ISSN: 2583-9888
DOI: https://doi.org/10.51219/JAIMLD/Moshood-Yahaya/06

URFORUM PUBLISHERS
CONNECTING WITH RESEARCH WORLD

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 1 & Iss: 1

Original Research Article

Maintenance of Aging Offshore Assets - A Digital Twin Approach

Moshood Yahaya*, Takao Maruyama, Alex Umagba, Sebastian Obeta

*Applied Artificial Intelligence and Data Analytics, School of Management, Faculty of Management, Law and Social Sciences, University of Bradford, Bradford, West Yorkshire

Citation: Yahaya, M., Maruyama, T., Umagba, A., & Obeta, S. (2023). Maintenance of Aging Offshore Assets - A Digital Twin Approach. *J Artif Intell Mach Learn & Data Sci*, 1(1), 49-66. DOI: https://doi.org/10.51219/JAIMLD/Moshood-Yahaya/06

***Corresponding author:** Moshood Yahaya, Applied Artificial Intelligence and Data Analytics, School of Management, Faculty of Management, Law and Social Sciences, University of Bradford, Bradford, West Yorkshire. Email: moshoodyahaya09@yahoo.com

Received: 19 January, 2023; **Accepted:** 24 March, 2023; **Published:** 31 March, 2023

Copyright: © 2023 Yahaya, M., et al,. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

A B S T R A C T

As the global demand for oil and gas continue to increase the focus of operators on offshore oil and gas production and by extension, put a huge strain on the continued reliability of offshore assets, due to aging challenges, the need to an adequate maintenance of aging assets has become ever so important. Extensive review of existing literature and industrial application of maintenance techniques revealed a huge gap in the application of machine learning predictive maintenance in detecting the failure rate of major offshore assets. Most importantly, pipelines, which not only serve as production tools, but also as separation and transportation purpose have been largely under researched, particularly with relation to corrosion attack. Therefore, this research aimed to analysis the prediction of corrosion rate in offshore pipelines, by answering research questions relating to the adequacy of applying machine learning to predicting corrosion rates, identifying the top features-based Pearson's Correlation Coefficient and their P-Values and identifying the best performing models for corrosion rate prediction, for the whole features, and equally applied on selected features. The base line traditional regression model was compared to more advanced Gradient Boosting Regressor, XGBoost Regressor and AdaBoost Regressor. Based on acquired results, Gradient Boosting Regressor performing best, followed by XGBoost Regressor and AdaBoost Regressor, while the least performance was from the traditional regression model. This result was the same for models trained on all the features and on selected features.

Keywords: Machine Learning, AI, Artificial Intelligence, Digital Twin Technology, Maintenance, Offshore, Oil and Gas, Assets, Asset Management, Remaining useful life

Section 1

Introduction

Background

The continued demand and the resulting obligations to continue produce oil and gas, has brought about increased offshore activities (Chen et al. 2014); Chukwunonso (2015). Hence, challenges emanating from asset ageing (due to the continued usage) presents significant issues to the offshore sector of the petroleum industry. In effect, significant number of offshore facilities are veering towards (or have surpassed) their nominal design life (Clausard 2006b; Ersdal et al. 2011). With the continued global reliance on oil and gas, numerous offshore assets are expected to be utilised beyond their design lives in the coming years. Therefore, asset life planning have become progressively more crucial, encouraging the need for resource efficient solutions that eases burdens on all stakeholders (Hudson 2010).

As assets approach their end of design life, significant risks persistently endanger its safety and reliability (Shukla and Karki 2016), as uninterrupted exposure to conditions of stress, as well as environmental effects will result to degradation. Ageing of assets can be profiled with bathtub curve as per Figure 1 below.

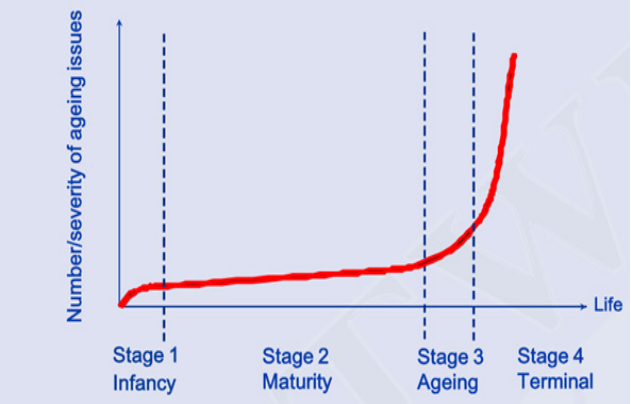


Figure 1: Bathtub Curve Representing Ageing (Wintle 2010)

Presented here-under is a coincise depiction of the different varying phases of asset age in relation to the severity of age related issues, as described by Wintle et al. (2012).

1. Inetegrity challenges noted at asset infancy stage are generally rectified through first comprehension investigations.
2. At maturity stage, assets are generally characterised low and fairly steady fault rate, which requires little or no care.
3. Assets at ageing stage, are anticipated to display increasing rate of failure, which generally wears away the systems design life (calling for a life extension strategy). To evaluate the remaining useful life at this stage, it is imperative that the rate and extent of failure is quantitatively established.
4. At the terminal stage, attention is largely focused on safetu as the asset is severely damaged. Although, asset will be continually managed, as long as reasonably possible.

Ageing

According to Ersdal et al. (2011), ageing describes the numerous circumstances that generally ensues, when structures, equipment and systems gets older. However, according to Gupta and Patel (2010), "Ageing is not about how old your equipment is, it is about what you know about its condition, and how that is changing over time". This definition was corroborated by Horrocks et al. (2010), who defined ageing as the deterioration or impairment of a system, usually but not always related to the time of usage.

While the service age of an asset is a major element of its output efficiency, it does not always decide the inadequacy of the asset. That a system is old does not infer that its efficiency has declined or that the asset is broken. Fundamentally, the level of aging of an asset predominantly centres on both the service conditions and the material sensitivity to those conditions (Novak and Podest 1987). These conditions are classified in respect to their technical and non-technical or external characteristics (Ersdal et al. 2011), as shown in the Table 1 below:

Table 1: Asset Ageing Related Issues (Ersdal et al. 2011)

Technical aspects:		Non-technical aspects:	
Time-dependent: - Fatigue - Corrosion - Clog - Hydrogen cracking - Material degradation - Hardening - Metal loss - Creep - Wear, etc.	Accumulated over time: - Overload - Damage from collision or impact - Geotechnical conditions of - Subsidence and Scour	Obsolescence: - New Technology - New requirements - New technology in conjunction with old technology - New requirements	Organizational issues: - Leaner organizations - Marginal tail production - Personnel with experience retire, a need for transfer of experience to new and younger staff

The technical aspects can be time-dependent and in other cases, ageing accumulated over time. The non-technical attributes can be in form of obsolescence of the equipment, or as a result of organizational challenes. This research work will focus on the time-dependent, technical aspect of aging.

Problem Statement

The vast majority of assets in the offshore oil and gas industry exceeding their design life of between 20 -25 years (Clausard 2006a; Animah and Shafiee 2017). This trend is anticipated to increase with the continued rise in the unrelenting desire towards oil and gas exploration and production (thereby increasing offshore activities) (Chukwunonso 2015). Additionally, the decrease in platform decommissioning and installations of new offshore structures, due to the enormous financial implication, has necessitated the continued use of assets beyond the service life (Chukwunonso 2015).

On a Global scale, about 30% of an estimated total of 6700 platforms have been operated for over 20 years (Patterson 2013), a greater part of this number have already surpassed their original design life expectancy. Also, majority of the 5000 fixed offshore production platforms, and about 40% of the world's mobile drilling platforms have operated well above their planned service life (Paik and Melchers 2008). In the United Kingdom, about half of the offshore have functioned further than their service life (Animah and Shafiee 2017).

Ageing of offshore oil and gas assets have adverse personnel, environmental, economic, and operational effect on the industry. According to the European Union Major Accident Reporting System (MARS), for a 26 years period from 1980, a total of 96 major accident was reported due to ageing (Gupta and Patel 2010). Additionally, in the EU hazardous industry, roughly 60% of major hazard loss of containment incidents centred on asset integrity, 50% of which are ageing related (Duncan 2012) and (Chukwunonso 2015). Furthermore, ageing plant was reported to be responsible for up to 28% of loss of containment incidences in Europe (Candreva and Houari 2013).

The financial impact of failures because of aging assets can be equally far reaching. Due to the dramatic increase in the decommissioning of aging offshore assets, and about an excess of 600 projects predicted for disposal globally, by the end of the year 2021, with 2,000 more offshore projects expected to be decommissioned between 2021 and 2040 (*IHS Markit* 2016), resulting in significant rise in expenditures. According to (*World Oil* 2016), spending on decommissioning projects is expected to increase from approximately $2.4 billion in 2015, to $13 billion-per-year by 2040, or a whooping increase of 540%.

Given the current outlook of the global oil and gas industry and global dependence of oil and as a unique source of energy, the oil and gas industry require a drastic implementation of an improved maintenance structure for the management of aging offshore assets. Therefore, the aims and objective, as well the proposed methodology of this research, which aims to deploy digital twin technology towards solving prevalent aging challenges in the offshore, is as presented below.

Rationale for the Research

As stated above, continuous dependent on oil and gas as the premium and most viable source of crude oil and the continue reliant and focus of operators on the offshore, has placed a huge strain on the usage of offshore assets. While organisations

continue to avoid decommissioning due to the associated costs, the impact of damages from aging related occurrences has increased the need for a deliberate approach to maintaining assets beyond the design life. Therefore, the motivation for carrying out this study are as highlighted below

1. Offshore assets are getting used beyond services life, due to continued demand and global dependency on crude oil and the cheapest source of energy.
2. Operators always try to avoid asset decommissioning due to its associated costs, this means that assets have to be used beyond the design life
3. While offshore aging assets have received vast attentions recently, only few researchers have focused on key failure modes like corrosion, with majority of focus on condition monitoring
4. The advent of digital twin and artificial intelligence has necessitated the need for an intelligent and smart approach to maintaining assets.

Aims

This aims of this research is to evaluate the evaluate the application of digital twin technology and artificial intelligence on the maintenance of aging offshore assets and develop a machine learning approach to maintaining them accordingly. To achieve this, detailed analysis will be carried out using different machine learning models and comparing their performance on the utilised dataset. Therefore, the objectives of this research are as follows:

1. To extensively explore current methodologies used in the managing aging and life extension of assets in the offshore. Additionally, limitations in the current methods will be identified and a suitable plan of corrective actions will be prescribed.
2. To develop a machine learning algorithm for predicting rate of failure of an asset, with specific focus on the prediction of corrosion rate in offshore pipelines
3. To evaluate the relevant machine learning models used and determine the best performing models.

Research Scope

With the vast nature of the offshore oil and gas industry and variety of assets in use, as well as abundance of machine learning approaches to solving aging challenges, solving aging problems, with an open scope is almost impossible in an MSc dissertation. Therefore, the scope of this research is limited to the prediction of corrosion rate in offshore pipelines.

Research Questions

Traditionally, maintenance of assets in the offshore is mainly knowledge based. This current approach has proven to be ineffective with the amount accidents, financial and human losses attributed to aging related asset failure. This has raised more questions than answers, ones that this research aims to address:

1. What is the effectiveness of the current maintenance methods and what are the associated perils?
2. What are the key conditions and contributing factors (Features) towards corrosion in the offshore environment?
3. Which machine learning model is most effective in predicting corrosion rates in offshore pipelines?
4. Can digital twin and machine learning be deployed towards improving the current standards in asset maintenance?

