People’s Democratic Republic of Algeria

الجمهورية الجزائرية الديمقراطية الشعبية

Ministry of Higher Education and Scientific Research

وزارة التعليم العالي و البحث العلمي

University of Kasdi Merbah Ouargla

Faculty of New Information Technologies and Communication

Department of Computer Science and Information Technology

Thesis Submitted in Candidacy for a Master’s Degree in Computer Science

Option: Industrial Computer Science

Presented by: Hadjoudj Izdihar, Legougui Ziad

Predictive AI Models for Detecting Pipeline Leaks in the Energy Industry

JURY MEMBERS:

Dr. OUSSAMA AIDI	UKM OUARGLA	PRESIDENT
Dr. LEILA AMRANE	UKM OUARGLA	EXAMINER
Dr. BASMA HAMROUNI	UKM OUARGLA	SUPERVISOR
Dr. KHADRA BOUANANE	UKM OUARGLA	CO-SUPERVISOR

ACADEMIC YEAR:

2023/2024

ACKNOWLEDGMENTS

*"First and foremost, I would like to express my deepest gratitude to the one above all of us, the omnipresent **Allah**, for answering my prayers and providing me with the strength to overcome challenges. Thank you, Lord, for guiding me and bestowing upon me the perseverance to see this journey through.*

I would also like to acknowledge and thank my beloved parents for their unwaver- ing support, love, and guidance. Your presence in my life has been a constant source of strength, and I am forever grateful for your sacrifices and belief in my abilities. Lastly, I would like to extend my appreciation to all those who have contributed in any way to the completion of this thesis. Thank you for being a part of my academic journey and for your unwavering belief in my potential.

*I extend my sincere appreciation to my supervisor **Dr. BASMA HAMROUNI** , for her unwavering support, sage advice, insightful criticisms, and constant encouragement. Her guidance has been invaluable, and her belief in my capabilities has motivated me to push beyond my limits, Similarly, **Dr. KHADRA BOUANANE** has been an invaluable source of support, offering unique perspective and expertise that greatly enriched my understanding and approach to this study. Together, their combined guidance has not only shaped this thesis but has also significantly contributed to my growth as a researcher and thinker.*

*To all my **Teachers** who have shared their knowledge and wisdom with me throughout my academic journey, I express my heartfelt thanks. Your dedication to education and passion for teaching have played a significant role in shaping my intellectual growth. I am grateful for the opportunities you have provided me to expand my horizons. Thanks for making me see this adventure through to the end. "*

IZDIHAR HADJOUDJ

ABSTRACT

Pipelines serve as critical infrastructure for transporting oil and gas, but any leaks in these systems can lead to severe outcomes, including fires, injuries, environmental pollution, and property destruction. Thus, maintaining the integrity of pipelines is paramount for ensuring a safe and sustainable energy supply. This thesis investigates the application of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) in enhancing leak detection within oil and gas pipeline systems, key to ensuring environmental safety and economic stability. Through a comprehensive review and data-driven methodologies, the study demonstrates how ML algorithms, including neural networks and deep learning models, significantly outperform traditional leak detection methods in accuracy and timeliness. Herein, the research introduces machine learning-based anomaly detection models proposed to solve the problem of oil and gas pipeline leakage. To address this, several machine learning and deep learning algorithms, namely, Random Forest, Support Vector Machine, K-Nearest Neighbor, Gradient Boosting, Decision Tree, Convolutional Neural Network, and Multi-Layer Perceptron, were employed to develop robust detection models for pipeline leaks. Among these, the Support Vector Machine algorithm, achieving an accuracy of 96.6%, notably outperformed other models, thereby confirming its efficacy as a highly accurate tool for detecting leakage in oil and gas pipelines.

Keywords : Artificial Intelligence, Machine Learning, Deep learning, Leak Detection, Oil and Gas Pipeline Systems, Environmental Safety, Economic Stability, Neural Networks, energy sector.

الملخص

تُعد خطوط الأنابيب بنية تحتية حاسمة لنقل النفط والغاز، لكن أي تسربات في هذه الأنظمة قد تؤدي إلى نتائج خطيرة، بما في ذلك الحرائق، الإصابات، تلوث البيئة، وتدمير الممتلكات. وبالتالي، فإن الحفاظ على سلامة خطوط الأنابيب أمر بالغ الأهمية لضمان توفير طاقة آمن ومستدام. تستقصي هذه الأطروحة تطبيق الذكاء الاصطناعي ،تعلم الآلة و تعلم العميق في تعزيز كشف التسربات داخل أنظمة خطوط أنابيب النفط والغاز، الأمر الذي يُعد مفتاحًا لضمان الأمان البيئي والاستقرار الاقتصادي. من خلال مراجعة شاملة ومنهجيات مستندة إلى البيانات، تُظهر الدراسة كيف أن خوارزميات تعلم الآلة، بما في ذلك الشبكات العصبية ونماذج التعلم العميق، تتفوق بشكل كبير على طرق كشف التسرب التقليدية من حيث الدقة وسرعة الاستجابة. تُقترح هنا نماذج كشف الشذوذ المستندة إلى تعلم الآلة لحل مشكلة تسربات خطوط الأنابيب للنفط والغاز. تم مقارنة مجموعة من خوارزميات تعلم الآلة والتعلم العميق، لتطوير نماذج لكشف تسربات الأنابيب. وقد تفوقت خوارزمية آلة الدعم الناقل ، بدقة تصل إلى ٦.٩٦ %، أداءً يفوق الخوارزميات الأخرى في كشف تسربات الأنابيب، مما أثبت كفاءتها كنموذج دقيق لكشف التسربات في خطوط أنابيب النفط والغاز.

الكلمات المفتاحية: الذكاء الاصطناعي، التعلم الآلي، التعلم العميق، كشف التسرب، النفط والغاز أنظمة خطوط الأنابيب ، السلامة البيئية ، الاستقرار الاقتصادي ، الشبكات العصبية ، قطاع الطاقة.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

DT	Decision Trees
RF	Random Forests
SVM	Support Vector Machines
LR	Logistic Regression
KNN	K-Nearest Neighbors
AI	Artificial Intelligence
ML	Machine learning
DL	Deep Learning

CHAPTER I

GENERAL INTRODUCTION

Pipelines are considered the most economical and advanced technology that is currently utilized for oil and gas transportation in the world, contributing to 57.5% global primary energy consumption . to support transportation of flammable and poisonous fluids such as crude oil, natural gas, and refined petroleum products [13]. They carry fluids in larger volume, safer way, and more environmental friendly compared to trucks . However, like any other equipment, pipelines can have various failures to some degree. One of which is studied in this thesis work that focused on leakage , Leakage in the pipelines can initiate the occurrence of progressive accidents, such as fluid spillage, fire, and explosion. The exposure of that accidents can lead to the injuries, even worst, fatalities, environmental and asset damages, bad reputations, financial distress, and more other negative impacts. Thus, ensuring their integrity and functionality is vital for the economy, the environment, and public safety. However, pipeline leaks pose a significant challenge, leading to severe environmental damage, economic losses, and safety hazards.

The detection and prevention of pipeline leaks have thus become paramount in the field of pipeline management. There are many widely agreed-upon risk assessment parameters, including third-party interference, corrosion, design, pressure, temperature, incorrect operation, and third-party. Among these, corrosion stands out as the most critical factor. Extensive research and expert analysis have consistently highlighted corrosion as a leading cause of pipeline leaks. Statistical data further corroborate this, showing a high incidence of leaks attributed to corrosion-related issues. Corrosion has been identified as the most influential factor contributing to pipeline leaks, a conclusion supported by numerous scientific studies, expert opinions, and statistical analyses of incident data. Many incident analyses have shown that leaking phenomena in pipelines are primarily caused by corrosion. Given the significant impact of corrosion, the study of pipeline leaks must inherently focus on understanding and mitigating corrosion. Effective pipeline management strategies must prioritize corrosion detection, monitoring, and prevention to ensure the safety, reliability, and longevity of pipeline infrastructure. Therefore, addressing corrosion is not just a technical necessity but a crucial aspect of comprehensive pipeline risk management.

Machine learning (ML) algorithms are widely recognized as the leading approach for developing predictive models in complex engineering, energy, and environmental problems [2]. ML significantly enhances predictive accuracy, reduces reliance on conventional and manual data analysis, enables autonomous information processing, and efficiently handles high-volume, high-velocity, and high-variety data [35]. Its growing popularity across various fields is also due to its remarkable ability to learn and construct predictive models from performance data, even when dealing with incomplete and empirical datasets [2]. ML algorithms are adept at addressing complex problems and can discern intricate patterns without needing prior knowledge of the relationships between independent and dependent variables [7].

In the context of engineering materials, such as oil and gas pipelines, ML algorithms are particularly advantageous due to their ability to accurately and quickly estimate mechanical properties at a lower cost compared to traditional modeling methods [5]. ML algorithms used for detecting defects in pipelines can be categorized based on the learning method employed: supervised, semi-supervised, unsupervised, or reinforcement learning [2].

The literature on corrosion prediction in pipelines can be divided into two main areas, Computer Vision and Deep Learning Techniques, and Numerical Data based Machine Learning and Deep Learning Methodologies for Corrosion Defect Prediction , Aljameel et al. [3] investigate machine learning-based anomaly detection models to address leaks in oil and gas pipelines, comparing five algorithms: RF, SVM, KNN, GB, and DT. Their study finds that the Support Vector Machine (SVM) algorithm performs best, achieving an accuracy of 97.4 , making it a highly effective model for detecting pipeline leaks. Seghier et al. [37] focus on predicting internal corrosion rates using robust ensemble learning techniques, with the Extreme Gradient Boosting model showing superior performance, demonstrating an RMSE of 0.031 mm/y and a performance index of 0.61. Luo et al. [27] developed an SVM-based model to predict gas pipeline corrosion rates using various pipeline parameters, providing new insights for risk management and maintenance. Naveed Aslam et al. [4] employ AI algorithms, including the DeWaard Model, Norsok Model, and Leak Rate Model, to predict corrosion and leak rates, using a combination of data consistency checks and a type 2 fuzzy logic subroutine for refinement.

This thesis expands upon previous studies of the numerical Data based Machine Learning and Deep Learning Methodologies for Corrosion Defect Prediction, a comparison of support vector machine (SVM), k-nearest neighbours (KNN), random forest (RF), gradient boosting (GB), and the decision tree (DT) algorithm, which were used and compared algorithms to detect pipeline leakage using industrial datasets is performed; significantly broadening the scope of comparison beyond the initial five algorithms. Specifically, the addition of Multi-Layer Perceptrons (MLP) and Convolutional Neural Networks (CNN) enhances the comparison and Evaluation methodology in terms of accuracy, precision, recall, F1-score, accuracy,and ROC-AUC, providing deeper insights into the capabilities of these advanced models in handling datasets and tasks. This approach allows for a more comprehensive evaluation of each algorithm's performance and suitability for various applications, especially in fields of energetic sector. By including MLP and CNN, the thesis aims to offer a robust assessment of both traditional and modern machine learning techniques, facilitating a better understanding of their practical implications and potential benefits in real-world scenarios.

01 Motivation

The impetus for this thesis is driven by the critical need to enhance the safety and efficiency of pipeline operations, which are pivotal for global energy distribution. Despite their significance, pipelines face persistent risks from corrosion, a major factor in leaks that lead to environmental, economic, and safety hazards. Conventional detection methods, often reactive and labor-intensive, fall short in addressing these issues preemptively. This thesis is motivated by the necessity to adopt advanced Machine Learning (ML) techniques that promise more timely and accurate predictions of potential pipeline failures. By conducting a thorough comparative analysis of various ML models, including innovative approaches like Multi-Layer Perceptrons (MLP) and Convolutional Neural Networks (CNN), this research aims to identify and refine the best strategies for real-time corrosion monitoring and leak detection. This work seeks to bridge gaps in current methodologies, leveraging cutting-edge technology to forge advancements in pipeline integrity management, thus supporting safer and more reliable energy infrastructure.

02 Research Objectives

This thesis sets out to achieve the following objectives To conduct a comprehensive review of existing pipeline leak detection methods, with a focus on the application of AI and ML techniques, To comparison of ML algorithms to detect pipeline leakage using industrial datasets is performed; and Evaluation methodology in terms of accuracy, precision, recall, F1-score, accuracy, and ROC-AUC is proposed.

03 Contributions

In this study, we conducted a theoretical study and an in-depth analysis of pipeline leaks, focusing on the factors that cause these incidents. This comprehensive theoretical study was carried out to understand the various causes of leaks, among which we identified corrosion as the most influential factor. This discovery allowed us to specifically target corrosion in our modeling approach.

- Subsequently, we formulated the pipeline leak problem as an anomaly detection task, with the primary objective being the prediction of these incidents. Anomaly detection is a suitable approach for identifying deviant behaviors in the data, which is crucial for preventing leaks before they become critical.
- For the first time, we applied MLP and CNN algorithms to a specific dataset chosen for this experiment. This approach is innovative because previous studies using the same dataset have focused on other algorithms, such as random forests or support vector machines. By introducing MLP and CNN, we not only expanded the range of techniques used for this type of problem but also aimed to improve predictive performance compared to existing scientific work
- Thus, this research contributes to the advancement of knowledge in pipeline leak detection by proposing more robust and potentially more accurate methods for predicting these critical events. By integrating deep learning techniques, we open new perspectives for proactive pipeline monitoring, enabling better leak prevention and more effective management of infrastructure integrity.

04 Thesis Outline

This thesis is organized into several chapters to systematically address the research objectives

- **Chapter 1: Introduction** Introduces the research problem, significance, objectives, and structure of the thesis.
- **Chapter 2: Pipeline Risk Factors and Leak Detection Techniques** This chapter introduces the concept of pipeline integrity, detailing the various risk factors that can jeopardize it, such as environmental impacts, material fatigue, and operational errors. It examines the potential consequences of these risks, particularly focusing on pipeline safety and the prevalence of leaks. The latter section of the chapter reviews existing methods for detecting leaks, categorizing them into hardware-based and software-based techniques. It also discusses the limitations of current leak detection technologies and sets the stage for exploring advanced methods in subsequent chapters.
- **Chapter 3: Machine Learning For Corrosion Prediction** This chapter defines corrosion and its detrimental effects on pipeline systems. It outlines the traditional methods used for corrosion detection and management, and introduces how AI and ML technologies are revolutionizing this field. The chapter includes a comprehensive review of relevant literature, showcasing studies and findings that highlight the advancements in AI and ML applications for corrosion detection. This discussion sets the groundwork for a deeper analysis of ML models in the following chapter
- **Chapter 4: Comparative Study of Machine Learning for Corrosion Detection** delves into a comparative analysis of various machine learning models used in the detection and prediction of corrosion in pipelines. It discusses the methodologies, data requirements, and effectiveness of each model, providing a detailed evaluation based on recent studies and experimental results. This chapter also examines the integration of these models into existing pipeline monitoring systems and assesses their practical implications.
- **Chapter 5: Results and Discussion** The final chapter synthesizes the key findings from each of the previous chapters, discussing their implications for the pipeline industry. It emphasizes the transformative potential of AI and ML technologies in enhancing pipeline safety and integrity. The chapter concludes with strategic recommendations for industry practitioners and outlines potential areas for future research in pipeline integrity monitoring and leak detection.
- **Chapter 6: Conclusion** Conclusion and Recommendations - Summarizes the research findings, discusses the limitations of the study, and suggests areas for future research

CHAPTER II

PIPELINE RISK FACTORS AND LEAK DETECTION TECHNIQUES

01 Introduction

Leakage failure risk in oil and gas pipelines denotes the possibility of pipelines developing leaks, leading to the release of oil, gas, or other hazardous materials. After many years of pipeline accident research, considering the balance between the accuracy of a risk assessment model and its usability, there are many widely agreed risk assessment parameter factors. These risk assessment model parameters are Third Party, Corrosion, Design, Pressure, Temperature and Incorrect Operation [25]. This chapter explores the critical risk factors that influence the integrity and operational safety of pipeline systems, with a particular emphasis on the risks associated with leaks. Advanced techniques and monitoring technologies employed to detect and mitigate these risks, such as smart sensors and automated control systems, are also discussed. A critical component of this chapter is the exploration of leak detection technologies. It covers traditional methods such as distributed fiber optics sensing and real-time data analytics.

02 Leakage risk oil and gas pipeline

Leakage failure risk in oil and gas pipelines refers to the potential for pipelines to develop leaks, which can result in the release of oil, gas, or other hazardous substances. This risk is influenced by various factors such as pipeline age, material, operational pressure, environmental conditions, and the presence of corrosion or mechanical damage [6]. Effective risk management involves regular inspections, maintenance, and the implementation of advanced monitoring technologies to detect and mitigate potential leaks, ensuring the safety and integrity of the pipeline infrastructure.

03 Oil And Gas Pipelines

Pipelines are closed systems that transport fluids commodities from one location in space and time to another. It includes all physical devices, components, computer systems, telecommunication systems and the pipe itself [32] Fundamentally, pipelines are simple. They connect a place of higher pressure to another of lower pressure. However, they can be added complexities. Equipment like pumps, compressors may be used to provide additional pressure increase. Tanks may provide temporary storage (even the pipe could act as a tank of sort), valves may be used to divert flow, prevent backflow and topology/ terrain may differ greatly

Generally, there are three types of pipelines:

- Gathering lines: They usually consist of low pressure, small pipelines that transport the raw natural gas from the wellhead to the processing plants
- Transmission lines: They usually consist of high pressure, large pipelines that transport natural gas from the processing plants to the centres of consumption
- Distribution lines: They are similar to gathering lines. They deliver gas to the final consumer.

04 Pipeline Risk Factors

When reviewing the factors that detect and manage risks associated with oil and gas pipeline leaks, it's crucial to understand both the risk factors themselves and the strategies to mitigate these risks.

04.1 Temperature

Temperature variations in pipelines can significantly impact their integrity and lead to potential leakages. As the temperature fluctuates, the materials making up the pipeline, typically steel or other metals, expand and contract. This thermal expansion and contraction can stress the pipeline and its joints, potentially causing micro-fractures or exacerbating existing faults within the pipeline structure. [19], This susceptibility to leakage underscores the need for careful design and maintenance protocols that consider thermal dynamics to ensure the longevity and safety of pipeline systems [26]

04.2 Corrosion

Corrosion is a significant contributor to pipeline deterioration and subsequent leaks, representing a serious threat to the integrity of oil and gas transmission systems. This process, whether chemical or electrochemical, leads to the loss of metal and structural weakening of pipelines. Corrosion typically occurs both internally, where the pipeline contents interact with the pipe material, and externally, where environmental factors such as soil composition and moisture play a crucial role. [33]

Corrosion rates are highly variable and depend on the local environmental conditions, material properties, and the chemical composition of the transported fluids. This variability makes it challenging to predict and manage, often leading to unexpected leaks or ruptures. Internally, corrosive fluids can cause pitting and crevice corrosion, which directly undermine the pipe's mechanical strength and leak tightness. Externally, factors such as soil acidity and moisture levels can accelerate corrosion, especially in buried pipelines, making them prone to developing leaks over time [28] Effective corrosion management involves regular monitoring using techniques like pigging and cathodic protection, along with the implementation of corrosion inhibitors. Moreover, the use of corrosion-resistant materials during the pipeline construction phase can significantly mitigate these risks. Overall, understanding and addressing the various aspects of corrosion is essential for maintaining pipeline integrity and preventing environmental and economic consequences associated with pipeline leak[41]

04.3 Pressure

Pressure variations within a pipeline system can significantly impact the integrity of the infrastructure, potentially leading to leaks. Fluctuations in pressure can arise from changes in operational conditions, such as variations in the flow rate, temperature changes, or mechanical disruptions. When the pressure within a pipeline exceeds or drops below-designed thresholds, it can strain the materials, leading to fatigue, cracking, or even rupture [9], pressure transients, which are sudden changes in pipeline pressure, can cause significant stress on the pipeline walls and joints, increasing the likelihood of leakage.

