive noise and artificial neural networks[J]. Energy Conversion and Management, 2018, 155: 188-200.

[28] Lu J, Yue J, Zhu L, et al. Variational mode decomposition denoising combined with improved Bhattacharyya distance[J]. Measurement, 2020, 151: 107283.

[29] Gao Y, Brennan M J, Joseph P F, et al. A model of the correlation function of leak noise in buried plastic pipes[J]. Journal of Sound and Vibration, 2004, 277(1-2): 133-148.

[30] Yazdekhasti S, Piratla K R, Atamturktur S, et al. Experimental evaluation of a vibration-based leak detection technique for water pipelines[J]. Structure and Infrastructure Engineering, 2018, 14(1): 46-55.

[31] Tan X, Fan L, Huang Y, et al. Detection, visualization, quantification, and warning of pipe corrosion using distributed fiber optic sensors[J]. Automation in construction, 2021, 132: 103953.

[32] Jamshidi V, Davarnejad R. Photon backscatter radiography application for the simulation of corrosion detection inside a pipeline: A novel proposal for 360 corrosion consideration in the pipelines[J]. Applied Radiation and Isotopes, 2021, 176: 109844.

[33] Fu Y W, Kang X W, Yu X X. Detection of localized corrosion in ferromagnetic metal pipe under insulation with pulsed eddy current testing[J]. J. Basic Sci. Eng, 2013, 21: 786-795.

[36] Gong C, Li S, Song Y. Experimental validation of gas leak detection in screw thread connections of galvanized pipe based on acoustic emission and neural network[J]. Structural Control and Health Monitoring, 2020, 27(1): e2460.

[34] Quy T B, Kim J M. Leak detection in a gas pipeline using spectral portrait of acoustic emission signals[J]. Measurement, 2020, 152: 107403.

[35] El-Zahab S, Abdelkader E M, Zayed T. An accelerometer-based leak detection system[J]. Mechanical Systems and Signal Processing, 2018, 108: 276-291.

[36] Yu X, Liang W, Zhang L, et al. Dual-tree complex wavelet transform and SVD based acoustic noise reduction and its application in leak detection for natural gas pipeline[J]. Mechanical Systems and Signal Processing, 2016, 72: 266-285.

[37] Ji J, Li Y, Liu C, et al. Application of EMD Technology in Leakage Acoustic Characteristic Extraction of Gas-Liquid, Two-Phase Flow Pipelines[J]. Shock and Vibration, 2018, 2018(1): 1529849.

[38] Diao X, Chi Z, Jiang J, et al. Leak detection and location of flanged pipes: An integrated approach of principle component analysis and guided wave mode[J]. Safety Science, 2020, 129: 104809.

[39] Kang J, Park Y J, Lee J, et al. Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems[J]. IEEE Transactions on Industrial Electronics, 2017, 65(5): 4279-4289.

[40] Shukla H, Piratla K. Leakage detection in water pipelines using supervised classification of acceleration signals[J]. Automation in Construction, 2020, 117: 103256.

[41] Guo G, Yu X, Liu S, et al. Leakage detection in water distribution systems based on time–frequency convolutional neural network[J]. Journal of Water Resources Planning and Management, 2021, 147(2): 04020101.

[42] Hu X, Zhang H, Ma D, et al. A tnGAN-based leak detection method for pipeline network considering incomplete sensor data[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-10.

[43] Hu X, Zhang H, Ma D, et al. Minor class-based status detection for pipeline network using enhanced generative adversarial networks[J]. Neurocomputing, 2021, 424: 71-83.

[44] Ye Z, Chen Z, Lu H, et al. Analysis of Parallel Misalignment of Gear Coupling in Rotor System Using EEMD-median Filter Method[C]//2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2021, 5: 1113-1118.

[45] Nouioua M, Bouhalais M L. Vibration-based tool wear monitoring using artificial neural networks fed by spectral centroid indicator and RMS of CEEMDAN modes[J]. The International Journal of Advanced Manufacturing Technology, 2021, 115(9): 3149-3161.

[46] Dragomiretskiy K, Zosso D. Variational mode decomposition[J]. IEEE transactions on signal processing, 2013, 62(3): 531-544.

[47] Xu Y, Luo M, Li T, et al. ECG signal de-noising and baseline wander correction based on CEEMDAN and wavelet threshold[J]. Sensors, 2017, 17(12): 2754.

[48] Bowen S, Huaqing W, Gang T, et al. Acoustic signal fault feature extraction method based on MCKD and CEEMDAN [J][J]. Journal of Fudan University (Natural Science), 2019, 3.

[49] Pichler K, Lughofer E, Pichler M, et al. Fault detection in reciprocating compressor valves under varying load conditions[J]. Mechanical Systems and Signal Processing, 2016, 70: 104-119.

[50] Kayaalp F, Zengin A, Kara R, et al. RETRACTED ARTICLE: Leakage detection

and localization on water transportation pipelines: a multi-label classification approach[J]. Neural Computing and Applications, 2017, 28(10): 2905-2914.

[51] Jin H, Zhang L, Liang W, et al. Integrated leakage detection and localization model for gas pipelines based on the acoustic wave method[J]. Journal of Loss Prevention in the Process Industries, 2014, 27: 74-88.