Section 2

Literature Review

Maintenance and Asset Integrity Management

Over time, maintenance has been an ever-evolving concept, with varying interpretation of its ideas. In the past, maintenance practices refers only to actions associated with equipment repair after breakdown (Onawoga and Akinyemi 2010). However, a more modern view have majority of researchers and authors in agreement, to define maintenance as a "set of activities required to keep physical assets in the desired operating condition or to restore them to this condition" (Pintelon and Parodi-Herz 2008). Maintenance management on the other hand, refers to the integration of all technical, administrative and supervisory activities, intended to monitor, control and retain an item, machine or process in, or restore it to a state, in which it can perform a required function (Achilla 2015).

While asset maintenance management itself is an old and mature discipline, elements like competition, productivity, customers, and technology has forced a continuous evolution of the practice. Evidently, increasing international industrial demands continues to threaten business survival, requiring that industries sustain full productive capacities while minimizing the required capital investment. From the maintenance perspective, this means exploiting asset reliability, by extending each individual component's life (ETI et al. 2006).

Over time industrial maintenance and asset management has evolved from the ancient regressive (delayed or no reaction) approach, through to the modern strategic method. This was best illustrated in Du Pont Corporation's study on the effectiveness of the maintenance operations, which identified the characteristics of the different evolutional stages of maintenance operations, as shown below.

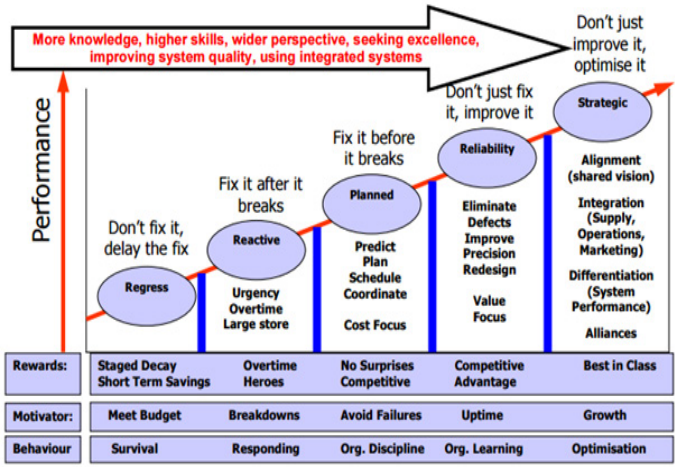


Figure 2: DuPont Stable Domain Model (Ledet 2016).

Due to increasing demands for improved asset management practices (Misra 2008), quest to eliminate/reduce asset downtime (Ledet 2016), demands for higher production performance (Blann 1999), the need for employee involvement (Thomas 2000) and continued demand for continuous improvement (Blann 1997), the field of asset management and maintenance has experienced immense growth and evolution in recent times. From the regressive (little or no reaction to failure) approach

(Ledet 2016), to the reactive methods where assets are only fixed after breakdown is experienced (Misra 2008; Onawoga and Akinyemi 2010). Further improvements brought about the planned maintenance approach where equipment is consistently examined and repaired to prevent breakdowns from occurring (Ledet 2016). Planned maintenance is typically divided into the Scheduled (Preventive Maintenance) and Condition-based maintenance (Predictive Maintenance).

Demands for higher production performance led to the movement to the more advanced Reliability Domain. Typically, Reliability centred maintenance (RCM) and total productive maintenance (TPM) are the main types of maintenance at the reliability domain (characterised by proactive maintenance methods (Blann 1999)) (ETI et al. 2006). According to (ETI et al. 2006), and (Sullivan et al. 2010), RCM can be defined as "a systematic proactive approach, used to define the maintenance needs of an asset in its functional context". With RCM, the system purposes, their failure mechanisms, and the criticality are cautiously investigated in order to deliver a decent footing for the maintenance program (Milje 2011).

The TPM is an asset management methodology that permits incessant and swift improvement of production processes, through the use of employee involvement and employee empowerment and closed-loop measurement of result (Thomas 2000). It is a profit focused, zero failure approach towards maintenance management, aimed at applying planned maintenance strategies in identifying and repairing equipment before deterioration arises (Achilla 2015).

The final domain is classified as a world class maintenance phase. The behavior required for world class performance domain is believed to be "Organizational Learning" as shown in the DuPont domain model. Although currently classified as the best maintenance methodologies (Blann 1997), TPM and/or RCM are not the final frontier, as even better performing maintenance approaches will continue to come up, through continuous improvement (Blann 1999). The concept of continuous improvement is seen as a journey, not the endpoint. To this end, researchers have continued to focus of methods of further improving the current "best practices" (through the incorporation of Digital Technologies) in order continuously minimise the challenges faced by operators as offshore assets approach the end of their useful lives.

Maintenance of Ageing Assets

Age creates a wealth of challenges for offshore operators. As assets ages, operators require a wide-ranging maintenance plan, to prolong the lifespan and efficiency of the asset, as well as to avert any failure that might lead to a major incident (halting production and putting workers at risk of serious injury). Significantly, as often addressed in books, articles and journals, an effective maintenance campaign involves putting in place a fully planned, coordinated and implemented maintenance strategy (ETI et al. 2006; Achilla 2015). It is usually not a case of sudden, unorganised actions.

Currently, very few studies have focused on maintenance decision-making beyond the original design life (Saxena et al. 2008; Tiddens et al. 2015), despite clear evidence alertness on the challenges inherent in ageing and life extension in recent research works (Saxena et al. 2008). Unfortunately, methods and quality of implementation of asset life extension frameworks remains a huge challenge facing the offshore oil and gas industry (Chukwunonso 2015; Animah and Shafiee 2018).

Several studies have been initiated on this topic. Incidentally, bulk of this has focused on the physical conditions and structural aspects of asset ageing management. Particularly, (Baker and Descamps 1999; Hörnlund et al. 2008) and (Hörnlund et al. 2011) undertook studies on the material related risk in ageing of offshore assets. Additionally, several other researchers focussed on structural integrity of offshore structures and installations (Ersdal 2005), (Sørensen and Ersdal 2008) , (Ersdal et al. 2008) jlC and (Galbraith and Sharp 2007b), while (Hart et al. 2009) worked extensively on the impact of ageing safety critical element in offshore. Furthermore, (Galbraith and Sharp 2007a) carried out a study on recommendations for design life extension regulations for ageing offshore production facilities.

Unfortunately, virtually all the aforementioned researchers focused on investigating the physical condition and structural integrity of assets, using inspection techniques (condition monitoring) and engineering analysis methods (Ersdal 2005; Galbraith and Sharp 2007b; Sørensen and Ersdal 2008; Hart et al. 2009; Ersdal et al. 2011). Unquestionably, developing a maintenance plan for ageing assets, demands a careful review of the design (of the asset) and projections of impending damage mechanisms, (including fatigue and corrosion) (Paik and Melchers 2008). However, according to (Hudson 2010), integrity management of ageing assets depends not only on the physical condition of the system, but also on the procedures adopted in dealing with the growing risk of failure. It is pretty much essential to strike an accurate equilibrium between asset management techniques and inspection, for improved management of ageing assets. This growth has led to the paradigm shift in asset management and monitoring, toward the application of digital technology (Animah and Shafiee 2017; Animah and Shafiee 2018; Errandonea et al. 2020), which aims to predict asset integrity and forecast damage from sensor acquired data (Errandonea et al. 2020).

Digital technology not only present an interesting prospect, but also a novel opportunity of developing a strategic (world class) solution to maintenance of offshore ageing asset (Bhowmik 2019). To push the boundaries of innovation and continuous improvement, digital technology providers and researchers are working on opening up new possibilities in the field of asset management through the application of Artificial Intelligence, Machine Learning, Advanced Statistics, Internet of Things and Digital Twins (Errandonea et al. 2020).

Digital Approach to Asset Life Extension

Developments in computing and information technology have greatly increased the potential implementation of digitalisation solutions to prevalent industrial challenges. Consequently, knowledge extraction from data science has attracted a lot of interests from many fields of research in recent years (Provost and Fawcett 2013). Artificial Intelligence, Machine Learning, Advanced Statistics, Internet of Things and Digital Twins and related strategies for optimal data management are good examples of such interests.

In recent times, digitalisation has presented a fundamental revolution in our everyday life and the offshore oil and gas industry is no exception. Digitisation of maintenance practices in the offshore environment has become more feasible with the astronomical increase in amount of accessible data (relating to aging assets) and the availability of fast paced potential (of computing systems) in providing optimum solutions in relation the manual analytical methods. According to (Renzi 2019) the global oil and gas industry has collected petabytes of data, much of which are left unused and unrefined.

Digital Twin technology, which is a digital model of a physical entity (Liu et al. 2012; El Saddik 2018; White et al. 2021), utilises the vast available data to integrate internet of things, machine learning, artificial intelligence and data analytics, to create virtual digital simulation of a physical entity, while updating and changing with their physical counterparts over time and continuous operation (Luo et al. 2019; White et al. 2021), to predict asset performance and conditions (Qi and Tao 2018).

Continuous improvement in standards has ensured that maintenance and asset management has evolved towards the strategic domain as shown in figure 2, with operators and researchers developing futuristic view to maintenance (Heng et al. 2009). This involves real time, time to failure tracking using sensors (Bhowmik 2019) and remaining useful life prediction, before failure is experienced (Qian et al. 2017)

Majority of early application of digital technology in maintaining offshore assets involves using specific artificial intelligence and machine learning (Tan et al. 2019) techniques to predict the reliability of an asset, through the prediction and estimation of the probability if failure (Sharma et al. 2017). Certain researchers like (Håbrekke et al. 2011; Animah and Shafiee 2017; Oliván 2017; Simm 2019; Stetco et al. 2019) also, extended the trend further by predicting time to failure and useful life of assets.

Similarly, numerous researchers have worked on the application of digital twin technology for prognostic analysis of offshore oil and gas assets. However, majority of the research fell short on two grounds, as some like (Liu et al. 2012) only considered digital twin application for simulating and monitoring asset performance, while others incorporated artificial intelligence techniques for prognostic (assessment of useful life), largely ignoring the diagnostic (fault detection) aspect of asset management and predictive modelling of specific failure modes (Sharma et al. 2017; Ochella et al. 2021). Although, (Altamiranda et al. 2009) developed a digital twin technology for the diagnosis of subsea processing system, the research focused only on fault detection and monitoring, and failed to include the evaluation of the remaining useful life of the asset or prediction of failure rates (prognosis). Additionally, the diagnosis technique by (Altamiranda et al. 2009) did not consider image recognition, which would have presented a more intelligent and accurate fault diagnosis.

In view of the above, while enormous research time and work has been invested in unearthing novel ways of maintaining offshore assets, existing research are limited in their application of digital twin technology and artificial intelligence in offshore asset management. As enumerated above, many attempts to proffering solutions to this challenge have fall short in applying image recognition technologies in diagnostic analysis of asset faults, despite the enormous research outlays in fault and condition monitoring. Additionally, as stated in the paragraph above, researchers have also fallen short in the application of machine learning approaches in evaluating prognosis of aging asset, for failure rate prediction. Strangely, these two concepts are already in wide usage in the aviation industry (Altay et al. 2014). Furthermore, majority of approaches to offshore maintenance are void of specific scope on particular asset types or focus on specific failure mode like corrosion, based on the above paragraph.

Consequently, this research is aimed at exploring the possibility of developing a machine learning approach for prognostic analysis of offshore oil and gas assets. This offshore asset life extension approach will be based on a scope limited to subsea/offshore pipelines processing units and the aim will be achieved by incorporating Artificial Intelligence technology with corrosion prediction datasets obtained from sensor readings from subsea/offshore pipelines. The analysis will incorporate the use of analytic insights in discovering the most important features and the model will be trained using numerous machine learning algorithms. The performance of the algorithms will be validated with relevant performance indices. In line with the identified gaps in this literature review, this project would have incorporated the use of image classification methods in performing diagnostic analysis, with the prognostic analysis. However, scarcity of required datasets will limit the scope of this work to the prognostic approach alone.