Leaks resulting from pressure variations are particularly concerning because they can rapidly escalate, causing significant environmental damage and safety hazards. To manage and mitigate these risks, pipeline systems are equipped with pressure monitoring and control systems. These systems utilize sensors and automatic shutdown valves to detect abnormal pressure changes and respond quickly to prevent a minor leak from becoming a major disaster [12], predictive maintenance strategies based on real-time data from pressure sensors can be used to anticipate potential leaks and perform necessary repairs before a failure occurs.

04.4 Third Party Damage

Third Party Damage (TPD) is defined as any pipeline failures that result from human errors which are not related to the pipeline itself [10]; According to failure records, TPD is now considered as the biggest threat to the reliability and safety of pipelines; TPD can be caused by internal or external forces, which include excavating, earth movement, and other damages caused by people[39]. Nowadays, excavating activities are the lead TPD failure parameters.

04.5 Incorrect Operation

Most forms of pipeline failure, such as leaking and rupture, can be attributed, to some grade, to human factors, which considered as human errors as well. The Human factors is a complex field that aims to understand the various aspects of human characteristics and job experience, job and task design, tool and equipment design, and work environment that can affect operations and overall system performance[10]. Although many factors can cause a pipeline failure or accident, based on statistical records, almost 80% of all accidents are results of human error [10]. It is a valuable effort to identify, measure, assess, and manage potential human error factors that can significantly decrease the risk of pipeline leak.

05 Leak Detection Techniques

In this section , we first looked at organizing the available leak detection methods. They could be classified based on their technical approach. There are two general way for leak detection: hardware-based methods and software-based methods [30]. figure II.1 illustrates classifying leak detection Techniques.

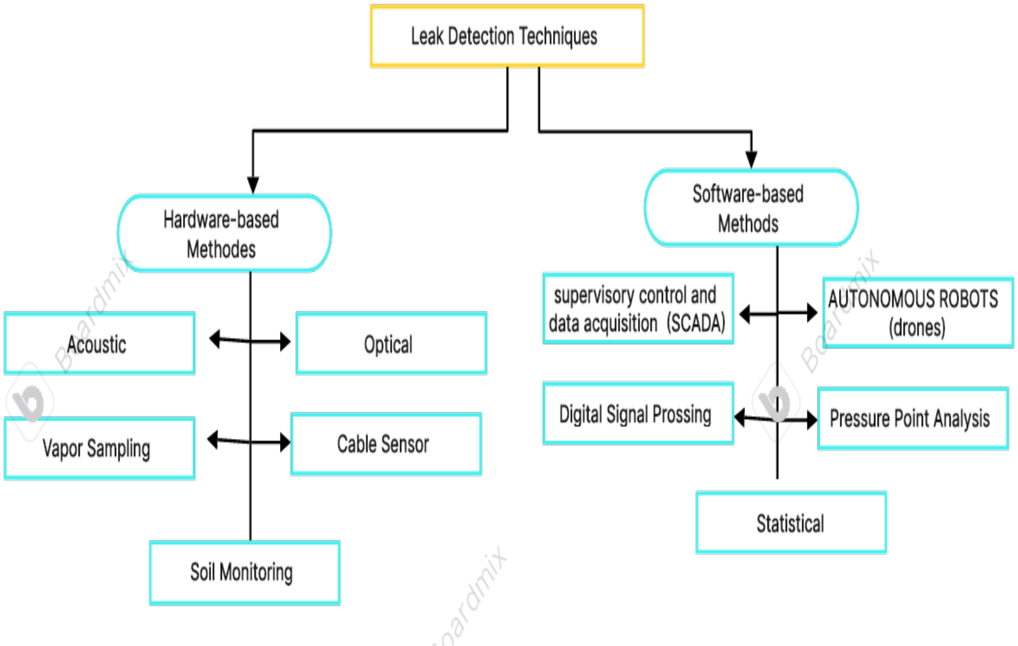


Figure II.1: Classification of Leak Detection Techniques
[30]

05.1 Hardware Based Leak Detection

Hardware-based methods for leak detection and localization detect the present of leaks from outside the pipeline by visual observation or by using appropriate equipment. These kinds of techniques are featured by a very good sensitivity to leaks and are very precise in finding the leak location. However, they are expensive and installation of their equipment is very complex task. As a result, their uses are restricted to places with high potential of risk like near rivers or nature protection areas or in conditions which pipe is transferring a hazardous material [30]. Examples of this method are acoustic leak detection, fiber optical sensing cable, vapor sensing cable and liquid sensing cable-based systems.

Acoustic leak detection

The principle of this method is based on the fact that when a leak happens, it produces an acoustic noise around the place of leakage. Acoustic sensors which are installed outside the pipe track and detect internal noise levels and create a baseline with specific features. The self-similarity of this signal is continuously analyzed by acoustic sensors. When a leak happens, a produced low-frequency acoustic signal is detected and investigated. If this signal „features differs from the baseline, an alarm will be activated [16]. The received signal is stronger near the leak site thus enabling leak localization. In the acoustic methods, the most common approach for detecting and localizing of leakage involves cross-correlation. In general, the technique is based on detecting the noise that occurs when a leak exists in the pipeline. The method works by placing sensor devices on both sides of the pipes where the leak is suspected. The sensors can be placed on the road surface or directly on a particular point such as fire hydrants as shown by Figure II.2

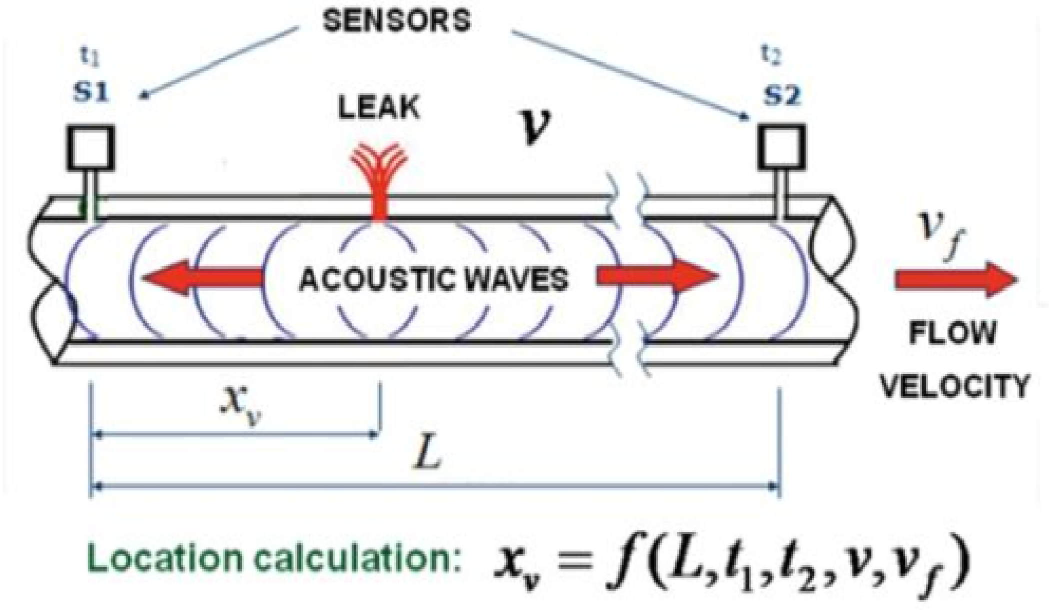


Figure II.2: Leak Detection Using Acoustic Sensors [16]

Fiber optic sensors

The fiber optic sensing leak detection method relies on the installation of a fiber optic cable all along the pipeline. Its principle is as a leak occurs in pipeline the substance inside the pipeline getsin touch with fiber cable. So, the temperature of the cable changes due to this contact. By measuring the temperature changes in fiber cable leak could be detected.

This technique is based on the Raman Effect or Optical Time Domain Reflectometry (OTDR). The laser light is scattered as the laser pulse spreads through the fiber as a result of molecular vibrations. So, the backscattered light carries the information of local temperature along the pipeline. Indeed, Raman backscattered light has two frequency shifted components: the Stokes and the Anti-Stokes components. The amplitude of the Anti-Stokes component varies dramatically with regard of temperature variations.Butthe amplitude of the Stokes component is not affected by temperature. Therefore some filtering is needed to isolate Anti-stoke components from stokes components[29]. The problem associated with this technique is low magnitude of backscattered light To overcome his issuehigh numerical aperture multimode fibers are used.

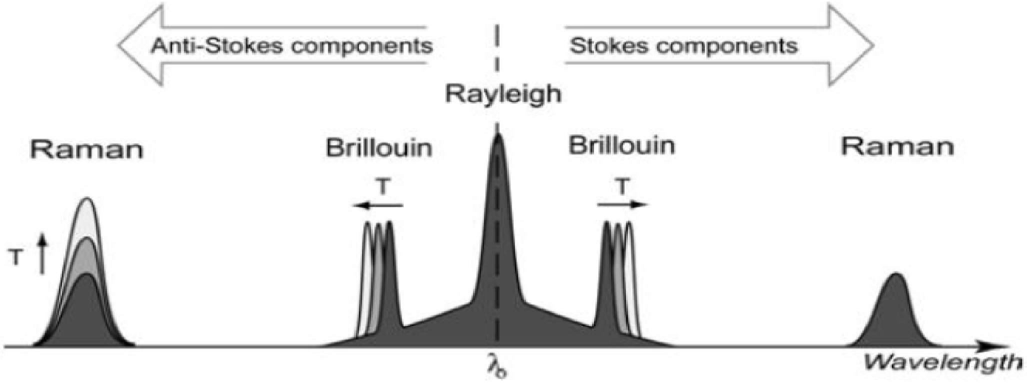


Figure II.3: Schematic Representation Of The Scattered Light Spectrum From A Single Wavelength Signal Propagating In Optical Fibers [17]

Vapor or liquid sensing tubes

The vapor or liquid sensing tube-based leak detection method involves the installation of a tube along the entire length of the pipeline. If a leak happens, the content of pipe gets in touch of tube. The tube is full of air in atmospheric pressure. Once the leak occurs, the leaking substance penetrates into the tube. First of all, to assess the concentration distribution in the sensor tube, a column of air with constant speed is forced into the tube. There are gas sensors at the end of the sensor tube. Every increase in gas concentration leads to a peak in gas concentration which its size is an indication of the size of the leak The detected line is equipped with an electrolytic cell. This cell diffuses an exact volume of test gas into the tube constantly. This along with air passes through the whole length of the sensor tube. When the test gas travels through the detector unit, it produces an end peak. So, the end peak is a sign of the whole length of the sensor tube. Leak localization is carried on by calculating the ratio of end peak arrival to leak peak arrival [17]. Figure II.4 indicates this technique.As a shortcoming of this method, it could be mentioned that its speed of leak detection is very low. In addition, it's not very practical for applying in long pipelines as the cost of its equipment is very high. The other drawback of vapor sensing tubes is the difficulty of their application in pipelines above ground or in deep sites.

Liquid sensing cables.

Liquid sensing cables are placed near to a pipeline and their main function is a representation of changes in transmitted energy pulses that has happened due to impedance differentials. Safe energy pulses are continually sent through the cable. As these energy pulses travel down the cable, reflections are returned to the monitoring unit and a ”map” of the reflected energy from the cable is stored in memory. The presence of liquids on the sensor cable, in sufficient quantities to ”wet” the cable, will alter its electrical properties. This alteration will cause a change of the reflection at that location. The alteration is then used

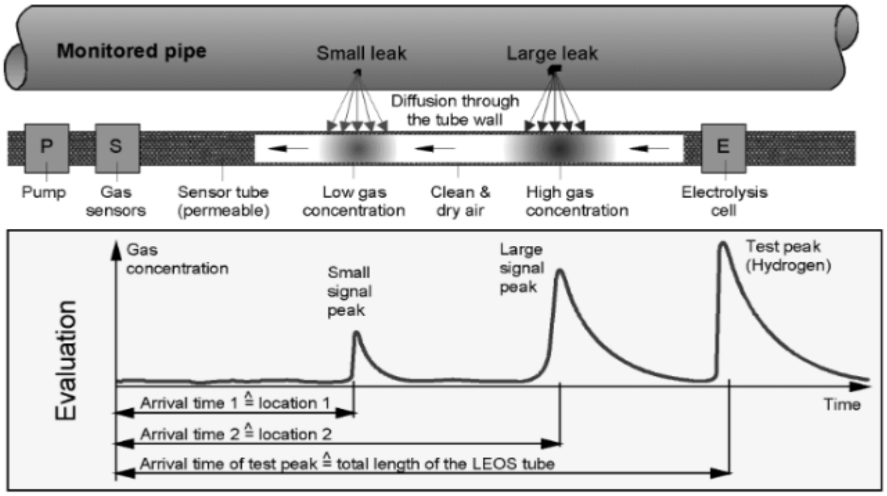


Figure II.4: Leak Detection And LocalizationUsing Vapor Sensing Tube [17]

to determine the location of a potential leak. For localization time delay between input pulse and reflected pulse are used [17].This method works will for multiple leak detection and localization for short pipelines

05.2 Software Based Systems

The internal method is based on the monitoring of internal pipeline parameters (pressure, flow and temperature). Generally, the effectiveness of the internal based methods depends on the uncertainties associated with the system's characteristics, operating conditions and collected data

supervisory control and data acquisition

In the oil and gas industry, the control center is an important part of operations and it is a command center for control of all the processes and monitoring of all the parameters. The control rooms deploy SCADA systems that are interfaced with displays and monitors The operations in the control centers include emergency shut down, and monitoring of equipment such as pumps and compressors. In many cases, the control centers still require human intervention to handle these operations and therefore have to be manned 24/7 [14] .

Real-time Monitoring and Control SCADA systems allow for continuous monitoring of field activities from centralized control centers. They let operators to remotely manage equipment such as valves and pumps, modifying operating settings in response to changing conditions or crises without having to be physically present on-site [18].. These systems are configured to automatically detect and alert operators about operational anomalies or failures, such as pressure drops, equipment malfunctions, or potential leaks. This capability is critical for prompt intervention, minimizing the risk of accidents and ensuring the safety of both the facility and personnel. Moreover, collect and store historical data which can be used for long-term trend analysis and predictive maintenance. This data helps in refining operational strategies and preventive maintenance schedules, enhancing the longevity and efficiency of equipment.[30]. Despite the high level of automation, many control centers still require human oversight to manage these complex operations, necessitating 24/7 staffing

Figure II.5: Example of supervisory control and data acquisition [14]

Pressure point analysis pressure

This method detects the occurrence of leaks by comparing the current pressure signal with a running statistical trend taken over a period of time along the pipeline by pressure monitoring and flow monitoring devices [18]. The principle of this method is based on the fact of pressure drop as a result of leak occurrence. Using an appropriate statistical analysis of most recent pressure measurements, a sudden change in statistic properties of pressure measurement such as their mean value is detected. If the mean of newer data is considerably smaller than the mean of older data, then a leak alarm is generated. This method may require sensitive high resolution but not necessarily very precise instrumentation. So, the lower overall installation costs are not very high. Furthermore, this method is able to identify the occurrence of leaks, but not necessarily the presence of them. Since this method use of pressure drop as a leak signature, it can yield false alarms as the pressure drop is not unique to the leak event.

Statistical

A statistical leak detection system uses advance statistical technique to analyze the flow rate, pressure and temperature measurements of a pipeline. This method is appropriate for complex pipe system as it can be monitored continuously for continual changes in the line and flow/pressure instruments. In addition, this technique could be used for leak localization. Using statistical analysis is also very easy and applicable in to different pipeline systems [30]. The main objective of this system is to minimize the rate of false alarm. It is also suitable for real-time application and has been successfully tested in oil pipeline systems [18]. The main disadvantage of statistical leak detection is that noise interferes in the statistical analyses, and some leaks were hidden in the noise which prevented them from being detected

Digital signal processing

Another method for leak detection is using digital signal processing techniques. The procedure of this method is that the response of the pipeline to a known input is measured over a period of time. Afterwards, this response is compared with the later measurements. Based on comparison of their signal's features like frequency response or wavelet transform coefficients a leak alarm could be generated. Similar to statistical methods this technique does not need a pipeline model. The problem associated with using this method for leak detection is only leak occurrence could be detected not leak presence unlessthe size of present leak increases considerably [30]

AUTONOMOUS ROBOTS (Drones)

The use of drones or UAV or UAS in the Oil and gas industry provides safety, efficiency, and is considered cost-effective and has been used extensively for various applications The use of drones has been used to complement other forms of surveillance technologies such as satellites, plane or helicopter imagery and ground digital acquisitions and observations. For instance, the use of UAV was shown to provide key input for reservoir modelling in analogue-producing fields which is useful for digital outcrop models of subsurface reservoirs. Some of the applications of the drones are illustrated in Figure II.6. [17]

They provide high-resolution imagery and real-time data, allowing for the early detection of leaks, corrosion, and other potential issues, which helps in preventing costly and hazardous incidents. Drones also reduce the need for manual inspections, significantly lowering the risk to human workers, especially in hazardous and hard-to-reach areas. [14]

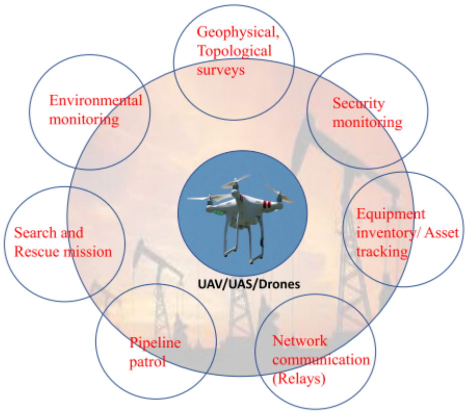

Geophysical, Topological surveys
Environmental monitoring
Security monitoring
Search and Rescue mission
Equipment inventory/ Asset tracking
UAV/UAS/Drones
Pipeline patrol
Network communication (Relays)

Figure II.6: Application of drones in oil and gas sector [14]

06 Leak Detection Techniques' Limitations

While these techniques are beneficial in leak detection, they fall short in predicting these leaks to prevent potential damages. Hence, we have turned to the use of artificial intelligence to obtain predictions well before the leak occurs, enabling the implementation of effective preventive measures and reducing associated risks. smaller leaks without additional sensitive technology. Digital signal processing provides thorough data analysis, though it demands significant computational power and expertise. Lastly, autonomous robots like drones offer accessibility to challenging locations but are constrained by factors like battery life and adverse weather conditions, requiring advanced navigation systems for effective operation.

While these techniques are beneficial in leak detection, they fall short in predicting these leaks to prevent potential damages. Hence, we have turned to the use of artificial intelligence to obtain predictions

well before the leak occurs, enabling the implementation of effective preventive measures and reducing associated risks.