[52] Santos R B, De Sousa E O, Da Silva F V, et al. Detection and on-line prediction of leak magnitude in a gas pipeline using an acoustic method and neural network data processing[J]. Brazilian Journal of Chemical Engineering, 2014, 31: 145-153.

[53] Cho J, Kim H, Gebreselassie A L, et al. Deep neural network and random forest classifier for source tracking of chemical leaks using fence monitoring data[J]. Journal of Loss Prevention in the Process Industries, 2018, 56: 548-558.

[54] Zhang D, Qian L, Mao B, et al. A data-driven design for fault detection of wind turbines using random forests and XGboost[J]. Ieee Access, 2018, 6: 21020-21031.

[55] da Cruz R P, da Silva F V, Fileti A M F. Machine learning and acoustic method applied to leak detection and location in low-pressure gas pipelines[J]. Clean Technologies and Environmental Policy, 2020, 22: 627-638.

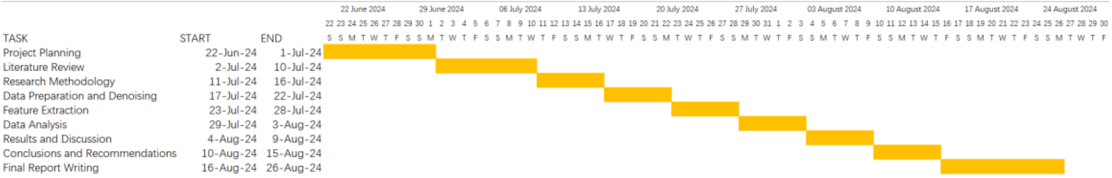
[56] Xiao R, Hu Q, Li J. A model-based health indicator for leak detection in gas pipeline systems[J]. Measurement, 2021, 171: 108843.

[57] Xiao R, Li J. Evaluation of acoustic techniques for leak detection in a complex low-pressure gas pipeline network[J]. Engineering Failure Analysis, 2023, 143: 106897.

[58] Peng Z, Jian J, Wen H, et al. Distributed fiber sensor and machine learning data analytics for pipeline protection against extrinsic intrusions and intrinsic corrosions[J]. Optics Express, 2020, 28(19): 27277-27292.

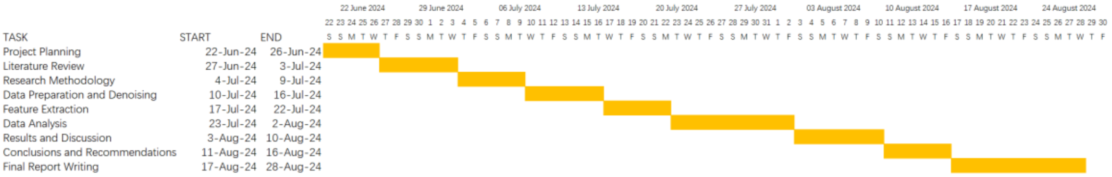
RISK ANALYSIS

1. GANTT CHART



TASK	START	END
Project Planning	22-Jun-24	1-Jul-24
Literature Review	2-Jul-24	10-Jul-24
Research Methodology	11-Jul-24	16-Jul-24
Data Preparation and Denoising	17-Jul-24	22-Jul-24
Feature Extraction	23-Jul-24	28-Jul-24
Data Analysis	29-Jul-24	3-Aug-24
Results and Discussion	4-Aug-24	9-Aug-24
Conclusions and Recommendations	10-Aug-24	15-Aug-24
Final Report Writing	16-Aug-24	26-Aug-24

Fig 1 Original Gantt chart



TASK	START	END
Project Planning	22-Jun-24	26-Jun-24
Literature Review	27-Jun-24	3-Jul-24
Research Methodology	4-Jul-24	9-Jul-24
Data Preparation and Denoising	10-Jul-24	16-Jul-24
Feature Extraction	17-Jul-24	22-Jul-24
Data Analysis	23-Jul-24	2-Aug-24
Results and Discussion	3-Aug-24	10-Aug-24
Conclusions and Recommendations	11-Aug-24	16-Aug-24
Final Report Writing	17-Aug-24	28-Aug-24

Fig 2 Real Gantt chart

2 INITIAL RISK ANALYSIS

1. Risk: Data Availability and Quality

Description: The success of the project heavily relies on the availability and quality of pipeline sensor data. There was a risk that the data could be incomplete, noisy, or unavailable, which could hinder the analysis and affect the accuracy of the results.

Impact: High

Likelihood: Medium

Mitigation Strategy:

Data Sourcing: Ensured early access to the data by collaborating closely with the project stakeholders.

Data Quality Checks: Implemented data quality checks and preprocessing steps to clean and validate the data.

Contingency Plan: Developed a backup plan to use simulated or historical data if real-time data was unavailable.

2. Risk: Technical Challenges in Signal Processing and Machine Learning

Description: The implementation of advanced signal processing techniques (CEEMDAN and BVD) and machine learning algorithms (KNN, SVM, XGBoost) might encounter technical difficulties, such as software bugs or algorithmic inefficiencies.