Section 3

Research Methodology

This section focuses on the research methodology employed for this research. It proceeds to establish the source of the acquired datasets and clearly define the components of the dataset. Furthermore, it outlines the different approaches employed toward predicting and maintaining aging issues in offshore assets, with specific focus on corrosion related challenges in pipelines. Finally, this chapter outlines the different methods used in validating our prediction models.

Research Strategy

In delivering the research aim, this paper employs the popular CRISP-DM framework, which is presented below.

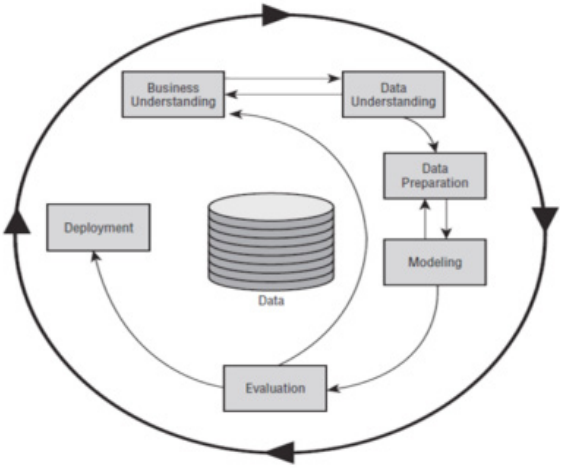


Figure 4: Overview of the Deployed Research Approach(Shafique and Qaiser 2014)

While this research follows the approach presented above, it is worth noting that certain elements of the Crisp-DM framework was not applied to this project. Therefore, the deployment of the result of this research was not carried, given the time constraints associated with the delivery of the project.

The Crisp-DM Model

Crisp-DM is the cross-industry standard and the widely accepted knowledge discovery process for data mining (Huber et al. 2019). The model provides a detailed sequence of activity to be followed for any data mining project. As shown in figure 4 above, the model starts with the requirement to develop a proper understanding of the expected business use case, which creates

a clear outline of questions intended to be answered (Schröer et al. 2021). Additionally, it sets the standard and framework for the understanding of the dataset and adequate preparation of data for modelling. Data preparation is generally regarded as the most important step in a machine learning project (Huber et al. 2019; Schröer et al. 2021). Subsequently, as presented in figure 4, the prepared data is modelled based on the designed business question, and deployed for testing with external data (Huber et al. 2019). For the sake of this research, the deployment aspect will not be discussed, as the model was not concluded to deployment.

Business Understanding

A detailed background of maintenance methods has already been presented in the **Chapter 1** of this project and as such, this section will streamline explanations in line with the specific project scope. As stated in the literature review above, the scope of this project is particularly limited to the evaluation of corrosion in offshore/subsea pipelines. In the quest to present an wholistic approach to monitoring and maintaining corrosion in pipeline, the project explored the prognostic elements of assess maintenance, as presented by (Butler 2012; Malekloo et al. 2021), a pictorial snapshot of which is presented below.

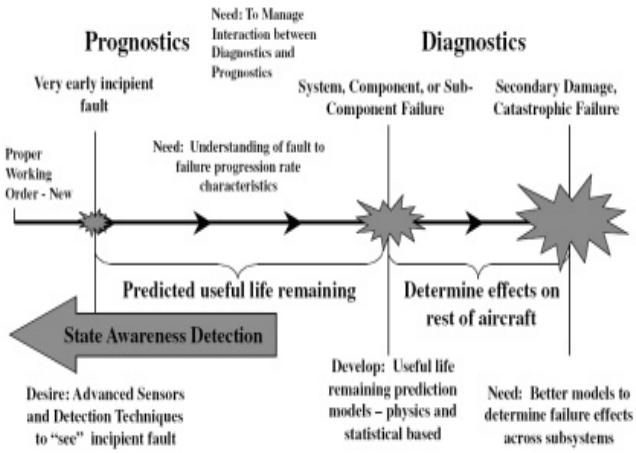


Figure 5: Overview of Prognostic and Diagnostic Approach to Asset Maintenance (Butler 2012).

While the Diagnostics analysis refers to the investigation of an asset's condition, possible problems or anomalies and or the exact operational situation of the asset, prognostics relates to the analysis or prediction the future, with reference to available pertinent data (Lee et al. 2014). Simply put, diagnostics analysis refers to the process finding and identifying the failure of an asset or otherwise, while prognostics refers to the process of predicting or estimating an asset's failure rate or the remaining useful life. A simplified graphical representation of the relationship between prognosis and diagnosis is presented below

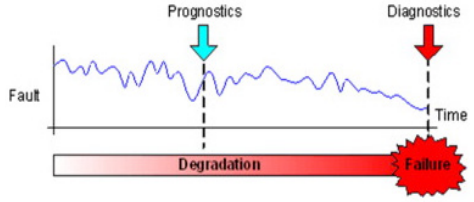


Figure 6: Relationship Between Diagnosis and Prognosis(Lee et al. 2014)

In view of the above and as already stated in the aims and objectives, as well as the literature review, this research will adopt a prognostic approach to predicting the failure of offshore assts through the training of numerous regression models for the prediction of corrosion rate in a pipeline.

Data Understanding

Given that the project explored the prognostic method of evaluating corrosion threats to many offshore pipelines. Datasets was sourced to answer pertinent research questions and hypothesis raised.

Data Limitation

Difficulties in obtaining datasets is a major limitation in this research. For research that promised to break uncatered grounds in the maintenance of aging assets in the offshore, it was particularly noted that oil and gas practitioners and industry players are reluctant in releasing datasets, even with offers for non-disclosure and service level agreements. The secretive approach to dataset can be put down to the sensitive nature of the industry. Nonetheless, datasets were obtained for the purpose of the research, although in far lesser quality and quantity than desired.

The research initially intended to combine diagnostic elements of monitoring the condition and operating situation of assets, to the prognostic approach eventually employed. This would have created a wholesome solution that identifies the exact operating condition of an asset and in turn, predict the remaining useful life of that asset. Difficulties in obtaining suitable datasets made this an impossible task.

Description of Data

Similar to the issue noted above, obtaining dataset for the prediction of corrosion rate of offshore pipeline proved difficult and was limited to exploration of past research for available data. Therefore, dataset used for the predictive analysis was obtained from (Chou et al. 2017). The data (presented in **Appendix 1**) refers to sensor readings from stainless steel, under different seawater environmental conditions. The data is made up of 6 columns (detailed below), for 46 sample spaces (rows).

Table 4: Environmental Factors Forming Part of the Corrosion Rate Dataset

Temperature (T)	**°C**
Dissolved oxygen (DO)	mg L^{-1}
Salinity (Sal)	Ppt
Solution pH (pH)	pH
Oxidation–Reduction Potential (ORP)	mV
Corrosion rate (Rate)	μAcm^{-2}

What the dataset lacked in quantity, it made it up with quality, as it is without any missing data, repeated row, relatively normally distributed and with only little cases of outliers as shown below.

Data Preparation/Pre-processing

According to (Zhang et al. 2003; Kwak and Kim 2017), data preparation of an analysis framework an essential stage in data analysis. According to (Coussement et al. 2017), data preparation, also known as the pre-processing stage in arguably the most important stage of an analytics job as it depends the ultimate success of a prediction.

Prior to data modelling, it is necessary to analyse the data quality and treat relevant mis normal where required. The obtained data is relatively clean, with little outliers and relatively

normal distribution as shown above. Additionally, there was no missing values. Therefore, typical data cleaning steps that involves the treatment of missing values and outliers was not required. However, due to the slight element of lack of normal distribution in some cases and the need to carry out feature selection, the type of data pre-processing technique used is presented below.

Feature Selection

According to (Coussement et al. 2017) data reduction is the processes of reducing the dimensionality of a dataset by selecting the relevant and most important features to the required prediction. This process drives the concept of feature selection. For the sake of this research, an extensive the Pearson's Correlation Coefficient and P-Value was used, for the purpose of detecting and selecting the most relevant features in the dataset.

Pearson's Correlation Coefficient

The Pearson's Correlation Coefficient is used to determine the strength of the linear correlation between a dependent and its corresponding independent variable (Mukaka 2012). The coefficient typically assigns a value between -1 and 1, depending on the type and level of existing correlation. Positive value of 1 signifies an absolute positive linear correlation, while negative 1 signifies a total negative linear correlation. Lack of correlation between two variables returns a value that is tending to zero.

P-Value

P-Value is a statistical tool used to evaluate the probability that the relationship existing between two variables is statistically significant (Goodman 2008). Conventional P-Value interpretation stipulates that the p-value of less than 0.001 points towards strong evidence of a significant correlation, while 0.05 refers to moderate proof of significant correlation (Goodman 2008; Halsey et al. 2015). Additionally, 0.1 suggest the existence of a possible week significant correlation, while 0.1 shows no evidence that the existing correlation is significant.

Data Normalisation

According to (Ferreira et al. 2019), data normalization or scaling is a key stage in data preparation, which is required to correct normalisation issues in numeric datasets. Datasets with normal distribution is hard to come by in real datasets, and the purpose of data normalisation is predetermined technique used to convert features, to ensure that all scaled attributes have the same level of impact. The Min-Max Scaler is used for this research.

Min-Max Scaler

This is a straightforward process that particularly fits features in a predetermined boundary with a predefined boundary (Patro and Sahu 2015). The equation for the Min-Max Scaler/ Normalisation is as presented below.

$$A'\left(\frac{A-\min value\ of\ A}{max\ value\ of\ A-min\ value\ of\ A}\right)*(D-C)+C \ldots\ldots\ldots \text{Equation 1}$$

A' contains Min-Max Normalized data one

If predefined boundary is[C,D]

If A is the range of original data

B is the mapped one data

Data Minning

Following the data preparation and cleaning stage, the cleaned datasets where analysed, the datasets was trained with different statistical approaches peculiar to the available datasets, whilst considering the application and use case.

The analysis of corrosion rate in offshore pipeline involved an extensive analysis of the acquired dataset. This includes a vast exploratory data and statistical analysis, correlation analysis for feature engineering and selection as well data modelling.

Exploratory Data Analysis (EDA)

According to (Komorowski et al. 2016), exploratory data analysis is a very critical step in the data mining process, following the acquisition and pre-processing of data (sometimes as part of the pre-processing phase). The EDA helps an analyst to get informed and actionable insights on a given dataset and could set the tone for a high performing model. The aim of the EDA is to examine data distribution, detect outliers and anomalies in the dataset and to determine the specific testing of the model hypothesis (Morgenthaler 2009).

In this project, data was extensively explored with numerous technics targeted at unearthing relevant anomalies in the dataset. Additionally, bivariate and univariate was conducted to explore the inherent relationship between the dependent variable (corrosion rate) and the independent variables.

Data Modelling

Following the selection of the features with the best correlations with the independent variable, and the exploratory data analysis, the model was built for the prediction of corrosion rate in offshore/subsea pipelines. For the purpose of this report, a regression analysis was carried out.

Regression Analysis

Regression refers to the mode of analysis that is typically deployed to determine the relationship existing between two or more variables, with intrinsic dependent and independent relation (Uyanık and Güler 2013). Regression analysis helps to provide answers ranging from the availability and extent of the correlation between the variables and the prospect of producing satisfactory predictions of the dependent variable (Uyanık and Güler 2013). According to (Tabachnick et al. 2007), there are two types of regressions, namely the univariate and the multivariate regression, otherwise known as the simple and the multiple regressions respectively. While the simple regression typifies the existence of a single dependent variable (otherwise known as the label) and a single independent variable. The multiple regression on the other hand, refers to a case of more than one independent variable, as shown below:

$$y = \beta x + \varepsilon \ldots\ldots\ldots\ldots \text{Equation 2}$$

$$y = \beta_0 + \beta_1 x_1 + \cdots \ldots + \beta_n x_n + \varepsilon \ldots\ldots\ldots\ldots \text{Equation 3}$$

y = Dependent variable
x = Independent
β = Parameter
ε = Error

As shown above, regression presents a generalisation of the classification problem, by outputting a "continuous value", as against predictable set as derived in binary and multiple classification problems (Awad and Khanna 2015). Invariably, the regression model is applicable for use cases where the forecasted or predicted output is in continuous form.