07 Advancing Corrosion Detection in Pipeline Management through AI-Enhanced Techniques

The limitations and high costs of the previously mentioned leak detection techniques have prompted experts to explore alternative methods for identifying leak risks. Research has shown that corrosion is the most critical damaging mechanism for pipelines. Both internal and external corrosion significantly impact the security and integrity of pipelines over time, necessitating continuous inspection [25]. This conclusion is also supported by experts in oil and gas industry. For this reason, effective pipeline management strategies prioritize periodic inspections of corrosion and measures to ensure the safety, reliability, and longevity of pipeline infrastructure.

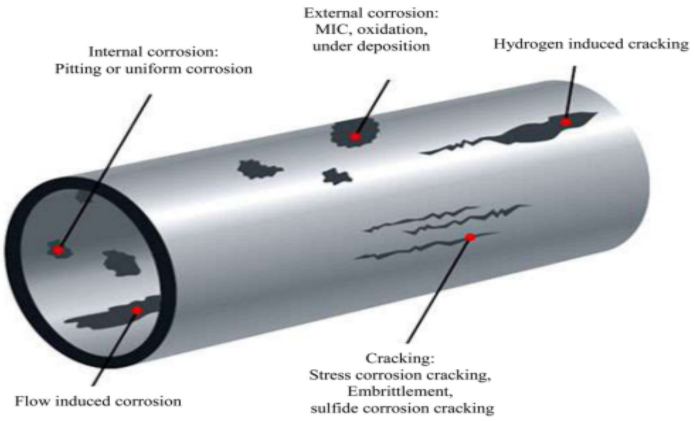


Figure II.7: Types of corrosion in pipeline
[25]

To verify the impact of corrosion and its correlation with pipeline leaks, we performed a statistical analysis on a dataset[1] encompassing various factors. Figure II.8 illustrates the correlation of different factors with pipeline leaks. We can notice the high incidence of leaks due to corrosion-related issues. This finding highlighted the need to concentrate our investigation on corrosion, given its prevalence among other contributing factors.

Given this significant impact of corrosion, our study of pipeline leaks must inherently focus on AI methods of detecting and predicting corrosion. Our research will leverage advanced artificial intelligence techniques, specifically multi-layer perceptrons (MLP) and convolutional neural networks (CNN), to develop predictive models that can accurately identify and predict corrosion rate. These models will utilize comprehensive datasets that include parameters like temperature, pressure, fluid composition, and other relevant factors. This work benefits significantly from using the level of corrosion as the target for classification, as it directly addresses the real-world conditions that are most critical. By classifying the severity of corrosion whether high, or low we provide a tangible target that aligns with the actual operational

[1] https://www.kaggle.com/datasets/usdot/pipeline-accidents

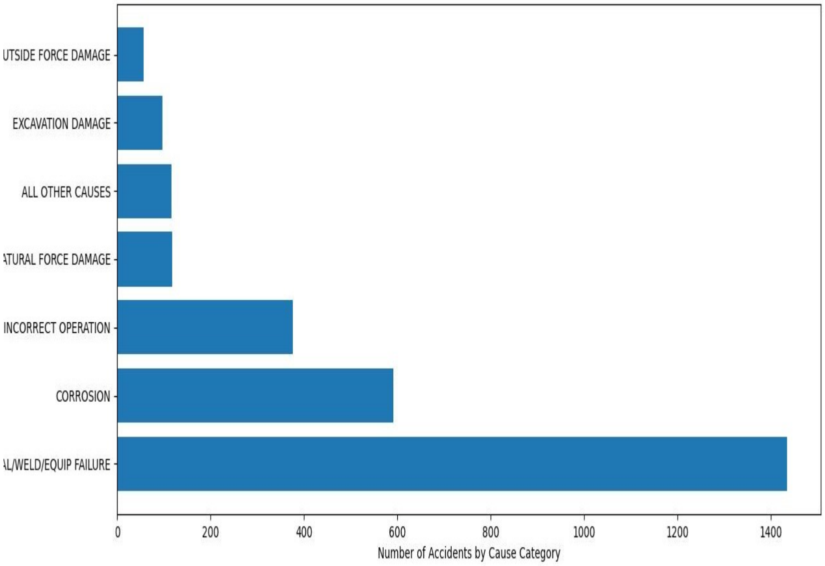


Figure II.8: Percentage Of Factors Detection

challenges faced by pipeline management.

08 Conclusion

This chapter has provided an overview of the methods used for detecting leaks in pipeline networks, distinguishing between hardware-based and software-based techniques. Hardware-based methods utilize specialized devices to detect leaks externally, which can be quite costly, particularly in extensive pipeline systems. Conversely, software-based methods leverage algorithms within software programs to continuously monitor critical parameters such as pressure, temperature, and flow rate, allowing for the detection of leaks based on anomalies in these data points.

While these techniques are beneficial in leak detection, they fall short in predicting these leaks to prevent potential damages. Hence, we have turned to the use of artificial intelligence to obtain predictions well before the leak occurs, enabling the implementation of effective preventive measures and reducing associated risks.

Additionally, the importance of corrosion assessment is highlighted as a vital factor in the decision-making process for selecting appropriate safety measures to prevent or address leaks.

The next chapter will delve into how AI techniques are increasingly being recognized as crucial tools in mitigating the risks associated with pipeline leaks. These techniques offer promising enhancements in detecting and managing potential pipeline failures, and the subsequent chapter will explore this in further detail.

CHAPTER III

MACHINE LEARNING FOR CORROSION PREDICTION

01 Introduction

This chapter seeks to classify the scientific literature according to topics, predictive criteria, and the range of Machine Learning (ML) methodologies employed for pipeline leaks detection. Organized into two main sections, it begins by introducing the necessary background on Artificial Intelligence and Machine Learning. The chapter then reviews empirical studies that have employed these technologies for pipeline leak detection, with a focus on the influence of corrosion factors in these predictive models. This study not only highlights the technological progress in pipeline monitoring but also deepens our understanding of the strengths and weaknesses of current ML applications in the oil and gas sector.

02 Artificial Intelligence

Since artificial intelligence was born in the 1940s, many researchers and projects about artificial intelligence have been done, and because of them, now it has become a greatly recognized field. There are many definitions of Artificial Intelligence; one of the most accepted definitions of AI is the capability and process of intelligent agents, which are capable of continuously learning the corresponding environment, perceiving and acting in certain activities. Artificial intelligence has the advantage of dealing with pervasive imprecision [23]. Over the years, artificial intelligence has evolved and generated other separate fields such as machine learning and deep learning

03 Machine Learning

With the technological developments in recent years, new terms have begun to emerge. Big data, Industry, and artificial intelligence are the most popular ones. Although machine learning is not as popular as these terms, it is a concept that has been on the rise. However, many questions surround machine learning. machine learning is applied in different sectors and applications today, and its use is increasing gradually [1]. ML is a sub-area of artificial intelligence. Information technology systems automatically learn patterns and relationships from data and gain without being explicitly programmed. ML has been successfully supported in business, investigation, and improvement for many years.

Furthermore, machine learning can automatically produce knowledge, train algorithms, identify relationships, and recognize unknown patterns. These identified patterns and relationships can be utilized to a new, unknown data set in order to make predictions and optimize processes. Unlike traditional software development, machine learning focuses on independent learning from data and information. Thus, machine learning technologies learn from data and create their own approach code on their own. These techniques will live in a particular situation and train themselves depending on the circumstances in which they will be.

Machine learning is categorized into three, namely :

- **Supervised learning:** In general, this type includes most of the problems in machine learning, which is characterised by looking at training samples. Each sample is entered as X so that it corresponds to a specific result, which is y. We need to train a model (mathematically is $x \rightarrow y$ relationship mapping f) in unknown samples x after giving, then we can obtain y predictions. If the prediction is a discrete value (often category types, such as spam/snail mail in the mail classification problem, such as whether a user will/will not buy a particular product), then it will be termed a classification problem. If the prediction result is a continuous value (e.g., apartment prices, stock prices, etc.), then this state will be termed a regression problem [1].

- **Unsupervised learning:** It is a form of learning in which information is categorized or not. Unsupervised learning finds hidden patterns in data. It uses them to infer from datasets entered into the system, without labelled data. Given that no classification has previously been done, the system can classify using data sets.

- **Semi-supervised Learning:** It is a form of learning that takes place between supervised learning and unsupervised learning. It is used for the same applications as supervised learning. Large amounts of unlabeled data and small amounts of labeled data are commonly used.

04 Machine learning Model

Support Vector Machine

SVM, a supervised learning approach, is one of the most popular and simplest ML techniques because its solutions are often perfect and unique [3]. A classification problem can be efficiently divided into two halves by a hyperplane, but SVM constructs two boundary lines with a certain distance between them so that the classification points can be easily divided linearly [20].

Support vectors are the nearest positive (2 blue) or negative (1 green) points, which are essentially the extreme points on either side. The dotted hyperplane in Figure III.1, which is parallel to the main hyperplane, was made possible by these support vectors. Moreover, the margin is the distance between these two dotted hyperplanes. To achieve a better classification result, SVM maximizes this margin distance. Now let's define two main terms which will be repeated again and again in this algorithm:

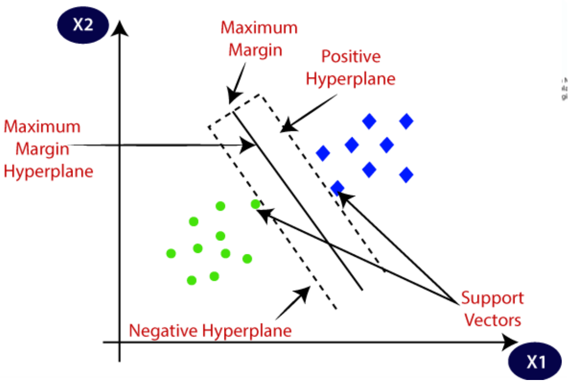


Figure III.1: margins [20]

Support Vectors These are the points that are closest to the hyperplane A separating line will be defined with the help of these data points. Margin It is the distance between the hyperplane and the observations closest to the hyperplane (support vectors). In SVM, a large margin is considered a good margin. There are two types of margins: hard margin and soft margin.

To make our discussion of SVMs easier we will be considering a linear classifier for a binary classification problem with labels y and features x. We'll use $y \in \{-1, 1\}$ to denote the class labels and parameters w, b:

$$f(x) = w^T x + b$$

- w: normal to the line.
- b: bias.

where sgn() is known as a sign function, which is mathematically represented by the following equation [**15**]:

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

The distance D of a data point x from the hyperplane is represented mathematically by the equation:

$$D = \frac{|w^T x + b|}{|w|}$$

Types of SVM Algorithms

- **Linear SVM** When the data is perfectly linearly separable only then we can use Linear SVM. Perfectly linearly separable means that the data points can be classified into 2 classes by using a single straight line(if 2D).
- **Non-Linear SVM** When the data is not linearly separable then we can use Non-Linear SVM, which means when the data points cannot be separated into 2 classes by using a straight line (if 2D) then we use some advanced techniques like kernel tricks to classify them. In most real-world applications we do not find linearly separable datapoints hence we use kernel trick to solve them.

SVM applies a kernel function to convert nonlinear inputs into linearly separable information. We saw a demonstration of this conversion in the section above. However, it can take a long time if we need to change millions of complex data items.

k-nearest neighbor

The k-nearest neighbor (KNN) is a algorithm that requires training data and a predefined k value to find the k nearest data points based on distance computation. If the k data points belong to different classes, the algorithm predicts the class of the unknown data to be the same as the majority class. The concept can be seen in Figure **??**.

The KNN algorithm employs various distance metrics to evaluate similarity between data points, including Euclidean, Standardized Euclidean, Mahalanobis, City Block, Minkowski, Chebyshev, Cosine, Correlation, Hamming, Jaccard, and Spearman distances. Each metric offers a different perspective on data similarity, thus affecting the algorithm's performance.

Evaluating the classification performance of the model on the test set helps in understanding how well the model performs in predicting the correct classes based on the learned distances. This evaluation is shown in Figure **??**.

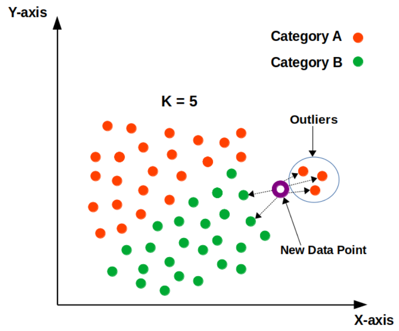


Figure III.2: KNN Demonstration of the k-nearest neighbor methodology [8]

Random forests

Random forests classifier is a popular classification way in machine learning. By constructing a great amount of decision trees, random forests classifier is strengthened. Decision trees, whose basic idea is that groups of weak learners come together and form a stronger learner, start with a root, keep growing its branches, and ultimately reach its terminal node called leaves [15]. The branches imported to the "tree" are features or processed information based on those features. Comparing to other algorithms, Random Forest Classifiers run efficiently on a large database with a relatively high accuracy due to its lower risk of overfitting. Random Forest is an advanced bagging technique instead of a boosting technique, which can help lead to "improvements for unstable procedures" (Breiman, 2001). By randomly splitting attributes, Random Forests decorrelate the decision trees (Figure III.3), leading to an improvement in the bagging techniques.

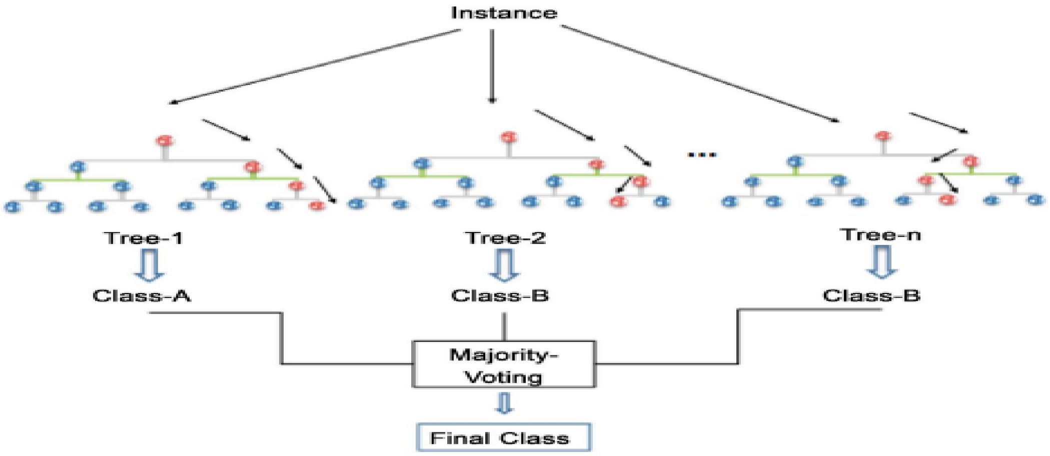


Figure III.3: Demonstration of the random Forest methodology. [15]

The random forest algorithm is one of the best among classification algorithms - able to classify large amounts of data with accuracy. It is an ensemble learning method for classification and regression that constructs a number of decision trees at training time and delivers the class that is the mode of the classes output by individual trees.

Random Forest Algorithm:

- For $b = 1$ to B:
  - Draw a bootstrap sample Z^* of size N from the training data.
  - Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached:
    - Select m variables at random from the p variables.
    - Pick the best variable/split-point among the m.
    - Split the node into two daughter nodes.
- Output the ensemble of trees $\{T_b\}_{b=1}^B$.

To make a prediction at a new point x:

Regression:

$$\hat{f}_{rf}^B(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x).$$

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random forest tree. Then,

$$\hat{C}_{rf}^B(x) = \text{majority vote}\{\hat{C}_b(x)\}_{b=1}^B.$$

In random forest classification method, many classifiers are generated from smaller subsets of the input data and later their individual results are aggregated based on a voting mechanism to generate the desired output of the input data set. This ensemble learning strategy has recently become very popular. Before RF, Boosting and Bagging were the only two ensemble learning methods used. RF has been extensively applied in various areas including modern drug discovery, network intrusion detection, land cover analysis, credit rating analysis, remote sensing and gene microarrays data analysis etc... [9][10]

There are two ways to evaluate the error rate. One is to split the dataset into training part and test part. We can employ the training part to build the forest, and then use the test part to calculate the error rate. Another way is to...

Decision Tree

A Decision tree is a classifier expressed as a recursive partition of the instance space. The Decision tree consists of nodes that form a Rooted Tree, meaning it is a Directed Tree with a node called root that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal node or test node. All other nodes are called leaves (also known as terminal nodes or decision nodes). [36]

In the decision tree, each internal node splits the instance space into two or more subspaces according to a certain discrete function of the input attribute's values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target value having a certain value. [36] Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path.

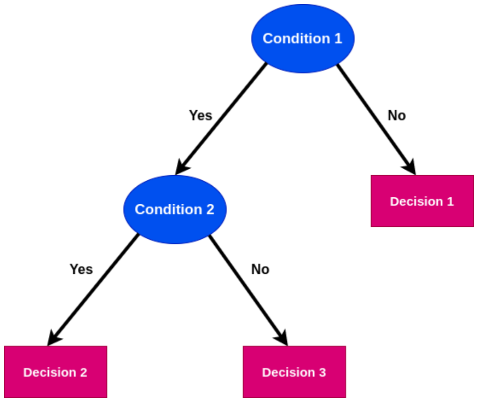


Figure III.4: Decision tree demonstration of the Random Forest methodology. [36]

Decision Tree Pseudo Code:

```
def decisionTreeLearning(examples, attributes, parent_examples):
    if len(examples) == 0:
        return pluralityValue(parent_examples)
        # return most probable answer as there is no training data left
    elif len(attributes) == 0:
        return pluralityValue(examples)
    elif (all examples classify the same):
        return their classification
    A = max(attributes, key=lambda a: importance(a, examples))
        # choose the most promising attribute to condition on
    tree = new Tree(root=A)
    for value in A.values():
        exs = [example for example in examples if example[A] == value]
        subtree = decisionTreeLearning(exs, attributes.remove(A), examples)
        # note implementation should probably wrap the trivial case
        # returns into trees for consistency
        tree.addSubtreeAsBranch(subtree, label=(A, value))
    return tree
```

Gradient Boosting

Gradient Boosting (GB) is a supervised algorithm used to build a predictive machine-learning model. In the process of integrating individual decision trees into the algorithm, a method called 'reinforcement' is used. reinforcement means developing a strong learner by merging several learning algorithms of weak learners into a single chain. DT in this algorithm represents weak learners. The model of this algorithm is characterised by high efficiency and accuracy because each tree inside it works to fix the errors of the tree that precedes it. However, the sequential increase of trees inside the algorithm improves its performance but slows the learning process [**26**]. In addition, the model relies on the loss function for residual detection. For example, the logarithmic loss is used in classification and regression tasks.