Impact: Medium

Likelihood: Medium

Mitigation Strategy:

Prototyping: Developed prototypes of key algorithms early in the project to identify potential issues.

Software Testing: Regularly tested the software tools (MATLAB, Python) used in the project to ensure they were functioning correctly.

Technical Support: Maintained access to technical support resources, such as forums, documentation, and expert consultations, to address any technical challenges promptly.

3. Risk: Time Management

Description: The project timeline was tight, with multiple complex tasks needing to be completed within a short timeframe. There was a risk that delays in any task could cascade and affect the overall project timeline.

Impact: High

Likelihood: Medium

Mitigation Strategy:

Gantt Chart Monitoring: Regularly updated the Gantt chart to track progress and adjust timelines as necessary.

Task Prioritization: Prioritized critical tasks and allocated additional resources to ensure they were completed on time.

Buffer Time: Included buffer time in the schedule for unexpected delays.

4. Risk: Data Security and Ethical Considerations

Description: The project involved handling sensitive data related to pipeline integrity, requiring strict data security and ethical considerations.

Impact: High

Likelihood: Low

Mitigation Strategy:

Data Anonymization: Anonymized any identifiable information within the data to protect privacy.

Secure Data Storage: Used secure, encrypted databases for data storage.

Ethical Compliance: Ensured compliance with relevant data protection regulations and obtained necessary approvals for data usage.

3 CHANGES IN RISKS AND STRATEGIES DURING THE PROJECT

1. Change in Risk: Data Quality Issues

New Risk: As the project progressed, it became apparent that the data quality was more variable than initially anticipated, with some datasets containing more noise and missing values.

Updated Mitigation Strategy:

Enhanced Data Cleaning: Implemented more robust data cleaning and preprocessing methods to handle the increased noise and missing values.

Data Augmentation: Used data augmentation techniques to artificially increase the size and diversity of the dataset, improving model training.

2. Change in Risk: Model Performance

New Risk: During the testing phase, it was found that some models did not perform as well as expected on real-world data.

Updated Mitigation Strategy:

Model Tuning: Performed hyperparameter tuning using techniques like GridSearchCV to optimize model performance.

Ensemble Methods: Considered using ensemble methods or hybrid models to improve classification accuracy and robustness.

4 CONCLUSION

The risk analysis and mitigation strategies outlined above played a critical role in managing the project's challenges. By identifying potential risks early and implementing targeted strategies, the project was able to proceed smoothly, with adjustments made as necessary to address any emerging issues. This proactive approach to risk management ensured that the project remained on track and achieved its objectives successfully.

APPLICATION FOR ETHICAL APPROVAL

In order for research to result in benefit and minimise risk of harm, it must be conducted ethically.

The University follows the OECD Frascati manual definition of **research activity**: “creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications”. As such this covers activities undertaken by members of staff, postgraduate research students, and both taught postgraduate and undergraduate students working on dissertations/projects.

The individual undertaking the research activity is known as the “principal researcher”.

This form must be completed and approved prior to undertaking any research activity.

SECTION A: About You (Principal Researcher)

1	Full Name:	Liwei Liu
2	Student Number:	2304725
3	Email address:	2304725@student.uwtsd.ac.uk
4	Programme of Study:	Software engineering and Artificial intelligence
5	Director of Studies/Supervisor:	Dr Gordon Dickers

SECTION B: Internal and External Ethical Guidance Materials

	Please list the core ethical guidance documents that have been referred to during the completion of this form (including any discipline-specific codes of research ethics, location-specific dodes of research ethics, and also any specific ethical guidance relating to the proposed methodology). Please tick to confirm that your research proposal adheres to these codes and guidelines. You may add rows to this table if needed.	
1	**UWTSD Research Ethics & Integrity Code of Practice**	☒
2	**UWTSD Research Data Management Policy**	☒
3		□

SECTION C: Details of Research Activity

1	Indicative title:	Machine Learning-Driven Corrosion Detection and Classification in Pipelines		
2	Proposed start date:	5.2024	Proposed end date:	9.2024

	Introduction to the Research (maximum 300 words in each section) **Ensure that you write for a <u>Non-Specialist Audience</u> when outlining your response to the three points below:** • *Purpose of Research Activity* • *Proposed Research Question* • *Aims of Research Activity* • *Objectives of Research Activity* Demonstrate, briefly, how **<u>Existing Research</u>** has informed the proposed activity and explain • *What the research activity will add to the body of knowledge* • *How it addresses an area of importance.*
3	**Purpose of Research Activity** The primary purpose of this research is to improve the detection and classification of pipeline corrosion. Corrosion in pipelines, especially those carrying oil and gas, poses significant risks including environmental hazards and potential economic losses. Traditional methods of corrosion detection often involve manual inspections or basic sensor technologies, which may not be sufficiently accurate or timely. This research aims to leverage advanced machine learning (ML) techniques to enhance the accuracy of corrosion detection, thereby improving safety and efficiency in the maintenance of industrial pipelines. (this box should expand as you type)
4	**Research Question** The research seeks to answer the following question: How can machine learning techniques be utilized to accurately detect and classify pipeline corrosion using data obtained from advanced sensor systems? (this box should expand as you type)
5	**Aims of Research Activity** The aim of this research is to develop a robust, data-driven system that enhances the detection and classification of pipeline corrosion. By integrating machine learning techniques with data from advanced sensor systems, the research aims to improve the accuracy and reliability of corrosion monitoring systems used in pipelines, ultimately contributing to safer and more efficient pipeline management. (this box should expand as you type)
6	**Objectives of Research Activity** To achieve the research aim, the following objectives have been established: Data Collection: Gather high-quality sensor data from pipelines, focusing on vibration data that may indicate corrosion. Data Processing: Apply advanced signal processing techniques to denoise the collected data, ensuring its suitability for further analysis. Feature Extraction: Identify and extract relevant features from the processed data that are indicative of the pipeline's condition.