In implementing the regression model, different regression algorithms like the Simple Linear Regression, Multiple Linear

Regression, Support Vector Regression, as well as the XGBoost and Adaboost Regressors were utilised and the respective level of performances compared, to determine the best performing and optimum algorithm.

The simple and multiple linear regressions are as described above, while a simple description of the Support Vector Regression, as well as the XGBoost, Gradient Boost and Adaboost Regressors is presented below.

Simple Linear Regression

According to (Kavitha et al. 2016), the simple linear regression model is a machine learning model that is based on a single independent variable. The relationship between variables can be exemplified by the extent of the correlation between the variables. A typical case of simple linear regression is a presented below.

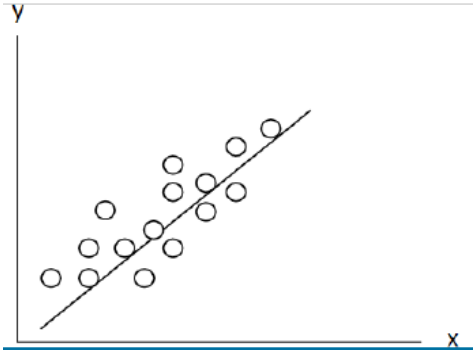
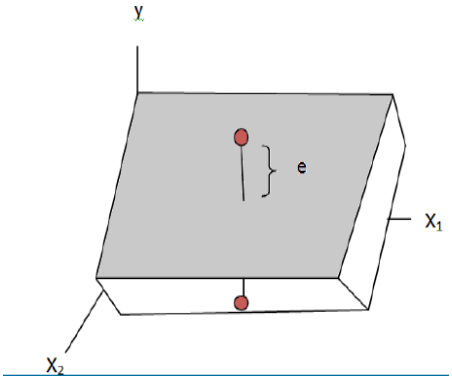


Figure 7: Simple Linear Regression (Kavitha et al. 2016).

Multiple Linear Regression

Multiple linear regression, also known as the multivariate regression, refers to the prediction module that involves the prediction of more than on independent variable, as shown in figure 9 below



XGBoost Regression

XGBoost is a gradient boosting machine learning library that uses recursive dual splitting method to choose the best split at every step, in order to arrive at the best predictive model (Zhang et al. 2020). According to (Zhang et al. 2020), due to its tree nature, the XGBoost is sensitive to outliers and is robust against overfitting, making it the preferred model for most data scientists.

Gradient Boost Regression

The Gradient Boost Regression model is one whose boosting iterations is centred on the efficient gradient descent (Wang and Mamo 2020). The boosting of regression trees potentially produces a strong and explainable technique for regression analysis.

Adaboost Regression

In its general form Adaboost combines weak learning algorithms (that is "boosted"), to achieve enhanced ensemble accuracy and produce a much improved and better classifier (Kummer and Najjaran 2014). Consequently, when applied to regression (otherwise known as Adaboost Regression), Adaboost permits for enhancements of the regressors and whilst adapting to specific case problems.

Model Evaluation and Performance Metric

In evaluating the performance of the respective algorithms stated above, the Coefficient of Determination (R-Square) and Root Mean Square Error were used to evaluate the performance of the models

Coefficient of Determination (R^2)

Coefficient of Determination (R2) is a performance metric which is typically described as the variance explained by a regression model. The Coefficient of Determination (R2) is expressed by the mathematical formular presented below. Generally, a Coefficient of Determination (R2) score of value closer to 1 typifies a very good analysis (Cameron and Windmeijer 1996; Miles 2014);

$$R^2 = 1 - \frac{RSS}{TSS} \ldots\ldots\ldots\ldots \textit{Equation 4}$$

R^2 = coefficient of determination
RSS = sum of squares of residuals
TSS = total sum of squares

Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is defined as the difference between forecasted values and observed values (Al-Omari 2015). It is usually a non-zero value, and zero indicates a perfect score, and a value closer to zero indicates better model performance (Chai and Draxler 2014).

$$RMSE = \sqrt{\frac{\sum_i^N (n_i + m_i)^2}{N}} \ldots\ldots \textit{Equation 5}$$

i = variable i

N = number of non-missing data points

n_i = actual observations time series

m_i = estimated time series

Mean Absolute Error (MAE)

This is one of the popular methods of relating predictions with eventual outcomes. It is a measure of error between matching observations conveying the identical phenomenon. It is regarded as the simplest measure of model accuracy and it is expressed as the average of the absolute error, as shown in the equation below (Kotz et al. 2005). It is a measure of how far away from the actual value, a predicted value is.

$$MAE = \frac{\sum_{i=1}^{n} |yi - xi|}{n} \ldots\ldots\ldots\ldots \textit{Equation 6}$$

MAE = Mean Absolute Error

y_i = Prediction

x_i = True Value

n = Total number of data points

Software Tools

Python Programming Language and Jupyter Notebook

Python is a popular multipurpose programming language used for web development and machine learning and artificial intelligence application. The use case for this project is for the analysis and modelling of outlined machine learning and artificial intelligence processes. Jupyter notebook on the other hand is a code writing application for the creation of machine learning models, in python programming language.

Research Ethics

The dataset for this research was obtained from open-source sources, from an openly available research publication, in the case of the numeric data. Therefore, there is no ethical complication from this research as is void of potential damage to any individual or organisation.

Section 4

Analysis and Discussion

Having extensively presented the relevant research methodology and approaches employed in this paper (in the last chapter), this chapter presents the results and findings of the research. This chapter will be presented in two folds. The first section (Descriptive Analysis) simply presents the results of our analysis, while the second part (Discussion and Analysis) fully discusses the findings and presents actionable insights. Finally, this section summarises the research contributions, in view of the questions raised in the chapter 1 of this report.

Descriptive Analysis

This section is used to present the findings and results, before eventual discussion, under the "Discussion and Analysis" section. The approach to the result presentation focuses on the diagnostic analysis first, and the prognostic analysis. As presented in the research methodology, this project equally explored the prognostic method of determining the corrosion rate of the offshore petroleum pipeline. This was done using various regression methods and the result is presented below

Data Pre-processing

For a dataset that is only made up of 46 rows, data pre-processing was a bit straightforward with outliers (Box plot presented in **Appendix 3**) in a few cases, while there was no missing values. The frequency distribution charts are as presented in the figure below.

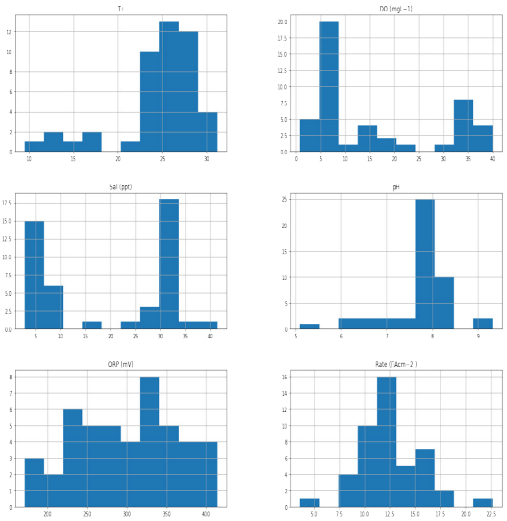


Figure 8: Check for Feature Distribution.

AS shown above, the features were normally distributed in most cases, except in just a few cases. With the salinity and Dissolved Oxygen, largely non-normal, while the PH and temperature are with outliers. Furthermore, the dataset is fully numeric, with no categorical data, eliminating the need for data encoding, both cases are shown in the diagram below. This made for a easy pre-processing exercise and the project immediately proceeded to the exploratory data analysis phase.

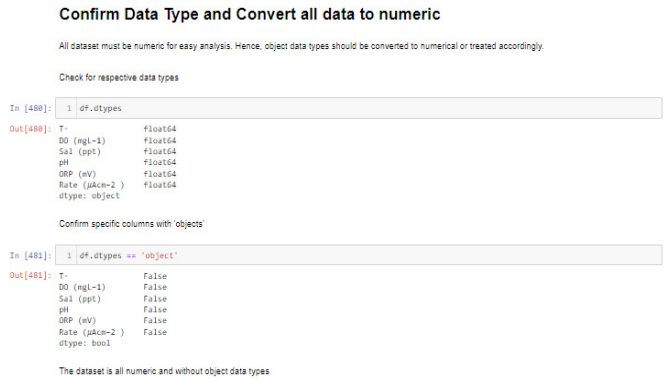


Figure 9: Check for Data Types.

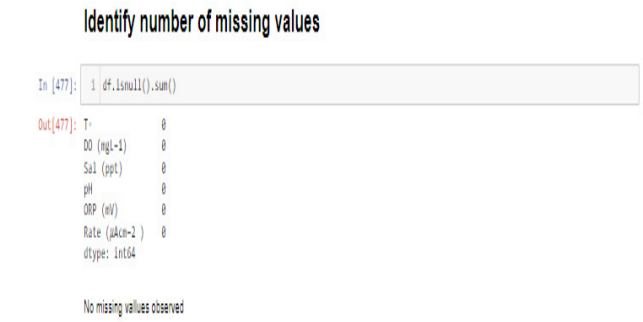


Figure 10: Check for Null Values.

Exploratory Data Analysis

Feature Analysis

Due to the limited amount of sample, the dataset presents a rare case of little statistics to be drawn. However, deeper view of the result reveal some exiting details. The statistical analysis of the dataset shows a relatively normally distributed sample, with the mean and median close in value, except for the "Dissolved Oxygen" column. Additionally, the data samples are more dense in the middle area, a signs of normally distributed samples.

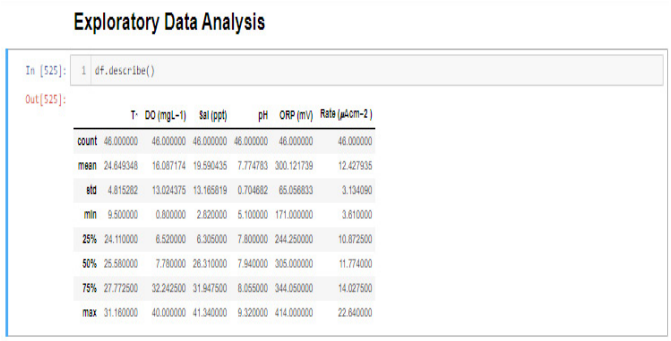


Figure 11: Statistical Analysis.

Virtually all of the datasets have unique values, which can make it hard to examine deeper relationships between features and the target label. Similarly, the label is largely limited in range, hence almost all values are uniquely represented. This can be seen below.

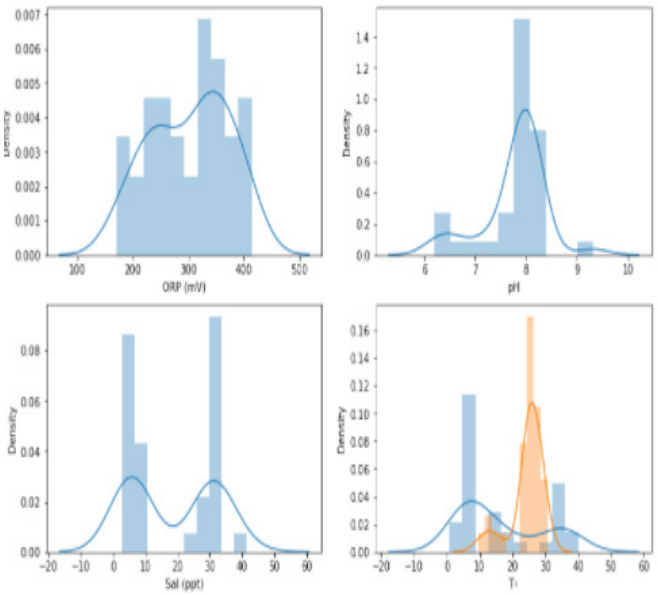


Figure 12: Density Graphs.