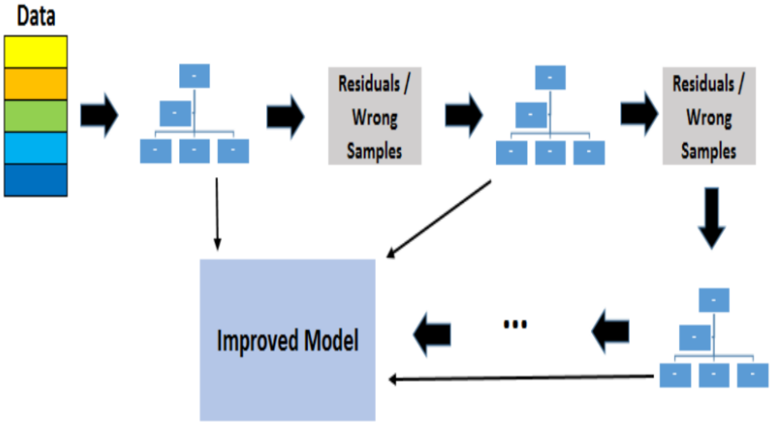


Figure III.5: Demonstration of the Gradient Boosting methodology. [15]

Multi-Layered Perceptron

A Multi-Layer Perceptron (MLP) is a network made up of perceptrons. It has an input layer that receives the input signal, an output layer that makes predictions or decisions for a given input, and the layers present in between the input layer and output layer is called hidden layer. There can be many hidden layers, the number of hidden layers can be changed as per requirement. In the proposed methodology for Speech emotion Recognition, the multi-layer Perceptron network will have one input layer [24], of (300,) and (40,80,40) hidden layers and one output layer. The input layer will take as input, the five features, that are extracted from the audio file. The extracted five features being, Mel Frequency Cepstral Coefficients, Mel Spectrogram Frequency, Chroma, Tonnetz and Contrast

The hidden layer uses an activation function to act upon the input data and to process the data. The activation function used is logistic activation function. The output layer brings out the information learned by the network as output. this layer classifies and gives output of the predicted emotion, according to the computation performed by the hidden layer. Fig.1 illustrates Multilayer Perceptron.

Multilayer perceptron is applied for supervised learning problems. The multi-layer perceptron is used for the purpose of classification. The MLP is made to train on the given dataset. The training phase enables the MLP to learn the correlation between the set of inputs and outputs. During training the MLP adjusts model parameters such as weights and biases in order to minimize the error. The MLP uses backpropaga-

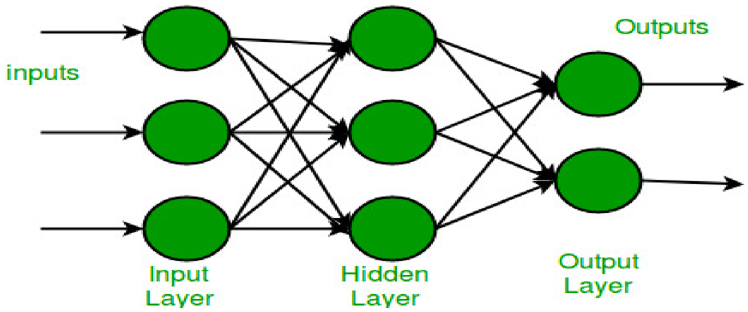


Figure III.6: Demonstration of Multi-Layer Perceptron methodology. [24]

tion [24], to make weight and bias adjustments relative to the error. The error can be calculated in many ways.

Convolutional Neural Network

CNN is a typical feedforward neural network with convolutional computation and deep structure. It is one of the representative algorithms of deep learning. CNN can perform translation invariant classification, CNN model can conduct supervised learning and unsupervised learning by parameter sharing of the convolutional kernel and the sparseness of inter-layer links. CNN models include some structural characteristics, such as local perceptual domain, weight sharing, and pooling [22]. Compared with conventional neural networks, the most distinguishing feature of CNN is the convolutional layer (feature extraction) and pooling layer (feature optimization and selection). The traditional CNN structure

- **Input Layer :** The input layer would take in 1D numerical data, such as timeseries sensor measurements or other 1D features related to corrosion. The data would need to be preprocessed and formatted to fit the input requirements of the CNN. In a mathematical format, this layer will receive input of N number of features. The value of n depends on the number of features in the dataset. For example, for the pipes corrosion dataset, the input layer receives 8 features, the final input can be written as follows [40]:

  $$X_i = \{f_n^i\}_{n \in N}$$

  where f_n^i represents the features of the input data. [22]

- **Convolutional Layer:**

  In this layer, the CNN would apply a set of learnable filters (convolution kernels) to the input data. The filters would be designed to extract relevant features or patterns from the 1D numerical data that could be indicative of corrosion. The convolutional operation would slide the filters across the input data, generating feature maps that highlight the presence of these patterns. In our proposed method, we have used a kernel of size 2, and generate 32 features as an output of the convolution block. Also, the convolutional layer applies activation function at the end of the process, which is ReLU in our case. This operation can be mathematically noted as following:

  $$y_0 = \sum_{k=-p}^{p} x_{-k} w_k$$

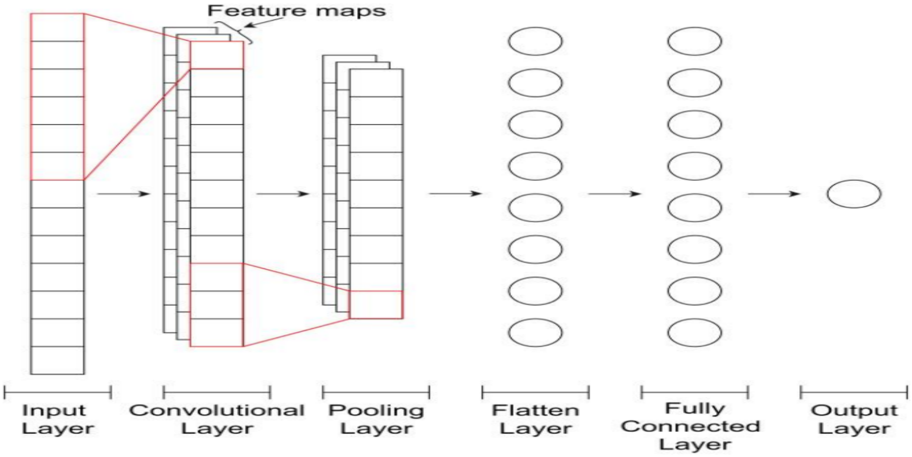
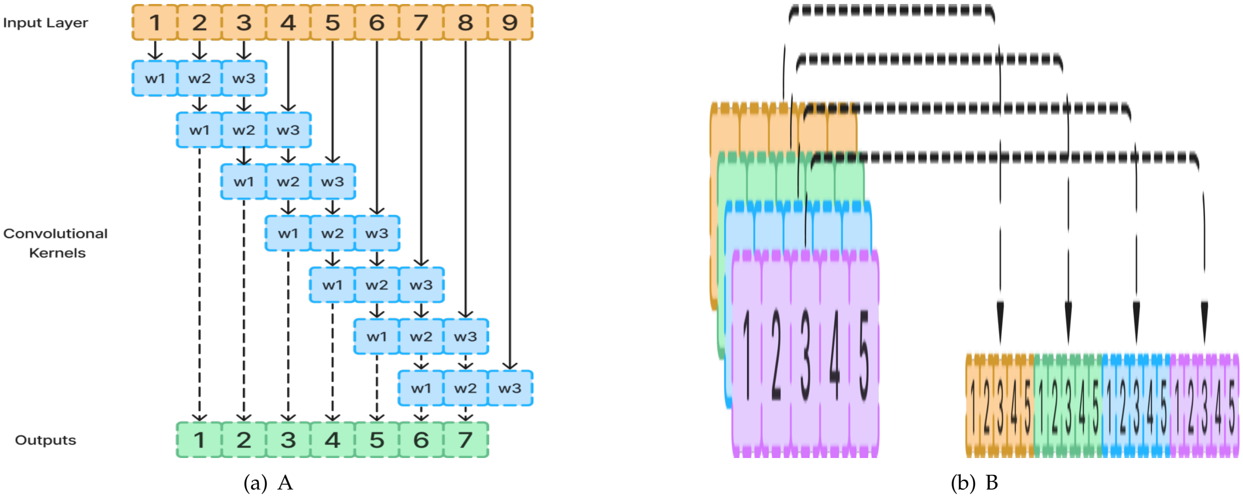


Figure III.7: The Convolutional operation in 1D input



(a) A

(b) B

Figure III.8: (A) The Convolutional operation in 1D input, (B) The pooling layer.

$$y_1 = \sum_{k=-p}^{p} x_{1-k} w_k$$

$$y_m = \sum_{k=-p}^{p} x_{m-k} w_k$$

Here, w_k represents the weights of the filter, and the sums are over the kernel window indexed by k from $-p$ to p. The indices of x shift according to the position of the filter being applied to the input.

- **Pooling Layer:**

  The pooling layer follows the convolutional layer, and its purpose is to reduce the spatial dimensions of the feature maps. This reduction is typically achieved through operations like max pooling or average pooling along the 1D input data, focusing on extracting the most important features. In our implementation, the pooling operation used was max pooling. Pooling helps to reduce the number of parameters in the network, enhancing efficiency and reducing the likelihood of overfitting. The simplest case of output from a layer with input size (N, L) and output (N, C, L_{out}) can be precisely described as:

$$\text{out}(N_i, C_j, k) = \max_{m=0,\ldots,\text{kernel\_size}-1} \text{input}(N_i, C_j, \text{stride} \times k + m)$$

Where the stride defines the movement of the sliding window to generate the next output, and k represents the pooling size.

- **Flattening Layer :** After the convolutional and pooling layers, the feature maps would be flattened into a 1D vector. This step prepares the data for the fully connected layers that follow.
- **Fully Connected Layer:**

  The flattened feature vector would be fed into one or more fully connected layers. These layers would learn non-linear combinations of the extracted features to make the final predictions or classifications related to corrosion. The term ”Dense Layers” refers to the fully interconnected layers found in neural networks. In essence, Dense Layers are created when all of the neurons in this area are fully connected to both the neurons in the previous layer and to each other. These dense layers contain 2 crucial components which are the Biases and the Weights [22], where each node will perform a mathematical operation between these components and the input values to get the final output values, these operations can be noted for each layer as follows [40]:

  $$Y = \sigma(XA^T + B)$$

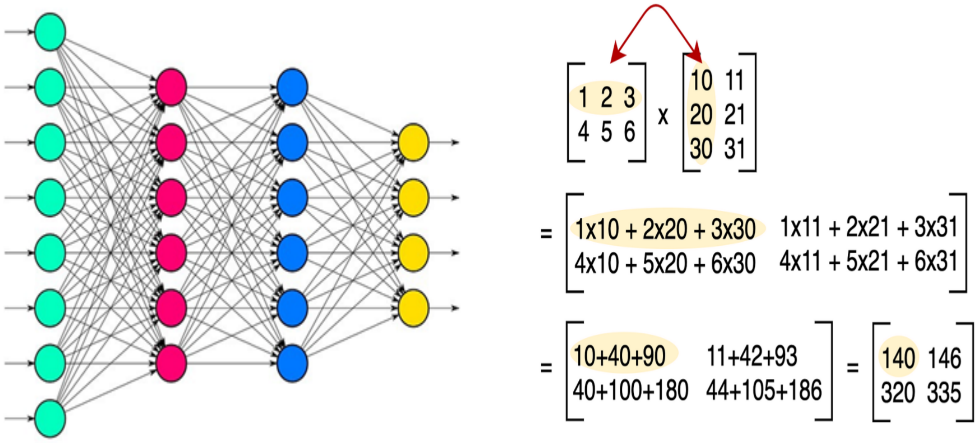
  where A represents the weights matrix, B is the biases vector, and σ denotes the activation function, such as ReLU.

  Where A is the weights matrix, and B is the Biases vector, and σ is the activation function. The operation between the input X and the weights matrix is the matrix multiplication operation. Example of the matrix multiplication can be seen in the following image, where the multiplication happens between two matrices and results in one matrix, in our case the input and output matrix are 1D vectors. For the activation function, we used the ReLU in most of the layers, where the ReLU activation function is defined as

  $$f(x) = \max(0, x)$$

  , where x is the input to the function. In other words, the ReLU function returns the input value if it is positive, and 0 if the input is negative. The ReLU function is widely used in the hidden layers of CNNs for several reasons:

  - Nonlinearity: The ReLU function introduces nonlinearity into the neural network, which allows the model to learn complex, nonlinear relationships in the data. This is important because most real-world problems involve nonlinear relationships between inputs and outputs.
  - Sparsity: The ReLU function tends to produce sparse activations, meaning that many of the neuron outputs will be exactly zero. This sparsity can help the network learn more efficiently and interpretable representations of the data.
  - Computational Efficiency: The ReLU function is computationally efficient to compute, as it simply involves a max operation. This makes it faster and easier to train deep neural networks compared to other, more complex activation functions.



(a) C

(b) D

Figure III.9: (C) The fully connected dense layers, (D) Example about the matrix multiplication

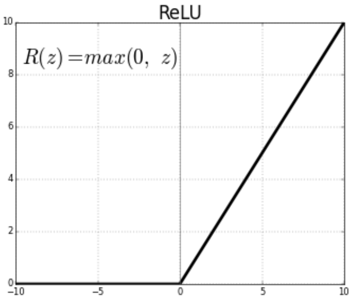


Figure III.10: The Relu activation function

  - Vanishing Gradient Problem: The ReLU function does not suffer from the vanishing gradient problem, which can occur with activation functions like the Sigmoid function. The vanishing gradient problem can make it difficult to train deep neural networks effectively.

- **Output Layer:**

  There are several ways to plot a function of two variables, depending on the information you are interested in. For instance, if you want to see the mesh of a function so it easier to see the derivative you can use a plot like the one on the left.

  The final layer of the CNN would produce the output, which is a binary classification (corrosion high or not) or a more granular prediction related to the extent or type of corrosion. In this layer, we used the sigmoid activation function, which is defined as

$$f(x) = \frac{1}{1 + e^{-x}}$$

  where x is the input to the function. The Sigmoid function maps the input values to the range (0,1)

  effectively squashing the input values and introducing nonlinearity. The Sigmoid function is often used in the output layer of a CNN, especially for binary classification problems, where the output represents the probability of the input belonging to one of two classes. The Sigmoid function is well-suited for this task because it produces output values between 0 and 1, which can be interpreted as probabilities [22]. This probability

defines the output class where after rounding the final values we get either 0 or 1 which refers to False and True respectively. The specific architecture and hyperparameters of the CNN, such as the number and size of the convolutional and fully connected layers, the choice of activation functions, and the optimization algorithm, would need to be carefully designed and tuned based on the 1D numerical data and the specific requirements of the corrosion detection problem. In our case we used the commonly and known used parameters for more accurate results [40], however, changing this parameter may lead to better or worst results.

05 Comparision between CNN for numerical data and CNN for images data

Convolutional Neural Networks (CNNs) have diverse applications, significantly differentiated by the type of data they process. CNNs configured for numerical data, such as 1D feature vectors or time-series data, are primarily engineered to identify temporal patterns and correlations. Their architecture emphasizes trend detection, anomaly identification, or other time-dependent characteristics, with pooling layers that effectively aggregate features along the 1D input dimension to preserve essential temporal information. In contrast, CNNs designed for image data focus on extracting spatially-dependent features like edges, textures, and object shapes from 2D spatial arrays. These networks utilize convolutional layers to explore complex spatial relationships and patterns The following table presents a comparison between the application of CNN for numerical data and image data .

Table III.1: Comparison of CNN Application for Numerical Data vs. Image Data

Feature	**CNN for Numerical Data**	**CNN for Image Data**
Data Structure	1D feature vectors or time-series data	2D spatial data (images)
Focus of Convolutional Layers	Extracting temporal patterns or correlations between features	Extracting spatially-dependent features like edges, textures, and object shapes
Design of Filters	Designed to detect trends, anomalies, or other time-dependent characteristics	Designed to exploit spatial relationships and patterns
Pooling Layers	Aggregates features along the 1D input dimension	Aggregates features across a 2D spatial grid
Flattening and Fully-Connected Layers	Learn higher-level representations focusing on relationships between numerical inputs	Learn higher-level, spatially-invariant features
Activation Functions	Might use different activations like Sigmoid in output for probability-like outputs	Commonly uses ReLU in hidden layers; Softmax or Sigmoid in output layer for classification
Optimization for Data Type	Architectural and hyperparameter choices adapted to capture temporal or 1D spatial patterns	Architectural choices tailored to capturing and classifying based on spatial structures

06 literature ML Of Corrosion Detection

Considering the range of research on detecting and predicting corosion in oil and gas pipelines utilizing machine learning and deep learning, various innovative approaches and methodologies have been explored. Below, we outline key studies and their findings:

06.1 Numerical data based Machine Learning and deep learning researches for Corrosion Defect Prediction

Aljameel et al [3] explore the application of machine learning-based anomaly detection models to address the pervasive issue of leaks in oil and gas pipelines. They conducted a comparative analysis using five prominent machine-learning algorithms: RF, SVM, KNN, GB, and DT. Each of these algorithms was tasked with developing models to effectively detect pipeline leaks. The results of their study revealed that the Support Vector Machine algorithm achieved an exceptional performance, registering an accuracy of 97.4 %, which was significantly higher than the other evaluated algorithms. This high level of accuracy demonstrates the SVM's capability to act as a highly efficient and reliable model for detecting leakage in oil and gas pipeline systems. The success of the SVM in this context underscores its potential as a preferred tool in the operational monitoring and maintenance of pipeline infrastructure, ensuring safety and minimizing the risk.

Seghier et al. [37] focuses on the implementation of robust ensemble learning techniques to predict internal corrosion rates in oil and gas pipelines. They use four advanced techniques: Random forest, adaptive boosting, gradient boosting regression tree, and extreme gradient boosting. Each model's performance is evaluated through k-fold cross-validation for robustness and generalizability. The Extreme gradient boosting model shows superior performance, with an RMSE of 0.031 mm/y and a performance index of 0.61, demonstrating exceptional predictive accuracy.

Another Study was conducted by Luo et al. [27] to develop a model based on SVM to predict the corrosion rates of gas pipelines. Known corrosion inspection data of oil pipelines have been used to train the model considering pressure, deposition rate, angle, the density of the gas, density of the liquid, liquid hold-up, liquid velocity, surface tension, pH value, fluid temperatures, inner wall surface temperature, flow regime, superficial velocity of gas, thermal conductivity of gas and maximum wall shear stress as inputs parameters. It was claimed that developed models provided a new thought for risk management, risk assessment and maintenance of oil pipelines. It was also stated that developed models could be useful for integrity management and quantitative assessment for long distance oil and gas pipelines.

Naveed Aslam et al. [4], use artificial intelligence algorithms like the DeWaard Model, Norsok Model, and Leak Rate Model to predict corrosion and leak rates in oil and gas pipelines. If data is consistent with previous data, the model predicts corrosion and leak rates. If data is inconsistent, a generic algorithm is used. The predictions are refined using a type 2 fuzzy logic subroutine algorithm. The method considers measured pipeline parameters and compares them against past data to predict corrosion and leakage sites. This method improves reliability by addressing uncertainties and measurement techniques [4].

Ismail et al [21] applied several machine learning techniques to address the challenges of manually processing large volumes of in-line inspection data for pipeline corrosion analysis. Specifically, the researchers tested decision tree, random forest, support vector machines, and logistic regression models for classifying pipeline defects according to the Pipeline Operator Forum's standards. The models were implemented using Python programming, and their performance was compared in terms of classification accuracy. The results showed that the decision tree classifier achieved the highest accuracy at 99.9%, outperforming the other machine learning approaches. This demonstrates the effectiveness of decision tree models in automating the classification of different types of pipeline corrosion defects, such as pitting, grooving, and slotting, which is a crucial step for determining corrosion growth rates and remaining pipeline life.