	Machine Learning Implementation: Develop and train machine learning models to classify the severity of pipeline corrosion based on the extracted features. System Evaluation: Test and evaluate the developed system using real-world data to ensure its effectiveness in accurately detecting and classifying corrosion. (this box should expand as you type)
	Proposed data collection methods (maximum 600 words) Provide a brief summary of all the methods that **may** be used in the research activity to collect data, making it clear what specific techniques may be used. If methods other than those listed in this section are deemed appropriate later, additional ethical approval for those methods will be needed. You do not need to justify the methods here, but should instead describe how you intend to collect the data necessary for you to complete your project.
7	In this research, the primary data collection method involves the use of advanced optical fiber sensors, specifically Fiber Bragg Grating (FBG) sensors, to monitor and collect data from pipelines. These sensors are strategically placed along an oil pipeline to capture vibration data, which is critical for detecting signs of corrosion. The data collection process will be conducted in collaboration with a petrochemical company in China, which has provided access to an operational pipeline for the purpose of this study. 1. Sensor Placement and Data Acquisition FBG Sensor Installation: The FBG sensors will be installed at various points along the pipeline. These sensors are selected for their high sensitivity to vibrations, which allows them to detect subtle changes in the pipeline's structural integrity. The placement of sensors will be determined based on critical points in the pipeline where corrosion is most likely to occur, such as bends, joints, and areas with previous maintenance history. Data Recording: Once installed, the FBG sensors will continuously monitor the pipeline, capturing vibration data that is indicative of the flow of liquid through the pipeline. This data is expected to reflect any anomalies in the pipeline’s structure, such as those caused by corrosion. The sensors will transmit the data to a central data processing unit in real-time. 2. Data Transmission and Storage Real-Time Data Transmission: The vibration data collected by the FBG sensors will be transmitted in real-time to a dedicated server via a secure network. This ensures that the data is immediately available for processing and analysis without delays. Data Storage: The collected data will be stored in a secure, encrypted database. This storage system will be capable of handling large volumes of data, as the sensors are expected to generate substantial amounts of information continuously over the monitoring period. Access to this database will be restricted to authorized personnel to ensure data security and integrity. 3. Data Preprocessing Noise Reduction: Before the data can be analyzed, it must be preprocessed to remove noise and other irrelevant signals that could interfere with accurate corrosion detection. This will involve using advanced signal processing techniques, specifically Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN). This method decomposes the vibration signals into intrinsic mode functions (IMFs) to isolate relevant components of the data.

Selection of Relevant Data: After noise reduction, the Bhattacharyya Variance Distance (BVD) algorithm will be used to select the most relevant IMFs for further analysis. This ensures that only the data most indicative of corrosion is retained, reducing the computational load and enhancing the accuracy of subsequent machine learning models.

4. Data Annotation

Manual Annotation: A subset of the collected data will be manually annotated by experts to identify known instances of corrosion. This annotated data will serve as a training set for the machine learning models.

Automated Annotation: In addition to manual annotation, automated techniques will be employed to label the remaining data based on patterns detected in the annotated subset. This will help in scaling the training process while maintaining accuracy.

5. Data Segmentation and Feature Extraction

Data Segmentation: The preprocessed data will be segmented into smaller time windows to facilitate detailed analysis. Each segment will represent a specific period of pipeline operation and will be analyzed individually.

Feature Extraction: Key features indicative of the pipeline's condition, such as mean frequency, median frequency, peak frequency, spectral density index, and the peak frequency of the Hilbert marginal spectrum, will be extracted from each data segment. These features will serve as inputs for the machine learning models used for corrosion detection and classification.

6. Machine Learning Dataset Preparation

Dataset Creation: The extracted features from the segmented data will be compiled into a comprehensive dataset. This dataset will be used to train, validate, and test various machine learning models developed for the classification of pipeline corrosion severity.

Data Splitting: The dataset will be split into training, validation, and test sets to ensure that the machine learning models are accurately trained and can generalize well to unseen data.

The described data collection methods will provide a comprehensive and high-quality dataset necessary for the successful completion of this research. If additional data collection methods become necessary as the research progresses, further ethical approval will be sought to ensure compliance with research standards.