Figure 13: Univariate Analysis of Datasets.

Label Analysis

Similar to the case of feature analysis, the extent of possible analysis of the label is limited due the small data size. Additionally, the label is made up of almost unique datasets, limiting the possible insight to be derived from the data. To extend the analysis, a cumulative frequency chart with a normally distributed shape achieved, as shown below.

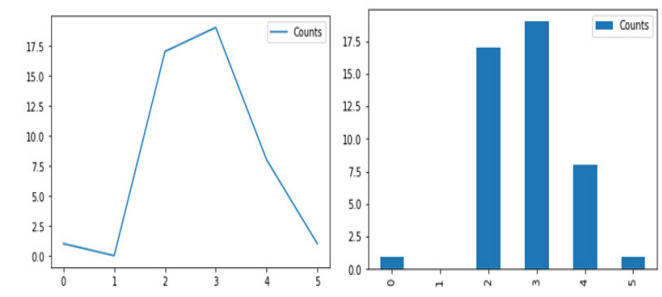


Figure 14: Frequency Distribution Line and Bar Charts.

Feature Selection

The project utilised the Pearson's Correlation Coefficient and the P-Value to establish the relationship between features and the target label, in order to identify the features that best correlates with the corrosion rate.

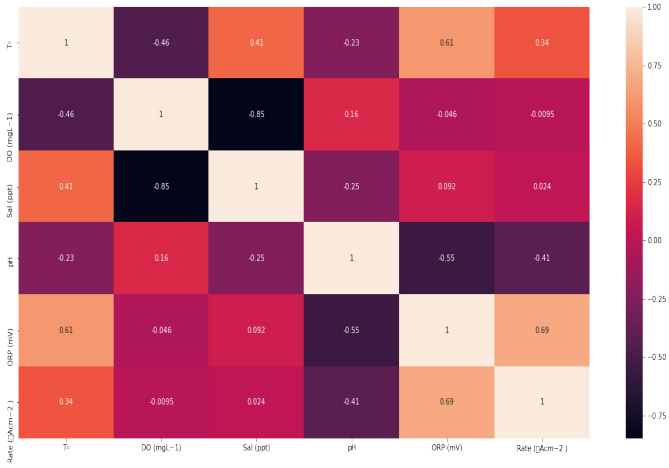


Figure 15: Pearson's Correlation Coefficient.

The correlation above shows that the corrosion rate of a pipeline positively correlates with the Temperature, Salinity and Oxidation–Reduction Potential (ORP). While the solution PH and Dissolved Oxygen, negatively affects the corrosion rate. Although, only the Temperature, PH (negatively) and Oxidation–Reduction Potential (ORP) shows evidence of strong correlation

Similarly, with the respect to the P-Value only the PH, Temperature and Oxidation–Reduction Potential (ORP) displays evidence of possibly significant correlation with the corrosion rate. Salinity, Oxidation–Reduction Potential (ORP) and Dissolved Oxygen all have P-Values that suggests that the possible correlation is insignificant. This was considered in the modelling of the data.

```
In [559]: for Label_Corr in ['T◦', 'DO (mgL-1)', 'Sal (ppt)', 'pH', 'ORP (mV)', 'Rate (μAcm-2 )']:
              pearson_coef, p_value = stats.pearsonr(df['Rate (μAcm-2 )'], df[Label_Corr])
              print("The Pearson Correlation Coefficient for", Label_Corr,"is", pearson_coef, " with a P-value of P = ", p_value)

The Pearson Correlation Coefficient for T◦ is 0.3395253187098665  with a P-value of P =  0.020974525218799743
The Pearson Correlation Coefficient for DO (mgL-1) is -0.008465244487141675  with a P-value of P =  0.950219578984178
The Pearson Correlation Coefficient for Sal (ppt) is 0.024129859964619506  with a P-value of P =  0.8735303458891717
The Pearson Correlation Coefficient for pH is -0.4135068995130606  with a P-value of P =  0.004279958842980986
The Pearson Correlation Coefficient for ORP (mV) is 0.6893532821344478  with a P-value of P =  1.1725629532424 3e-07
The Pearson Correlation Coefficient for Rate (μAcm-2 ) is 1.0  with a P-value of P =  0.0
```

Figure 16: Pearson's Correlation Coefficient and P-Value.

Discussion and Analysis

This section of this chapter presents the result this research for extensive discussion and analysis. As is the trend already in this report, this will focus on the analysis of the result of the prognostic corrosion rate analysis, for extensive discussion. The prognostic analysis for corrosion rate prediction was carried out using a number of algorithms. The discussion of results will follow the same framework as presented by the methodology, for the sake of ease.

Data Pre-processing

The limitation in the volume of the dataset ensured that analysis is carried out with little data pre-processing done. While there was no case of missing values, hence the data was not treated accordingly. Additionally, while outliers were observed in a few of the columns, treatment of outliers was specifically avoided for a couple of reasons. Firstly, the dataset contains only 46 rows, removing or deleting affected rows will further limit the already thin research scope. Additionally, applying

relevant treatment methods like Capping/Flooring, Sigma and the Exponential Smoothing approaches (Tiwari et al. 2007), will alter the accuracy and integrity of the dataset, particularly with the limited sample space.

Feature Selection

As stated in the methodology section, a positive Pearson's Correlation Coefficient tending towards 1 signifies the existence of a strong positive correlation, while a negative value towards one shows a negatively strong correlation. The value tending towards evidences the lack of correlation. Additionally, the P-Value of less than 0.001 shows that there is strong evidence of strong correlation, while the corresponding value less than 0.05 shows moderate evidence of a strong correlation. This forms the baseline for this analysis, that is P-Value above 0.05 is regarded as low evidence of possible strong correlation. The table below shows the scores used to guide the feature selection process.

Table 5: Feature Selection Analysis - Pearson's Correlations and P-Value

Features	P-Value	Pearson's Correlation
Temperature	0.0210	0.3395
PH	0.0043	- 0.4136
Salinity	0.8735	0.0241
Oxidation – Reduction Potential	1.17E-07	0.6894
Dissolved oxygen	0.9502	- 0.0095

In view of the above, both feature selection methods used (the Pearson's Correlation Coefficient and the P-Value), clearly produced a consistent result with both methods identifying the PH, Temperature and Oxidation–Reduction Potential (ORP) as the features with the topmost level of correlation and probability of significant correlation.

Data Modelling

Following the feature selection analysis and the identification of features with the strongest evidence of correlation with the corrosion rate, an approach to data modelling was devised. In training the prediction model, the top features were selected for training, and the respective performance was compared to the result of the simple linear regression (for each feature against the label), the model trained on the whole features and finally, model trained on features selected from professional knowledge and experience.

Data Assumptions

It is assumed that the features have a linear relationship with the target label and this was validated with the univariate regression of each feature against the label, as presented below.

Simple Linear Regression

In order to analyse the impact of respective features on the dependent variable, the first approach employed was to conduct simple linear regression using each of the independent variable against the feature (corrosion rate).

Temperature

The first variable to be analysed is Temperature. The model produced a very suitable line of best and the linear correlation corroborates (Konovalova 2021)'s position that increasing value of temperature, leads to an increased threat of corrosion on steels

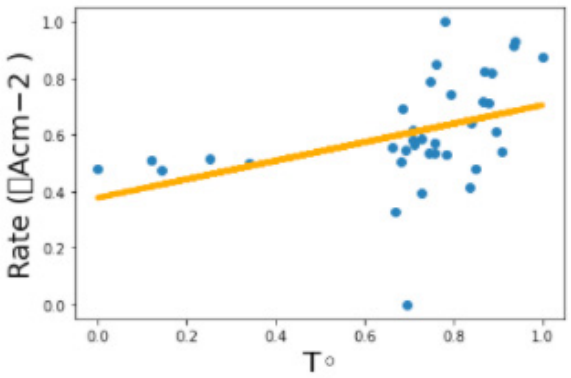


Figure 17: Simple Linear Regression of Temperature vs Corrosion Rate.

PH

The next feature analysed is the PH of the oceanwater. The regression produced a negative linear regression line, with a good line of best fit, as shown below. This shows that corrosion rates increases with decreasing value of water PH, which corroborates (Millette and Mavinic 1988)'s analysis.

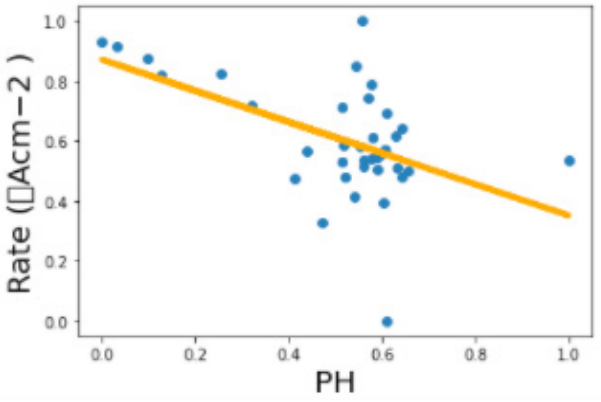


Figure 18: Simple Linear Regression of PH vs Corrosion Rate.

Salinity

The regression line for salinity, which refers to the salt content of water, is nearly flat, with almost a straight-line (neutral) behaviour. This totally contradicts the expert opinion of (Zakowski et al. 2014), who reported that areas with high salinity, typically pose high corrosion threats to steels, when compared to areas with low salinity.

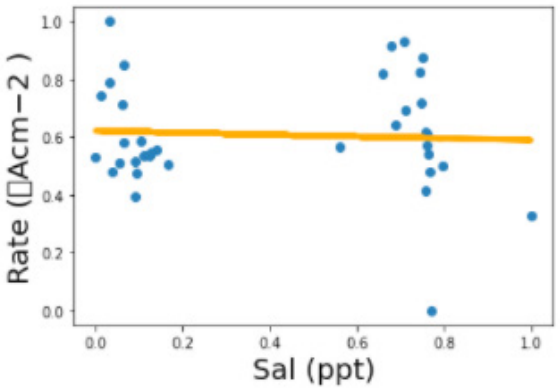


Figure 19: Simple Linear Regression of Salinity vs Corrosion Rate

Oxidation – Reduction Potential (ORP)

Similar to trends recorded in the case of Temperature, the ORP possesses a positive linear relationship with the corrosion rate. This infers that corrosion rate increases with increasing

ORP and vice versa. This position matches the result of (Lee et al. 2021)'s report, which predicted high corrosion rates for corresponding high value of ORP

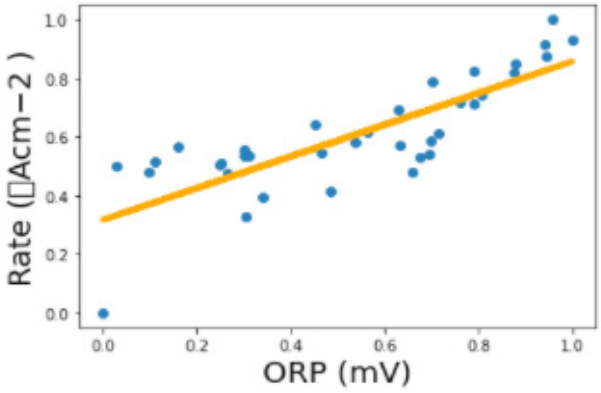


Figure 20: Simple Linear Regression of ORP vs Corrosion Rate.