Abiral et al . [34] presents a study focused on assessing the corrosive nature of soil samples collected from 6 sites along the Budhanilkantha-Maharajganj roadway in Nepal. Soil samples were taken from depths of 0.3 to 1.5 meters and analyzed for various parameters using standard methods. The measured soil properties included pH, moisture content, electrical resistivity, redox potential, chloride ion concentration, and sulfate ion concentration. Based on these measured values, the authors applied an empirical corrosion rating model to classify the soils into different corrosivity groups, ranging from mildly corrosive to less corrosive, with respect to their potential impact on buried galvanized steel and cast iron water pipelines. The authors found good positive or negative correlations between the soil properties, indicating their individual contributions to the overall soil corrosivity. The successful application of the empirical corrosion rating model led the authors to propose its potential use for creating corrosive land maps to guide water pipeline infrastructure planning and management in urban areas of Nepal.

Bingyan et al. [11] propose a comprehensive approach to analyzing and modeling pipeline corrosion defects using Infrared Light Infrared (ILI) data. They start by analyzing the raw ILI data to visualize key features like corrosion depths and a number of corrosions detected. They then use a hierarchical clustering method to classify defects into severity levels based on corrosion depth and repair factor, considering interaction effects between adjacent corrosions. They then use machine learning algorithms to explore the relationship between the location parameters of adjacent corrosions and their severity levels. They also extract critical information from the raw ILI data across multiple inspection periods, filtering out maximum corrosion depths and density for long-term growth and failure prediction. Finally, they establish stochastic growth models to forecast the evolution of corrosion defects over time, crucial for pipeline integrity management.

06.2 Images data set based deep learning researches for Corrosion Defect Prediction

Will Nash et al. [31] presents a deep learning-based approach for pixel-level corrosion detection, along with three Bayesian variants that provide uncertainty estimates to better inform decision-making. Corrosion is a significant economic problem, costing 3-4% of GDP annually, and automated detection using deep learning has been an area of research. However, the lack of publicly available corrosion image datasets has hindered progress. The authors' previous work using a Fully-Convolutional-Network (FCN) model achieved limited performance, with issues like false positive detections on out-of-distribution data. The Bayesian variants introduced in this research include variational inference, Monte Carlo dropout, and an ensemble method - each of which replaces deterministic weights with distributions to output not just the predicted class map, but also estimates of epistemic and aleatoric uncertainty. Experiments on a new dataset of 225 corrosion images validate the improved performance and uncertainty estimates provided by the Bayesian models compared to the original deterministic deep learning approach, demonstrating the importance of quantifying uncertainty for practical deployment of these corrosion detection systems .

Bastian et al. [6] have developed a computer vision-based method for detecting corrosion in pipelines. They use a large dataset of optical images and a custom-designed CNN to classify corrosion levels with high accuracy. The network discriminates between corroded and non-corroded images and identifies corroded areas with precision. This approach surpasses traditional manual inspections and non-vision-based non destructive evaluation techniques in efficiency and cost, offering a promising alternative for pipeline maintenance and safety. The Custom CNN model has 200 times parameters than VGGNet and 32 times fewer than ZFNet.

Shirsath et al. [38] investigates the use of supervised machine learning and deep learning techniques for automated corrosion detection. It focuses on two key visual attributes of corrosion: color and texture. The first method employs a traditional computer vision approach, using a color tracking algorithm to detect corrosion based on color changes in images. The second method utilizes deep learning with a convolutional neural network (CNN) architecture, employing transfer learning to build a binary classification model that can detect corrosion based on texture. The third approach treats corrosion detection as an object detection problem, using a Single Shot Detector (SSD) deep learning model, also leveraging transfer learning, to identify and localize instances of corrosion in real-world images. To support the development and evaluation of these methods, the researchers created two datasets - one consisting of laboratory-generated images of corroded metal surfaces, and another containing real-world images of corroded compartments from bulk carrier inspections. The study found that all three approaches were capable of detecting corrosion, with the deep learning techniques outperforming the traditional color-based method. Additionally, the object detection approach using the SSD model was determined to be the most suitable for handling real-world corrosion detection scenarios.

In this chapter, we have exposed the related works of prediction corrosion. These studies have notably leveraged images and numerical datasets. The numerical data include critical parameters such as temperature, pressure, and production rates. These variables are essential for predicting the corrosion rate of pipelines.

However, despite the robustness of the techniques used so far, they do not fully exploit the recent advancements in artificial intelligence, such as multi-layer perceptron (MLP) networks and convolutional neural networks (CNNs). These advanced methods could significantly enhance the accuracy of predictions by utilizing the richness of datasets containing these important parameters.

Table III.2: Prediction Models for Oil and Gas Pipelines with Corrosion Factors

Reference	ML Technique Used	Input Parameters	Output Evaluated	Remarks
Seghier, Mohamed El Amine Ben et al. [37]	Ensemble Learning (Random Forest, Adaptive Boosting, Gradient Boosting Regression Tree, Extreme Gradient Boosting)	Presseure, temperature, CO2 and fluid Flow	Internal corrosion rates with RMSE of 0.031 mm/y and PI of 0.61	Utilizes robust ensemble learning models. Extreme Gradient Boosting showed superior performance. Statistical and graphical analyses used for evaluation.
Naveed Aslam et al. [4]	AI with Generic Algorithm, Neural Networks, Type 2 Fuzzy Logic	Various parameters including flow rates, pressure, gas composition, and past data	Corrosion and leak rates	Uses AI to compare current data against historical data to predict corrosion and leaks. Includes fuzzy logic for refining predictions. Continuous update of predictive model enhances reliability and accuracy.
Luo et al. (2013) [27]	SVM	Pressure, deposition rate, angle, density of the gas, density of liquid, liquid hold-up, liquid velocity, surface tension, pH value, fluid temperatures, inner wall surface temperature, flow regime, superficial velocity of gas, thermal conductivity of gas, maximum wall shear stress	Corrosion rate of gas pipeline	SVM results in adequate models that can be used for predicting corrosion rates based on the inputs in this study.
Bastian et al. (2019) [6]	Deep Neural Network (CNN)	Images dataset collected from oil and gas pipelines	Detect the level of corrosions	DNN demonstrated high accuracy in developing models to successfully identify corroded regions on pipelines.
Aslam (2018)	ANN, GA, FL	External parameters (external temperature, weather patterns, elevation) and internal stress parameters (gas composition, hydrocarbon composition, velocity, flow rate)	Leak and corrosion prediction	Data collected from field measurements. The developed HML model proved to be practical and accurate for predicting leak and corrosion in pipelines.
Sumayh S. Aljameel et al. [3]	RF, SVM, k-NN, GB, DT	Temperature ,Pressure , MMCFD Gas , BOPD, BWPD and CO2 MOL	SVM achieved the highest accuracy at 97.4%	SVM achieved the highest accuracy at 97.4%. Demonstrates the effectiveness of using machine learning algorithms to detect minor leaks in pipelines.
Ismail et al. **[ismail2017decision]**	Decision Tree, Random Forest, SVM, Logistic Regression	ot reported in year three dataset; absolute distance of the defect starting point from the origin (in meters), defect starting point circumferential location (in degrees), absolute distance of the defect endpoint and other factors	Classification of pipeline defects	Decision tree model showed the highest accuracy at 99.9%, effectively classifying various types of pipeline corrosion defects.
Will Nash et al. [31]	Deep Learning FNC	Pixel-level data from corrosion images	Corrosion detection with uncertainty estimates	Introduces Bayesian approaches to quantify uncertainty in corrosion detection, enhancing reliability and decision-making. Demonstrates significant improvements over previous deterministic models.
Bingyan et al. [11]	Hierarchical Clustering, Machine Learning Algorithms (not specified)	Raw ILI data: corrosion depths, number of corrosions, location parameters of adjacent corrosions	Classification of defect severity, long-term growth and failure predictions	Proposes a comprehensive approach utilizing ILI data. Begins with data visualization, uses clustering to classify defects, and applies ML algorithms to analyze relationships and predict future corrosion progression. Critical for integrity management.

07 Conclusion

This chapter has thoroughly examined the pivotal role of machine learning in enhancing the detection and management of corrosion within oil and gas pipeline infrastructure. It has demonstrated the crucial importance of integrating diverse operational and environmental factors into predictive models using various machine learning approaches. These approaches, including ensemble learning techniques, Support Vector Machines, and broader AI algorithms, utilize extensive datasets to effectively predict corrosion rates and potential failures in pipeline systems.

The limitations of existing methodologies are becoming more apparent as the complexity of pipeline networks increases. Traditional machine learning approaches, while effective, often rely on simpler models that may not capture the intricate relationships and patterns present in large and complex datasets. This shortcoming can lead to less accurate predictions, which in turn affects the reliability of corrosion detection and the timely maintenance of pipeline infrastructure.

Furthermore, the high costs associated with pipeline leaks underscore the need for more precise and reliable prediction models. Leaks can cause significant environmental damage, economic loss, and safety hazards. The ability to accurately predict corrosion can lead to better preventive maintenance strategies, reducing the occurrence of leaks and their associated costs.

To address these challenges, our research proposes to harness the latest advances in artificial intelligence, specifically MLPs and CNNs. These techniques are capable of processing more complex datasets and capturing nonlinear interactions between variables, which are often missed by traditional methods. MLPs, with their layered architecture, can model complex patterns through deep learning, while CNNs, known for their effectiveness in image processing, can be adapted to handle spatial data and detect subtle features indicative of corrosion.

CHAPTER IV

COMPARATIVE STUDY OF MACHINE LEARNING FOR CORROSION DETECTION

01 Introduction

This chapter presents the findings from the comprehensive analysis conducted to evaluate different predictive modeling techniques for corrosion categorization within oil and gas pipeline systems. Given the critical nature of accurately predicting corrosion to prevent operational failures and ensure safety, this study employed seven several advanced machine learning models, each with unique characteristics and capabilities. The models evaluated include K-Nearest Neighbour , Gradient Boosting , Decision Tree , Random Forest , support vector machine , multilayer perception and Convolutional Neural Networks

The objective of this analysis was to identify which model demonstrates the highest accuracy, precision, recall, and F1-score in predicting high and low corrosion scenarios. Each model was rigorously trained and tested using a well-curated dataset, and the results were systematically recorded and analyzed. The performance metrics of these models are critical, as they directly impact the reliability of corrosion predictions, which in turn, influence maintenance strategies and operational efficiencies in the oil and gas industry.

The following sections will detail the performance of each model, discuss their strengths and weaknesses, and provide a comparative analysis to guide the selection of the most appropriate modeling technique for effective corrosion management. Through this evaluation, this study aims to contribute valuable insights to the field of predictive maintenance and enhance the technological approaches used in the management of pipeline integrity.

02 Methodology

We delve into the methodologies employed to assess and enhance the predictive accuracy of various machine-learning classifiers of corrosion detection. It begins with a detailed explanation of our initial data handling process, emphasizing the crucial role of preprocessing in preparing the dataset for effective model training. This involves techniques such as data cleansing, normalization, and the transformation of features to ensure that the classifiers receive high-quality inputs.

Following the data preparation, we introduce a decision-making node that evaluates whether the conditions observed in the dataset suggest high or low corrosion risk. This binary decision is pivotal as it determines the subsequent analysis path, either proceeding with deeper investigation or bypassing further unnecessary computations on non-corrosive instances.

The core of this research focuses on a comparative analysis of several sophisticated classifiers, including SVM, KNN, CNN, GB, DT, RF and RT. Each classifier is rigorously evaluated across multiple performance metrics such as accuracy, precision, recall, and the F-score. These metrics serve as benchmarks to gauge each model's effectiveness in identifying and predicting different levels of corrosion based on the preprocessed data.

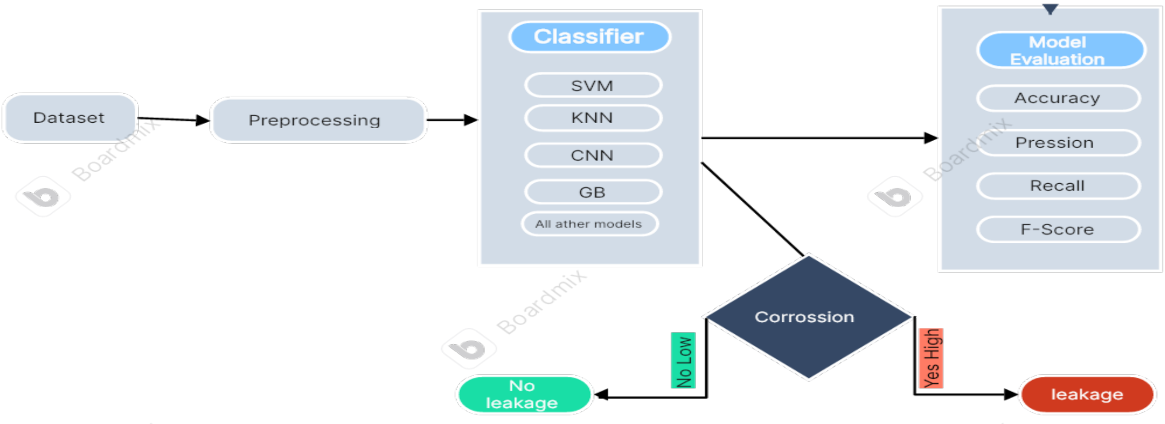


Figure IV.1: Visual Illustration of the Methodology

A comprehensive flowchart illustrates IV.1the sequence of operations from dataset acquisition through to model evaluation, providing a visual representation of the methodology. This flowchart is not only integral for understanding the systematic approach to our analysis but also aids in the replication and validation of our results by other researchers in the field.

03 Data Collection

A publicly accessible, open-source dataset hosted on GitHub [Click here to visit dataset.com] was employed for the analyses presented in this study. This dataset, originally designed for regression analysis targeting corrosion defects, contains 10,293 instances across eight features, including critical variables such as wellhead temperature, pressure, and gas composition. These variables are crucial for assessing the structural integrity of pipeline systems. Before its application, the dataset underwent several preprocessing steps such as normalization and outlier removal to ensure the accuracy and reliability of subsequent analyses. The versatility of the dataset allows for its application in classification tasks, making it an invaluable resource for predictive maintenance within the oil and gas industry.

Table IV.1: Features Description

Feature	**Description**
Wellhead Temperature (°C)	Temperature at the pipeline wellhead, critical for evaluating material stress and corrosion potential.
Wellhead Pressure (psi)	Pressure inside the wellhead, indicative of stress levels on pipeline materials.
MMCFD Gas	Daily gas production measured in million standard cubic feet, essential for assessing production efficiency.
BOPD (Barrels of Oil Per Day)	Daily oil production rate, crucial for operational planning and efficiency analysis.
BWPD (Barrels of Water Per Day)	Water output rates, influencing corrosion within the pipelines.
BSW (Basic Solid and Water)	Ratio of solid particles to water in the oil, pertinent to processing needs and corrosion assessment.
CO2 Mol. (Molecular Mass of CO2)	Concentration of carbon dioxide, a critical factor in corrosion dynamics.
Gas Gravity	Relative density of gas to air, important for understanding the composition and its impact on materials.
CR (Corrosion Rate)	Direct measurement of the corrosion rate, vital for maintenance scheduling and infrastructure longevity.

Features Description

The dataset employed in this research comprises a comprehensive array of features essential for analyzing the operational dynamics and structural integrity of oil and gas pipelines. The detailed description of each feature is outlined below, emphasizing their relevance in assessing corrosion and pipeline performance show table IV.1. The detailed descriptions of key features such as Wellhead Temperature and Wellhead Pressure, along with their respective data distributions, are outlined below to illustrate their roles in operational analysis and predictive maintenance strategies.

Pressure conditions are similarly categorized and analyzed through visual representations in pie and bar charts. The data is segmented into five pressure ranges: 0-500 psi, 501-1000 psi, 1001-1500 psi, 1501-2000 psi, and 2001-2500 psi. The bar chart complements this by showing the number of occurrences in each category, with a noticeable concentration in the 1001-1500 psi range IV.2. Such analysis is essential for understanding the pressure stresses exerted on the pipeline materials and for assessing potential vulnerability to pressure-induced failures or corrosion.

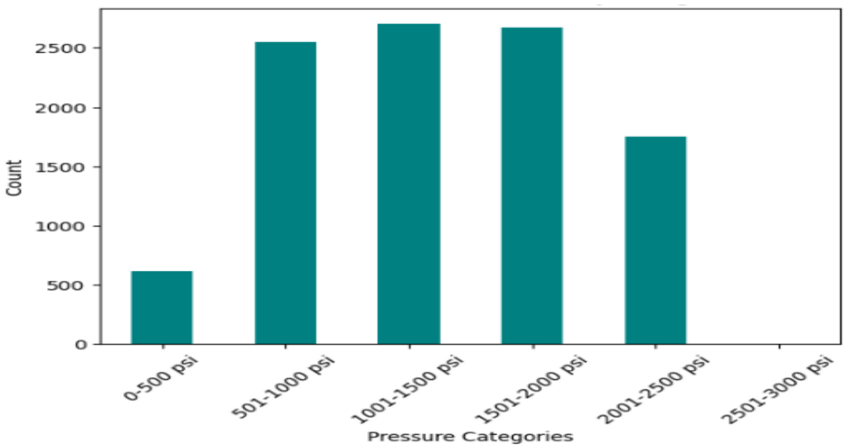


Figure IV.2: pressure Distribution by categories

The distribution of Million Cubic Feet per Day (MMCFD) categories offers valuable insights into operational volumes where leak risks might be more prevalent or critical.

- The 0-5 MMCFD category may have fewer leak incidents due to simpler infrastructure and less pressure, potentially reducing leakage risk. However, consistent monitoring is crucial for early detection and mitigation, as leaks still pose significant environmental and safety risks.
- The high frequency of data points in 6-10 MMCFD and 11-15 MMCFD ranges suggests common operational volumes, potentially increasing leak likelihood. Facilities operating in these ranges should use robust safety and monitoring systems to mitigate the environmental and operational consequences of leaks.
- The 16-20 MMCFD, despite its lower frequency of operations, poses significant risks due to leaks. Higher flows can cause environmental damage and safety hazards. The lower frequency suggests specialized operations requiring additional safeguards and advanced monitoring technologies

From an oil and gas safety perspective, understanding these MMCFD distributions helps in prioritizing risk management efforts. Facilities operating in the higher volume categories, despite their fewer numbers, may necessitate more stringent leak detection and control measures due to the potential severity of leaks.

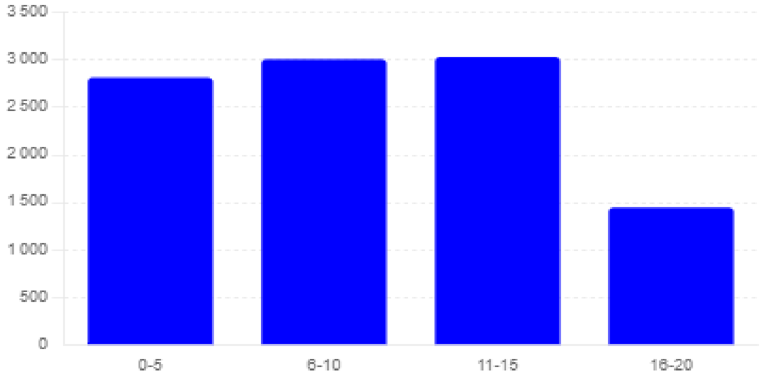


Figure IV.3: Distribution-Gas Of MMCFD Categories

The distribution breakdown of barrels of oil produced per day (BOPD) across different production categories are shown in depth in the graphic representation that is provided. A typical production range for the majority of activities is indicated by the frequency chart, which shows a constant number of occurrences throughout the categories ranging from 0-500 to 1501-2000 barrels per day. Significantly, the frequency of production levels beyond 2000 barrels per day declines sharply, indicating that very large outputs are comparatively uncommon.