(this box should expand as you type)

SECTION D: Scope of Research Activity

	Will the research activity include:	**YES**	**NO**
1	Use of a questionnaire or similar research instrument?	☐	☒
2	Use of interviews?	☐	☒
3	Use of focus groups?	☐	☒
4	Use of participant diaries?	☐	☒
5	Use of video or audio recording?	☒	☐
6	Use of computer-generated log files?	☐	☒
7	Participant observation with their knowledge?	☐	☒
8	Participant observation without their knowledge?	☐	☒

9	Access to personal or confidential information without the participants' specific consent?	☐	☒
10	Administration of any questions, test stimuli, presentation that may be experienced as physically, mentally or emotionally harmful / offensive?	☐	☒
11	Performance of any acts which may cause embarrassment or affect self-esteem?	☐	☒
12	Investigation of participants involved in illegal activities?	☐	☒
13	Use of procedures that involve deception?	☐	☒
14	Administration of any substance, agent or placebo?	☐	☒
15	Working with live vertebrate animals?	☐	☒
16	Procedures that may have a negative impact on the environment?	☐	☒
17	Other primary data collection methods. Please indicate the type of data collection method(s) below.	☒	☐
	Details of any other primary data collection method: or this research, the primary data collection methods will focus on gathering vibration data from pipelines using advanced sensor technology. Below are the specific types of data collection methods that will be employed: Sensor-Based Data Collection: Type: Vibration Data Monitoring via Fiber Bragg Grating (FBG) Sensors. Purpose: To monitor real-time structural integrity of pipelines by detecting vibrations caused by the flow of liquids, which may indicate corrosion. Process: FBG sensors are installed along the pipeline, capturing continuous vibration data. The data is transmitted to a central server for storage and further analysis. Manual Data Annotation: Type: Expert Annotation of Sensor Data. Purpose: To create a labeled dataset that identifies known instances of corrosion, which will be used to train machine learning models. Process: Experts will manually review a subset of the sensor data to annotate it with labels indicating the presence or absence of corrosion. Automated Data Labeling: Type: Automated Annotation Based on Patterns Identified in Expert-Labeled Data. Purpose: To expand the labeled dataset by applying automated techniques to label additional data, facilitating the training of machine learning models. Process: Algorithms will be used to detect and label similar patterns in the remaining data, based on the manually annotated subset. If any additional primary data collection methods are deemed necessary during the research, they will be clearly documented and additional ethical approval will be sought to ensure compliance with all research standards and guidelines. (this box should expand as you type)		

If you have ticked NO to every question then the research activity is (ethically) low risk and you may skip section E and continue to section F.

If YES to any question, then no research activity should be undertaken until full ethical approval has been obtained.

SECTION E: Intended Participants

	Who are the intended participants:	**YES**	**NO**
1	Students or staff at the University?	☐	☒
2	Adults (over the age of 18 and competent to give consent)?	☐	☒
3	Vulnerable adults?	☐	☒
4	Children and Young People under the age of 18? (Consent from Parent, Carer or Guardian will be required)	☐	☒
5	Prisoners?	☐	☒
6	Young offenders?	☐	☒
7	Those who could be considered to have a particularly dependent relationship with the investigator or a gatekeeper?	☐	☒
8	People engaged in illegal activities?	☐	☒
9	Others. Please indicate the participants below, and specifically any group who may be unable to give consent.	☐	☒
	Details of any other participant groups: Complete this only if your participants cannot give consent. This includes animals (this box should expand as you type)		

	Participant numbers and source Provide an estimate of the expected number of participants. How will you identify participants and how will they be recruited?	
10	How many participants are expected?	Ballpark figures are fine, but make sure that you explain how you will identify and contact your participants. No participants are involved as this research does not include human or animal subjects. *(this box should expand as you type)*
11	Who will the participants be?	Not applicable. *(this box should expand as you type)*

12	How will you identify the participants?	Not applicable, as the study does not involve direct participants. *(this box should expand as you type)*

	Information for participants:	**YES**	**NO**	**N/A**
13	Will you describe the main research procedures to participants in advance, so that they are informed about what to expect?	☐	☐	☒
14	Will you tell participants that their participation is voluntary?	☐	☐	☒
15	Will you obtain written consent for participation?	☐	☐	☒
16	Will you explain to participants that refusal to participate in the research will not affect their treatment or education (if relevant)?	☐	☐	☒
17	If the research is observational, will you ask participants for their consent to being observed?	☐	☐	☒
18	Will you tell participants that they may withdraw from the research at any time and for any reason?	☐	☐	☒
19	With questionnaires, will you give participants the option of omitting questions they do not want to answer?	☐	☐	☒
20	Will you tell participants that their data will be treated with full confidentiality and that, if published, it will not be identifiable as theirs?	☐	☐	☒
21	Will you debrief participants at the end of their participation, in a way appropriate to the type of research undertaken?	☐	☐	☒
22	If NO to any of above questions, please give an explanation			
	You should be able to tick YES for all of these questions. If not, then explain why not in this box. The project focuses on data collection from pipeline sensors to monitor and analyze corrosion using signal processing and machine learning techniques. Since no individuals are involved in providing data, being observed, or participating in any research activities, there is no need to inform participants, obtain consent, or address issues related to withdrawal, confidentiality, or debriefing. *(this box should expand as you type)*			