Dissolved Oxygen

The result of the regression analysis for dissolved oxygen (which refers to the volume of oxygen contained in water), shows a neutral relationship (see figure below) with the corrosion rate of a steel metal, as it produced a near horizontal line across the dissolved oxygen axis. This implies that it has a neutral effect on the rate of corrosion produced from the metal. This position negates the view held by (Jung et al. 2011) that the corrosion rate in steels increases with increase in dissolved oxygen, as it attacks the passive film protecting the metal.

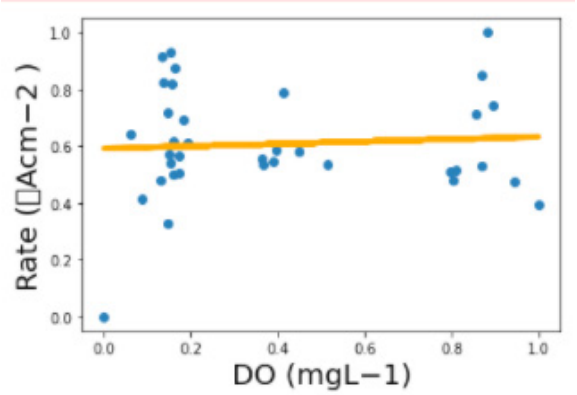


Figure 21: Simple Linear Regression of ORP vs Corrosion Rate.

Model Performance – Simple Linear Regression

Following the analysis of the fitting of the linear regression model for each of the features, it is essential to analyse the model performance, using the validation and performance metrices (R-Square, Root Mean Squared Error and Mean Absolute Error) as stated in the methodology. The table below presents the respective values of the different metrices, for the respective features.

Table 6: Model Performance for Different Features.

	Training Accuracy	Test Accuracy	R^2	Root Mean Squared Error	Mean Absolute Error
Temperature	0.1616	- 0.9176	- 0.9180	0.3840	0.3410
PH	0.2747	- 0.6738	- 0.6740	0.3590	0.3310
Salinity	0.0033	- 1.1795	- 1.1790	0.4100	0.3800
Oxidation–Reduction Potential	0.6559	- 0.4763	- 0.4760	0.3370	0.3150
Dissolved oxygen	0.0047	- 1.2058	- 1.2060	0.4120	0.3820

While, the models performed poorly on the individual features as shown by the training accuracy, the model had its highest training accuracy on the ORP, and the R-Squared was negative for all of the features. Normally, a R-Squared value close to the value of 1 suggests a well performing model as stated in the methodology section of this report. While the lowest obtainable value of R-Squared is zero, in practical sense, it is not impossible to obtain negative value. According to (Chicco et al. 2021), a negative R-Squared can be obtained if the regression line is worse than adopting the mean value. This is usually down to poor and unreliable dataset, which is the case in this research as the limited volume of the dataset ensures that the model is not as adequately trained as a larger dataset would afford. Additionally, the size of the data limited the data preparation and pre-processing approaches, which is regarded as the most important step in a machine learning project.

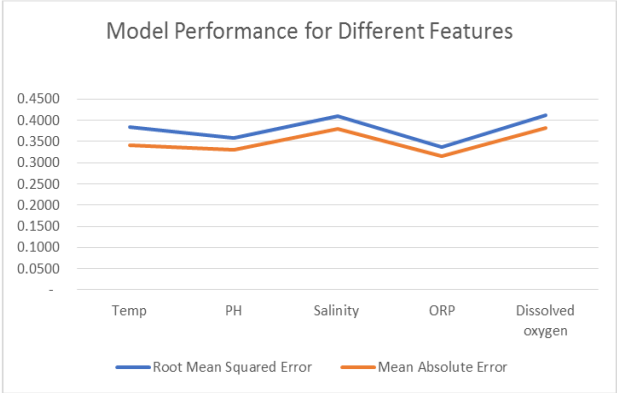


Figure 22: Model Performance for Different Features

Meanwhile, the root mean squared error whose score when inclined towards zero, signifies that the model is performing well. On this evidence, the model performed better in predicting corrosion rate with the ORP. Similarly, using the mean absolute error shows that the model performed better in predicting with the ORP. The mean absolute error measure the difference between the predicted and actual values, with the lower value indicating a better performance. The figure above present a plot of the mean absolute error and mean squared error values for the respective features.

Multiple Linear Regression

As stated in the methodology above, the analyses carried out for this research extends from the simple linear regression described above, to exploring model performance for multiple regression for the following cases.

1. Multiple Linear Regression using all features
2. Multiple Linear Regression using features with string correlations
3. Multiple Linear Regression using feature from industrial expertise and knowledge

The multiple regression analysis was carried out using a variety of regression algorithms like the Multiple Linear Regression, AdaBoost Regressor, Gradient Boosting Regressor, XGBoost, all of which has been explained in the methodology

Multiple Linear Regression Using All Features

The multiple regression analysis was done with the above algorithms and the performance was analysed with the same set of indices as in the simple linear regression. The first attempt at the multiple regression was done using all available features,

without considering the results of our feature selection. The model performance summary is as presented in the table below.

Table 7: Multiple Linear Regression - Model Performance for All Features

	Training Accuracy	Test Accuracy	R^2	Root Mean Squared Error	Mean Absolute Error
Multiple Linear Regression	0.7475	- 0.7462	- 0.7460	0.3670	0.3430
AdaBoost Regressor	0.9673	- 0.7870	- 0.7870	0.3710	0.3661
Gradient Boosting Regressor	0.9998	- 0.3197	- 0.3197	0.3188	0.2951
XGBoost Regressor	0.9814	- 0.3364	- 0.3364	0.3208	0.2924

Similar to issue with simple linear regression, the R-Squared returned completely negative values for all algorithms used. This equally points to the deficiency in the reliability of the dataset and restrictions against adequate and proper data preparation and pre-processing. The tendency to return a negative value is on of the reasons that researchers have declared R-Squared being too limited in usefulness, to be regarded as an effective measure of variance in prediction (de Heus 2012). According to Chicco et al. (2021), a negative R-Squared can be obtained if the regression line is worse than adopting the mean value

Beyond the R-Squared, the training accuracy of the model suggests that the Gradient Boosting Regressor produced the best performance of all the models used, with XGBoost coming second and Adaboost coming third. The generic multiple linear regression performed the poorest of the four algorithms used. The chart below shows the relationship between the performance of the different models.

Similarly, Gradient Boosting Regressor has the lowest Root Mean Squared and Mean Absolute Errors, showing that it performed better on test dataset, as it did on the training dataset (with the training accuracy.

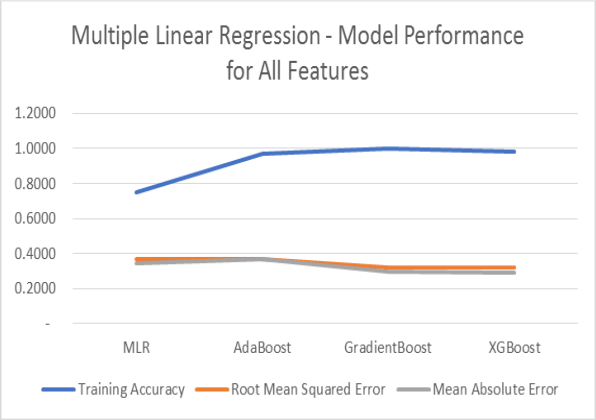


Figure 23: Multiple Linear Regression - Model Performance for All Features

Multiple Linear Regression Using Top Features

Although, based on the performance matrix for the multiple linear regression for all features, three of the four algorithms used for the multiple linear regression analysis shows impressive accuracy and reduced error, it is imperative to try the models on the features with the best correlation. According to the feature selection result, the Temperature, PH and Oxidation–Reduction Potential (ORP) presented the best correlations with the corrosion rate. Therefore, a multiple linear regression model was carried out, using the same algorithms and performance metrices as above.

As shown in the table below, negative R-Squared was recorded for all of the models following similar trends with early analysis. Additionally, The Gradient Boosting Regressor presents the best training accuracy as it is in the case of multiple regression for all features. However, the Root Mean Squared and Mean Absolute Errors are highest for the Gradient Boosting Regressor, contrary to the result from the analysis of the whole features.

Table 8: Multiple Linear Regression - Model Performance for Top Features.

	Training Accuracy	Test Accuracy	R^2	Root Mean Squared Error	Mean Absolute Error
Multiple Linear Regression	0.7127	- 0.7242	- 0.7240	0.3640	0.3470
AdaBoost Regressor	0.9357	- 0.9446	- 0.9446	0.3870	0.3809
Gradient Boosting Regressor	0.9996	- 1.0759	- 1.0759	0.3999	0.3957
XGBoost Regressor	0.9801	- 0.9021	- 0.9021	0.3828	0.3747

The generic regression model turned out to have the lowest error indices. The relationship between the performance metrices and the respect models, are as presented in the figure below.

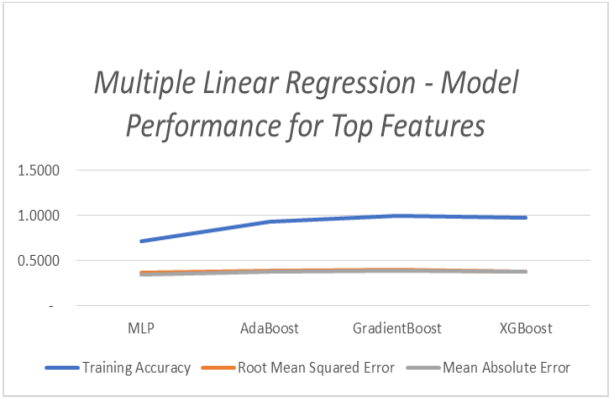


Figure 24: Multiple Linear Regression - Model Performance for Top Features

Multiple Linear Regression Using Feature from Industrial Knowledge

Research reporting from (Jung et al. 2011; Chou et al. 2017) confirmed the constituent of the dataset in this study as they revealed that the most important elements used for the prediction of corrosion rates in underwater cases are Temperature, Salinity and Oxidation–Reduction Potential (ORP), Dissolved Oxygen and PH. However, the domain knowledge description of the effect of a couple of these variables, differs from the correlation trend as described with the lines of best fits, earlier explained. While Dissolved Oxygen and Salinity are supposed to increase with increasing effect on corrosion. However, the lines of best fits suggest that the data distribution has a neutral effect on corrosion rate.

Therefore, the feature selection done by relating domain industry knowledge with the data behaviour, leaves us with PH, Temperature and ORP as the selected features. This is exactly

the same features selected through correlation analysis. This validates the feature selection approach and the analysis for feature selected from correlation analysis and that done with domain knowledge presents the same result.

Summary of Results and Analysis

Summarily, this chapter started off by presenting the research results with particular attention to descriptive analysis. The dataset utilised, while relatively clean, with no missing values and categorical data, little cases of outliers and relatively normal distribution, the limitation in sample size presented a huge challenge to the analysis. The discussion aspect of the chapter confirms this challenge as the R-Squared result are negative values all through the analysis, a result attributed to poor data reliability. Model accuracy was generally poor for models trained on the individual features. Meanwhile, with multiple regression, Gradient Boosting model achieved the best accuracy all through the analysis, and the baseline traditional regression model performed poorest. Additionally, the model performances dropped with the application on selected features. This is not entirely a bad situation, as feature selection is not only aimed at increasing model accuracy (sometimes have slight reduction effect as in the case of this research), but also to reduce noisy and unwanted features, so as to reduce overfitting, as relayed by (Shardlow 2016).

Section 5

Recommendations and Conclusion

Aging condition of offshore assets has presented enormous challenges to, not just the oil and gas operators, but the whole world, as global dependency on oil and gas as the premium source of energy continued to present itself. This continued dependence and high has strained the offshore assets as producers are obliged to continue to produce. While decommissioning is clearly not an immediate option for operators, due to its associated cost, the need for adequate maintenance of offshore assets is ever so necessary.