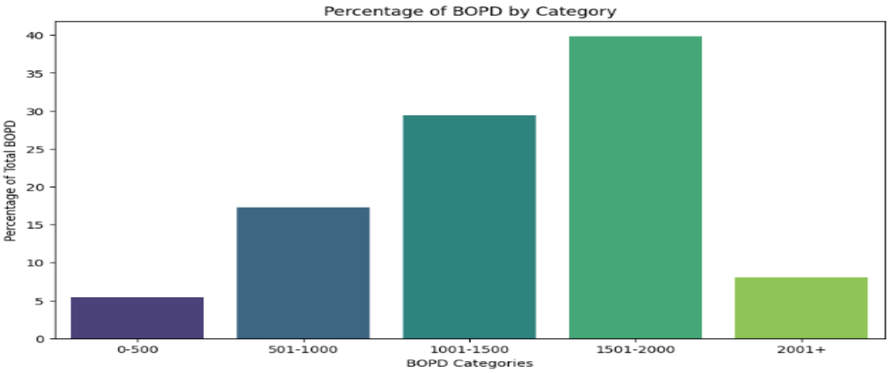


Figure IV.4: BOPD Distribution by categories

The temperature distribution within the pipeline system is visualized bar charts, categorizing the operational temperatures into four main intervals: $41 - 49°C$, $50 - 57°C$, $58 - 65°C$, and $66 - 74°C$. This visualization helps in identifying the most common operational temperatures and assists in evaluating how these conditions may influence corrosion rates and material integrity.

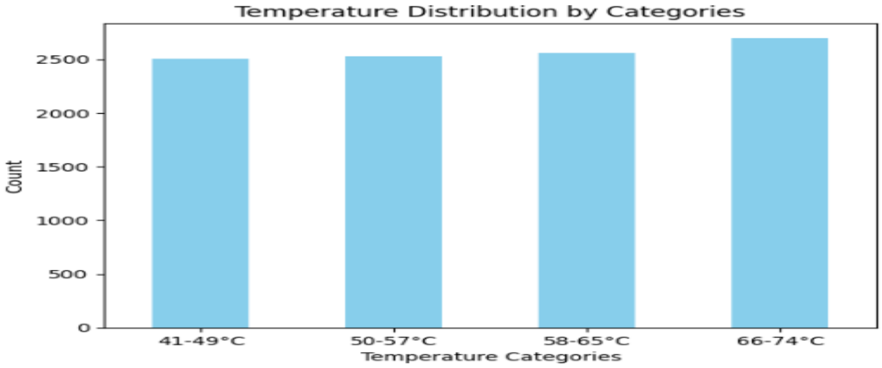


Figure IV.5: Temperature Distribution by categories

Data Preprocessing

Data Loading and Inspection This research makes use of a solid dataset that was taken from the operating data of oil and gas pipelines,,. It includes a wide range of critical parameters, including daily production metrics for gas (MMCFD-gas), oil (BOPD), and water (BWPD), as well as wellhead temperature (°C) and wellhead pressure (psi). Basic Solid and Water Content (BSW), CO2 content, Gas Gravity, and Corrosion Defects are additional variables monitored, resulting in a dataset of 10,292 entries. Python was used for the first explorations in a Google Collab Notebook environment.

Table IV.2: Summary Statistics of Oil Well Operations

Parameter	Count	Mean	Std	Min	25%	50%	75%	Max
Wellhead Temp. (C)	10292	57.35	9.43	41.07	49.22	57.36	65.41	73.87
Wellhead Press (psi)	10292	1361.74	559.27	382.00	880.00	1364.90	1848.25	2317.23
MMCFD-gas	10292	8.85	4.97	0.23	4.57	8.88	13.09	17.54
BOPD	10292	1103.56	565.39	129.47	611.64	1106.08	1589.71	2087.43
BWPD	10292	4636.56	2685.80	40.61	2295.52	4591.99	6997.44	9314.26
BSW (%)	10292	44.87	25.71	0.13	22.89	45.08	67.21	89.26
CO2 mol	10292	2.52	1.04	0.68	1.61	2.52	3.41	4.30
Gas Gravity	10292	0.82	0.06	0.71	0.77	0.82	0.87	0.93
CR-corrosion defect	10292	0.21	0.04	0.00	0.19	0.21	0.23	0.41

The Wellhead Temperature, with an average of 57.35°C and a standard deviation of 9.43°C, shows moderate variability, indicating diverse geothermal conditions at the pipeline. Similarly, the Wellhead Pressure averages at 1361.74 psi, yet the wide range from 382.00 to 2317.23 psi reflects significant differences in subsurface pressures, which may impact extraction efficiency and safety measures. the gas production, measured in MMCFD, and oil production, measured in barrels per day, exhibit substantial standard deviations, highlighting the varying productivity of the wells. Notably, water production significantly exceeds oil production, suggesting prevalent water intrusion or high water cut in the extracted fluids, which is critical for planning water handling and treatment facilities, as depicted in Table IV.2 The Basic Solid and Water content, and CO2 molecular percentage at standard conditions, are crucial for assessing corrosion risk and pipeline integrity. A higher variability in these parameters could indicate fluctuating levels of impurities and gas compositions, affecting the material selection and corrosion management strategies.

Lastly, the Gas Gravity and Corrosion Defect measurements provide insights into the chemical characteristics of the extracted gas and the integrity of the pipeline components, respectively. The slight variation in Gas Gravity suggests a relatively consistent gas composition across the dataset. In contrast, the Corrosion Defect index, though averaging low, shows potential hotspots for maintenance prioritization.

Label Binarization

It is a process in machine learning and data preprocessing where you convert categorical data into a format that can be easily used by algorithms, typically by transforming labels into a binary format. This technique is particularly useful when dealing with categorical target variables in classification problems , In our methodology, we analyze the 'CR-corrosion defect' parameter, which quantifies corrosion defects within oil and gas pipeline components. Originally a continuous variable, it was transformed into a binary format to simplify the analysis and enhance the predictive modeling and risk assessment process. This transformation involved setting a predetermined threshold of 0.211. Values above this threshold are classified as 'high', indicating a significant risk of corrosion, whereas values below are deemed 'low', signifying minimal risk.

This led to the creation of a balanced dataset, where the 'high' category comprises 5,491 samples and the 'low' category contains 4,801 samples, as illustrated in Figure IV.IV.6. This figure underscores the effectiveness of our label binarizing approach in the context of our broader study objectives, highlighting the potential corrosion challenges in pipeline management ,

Correlation Matrix

The analysis of pipeline integrity data shows correlations between corrosion defects and operational parameters. A significant negative correlation of -0.37 with Wellhead Pressure suggests higher pressures in the pipeline system lead to fewer or less severe corrosion defects, possibly due to materials resistance or reduced corrosive interactions.

Conversely, a positive correlation of 0.22 with MMCFD indicates that higher rates of gas throughput could exacerbate corrosion, likely due to increased flow rates and subsequent mechanical wear or the impact of gas composition on the internal surfaces of the pipelines. Furthermore, minor positive correlations with BWPD and BOPD suggest an increase in corrosion incidents associated with higher outputs of these fluids.

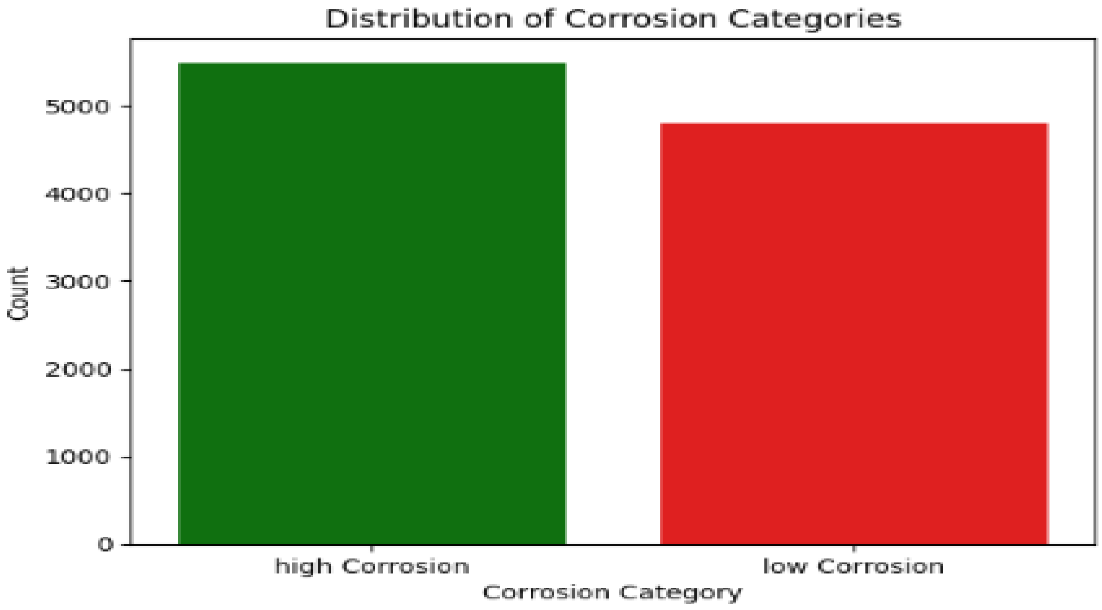


Figure IV.6: Class distribution

Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features in data. In machine learning, feature scaling is crucial because algorithms that compute distances between data points (such as k-nearest neighbors and support vector machines) are sensitive to the magnitude of the features. The main aim of feature scaling is to ensure that no single feature can dominate others in terms of its scale, thus giving each feature equal importance, which can significantly improve the performance of the model.

Normalization and standardization are two common scaling techniques. Normalization, or min-max scaling, adjusts the data values so that they fall within a specified range, typically [0,1]. Normalisation uses a general formula [3] give:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, x' is the new value, x is the original values, $\min(x)$ and $\max(x)$ are the minimum and the maximum values of the feature, respectively.

making it useful for algorithms that require data to be in a bounded interval. Standardization transforms the data to have zero mean and a variance of one, making it suitable for algorithms that assume data is normally distributed. Employing these techniques helps in speeding up the convergence of learning algorithms by providing a level playing field.

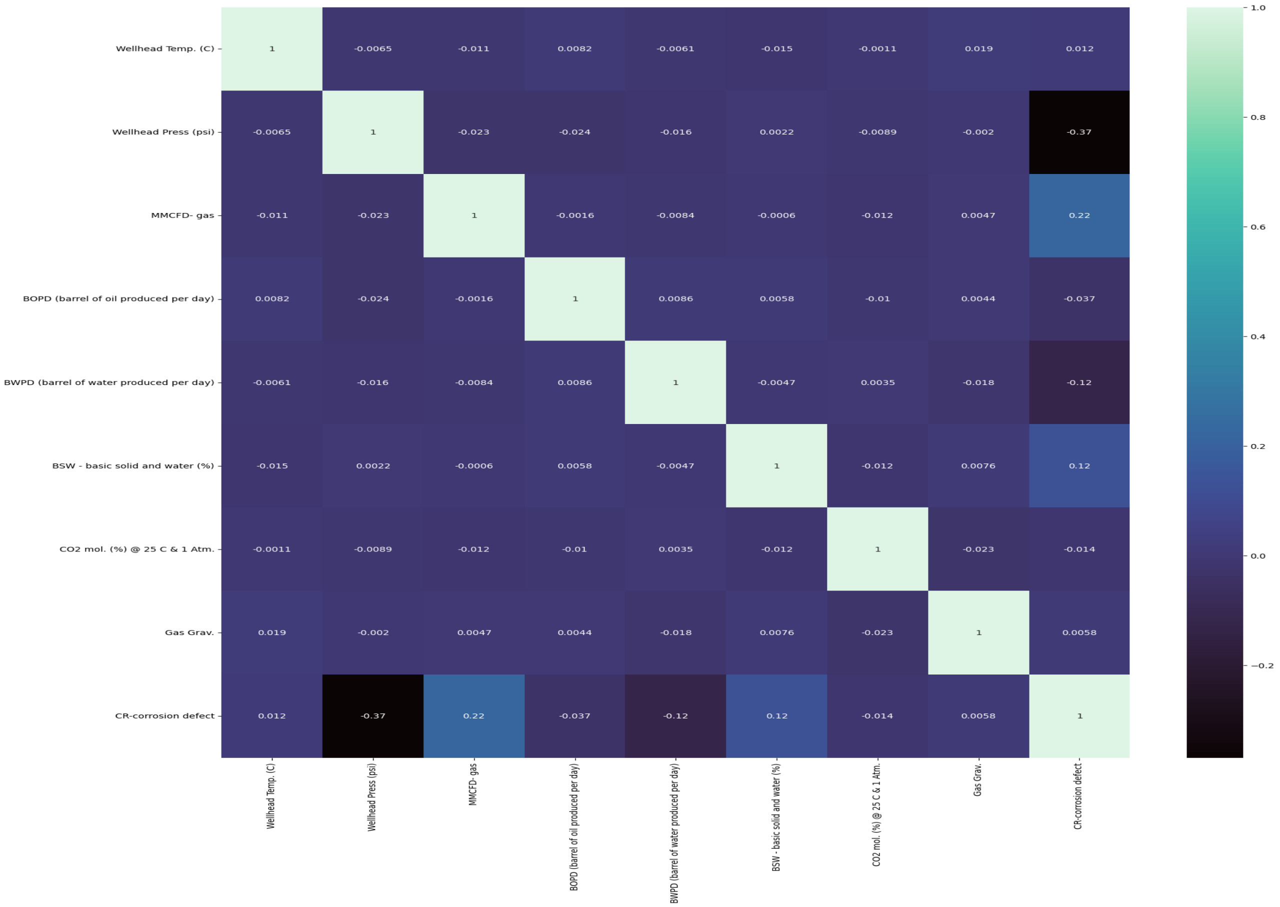


Figure IV.7: Correlation Matrix.

04 Implementation

- **K-Nearest Neighbors** Selecting the optimal value of k, To measure the similarity between target and training data points, Euclidean distance is used. Distance is calculated between each of the data points in the dataset and target point. Finding Nearest Neighbors In the classification problem, the class labels of are determined by performing majority voting. The class with the most occurrences among the neighbors becomes the predicted class for the target data point. k is set to the square root of the number of samples in the training dataset. It's a starting point, but it might not always provide the best results. If the total number of samples n is 10,292, then k would be approximately $\sqrt{10292} \approx 101$.

Neighbors	Accuracy
K = 101	87.6
K = 65	87.4
K = 75	87.9
K = 21	90
K = 10	89
K = 5	89

Table IV.3: Accuracy scores for different numbers of neighbors

– **Gradient Boosting** The process involves establishing a Gradient Boosting Classifier with appropriate parameters, n_estimators = 200, learning_rate = 0.1, and max_depth = 3. The model is then trained on scaled and processed training data. The model is then used to predict corrosion levels on the test dataset, and its performance is assessed using metrics like accuracy, precision, recall, F1-score, and ROC-AUC curve. The model's feature importance analysis is also conducted to guide data collection and preventive measures in pipeline management. The model is then deployed into a production environment for real-time predictions.

– **Random Forest Classifier** The model is configured with n_estimators = 100, and trained on the processed training data. Performance metrics are evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. A confusion matrix is generated to visually assess performance and feature importance.

– **Support Vector Machine** we utilize a Support Vector Machine (SVM) with a polynomial kernel to predict corrosion categories in pipeline systems. The SVM model is configured with a polynomial kernel (kernel=poly) to handle the non-linear patterns often observed in corrosion data. The degree of the polynomial is set to 3 (degree=3), allowing the model to capture more complex relationships without becoming too computationally intensive. The gamma parameter is set to 'scale' which automatically adjusts it based on the number of features, ensuring that the model is not overly sensitive to the scale of the data. Additionally, a coefficient of 0.1 (coef0=0.1) is used to control the model's independence term, which can significantly influence the decision boundary in higher-dimensional spaces.

– **Multi-layer Perceptron** we implemented a Multilayer Perceptron (MLP) Classifier to address complex classification challenges. The MLP was designed with a specific architecture comprising two hidden layers, with 6 neurons in the first layer and 5 in the second. This design was chosen to provide a balanced approach to learning, capturing essential patterns in the data without overly complicating the model structure, which can lead to overfitting. The learning rate was set to 0.01, a decision aimed at achieving a stable and efficient convergence during training by controlling how much the model adjusts its weights in response to the error each iteration. Additionally, we set the random_state to 101 to ensure reproducibility of the results, enabling consistent initialization of the weights. The MLP Classifier's configuration was carefully selected to optimize performance on our dataset, balancing complexity and learning capability to enhance predictive accuracy

  The Multilayer Perceptron (MLP) uses an input layer and two hidden layers for binary classification tasks. Its output layer, consisting of a single neuron with a sigmoid activation function, outputs a probability indicating the likelihood of input data belonging to the positive class. This architecture enables robust performance in binary classification tasks to predicting pipeline corrosion, making nuanced distinctions crucial for accurate predictions

- **Convolutional Neural Network** we utilized a Convolutional Neural Network (CNN), renowned for its proficiency in processing structured arrays of data, such as images or time-series. The initial configuration of our CNN model comprised 32 filters with a kernel size of 2, which are crucial parameters that determine the model's ability to extract fine-grained features from the input data. This setup ensures that the model captures both the low-level details and high-level features, essential for accurate predictions. Subsequent layers of the network included two densely connected layers with 128 and 64 neurons, respectively, providing the model with the capability to learn complex patterns effectively.

Table IV.4: Model Parameters for Various Classification Models

Model	**Parameters**
K-Nearest Neighbors	n_neighbors=21
Gradient Boosting	n_estimators=200, learning_rate=0.1, max_depth=3
Random Forest	n_estimators=100
Support Vector Machine	kernel=poly, degree=3, gamma=scale, coef=0.1
MLP Classifier	hidden_layers=(6, 5), lr=init=0.01, random_state= 101
Convolutional Neural Network	filters=32, kernel_size=2, layers=(128, 64), optimizer=Adam

05 Conclusions

This chapter compares seven models for predicting corrosion defects in oil and gas infrastructure. The models were evaluated for their efficacy in predicting risks and ensuring safety and integrity of operations. The goal is to identify the most reliable model.

The comparison of models aimed to predict corrosion-related incidents with high accuracy, enhancing protective measures for oil and gas infrastructure. It highlighted strengths and weaknesses, setting the stage for selecting the optimal approach for real-world applications. The next chapter will detail the analysis results and discuss the implications of the findings.

CHAPTER V

RESULTS AND DISCUSSION

01 Introduction

In this chapter, we present a comprehensive analysis and discussion of the results obtained from evaluating various machine learning models on their classification performance. The focus is on understanding the effectiveness of these models in predicting corrosion categories within our dataset. By leveraging a range of performance metrics, including Precision, Recall, F1-Score, and Accuracy, we aim to provide a clear comparative insight into each model's strengths and weaknesses.

The analysis covers several machine learning algorithms, such as K-Nearest Neighbors , Gradient Boosting, Decision Trees, Random Forest , Support Vector Machines, Multi-Layer Perceptron, and Convolutional Neural Networks. Through detailed examination, we highlight the superior performance of SVM and CNN models, which exhibit remarkable precision and recall across both classes, indicating their robustness in handling the classification tasks. Conversely, models like Decision Trees demonstrate lower accuracy, shedding light on their limitations in this context.