	Information for participants:	**YES**	**NO**	**N/A**
24	Will participants be paid?	☐	☐	☒
25	Is specialist electrical or other equipment to be used with participants?	☐	☐	☒
26	Are there any financial or other interests to the investigator or University arising from this study?	☐	☐	☒
27	Will the research activity involve deliberately misleading participants in any way, or the partial or full concealment of the specific study aims?	☐	☐	☒
28	If YES to any question, please provide full details			

	You should be able to tick NO for most of these questions. For any cases that you have ticked YES then provide details in this box. If you are using cameras/voice recorders to record interviews then please state that in this box. *(this box should expand as you type)*

SECTION F: Anticipated Risks

1	**Risks to participants** For example: sector-specific health & safety, emotional distress, financial disclosure, physical harm, transfer of personal data, sensitive organisational information. If you have identified in section D that there are no participants then enter N/A and go skip to question 3.	
	Risk to participants: There are always risks. Do not write N/A unless you have no participants. N/A *(this box should expand as you type)*	*How you will mitigate the risk to participants:* N/A *(this box should expand as you type)*
2	If research activity may include sensitive, embarrassing or upsetting topics (e.g. sexual activity, drug use) or issues likely to disclose information requiring further action (e.g. criminal activity), give details of the procedures to deal with these issues, including any support/advice (e.g. helpline numbers) to be offered to participants. Note that where applicable, consent procedures should make it clear that if something potentially or actually illegal is discovered in the course of a project, it may need to be disclosed to the proper authorities	
	This research activity does not involve any sensitive, embarrassing, or upsetting topics such as sexual activity, drug use, or any other issues that are likely to disclose information requiring further action, such as criminal activity. The focus of the research is on the technical aspects of pipeline corrosion detection using signal processing and machine learning techniques, with no involvement of human subjects or personal data. As a result, there are no procedures required to deal with sensitive issues, and no support or advice (e.g., helpline numbers) needs to be offered to participants, as there are no participants involved in this study. Additionally, since the research does not involve the collection or analysis of any information that could potentially be illegal or require disclosure to authorities, the consent procedures related to such disclosures are not applicable in this case. The research strictly adheres to ethical guidelines relevant to non-participant technical studies, ensuring that all activities are conducted in a manner that poses no risk of harm or distress to individuals. *(this box should expand as you type)*	
3	**Risks to the investigator** For example: personal health & safety, physical harm, emotional distress, risk of accusation of harm/impropriety, conflict of interest	
	Risk to the investigator: There are always risks. Do not write NA. While conducting this research, the primary risks to the investigator include:	*How you will mitigate the risk to the investigator:* Personal Health & Safety: The investigator will adhere strictly to all safety protocols established by the industrial facility, including wearing

	Personal Health & Safety: There may be potential health and safety risks when working with or near industrial pipelines, particularly if on-site data collection or inspections are required. Physical Harm: There is a risk of physical injury, especially when handling equipment, or if working in hazardous environments such as industrial facilities where the pipelines are located. Emotional Distress: While unlikely, the investigator might experience stress due to the technical challenges and pressure associated with meeting research deadlines. Risk of Accusation of Harm/Impropriety: Though minimal, there is a potential risk of being accused of causing damage or interference during the setup or data collection processes. Conflict of Interest: There might be a conflict of interest if the research outcomes benefit the investigator personally or if there are ties to the organization providing the pipeline data. *(this box should expand as you type)*	appropriate personal protective equipment (PPE) such as hard hats, gloves, and safety glasses. The investigator will also undergo any necessary safety training before entering the facility. Physical Harm: The investigator will minimize exposure to hazardous environments by conducting as much of the research as possible off-site, utilizing remote data collection methods. When on-site presence is required, the investigator will follow all safety guidelines and avoid high-risk areas unless accompanied by a trained safety officer. Emotional Distress: The investigator will manage workload effectively to reduce stress, ensuring adequate breaks and time for rest. Support from peers and supervisors will be sought when technical challenges arise, and any signs of stress or emotional strain will be addressed promptly. Risk of Accusation of Harm/Impropriety: The investigator will maintain clear communication with all stakeholders involved in the research, ensuring that all activities are documented and conducted transparently. Proper permissions and authorizations will be obtained before any interaction with the equipment or data. Conflict of Interest: The investigator will declare any potential conflicts of interest at the outset and will ensure that the research remains objective and unbiased. Any ties to the organization providing data will be transparently disclosed, and steps will be taken to prevent these ties from influencing the research outcomes. *(this box should expand as you type)*
4	**University/institutional risks** For example: adverse publicity, financial loss, data protection	
	Risk to the University: There are always risks. Do not write NA. Adverse Publicity: There is a risk of negative publicity if the research results are misinterpreted or if any issues arise during the research that reflect poorly on the university.	*How you will mitigate the risk to the University:* Adverse Publicity: The research will be conducted with the utmost professionalism and transparency. Clear communication strategies will be in place to ensure that the university's involvement is positively portrayed. Additionally, any publications or public presentations of the