Extensive literature research was performed to establish the current technological trend in maintaining offshore assets and limitations where drawn from existing literatures and industrial application. These limitations revealed an obvious gap, on the lack of adequate research on the application of machine learning on digital twin recorded sensor data, for the detection of corrosion rates (a major aging failure mode) in offshore pipelines. This research set out to fill this study gap and aimed to answer pertinent research questions and hypothesis.

This led to an extensive analysis of sensor recorded datasets, and the result was presented with initial attention on descriptive analysis. The analysis revealed that the dataset utilised, while relatively clean, with no missing values and categorical data, little cases of outliers and relatively normal distribution, the limitation in sample size presented a huge challenge to the analysis. The discussion and analysis of research finding confirmed this challenge (limitation of data) as the R-Squared result are negative values all through the analysis, a result attributed to poor data reliability. This revealed datasets with limited sample size can potentially affect model performances, corroborating the result of (Cui and Gong 2018; Moghaddam et al. 2020)'s analysis.

Model accuracy was generally poor for models trained on the individual features. Meanwhile, with multiple regression, Gradient Boosting model achieved the best accuracy all through the analysis, and the baseline traditional regression model performed poorest. Additionally, the model performances dropped with the application on selected features. This is not entirely a bad situation, as feature selection usually serve several purposes. Why it can lead to improvement of model accuracy, research shows that feature selection can also bring about slight decrease in model accuracy (as in the case of this research) (Shardlow 2016), while reducing noisy and unwanted features, so as to reduce overfitting, as relayed by (Shardlow 2016). Despite the recorded limitation with the dataset, and the challenges it introduced to this research, the analysis still proffered relevant answers to the raised research questions, as presented below.

Research Contribution

This chapter extensively analysed the acquired dataset and presented variety of machine learning models in predicting corrosion rate in offshore pipelines. Varieties of regression models were compared with the benchmark traditional multiple linear regression. The analysis shows that

Gradient Boosting Regressor, XGBoost Regressor, AdaBoost Regressor (in this order) produced better performances than the traditional multiple regression model, when trained with the whole features. Additionally, feature selection was carefully carried out by exploring the correlation between the label and features, to establish the best features, for improved model performance. Similarly, the same these models (as stated above for full features) performed better than traditional multiple regression model when the top features were considered. Consequently, this research has provided answers for the research questions raised, as presented below:

Research Question 1: What is the effectiveness of the current maintenance methods and what are the associated perils?

The review of existing literature revealed that bulk of the existing research is focused on conditional monitoring, structural modelling, with rare attention to predictive machine learning modelling. Additionally, little focus or attention is given to specific failure modes (like corrosion) and asset types like the subsea/offshore pipelines.

Research Question 2: What are the key conditions and contributing factors (Features) towards corrosion in the offshore environment?

The key conditions and contributing factors like Temperature (T), Dissolved oxygen (DO), Salinity (Sal), Solution pH (pH), Oxidation–Reduction Potential (ORP) made up the dependent variables (features) in the dataset. To select the top features, feature selection was carried out using the Pearson's Correlation Coefficient and P- Value, for effective and accurate selection. Additionally, the industry expertise and experience was utilised to compliment the Pearson's Correlation Coefficient and the P-Value. Both analysis selected Temperature, PH and Oxidation–Reduction Potential as the top features and model training was carried out with just these feature, and compared with the model result from the analysis using the whole features. The performance of the models was in the same order (with Gradient Boosting Regressor performing best, followed by XGBoost Regressor, AdaBoost Regressor and traditional regression model) when trained on all of the features and when done on the top features alone. However, the models had better training accuracies when trained on all feature, compared to a slightly lower accuracy on the selected feature. Additionally,

the validation errors are lower for the model training on the full features, compared to higher errors for training done on the selected features. This point to possible challenges relating to the size and reliability of the datasets, as earlier mentioned.

Research Question 3: Which machine learning model is most effective in predicting corrosion rates in offshore pipelines?

The results and analysis presented in this research has shown poor performances of the baseline traditional machine learning model, in comparison to more advance models used. According to the results, the Gradient Boosting Regressor performed best, when applied on the full datasets without feature selection, as well as when top features are selected. The XGBoost Regressor performed next in both cases, while AdaBoost Regressor came next, with the least performance coming from the baseline model.

Research Question 4: Can digital twin and machine learning be deployed towards improving the current standards in asset maintenance?

The presented results and discussion shows the potential of achieving enormous accuracy in predicting corrosion rate in offshore pipelines, with conventional regression model delivering around 70% accuracy, while specialised models like the Gradient Boosting Regressor, XGBoost Regressor, AdaBoost Regressor produced accuracy of about 95% with all available features, and also with top selected features.

This points to the fact that there is enormous potential in the application of machine learning techniques on digital twin obtained sensor data, in predicting corrosion rate in offshore oil and gas pipelines. However, this can be achieved with certain adjustment to the data acquisition process, for better data reliability and improved prediction model building.

Recommendation

This research laid bare a huge challenge presented to the oil and gas industry. The initial focus of the research aim to incorporate a diagnostic approach to asset management, to the performed predictive analysis. However, this proved abortive, with the lack of relevant datasets for the purpose. Additionally, dataset for the predictive model which eventually carried out, equally proved difficult to obtain. Hence the use of a dataset, with limited sample space, which drastically affected the quality of the result, evidenced by the negative R-Square values all through, and low performance of the selected features.

Therefore, it is recommended that a government regulation is enacted to enforce transparency amongst oil companies, to encourage growth and innovations, through the release of adequate data, for the promotion of growth focused research activities as this.

Future Work

Despite the exceptional research work delivered with this project, the limitation in data availability opens up potential area of research focus. Firstly, it would be exciting to perform similar analysis as it's been carried out in this research on a larger data sample, to further validate the findings and observations derived from this work. Additionally, the diagnostic analysis (which will utilise computer vision technologies in identifying asset failure) present a novel research potential toward improved maintenance of offshore assets. This is a concept that is already in wide use in the aviation industry, with the lack of availability of relevant data, limiting its application in the offshore oil and gas industry. Furthermore, this diagnostic technology can be incorporated with the prognostic approach presented in this research, to achieve an holistic solution to solving aging asset challenges, by firstly diagnosing the asset condition from it physical state, and analysing its failure rate and remaining useful life, through prognostic analysis.

References

1. Achilla, M. E. (2015). Development of Risk Based Maintenance Strategy For The Multipurpose Petroleum Pipeline Systemin Kenya. Nairobi, *Jomo Kenyatta University of Agriculture and Technology*. Retrived from http://ir.jkuat.ac.ke/handle/123456789/1800
2. Al-Omari, A. I. (2015). New entropy estimators with smaller root mean squared error. *Journal of Modern Applied Statistical Methods, 14* (2), 10. DOI: 10.22237/jmasm/1446350940. Retrieved from https://digitalcommons.wayne.edu/jmasm/vol14/iss2/10/
3. Altamiranda, E., Kiaer, L. & Hu, X. (2009). Condition monitoring and diagnosis for subsea control systems. A subsystem prototype. *OCEANS 2009-EUROPE, IEEE*. https://doi.org/10.1109/OCEANSE.2009.5278311
4. Altay, A., Ozkan, O., & Kayakutlu, G. (2014). Prediction of aircraft failure times using artificial neural networks and genetic algorithms. *Journal of Aircraft 51* (1), 47-53. http://dx.doi.org/10.2514/1.C031793
5. Animah, I., & Shafiee, M. (2017). Condition assessment, remaining useful life prediction and life extension decision making for offshore oil and gas assets. Journal of Loss Prevention in the Process Industries. (10.1016/j.jlp.2017.04.030.).
6. Animah, I., & Shafiee, M. (2018). Condition assessment, remaining useful life prediction and life extension decision making for offshore oil and gas assets. *Journal of loss prevention in the process Industries 53*, 17-28. https://doi.org/10.1016/j.jlp.2017.04.030
7. Awad, M., & Khanna, R. (2015), *Support vector regression.* In: Efficient learning machines, Apress, Berkeley, CA, Springer, 67-80. https://doi.org/10.1007/978-1-4302-5990-9_4
8. Baker, J. M., & Descamps, B. (1999). Reliability-based methods in the inspection planning of fixed offshore steel structures. *Journal of Constructional Steel Research 52*(1), 117–131. https://doi.org/10.1016/S0143-974X(99)00031-0
9. Bhowmik, S. (May 6–9, 2019). Digital twin of subsea pipelines: conceptual design integrating IoT, machine learning and data analytics. *Offshore Technology Conference*. OnePetro. https://doi.org/10.4043/29455-MS
10. Blann, D. R. (1997). *Proactive Maintenance as a Strategic Business Advantage*. Chicago Plant Services Magazine.
11. Blann, D. R. (1999). Action Teams: *Reliability Improvement on the Front Line*. California: Marshall Institute, Inc.
12. Butler, S. (2012). *Prognostic algorithms for condition monitoring and remaining useful life estimation*. National University of Ireland Maynooth. Retrived from https://mural.maynoothuniversity.ie/3994/
13. Cameron, A. C., & Windmeijer, F. A. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics, 14*(2), 209-220. Retrieved from https://econpapers.repec.org/RePEc:fth:caldav:93-24
14. Candreva, F., & Houari, M. (2013). Plant Screening for Ageing Impact in the Process Industry. Chemical Engineering Transactions, 31, 253-258. https://doi.org/10.3303/CET1331043
15. Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development, 7*(3),

1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

16. Chen, H., Stavinoha, S., Walker, M., Zhang, B., & Fuhlbrigge, T. (2014). Opportunities and challenges of robotics and automation in offshore oil & gas industry. *Intelligent Control and Automation, 5*, 136-145. doi: 10.4236/ica.2014.53016

17. Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science, 7*, e623. https://doi.org/10.7717/peerj-cs.623

18. Chou, J.-S., Ngo, N.-T. & Chong, W. K. (2017). The use of artificial intelligence combiners for modeling steel pitting risk and corrosion rate. Engineering Applications of Artificial Intelligence 65, 471-483. http://dx.doi.org/10.1016/j.engappai.2016.09.008

19. Chukwunonso, O. (2015). *Ageing of Offshore Assets: Issues and Challenges Bedfordshire: Cranfield University*.

20. Clausard, C. (2006a). Pipeline Integrity Management for Aging Offshore Pipelines. *Aberdeen: Pigging Products and Services Association*. Retrieved from https://ppsa-online.com/papers/2006-Aberdeen-7-Clausard.pdf

21. Clausard, C. (2006b). Pipeline integrity management strategy for aging offshore pipelines. *MACAW Engineering Limited*, UK 15.

22. Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems 95*, 27-36. https://doi.org/10.1016/j.dss.2016.11.007

23. Cui, Z., & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *Neuroimage 178*, 622-637. https://doi.org/10.1016/j.neuroimage.2018.06.001

24. de Heus, P. (2012). R squared effect-size measures and overlap between direct and indirect effect in mediation analysis. *Behavior Research Methods 44*(1), 213-221. https://doi.org/10.3758/s13428-011-0141-5

25. Duncan , A. (2012). *KP4 Ageing and Life Extension Project Interim Report*. Aberdeen: Health and Safety Executive.

26. El Saddik, A. (2018). Digital twins: The convergence of multimedia technologies. *IEEE multimedia 25*(2), 87-92. https://doi.org/10.1109/MMUL.2018.023121167

27. Errandonea, I., Beltrán, S., & Arrizabalaga, S. (2020). Digital Twin for maintenance: A literature review. *Computers in Industry 123*, 103316. https://doi.org/10.1016/j.compind.2020.103316

28. Ersdal, G. (2005). *Assessment of existing structures for life extension. Stavanger*, University of Stavanger. http://dx.doi.org/10.1115/OMAE2008-57451

29. Ersdal , G., Hornlund, E., & Spilde, H. (2011). *Experience From Norwegian Programme On Ageing And Life Extension. Rotterdam*. ASME and Government of Norway. http://dx.doi.org/10.1115/OMAE2011-50046

30. Ersdal, G., Kvitrud, A., Jones, W., & Birkinshaw, M. (2008). Life extension for mobile offshore units require robust management: How old is too old? *Journal of International Association of Drilling Contractors*.