02 Performance Evaluation Metrics

It is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.

It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves.

Let's understand TP, FP, FN, TN in terms of pregnancy analogy.

Accuracy represents the percentage of the truly predicted samples among all the samples in the testing set [3]

Predicted Values \ Actual Values	**Positive (1)**	**Negative (0)**
Positive (1)	TP	FP
Negative (0)	FN	TN

Table V.1: Confusion Matrix

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision represents the percentage of the truly predicted samples of the positive class among all the positive predictions [3]

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall (also known as sensitivity) represents the percentage of the positive samples that were correctly predicted among all the real positive samples [3]

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity represents the percentage of the negative samples that were correctly predicted among all the real negative samples [3]

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

03 Results

Comparative Analysis of Machine Learning Models on Classification Performance

The classification report table provides a comprehensive view of the performance metrics for various machine learning models, including KNN, GB, DT, RF, SVM, ANN, MLP, and CNN. Notably, SVM and CNN models exhibit superior performance across all metrics, showcasing their robustness in handling the task. Specifically, SVM achieved the highest overall accuracy and F1-scores of 0.96 for both classes, highlighting its exceptional precision and recall balance. Similarly, CNN also performed exceptionally well, with an accuracy of 0.94 and very high precision and recall values, suggesting its effectiveness in feature extraction and classification in complex datasets.

Conversely, the DT models underperformed in comparison, with DT achieving the lowest accuracy of 0.83 . Additionally, the traditional models like KNN, GB, DT, and RF showed competitive but varied results. RF stood out among these with an accuracy of 0.91, indicating its strength in managing overfitting through ensemble techniques.

	Low Corrosion			High Corrosion			Accuracy
	Precision	Recall	F1-score	Precision	Recall	F1-score	
KNN	0.91	0.86	0.88	0.89	0.93	0.91	0.89
GB	0.91	0.87	0.89	0.9	0.93	0.91	0.9
DT	0.82	0.82	0.82	0.85	0.85	0.85	0.83
RF	0.91	0.90	0.91	0.92	0.92	0.92	0.91
SVM	0.96	0.97	0.96	0.97	0.97	0.97	**0.96**
MLP	0.93	0.96	0.94	0.95	0.91	0.93	0.93
CNN	0.92	0.96	0.94	0.96	0.92	0.94	**0.94**

Table V.2: Performance metrics for the various Models

04 Discussion

KNN

In this study, the performance of a KNN classifier with 21 neighbors was assessed for its capability to predict corrosion categories . The results obtained from the confusion matrix indicated that the classifier achieved an overall accuracy of approximately 89.8%. Specifically, the classifier correctly identified low corrosion instances with a precision of 0.91 and a recall of 0.86, resulting in an F1-score of 0.88. High corrosion instances were predicted with a precision of 0.89 and a recall of 0.93, yielding an F1-score of 0.91. The normalized confusion matrix further revealed that the model successfully predicted high corrosion categories 93% of the time and low corrosion categories 86% of the time. These findings suggest that the KNN classifier is effective at distinguishing between high and low corrosion scenarios, demonstrating strong potential for practical application in predictive maintenance and operational adjustments to mitigate corrosion impacts in oil and gas pipeline .

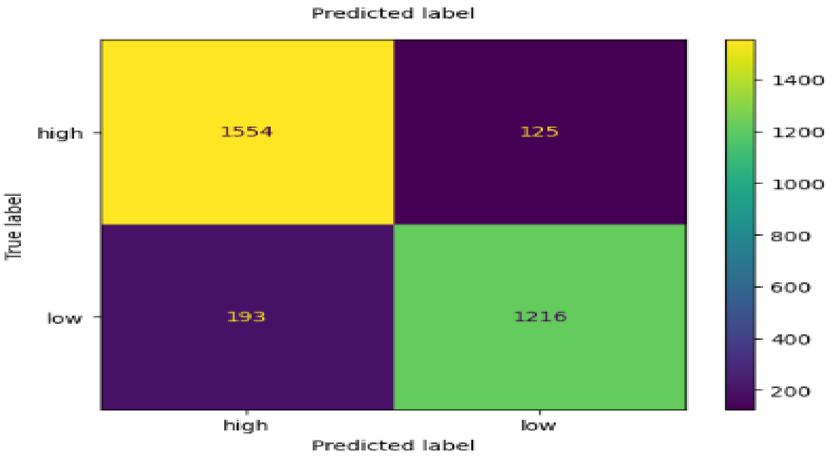
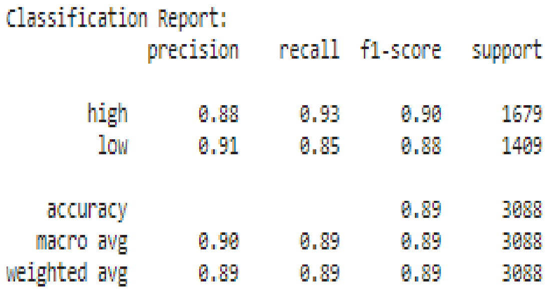


Figure V.1: Confusion Matrix KNN



```
Classification Report:
              precision    recall  f1-score   support

        high       0.88      0.93      0.90      1679
         low       0.91      0.85      0.88      1409

    accuracy                           0.89      3088
   macro avg       0.90      0.89      0.89      3088
weighted avg       0.89      0.89      0.89      3088
```

Figure V.2: Classification Report KNN

Gradient Boosting

The Gradient Boosting Classifier was employed to predict corrosion categories, utilizing a configuration of 200 estimators, a learning rate of 0.1, and a maximum depth of 3. The classifier demonstrated excellent performance, achieving an overall accuracy of 90.45%. The analysis of the confusion matrix reveals that the model successfully identified 1230 true negatives and 1563 true positives, while misclassifying 179 as false positives and 116 as false negatives. This indicates a strong predictive capability, particularly in correctly identifying instances of high corrosion. The precision rates stood at 91% for low corrosion and 90% for high corrosion scenarios, with recall rates at 87% and 93% respectively. The corresponding F1-scores were 89% for low corrosion and 91% for high corrosion, reflecting the model's balanced accuracy in classifying both corrosion categories. These metrics highlight the classifier's robustness and its potential utility in practical applications for monitoring and preventing corrosion in oil wells.

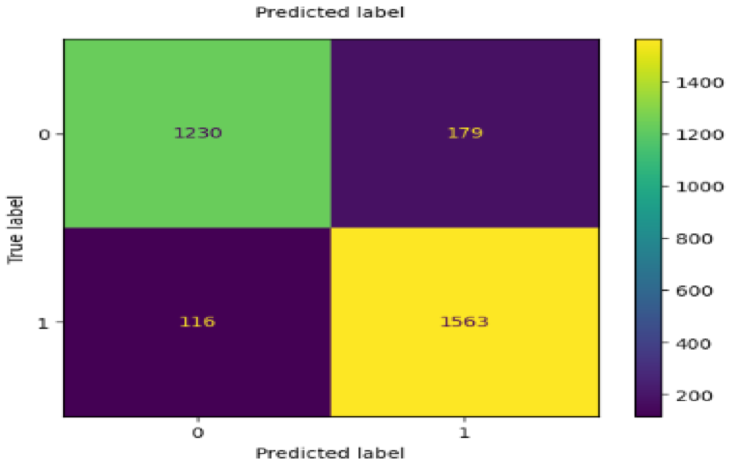
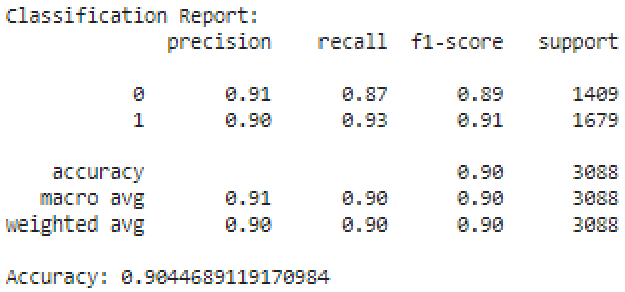


Figure V.3: Confusion Matrix GB



```
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.87      0.89      1409
           1       0.90      0.93      0.91      1679

    accuracy                           0.90      3088
   macro avg       0.91      0.90      0.90      3088
weighted avg       0.90      0.90      0.90      3088

Accuracy: 0.9044689119170984
```

Figure V.4: Classification Report GB

Random Forest Classifier

The Random Forest Classifier, with 100 estimators, exhibited exemplary performance in predicting corrosion categories, achieving an impressive overall accuracy of 91.52%. The detailed analysis of the confusion matrix revealed that the model correctly predicted 1275 cases as low corrosion (true negatives) and 1551 cases as high corrosion (true positives), while misclassifying 134 as false positives and 128 as false negatives. Such results demonstrate the model's high precision (91% for low corrosion and 92% for high corrosion) and recall (90% for low corrosion and 92% for high corrosion), with corresponding F1-scores of 91% and 92%, respectively. These statistics highlight the Random Forest Classifier's robust capability to discern between different corrosion levels effectively, making it a reliable tool for predictive maintenance in oil well operations. The normalized confusion matrix, which visually supports these findings, underscores a strong true positive rate of 92% and a true negative rate of 90%, affirming the classifier's accuracy and practical utility in real-world applications.

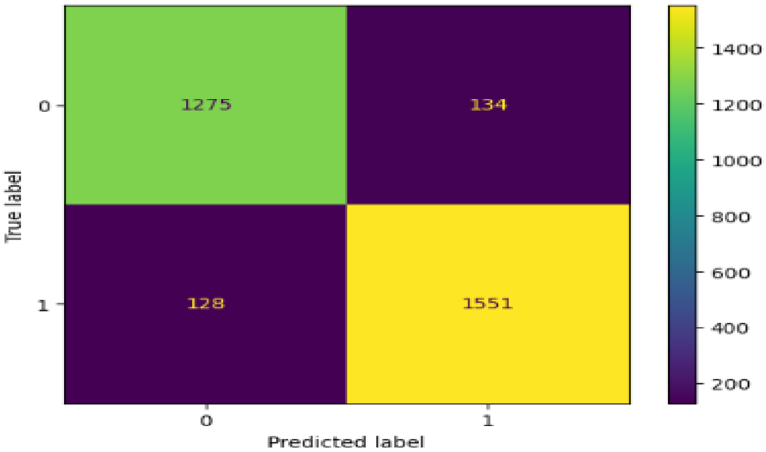
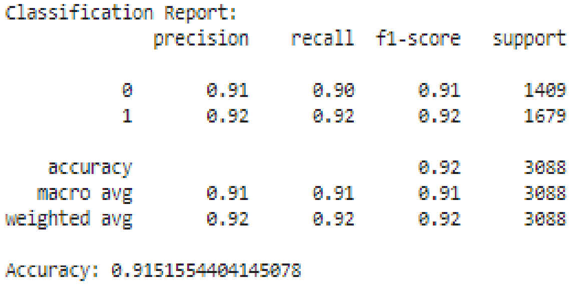


Figure V.5: Confusion Matrix RF



```
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.90      0.91      1409
           1       0.92      0.92      0.92      1679

    accuracy                           0.92      3088
   macro avg       0.91      0.91      0.91      3088
weighted avg       0.92      0.92      0.92      3088

Accuracy: 0.9151554404145078
```

Figure V.6: Classification Report RF

Support Vector Machine (SVM)

SVM classifier with a polynomial kernel was implemented to predict corrosion categories, achieving a remarkable overall accuracy of 96.97%. The analysis revealed a strong predictive capability, with the model accurately identifying 1361 low corrosion and 1625 high corrosion cases. The misclassifications were notably low, with only 48 false positives and 54 false negatives. This performance demonstrates the SVM's exceptional ability to discern between the two corrosion states effectively.

Precision metrics were impressive, with the classifier scoring 96% for low corrosion and 97% for high corrosion. The recall was equally robust, with 97% for high corrosion and 96% for low corrosion, leading to F1-scores of 96% and 97%, respectively. These statistics illustrate the model's consistent accuracy and reliability across different corrosion conditions.

The normalized confusion matrix, which provides a visual representation of the model's accuracy, showed that the SVM classifier successfully predicted high corrosion scenarios with 97% accuracy and low corrosion scenarios with 96% accuracy. This confirms the SVM's robustness and its practical applicability in accurately predicting corrosion levels, making it an invaluable tool for proactive corrosion management in industrial settings.

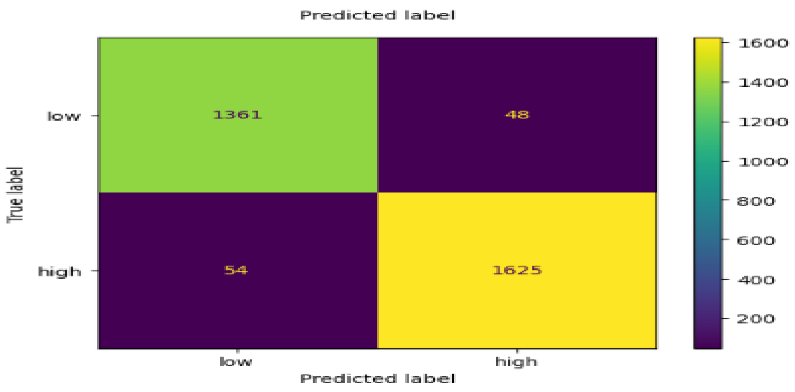
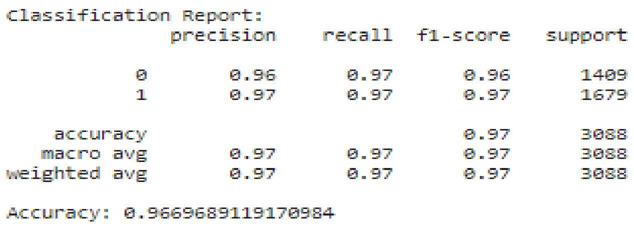


Figure V.7: Confusion Matrix SVM



```
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.97      0.96      1409
           1       0.97      0.97      0.97      1679

    accuracy                           0.97      3088
   macro avg       0.97      0.97      0.97      3088
weighted avg       0.97      0.97      0.97      3088

Accuracy: 0.9669689119170984
```

Figure V.8: Classification Report SVM

MLPClassifier

The MLPClassifier demonstrates excellent performance with an overall accuracy of approximately 93.91%, supported by high precision and recall values across both corrosion categories. Precision scores of 0.93 for low corrosion and 0.95 for high corrosion, alongside recall rates of 0.96 and 0.91 respectively, indicate the model's strong capability in both identifying actual cases of corrosion and minimizing false positives. The F1-scores, closely aligning with precision and recall at 0.94 for low corrosion and 0.93 for high corrosion, reflect a well-balanced model. The confusion matrix further substantiates the model's efficacy, with a substantial majority of true positives (1308) and true negatives (1592), while maintaining relatively low false positives (64) and false negatives (124). This robust performance suggests that the model is highly effective and reliable for predicting corrosion categories, making it a valuable tool in preventative maintenance strategies to avoid costly failures and enhance operational safety.

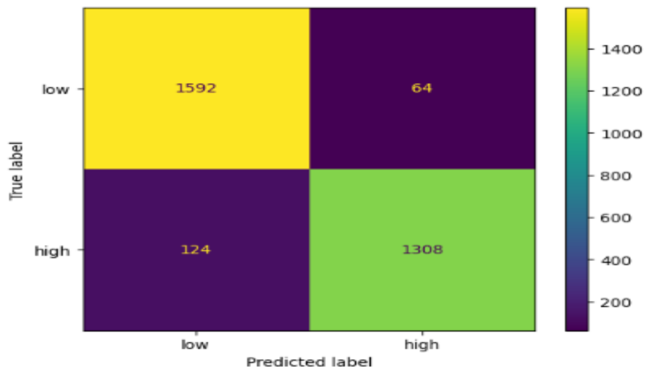
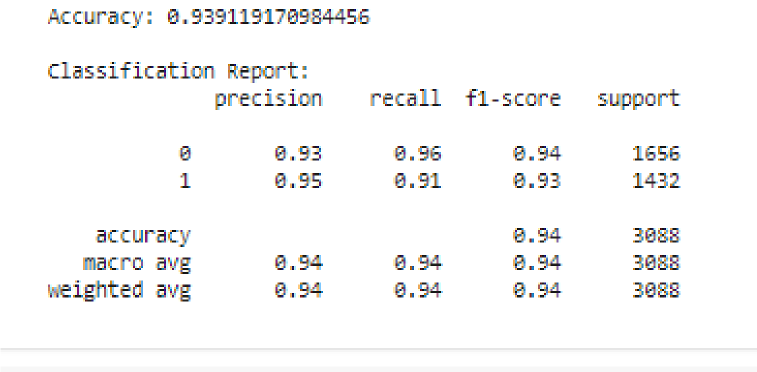


Figure V.9: Confusion Matrix MLP(ANN)



```
Accuracy: 0.939119170984456

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.96      0.94      1656
           1       0.95      0.91      0.93      1432

    accuracy                           0.94      3088
   macro avg       0.94      0.94      0.94      3088
weighted avg       0.94      0.94      0.94      3088
```

Figure V.10: Classification Report MLP(ANN)

Convolutional Neural Network

In a significant advancement in predictive accuracy for corrosion categories, the implementation of a Convolutional Neural Network model demonstrated outstanding performance. The model achieved a remarkable accuracy of 94.17%, with precision at 93.48% and a recall of 95.70%, leading to an F1 score of 94.58%. These metrics are indicative of the model's robust capability to identify and classify corrosion conditions accurately.

The confusion matrix, detailed in the analysis, visually represents the model's effective discrimination between 'Non-Corroded' and 'Corroded' states, affirming the precision of its predictive ability. This performance underscores the CNN model's application potential, leveraging its architectural strengths to handle complex pattern recognition tasks associated with corrosion data effectively.

The model was structured with an initial convolution layer followed by a max pooling layer and a sequence of dense layers, optimized using an Adam optimizer with a learning rate of 0.001. This configuration not only facilitated excellent learning dynamics but also ensured the model was sufficiently generalizable, maintaining high reliability across validation datasets.