	Financial Loss: If the research encounters significant issues or delays, there could be a financial impact due to wasted resources or the need for additional funding. Data Protection: The handling of sensitive or proprietary data from the petrochemical company poses a risk of data breaches or violations of data protection laws, which could lead to legal consequences and damage the university's reputation. *(this box should expand as you type)*	research will be reviewed to ensure accuracy and alignment with the university's values. Financial Loss: The research project will be carefully managed to stay within budget and on schedule. Regular progress reports will be provided to university administrators to monitor the project's financial health. Contingency plans will be in place to address any unforeseen challenges that could impact the budget. Data Protection: The university will implement strict data protection protocols in line with relevant legislation, such as the General Data Protection Regulation (GDPR). Data will be securely stored and only accessible to authorized personnel. Any data shared with third parties will be anonymized to protect the identities and proprietary information of the petrochemical company involved in the study. *(this box should expand as you type)*
5	**Environmental risks** For example: accidental spillage of pollutants, damage to local ecosystems	
	Risk to the environment: You may write NA if there are no research-related environmental risks. Driving to the university does not count as a risk. N/A *(this box should expand as you type)*	*How you will mitigate the risk to environment:* N/A *(this box should expand as you type)*

SECTION G: Feedback, Consent and Confidentiality

If you have identified in section D that there are no participants then enter skip this section and continue to section H.

1	**Feedback** What de-briefing and feedback will be provided to participants, how will this be done and when?
	You don’t need to email your participants with your final report. A good alternative is to set up an email address that they will be able to contact for further details or results. *(this box should expand as you type)*
2	**Informed consent** Describe the arrangements to inform potential participants, before providing consent, of what is involved in participating. Describe the arrangements for participants to provide full consent before data collection begins. If gaining consent in this way is inappropriate, explain how consent will be obtained and recorded in accordance with prevailing data protection legislation.

	If you are using a paper questionnaire then you should have the participants sign an appropriate consent form. These forms will count as personal data and should be noted as such in section J. If you are using an online questionnaire, then you should have a screen before the questions start that acts as a consent form, informing participants that by clicking on the NEXT button they are providing consent. *(this box should expand as you type)*
3	**Confidentiality / Anonymity** Set out how anonymity of participants and confidentiality will be ensured in any outputs. If anonymity is not being offered, explain why this is the case.
	Do not collect names unless you really need them. Do not name participants or organisations in any research publications (including the thesis) without their explicit permission. *(this box should expand as you type)*

SECTION H: Data Protection and Storage

	Does the research activity involve personal data (as defined by the General Data Protection Regulation 2016 “GDPR” and the Data Protection Act 2018 “DPA”)?	**YES**	**NO**
1	***“Personal data”*** *means any information relating to an identified or identifiable natural person (‘data subject’). An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. Any video or audio recordings of participants is considered to be personal data.*	☐	☒
	If YES, provide a description of the data and explain why this data needs to be collected:		
2	This includes audio/video data of participants, but can also include IP addresses and usernames. Names, addresses and emails also count, as do consent forms. *(this box should expand as you type)*		
	Does it involve special category data (as defined by the GDPR)?	**YES**	**NO**
3	***“Special category data”*** *means sensitive personal data consisting of information as to the data subjects’ –* *(a) racial or ethnic origin,* *(b) political opinions,* *(c) religious beliefs or other beliefs of a similar nature,* *(d) membership of a trade union (within the meaning of the Trade Union and Labour Relations (Consolidation) Act 1992),* *(e) physical or mental health or condition,* *(f) sexual life,*	☐	☒

	(g) genetics, *(h) biometric data (as used for ID purposes),*		
	If YES, provide a description of the special category data and explain why this data needs to be collected:		
4	What counts as ‘sensitive’ will differ between cultures. Any information on behaviour that is not in accordance with cultural norms would count as sensitive personal data. *(this box should expand as you type)*		

	Will data from the research activity (collected data, drafts of the thesis, or materials for publication) be stored in any of the following ways?	**YES**	**NO**
5	Manual files (i.e. in paper form)?	☐	☒
6	University computers?	☐	☒
7	Private company computers?	☐	☒
8	Home or other personal computers?	☒	☐
9	Laptop computers/ CDs/ Portable disk-drives/ memory sticks?	☒	☐
10	“Cloud” storage or websites?	☐	☒
11	Other – specify:	☐	☒
12	For all stored data, explain the measures in place to ensure the security of the data collected, data confidentiality, including details of backup procedures, password protection, encryption, anonymisation and pseudonymisation:		
	If possible, save your data on computers that are secure and regularly backed up. Many cloud services only provide GDPR-compliant storage for business customers. An example of suitable text is given below. *All data will be kept in password protected cloud storage on the University Office 365 system which will not be shared. Audio/visual data will be transcribed and would be shown to participants to check accuracy of reporting. Any USB sticks used to store or transfer data will be password protected. All participants will be given a unique identifier to ensure confidentiality and this list will be kept securely in the password protected folder.* All data will be securely stored in password-protected cloud storage provided by the University’s Office 365 system, which is compliant with GDPR regulations. Access to this data will be restricted to authorized personnel involved in the research project. Any physical storage devices used, such as USB sticks, will also be password protected to ensure additional security during data transfer. To further ensure data confidentiality, all participants will be assigned a unique identifier, which will be used in place of their personal information. This anonymization process will ensure that the data cannot be directly linked back to individual participants. The list of identifiers and their corresponding participant details will be stored separately in a secure, password-protected folder. Additionally, any sensitive data, such as audio or visual recordings, will be encrypted before storage. Transcriptions of any recorded data will be provided to participants for review to ensure the		

	accuracy of the information captured. Regular backups of all data will be performed to prevent data loss, and these backups will be stored in the same secure, password-protected environment. By implementing these security measures, the confidentiality and integrity of all collected data will be maintained throughout the research project. *(this box should expand as you type)*

	Data Protection		
	Will the research activity involve any of the following activities:	**YES**	**NO**
13	Electronic transfer of data in any form?	☒	☐
14	Sharing of data with others at the University outside of the immediate research team?	☐	☒
15	Sharing of data with other organisations?	☐	☒
16	Export of data outside the UK or importing of data from outside the UK?	☐	☒
17	Use of personal addresses, postcodes, faxes, emails or telephone numbers?	☐	☒
18	Publication of data that might allow identification of individuals?	☐	☒
19	If YES to any question, please provide full details, explaining how this will be conducted in accordance with the GDPR and Data Protection Act (2018) (and/or any international equivalent):		
	This includes data such as drafts of your thesis as well as experimental or survey data. An example of suitable text is given below. *All data will be encrypted and kept in password protected cloud storage on the University Office 365 system which will not be shared. Any USB sticks used to store or transfer data will be password protected. All data transfers will be encrypted and password protected. All participants will be given a unique identifier to ensure confidentiality and this list will be kept securely in the password protected folder. The data will be stored until the completion of the project and then deleted. In accordance with the DPA2018, participants will have the right to ask to see what data is held relating to them, and this data will be deleted immediately if the participant requests this, in which case the data will not be used in the project.* Data Encryption and Security: All data transferred electronically will be encrypted using secure protocols such as HTTPS and SFTP. Additionally, data will be stored in password-protected cloud storage on the University Office 365 system, ensuring that data integrity and confidentiality are maintained in compliance with GDPR. *(this box should expand as you type)*		
20	List all who will have access to the data generated by the research activity:		
	Normally the principal researcher, possibly also the supervisor and, if the project has an industrial partner, a representative of that partner. Possibly also external examiner or second marker? *(this box should expand as you type)*		
21	List who will have control of, and act as custodian(s) for, data generated by the research activity:		
	Usually the principal researcher. Principal Researcher: Will have full access to the data for the purposes of analysis and publication.		

	Research Supervisor: May access the data for supervisory and quality assurance purposes. University IT Personnel: Might access the data only in case of technical support or maintenance within the scope of their official duties and under strict confidentiality agreements. *(this box should expand as you type)*	
22	Give details of data storage arrangements, including security measures in place to protect the data, where data will be stored, how long for, and in what form.	
	All data will be encrypted and kept in password protected cloud storage on the University Office 365 system which will not be shared. Any USB sticks used to store or transfer data will be password protected, and will be reformatted at the end of the project in order to destroy the data. The data will be stored until the completion of the project and then deleted. storage, which is password-protected. Additional local copies, if necessary, will be stored on encrypted and password-protected devices. Duration of Storage: Data will be retained until the completion of the research project, after which it will be securely deleted unless further retention is required by law or university policy. Data Deletion: Upon the project's completion, data stored on any portable devices such as USB sticks will be securely deleted by reformatting the devices. *(this box should expand as you type)*	
22	Confirm that you have read the UWTSD guidance on data management (see https://www.uwtsd.ac.uk/library/research-data-management/)	☒
23	Confirm that you are aware that you need to keep all data until after your research has completed or the end of your funding	☒

SECTION I: Declaration

	The information which I have provided is correct and complete to the best of my knowledge. I have attempted to identify any risks and issues related to the research activity and acknowledge my obligations and the rights of the participants. In submitting this application I hereby confirm that I undertake to ensure that the above named research activity will meet the University's Research Ethics and Integrity Code of Practice which is published on the website: https://www.uwtsd.ac.uk/research/research-ethics/		
1	**Signature of applicant:**	Liwei Liu	**Date: 27/8/2024**
2	Director of Studies/Supervisor:	Dr Gordon Dickers	**Date:27/8/2024**
3	Signature:	Liwei Liu	

FOR INTERNAL USE ONLY:

	Ethical approval given

1	**Signature of assessor:**		**Date:**
2	Name:		
3	Role:		

GLOSSARY

FBG (Fiber Bragg Grating): A type of optical fiber sensor that reflects specific wavelengths of light and is used for measuring strain, temperature, and other physical parameters in pipelines.
EMD (Empirical Mode Decomposition): A technique used to decompose a signal into its intrinsic mode functions without relying on a predefined basis function, useful in analyzing non-linear and non-stationary signals.
IMF (Intrinsic Mode Function): A function derived from a signal through empirical mode decomposition, representing a simple oscillatory mode inherent in the signal.
CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise): An advanced signal processing technique used to decompose complex signals into simpler intrinsic mode functions (IMFs), helping to reduce noise and enhance feature extraction.
BVD (Bhattacharyya Variance Distance): A method used in signal processing to measure the similarity between two probability distributions, often used to select effective components in signal reconstruction.