The Receiver Operating Characteristic (ROC) curves for the models—GB, DT, KNN, RF, SVM, MLP ANN, and CNN—display varying degrees of performance. Notably, the SVM model exhibits a re-

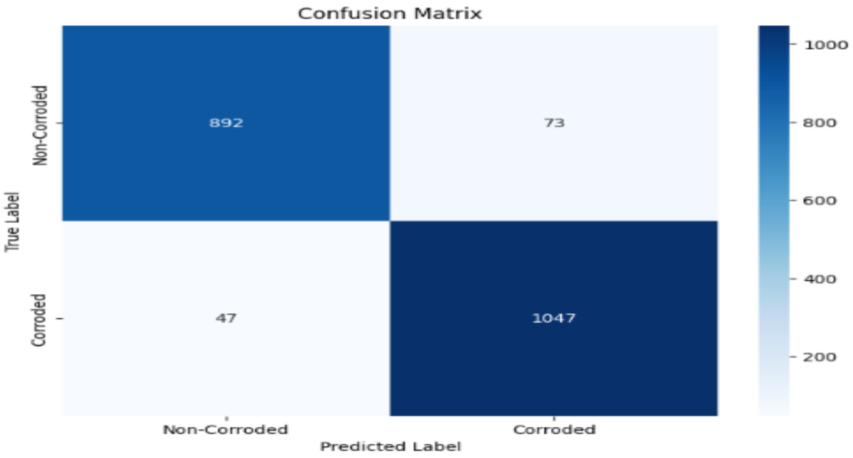
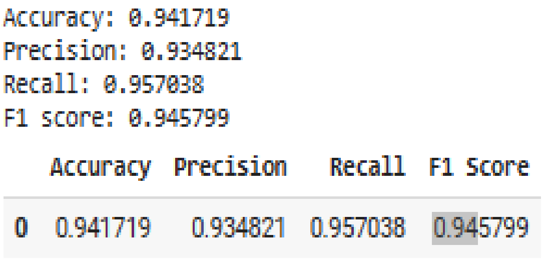


Figure V.11: Confusion Matrix CNN



```
Accuracy: 0.941719
Precision: 0.934821
Recall: 0.957038
F1 score: 0.945799
    Accuracy  Precision    Recall  F1 Score
0   0.941719   0.934821  0.957038  0.945799
```

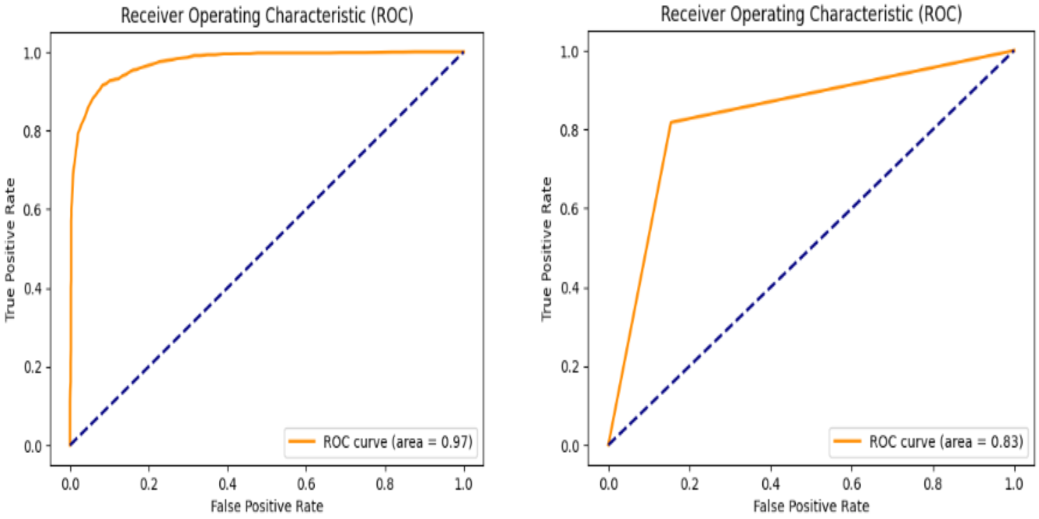
Figure V.12: Classification Report CNN

markably high Area Under the Curve (AUC), indicative of its superior discriminative ability between classes of leaks. This is corroborated by its outstanding performance metrics, achieving a precision of 0.96 and 0.97 for Class 1 and Class 2, respectively, matched by equally high recall and F1-scores, culminating in an accuracy of 0.96.

In contrast, DT model shows a more modest performance, with its AUC noticeably lower, suggesting a less effective capability in distinguishing between the classes. This is reflected in its lower precision, recall, F1-score, and accuracy, especially for Class 1, which significantly lags behind other models.

The performance of the RF and GB models are quite balanced, with both showing commendable precision and recall across the two classes. The MLP ANN and CNN, which are more complex models, also demonstrate strong overall metrics, closely competing with the high accuracy of SVM, underscoring their potential in handling complex patterns in the data associated with leaks.

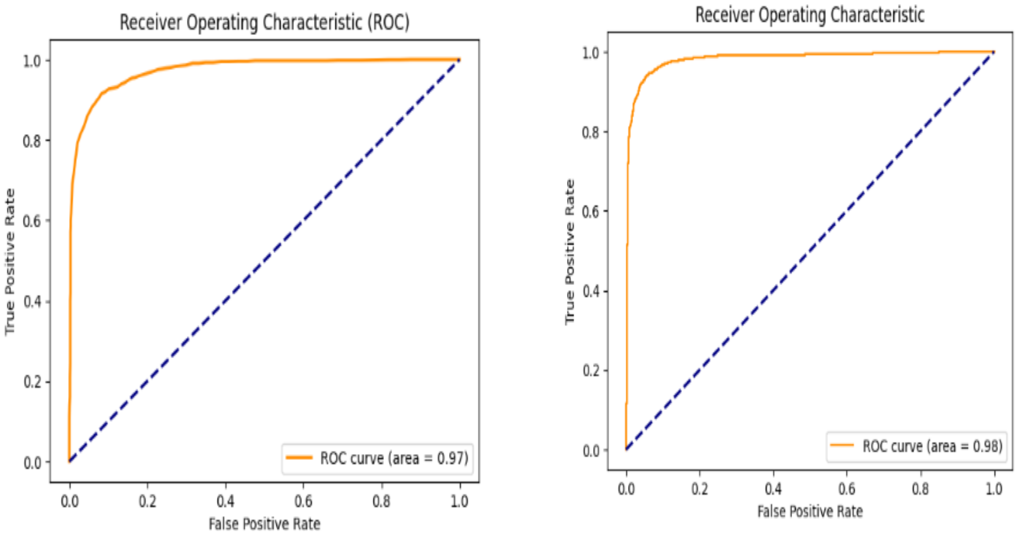
Overall, the SVM model stands out as the most potent model for the detection of oil and gas leaks, with its excellent predictive accuracy and the ability to maintain high performance across various thresholds. This analysis suggests that while more complex models like CNN and MLP ANN perform well, simpler models like SVM can provide robust results with potentially lower computational costs and quicker deployment in practical settings



(a) a

(b) b

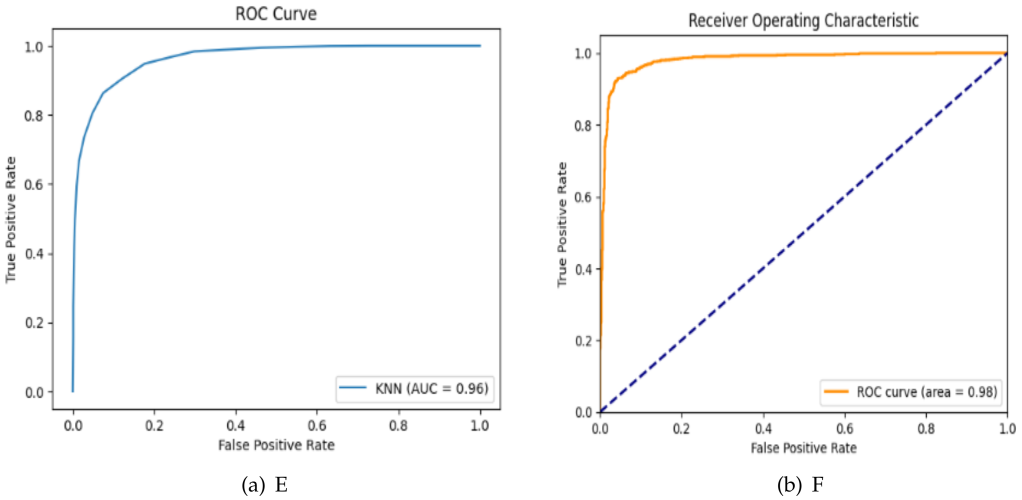
Figure V.13: (a) The ROC curve GB (b) The ROC curve DT



(a) C

(b) D

Figure V.14: (C) The ROC curve RD (D) The ROC curve MLP



(a) E

(b) F

Figure V.15: (E) The ROC curve RF (F) The ROC curve CNN

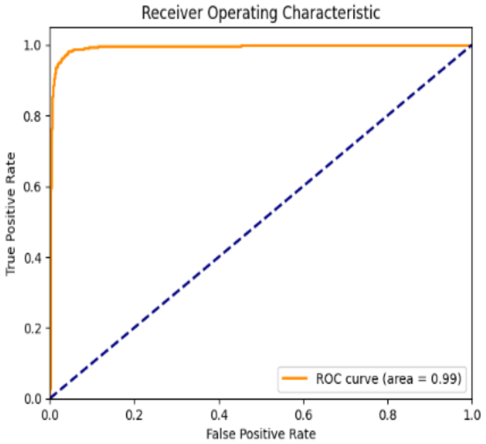


Figure V.16: (G) The ROC curve SVM

05 Conclusions

The analysis revealed that machine learning models, particularly Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Convolutional Neural Networks (CNN), Gradient Boosting (GB), and Decision Trees (DT), can be effectively utilized to predict and manage corrosion levels in pipelines. Among these, certain models stood out in terms of their predictive accuracy, precision, recall, and F-score, the Next chapter explains indicating their potential to be deployed in real-world scenarios to predict pipeline integrity threats

These findings suggest that for critical applications such as corrosion prediction, where precision and recall are crucial, advanced models like SVM and CNN are preferred due to their ability to handle complex and nonlinear data relationships effectively. Each model's selection should consider the specific requirements and characteristics of the dataset, including the balance between classes and the complexity of the data features

CHAPTER VI

CONCLUSION

In conclusion, this thesis the culmination of rigorous research and extensive analyses conducted on the performance of various machine-learning models in detecting oil and gas pipeline corrosion. The classifiers examined include KNN, GB, DT, RF, SVM, ANN, MLP, and CNN, with a detailed assessment provided through a classification report table that lays out the performance metrics for these models.

The SVM, CNN, and MLP models have emerged as the front-runners in this evaluation, demonstrating superior performance across all metrics. The SVM model, in particular, has shown exceptional proficiency, achieving the highest overall accuracy and F1 scores at 0.96 for both classes. This underscores its balanced precision and recall capabilities, making it highly suitable for practical applications in pipeline integrity management.

Conversely, the CNN model also excelled, with an accuracy of 0.94, and demonstrated its effectiveness in feature extraction and classification within complex datasets. These findings suggest that deep learning models like CNN are well-equipped to handle intricate patterns and anomalies in pipeline data.

However, not all models performed equally. The DT models, while promising, underperformed with lower accuracy rates. For MLP, the lowest accuracy was noted at 0.83, indicating potential inefficiencies in network architecture or training processes, such as overfitting or underfitting. Meanwhile, traditional models like KNN, GB, and RF showcased competitive but varied results. Notably, RF distinguished itself with an accuracy of 0.91, attributing its robustness to its ensemble technique that effectively manages overfitting.

Looking ahead, further research is recommended to explore the integration of more advanced deep learning techniques and the inclusion of larger, more diverse datasets from real-world industrial settings. This approach could enhance model robustness and applicability, ensuring more reliable corrosion detection and thereby minimizing the risks associated with pipeline leaks. The continued evolution of machine learning algorithms will also potentially allow for more nuanced and adaptive models that can respond dynamically to the ever-changing conditions of oil and gas pipelines

The implementation of these advanced machine learning models promises significant benefits for the oil and gas industry. By improving the accuracy and reliability of corrosion detection, companies can better anticipate maintenance needs, prevent hazardous spills, and optimize operational efficiency. This proactive approach not only safeguards the industry's assets but also protects the surrounding environment from potential damage.

In conclusion, this project not only highlights the effectiveness of various machine learning models in a critical area of industry need but also sets the stage for future innovations that could further revolutionize the field of pipeline maintenance.

BIBLIOGRAPHY

[1] Karan Aggarwal et al. “Has the future started? The current growth of artificial intelligence, machine learning, and deep learning”. In: *Iraqi Journal for Computer Science and Mathematics* 3.1 (2022), pp. 115–123.

[2] Abdulnaser M Al-Sabaeei et al. “Prediction of oil and gas pipeline failures through machine learning approaches: A systematic review”. In: *Energy Reports* 10 (2023), pp. 1313–1338.

[3] Sumayh S Aljameel et al. “An anomaly detection model for oil and gas pipelines using machine learning”. In: *Computation* 10.8 (2022), p. 138.

[4] : Naveed Aslam. “ARTIFICIAL INTELLIGENCE BASED ALGORITHM FOR PREDICTING PIPELINE LEAK AND CORROSION DETECTION”. In: *United States* 2018 / 0365555 A1.1 (2018), p. 5.

[5] Roland Barthes. *Le Neutre: Cours au Collège de France (1978)*. Seuil, 2023.

[6] Blossom Treesa Bastian et al. “Visual inspection and characterization of external corrosion in pipelines using deep neural network”. In: *NDT & E International* 107 (2019), p. 102134.

[7] Ali Behnood and Emadaldin Mohammadi Golafshani. “Artificial intelligence to model the performance of concrete mixtures and elements: a review”. In: *Archives of Computational Methods in Engineering* 29.4 (2022), pp. 1941–1964.

[8] Kittipong Chomboon et al. “An empirical study of distance metrics for k-nearest neighbor algorithm”. In: *Proceedings of the 3rd international conference on industrial application engineering*. Vol. 2. 2015, p. 4.

[9] Hin Chu et al. “Comparative replication and immune activation profiles of SARS-CoV-2 and SARS-CoV in human lungs: an ex vivo study with implications for the pathogenesis of COVID-19”. In: *Clinical Infectious Diseases* 71.6 (2020), pp. 1400–1409.

[10] Ivan S Cole and DJCS Marney. “The science of pipe corrosion: A review of the literature on the corrosion of ferrous metals in soils”. In: *Corrosion science* 56 (2012), pp. 5–16.

[11] Bingyan Cui and Hao Wang. “Analysis and prediction of pipeline corrosion defects based on data analytics of in-line inspection”. In: *Journal of Infrastructure Preservation and Resilience* 4.1 (2023), p. 14.

[12] Cristobal De Brey et al. “Status and Trends in the Education of Racial and Ethnic Groups 2018. NCES 2019-038.” In: *National Center for Education Statistics* (2019).

[13] Bob Dudley et al. “BP statistical review of world energy 2016”. In: *British Petroleum Statistical Review of World Energy, Bplc. editor, Pureprint Group Limited, UK* (2019).

[14] Olakunle Elijah et al. “A survey on industry 4.0 for the oil and gas industry: Upstream sector”. In: *IEEE Access* 9 (2021), pp. 144438–144468.

[15] Yijie Fu. “Combination of random forests and neural networks in social lending”. In: *Journal of Financial Risk Management* 6.4 (2017), pp. 418–426.

[16] Helmut V Fuchs and Rainer Riehle. “Ten years of experience with leak detection by acoustic signal analysis”. In: *Applied acoustics* 33.1 (1991), pp. 1–19.

[17] Gerhard Geiger, Daniel Vogt, and Ralf Tetzner. “State-of-the-art in leak detection and localization”. In: *Oil Gas European Magazine* 32.4 (2006), p. 193.

[18] MF Ghazali et al. “Comparative study of instantaneous frequency based methods for leak detection in pipeline networks”. In: *Mechanical Systems and Signal Processing* 29 (2012), pp. 187–200.

[19] Phil B Goodwin. “A review of new demand elasticities with special reference to short and long-run effects of price changes”. In: *Journal of transport economics and policy* (1992), pp. 155–169.

[20] Md Imran Hossain. “Support Vector Machine*”. In: *Frankfurt University of Applied Sciences. Frankfurt. Research for Master of Science in High Integrity Systems* (2022).

[21] Mohd Fadly Hisham Ismail et al. “Machine-learning-based classification for pipeline corrosion with monte carlo probabilistic analysis”. In: *Energies* 16.8 (2023), p. 3589.

[22] R Jayawardana and T Sameera Bandaranayake. “Analysis of optimizing neural networks and artificial intelligent models for guidance, control, and navigation systems”. In: *International Research Journal of Modernization in Engineering, Technology and Science* 3.3 (2021), pp. 743–759.

[23] Tyler L Jaynes. “Legal personhood for artificial intelligence: citizenship as the exception to the rule”. In: *AI & SOCIETY* 35.2 (2020), pp. 343–354.

[24] Jerry Joy et al. “Speech emotion recognition using neural network and MLP classifier”. In: *Ijesc* 2020 (2020), pp. 25170–25172.

[25] M Karami. “Review of corrosion role in gas pipeline and some methods for preventing it”. In: (2012).

[26] Claus Kjøller et al. “Novel experimental/numerical approach to evaluate the permeability of cement-caprock systems”. In: *International Journal of Greenhouse Gas Control* 45 (2016), pp. 86–93.

[27] Zhengshan Luo, Xiaoli Hu, and Yang Gao. “Corrosion research of wet natural gathering and transportation pipeline based on SVM”. In: *ICPTT 2013: Trenchless Technology*. 2013, pp. 964–972.

[28] Robert E Melchers. “The effect of corrosion on the structural reliability of steel offshore structures”. In: *Corrosion science* 47.10 (2005), pp. 2391–2410.

[29] Ashim Mishra and Ashwani Soni. “Leakage detection using fibre optics distributed temperature sensing”. In: *6th Pipeline Technology Conference*. Vol. 2011. 2011.

[30] Pal-Stefan Murvay and Ioan Silea. “A survey on gas leak detection and localization techniques”. In: *Journal of Loss Prevention in the Process Industries* 25.6 (2012), pp. 966–973.

[31] Will Nash, Liang Zheng, and Nick Birbilis. “Deep learning corrosion detection with confidence”. In: *npj Materials degradation* 6.1 (2022), p. 26.

[32] Ruben Orihuela, Christopher A McPherson, and Gaylia Jean Harry. “Microglial M1/M2 polarization and metabolic states”. In: *British Journal of pharmacology* 173.4 (2016), pp. 649–665.

[33] RN Parkins. “Stress corrosion cracking”. In: *Uhlig's Corrosion Handbook* (2011), pp. 171–181.

[34] Abiral Poudel et al. “A classification approach for corrosion rating of soil to buried water pipelines: a case study in Budhanilkantha-Maharajganj roadway areas of Nepal”. In: *World Journal of Applied Chemistry* 5.3 (2020), pp. 47–56.

[35] Andika Rachman and RM Chandima Ratnayake. “Corrosion loop development of oil and gas piping system based on machine learning and group technology method”. In: *Journal of Quality in Maintenance Engineering* 26.3 (2019), pp. 349–368.

[36] Lior Rokach and Oded Maimon. “Top-down induction of decision trees classifiers-a survey”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 35.4 (2005), pp. 476–487.

[37] Mohamed El Amine Ben Seghier, Daniel Höche, and Mikhail Zheludkevich. “Prediction of the internal corrosion rate for oil and gas pipeline: Implementation of ensemble learning techniques”. In: *Journal of Natural Gas Science and Engineering* 99 (2022), p. 104425.

[38] Komal Shirsath et al. “DEEP LEARNING FOR AUTOMATED CORROSION DETECTION USING SUPERVISED MACHINE LEARNING & CNN ALGORITHM”. In: ().

[39] Wentao Wu. “Oil and gas pipeline risk assessment model by fuzzy inference systems and artificial neural network”. PhD thesis. Faculty of Graduate Studies and Research, University of Regina, 2015.

[40] Zuhaira Muhammad Zain et al. “Software Defect Prediction Harnessing on Multi 1-Dimensional Convolutional Neural Network Structure.” In: *Computers, Materials & Continua* 71.1 (2022).

[41] Rongjun Zuo. “Biofilms: strategies for metal corrosion inhibition employing microorganisms”. In: *Applied microbiology and biotechnology* 76.6 (2007), pp. 1245–1253.