31. ETI, M. C., OGAJI, S. O. T., & PROBERT, S. D. (2006). Development and implementation of preventive-maintenance practices in Nigerian industries. *Applied Energy, 83*(10), 1163-1179. https://doi.org/10.1016/j.apenergy.2006.01.001

32. Ferreira, P., Le, D. C. & Zincir-Heywood, N. (2019) Exploring feature normalization and temporal information for machine learning based insider threat detection. *2019 15th International Conference on Network and Service Management (CNSM). IEEE*. http://dx.doi.org/10.23919/CNSM46954.2019.9012708

33. Galbraith, D., & Sharp , J. (2007a). *Recommendations for design life extension regulations*. Aberdeen, POSEIDON INTERNATIONAL LTD.

34. Galbraith, D., & Sharp, J. (2007b). *Specialist Support on Structural Integrity Issues*. Aberdeen, POSEIDON INTERNATIONAL LTD.

35. Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. Seminars in hematology. Vol. 45. Elsevier. https://doi.org/10.1053/j.seminhematol.2008.04.003

36. Gupta, R., & Patel, R. (2010). *Ageing Management & Residual Life Assessment of Heavy Water Plants*. 2010 2nd International Conference on Reliability, Safety and Hazard-Risk-Based Technologies and Physics-of-Failure Methods (ICRESH). IEEE.

37. Håbrekke, S., Hokstad, P., & Ersdal, G. (2011). A*geing and life extension for safety systems on offshore facilities*. Advances in Safety, Reliability and Risk Management: ESREL 2011, 247.

38. Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature methods 12*(3), 179-185. https://doi.org/10.1038/nmeth.3288

39. Hart, K., Ersdal, G., Wintle, J., Smith, S., Sharp, J., Galbraith, D., & Terry, E. (2009). *Assessing the Impact of Ageing Safety Critical Elements in Offshore Installations and How the Ageing Processes Impact the Role of SCES to Act as Barriers to Major Accidents*. IChemE.

40. Heng, A., Zhang, S., Tan, A. C., & Mathew, J. (2009). Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical systems and signal processing, 23*(3), 724-739. https://doi.org/10.1016/j.ymssp.2008.06.009

41. Hörnlund, E., Ersdal, G., Hinderaker, R. H., Johnsen , R., & Sharp, J. (2011). Material Issues in Ageing and Life Extension. A*SME 2011 30th International Conference on Ocean, Offshore and Arctic Engineering 3*, 261-267. https://doi.org/10.1115/OMAE2011-49363

42. Hörnlund, E., Sævik, O. y., Hinderaker, R. H. & Ersdal, G. (2008). Ageing of Materials. *ASME 2008 27th International Conference on Offshore Mechanics and Arctic Engineering 5*(1), 285-292.

43. Horrocks, P., Mansfield, D., Parker, K., Thomson, J., Atkinson, T., Worsley, J., House, W., & Park, B. (2010). *Managing ageing plant*. HSE, Warrington, UK, Tech. Rep 823.

44. Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. Procedia Cirp 79, 403-408. https://doi.org/10.1016/j.procir.2019.02.106

45. Hudson, B. G. (2010). Extending the life of an ageing offshore facility. Abu Dhabi. Society of Petroleum Engineers.

46. IHS Markit. (2016). Place Published [Medium]: Updated Last Update Date. http://news.ihsmarkit.com/press-release/energy-power-media/decommissioning-aging-offshore-oil-and-gas-facilities-increasing-si

47. Jung, H., Kwon, K., Lee, E., Kim, D., & Kim, G. (2011). Effect of dissolved oxygen on corrosion properties of reinforcing steel. *Corrosion engineering, science and technology, 46*(2), 195-198. DOI: 10.1179/1743278210Y.0000000008.

48. Kavitha, S., Varuna, S., & Ramya, R. (2016). A comparative analysis on linear regression and support vector regression. 2016 online international conference on green engineering and technologies (IC-GET). IEEE. https://doi.org/10.1109/GET.2016.7916627

49. Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory data analysis. Secondary analysis of electronic health records, 185-203. https://doi.org/10.1007/978-3-319-43742-2_15

50. Konovalova, V. (2021). The effect of temperature on the corrosion rate of iron-carbon alloys. *Materials Today: Proceedings 38*, 1326-1329. https://doi.org/10.1016/j.matpr.2020.08.094

51. Kotz, S., Balakrishnan, N., Read, C. B., & Vidakovic, B. (2005). *Encyclopedia of Statistical Sciences*, volume 1. John Wiley & Sons.

52. Kummer, N., & Najjaran, H. (2014). Adaboost. MRT: Boosting regression for multivariate estimation. Artif. Intell. Res. 3 (4), 64-76. https://doi.org/10.5430/air.v3n4p64

53. Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean journal of anesthesiology 70*(4), 407. https://doi.org/10.4097/kjae.2017.70.4.407

54. Ledet, W. (2016). Reliablity Web [Medium].Place Published: Updated Last Update Date. http://reliabilityweb.com/articles/entry/the_abcs_of_failure_getting_rid_of_the_noise_in_your_system

55. Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., & Siegel, D. (2014). Prognostics and health management design for rotary machinery systems - Reviews, methodology and applications. *Mechanical systems and signal processing, 42*(1-2), 314-334. https://doi.org/10.1016/j.ymssp.2013.06.004

56. Lee, S., Narayana, P., Seok, B. W., Panigrahi, B., Lim, S.-G., & Reddy, N. (2021). Quantitative estimation of corrosion rate in 3C steels under seawater environment. *Journal of Materials Research and Technology, 11*, 681-686. https://doi.org/10.1016/j.jmrt.2021.01.039

57. Liu, J., Wang, W., Ma, F., Yang, Y., & Yang, C. (2012). A data-model-fusion prognostic framework for dynamic system state forecasting. *Engineering Applications of Artificial Intelligence, 25*(4), 814-823. https://doi.org/10.1016/j.engappai.2012.02.015

58. Luo, W., Hu, T., Zhang, C., & Wei, Y. (2019). Digital twin for CNC machine tool: modeling and using strategy. *Journal of Ambient Intelligence and Humanized Computing, 10*(3), 1129-1140. https://doi.org/10.1007/s12652-018-0946-5

59. Malekloo, A., Ozer, E., AlHamaydeh, M., & Girolami, M. (2021). Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights. Structural Health Monitoring, 14759217211036880. https://doi.org/10.1177/14759217211036880

60. Miles, J. (2014). *R squared, adjusted R squared*. Wiley StatsRef: Statistics Reference Online.

61. Milje, R. (2011) Engineering methodology for selecting Condition Based Maintenance. University of Stavanger: Stavanger. https://doi.org/10.1002/9781118445112.stat06627

62. Millette, L., & Mavinic, D. S. (1988). The effect of pH adjustment on the internal corrosion rate of residential cast-iron and copper water distribution pipes. *Canadian Journal of Civil Engineering 15*(1), 79-90. https://dx.doi.org/10.14288/1.0062626

63. Misra, K. B. (2008). *Maintenance engineering and maintainability: An introduction*. Handbook of performability engineering, 755-772. DOI: 10.1007/978-1-84800-131-2_46

64. Moghaddam, D. D., Rahmati, O., Panahi, M., Tiefenbacher, J., Darabi, H., Haghizadeh, A., Haghighi, A. T., Nalivan, O. A., & Bui, D. T. (2020). The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers. *Catena 187*, 104421. https://doi.org/10.1016/j.catena.2019.104421

65. Morgenthaler, S. (2009). Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics 1*(1), 33-44. https://doi.org/10.1002/wics.2

66. Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal, 24*(3), 69-71. PMCID: PMC3576830, PMID: 23638278

67. Novak , S., & Podest, M. (1987). *Nuclear power plant ageing and life extension: Safety aspects*. vienna: International Atomic Energy Agency (IAEA).

68. Ochella, S., Shafiee, M., & Sansom, C. (2021). Adopting machine learning and condition monitoring PF curves in determining and prioritizing high-value assets for life extension. *Expert Systems with Applications 176*, 114897. https://doi.org/10.1016/j.eswa.2021.114897

69. Oliván, A. D. (2017). *Machine learning for data-driven prognostics: methods and applications*. Universidad Politécnica de Madrid. https://doi.org/10.20868/UPM.thesis.48053

70. Onawoga, D. T., & Akinyemi, O. O. (2010). Development of Equipment Maintenance Strategy for Critical Equipment. *The Pacific Journal of Science and Technology 11*(1), 328-342.

71. Paik, K. J., & Melchers, E. R. (2008). Condition Assessment of Aged Structures. *Cambridge: Woodhead Publishing Limited and CRC Press LLC*.

72. Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462. Retrieved from https://arxiv.org/ftp/arxiv/papers/1503/1503.06462.pdf

73. Patterson, R. (2013). Structural Integrity of Aging Assets and Effects of Asset Life Extension. *Claxton Engineering*.

74. Pintelon , L., & Parodi-Herz, A. (2008). Maintenance: An Evolutionary Perspective. *New York City, Springer*. DOI: 10.1007/978-1-84800-011-7_2

75. Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data 1*(1), 51-59. https://doi.org/10.1089/big.2013.1508

76. Qi, Q., & Tao, F. (2018). Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. *Ieee Access 6*, 3585-3593.

77. Qian, Y., Yan, R., & Gao, R. X. (2017). A multi-time scale approach to remaining useful life prediction in rolling bearing. *Mechanical Systems and Signal Processing, 83*, 549-567.

78. Renzi, D. (2019). Evolution of digital twins for floating production systems. *World Oil Magazine, 240*(11).

79. Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., & Schwabacher, M. (2008). Metrics for evaluating performance of prognostic techniques. *2008 international conference on prognostics and health management. IEEE*. http://dx.doi.org/10.1109/PHM.2008.4711436

80. Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science 181, 526-534. https://doi.org/10.1016/j.procs.2021.01.199

81. Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research 12*(1), 217-222.

82. Shardlow, M. (2016). An analysis of feature selection techniques. *The University of Manchester, 1*(2016), 1-7.

83. Sharma, P., Hamedifar, H., Brown, A., & Green, R. (2017). The dawn of the new age of the industrial Internet and how it can radically transform the offshore oil and gas industry. Offshore Technology Conference. OnePetro. https://doi.org/10.4043/27638-MS

84. Shukla, A., & Karki, H. (2016). Application of robotics in offshore oil and gas industry—A review Part II. *Robotics and Autonomous Systems 75*, 508-524. https://doi.org/10.1016/j.robot.2015.09.013

85. Simm, I. (2019). Oil Voice [Medium].Place Published: Updated Last Update Date. https://oilvoice.com/Press/32043/Machine-learning-technology-to-extend-offshore-asset-life

86. Sørensen, D. J. and Ersdal, G. (2008) Safety and inspection planning of older installations. *Journal of Risk and Reliability 22*(3), 403-418.

87. Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes , M., Nenadic, G. & Keane, J. (2019). Machine learning

methods for wind turbine condition monitoring: A review. *Renewable Energy, 133*, 620- 635. https://doi.org/10.1016/j.renene.2018.10.047

88. Sullivan, G. p., Pugh, R., Melendez, A. P., & Hunt, W. D. (2010). Operations and Maintenance Best Practices- A Guide to Acgievig Operational Efficiency (Release 3.0). *Tennessy: US Department of Energy*.
89. Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics*. Vol. 5. Pearson Boston, MA.
90. Tan, Y., Niu, C., Tian, H., Hou, L., & Zhang, J. (2019) A one-class SVM based approach for condition-based maintenance of a naval propulsion plant with limited labeled data. *Ocean Engineering, 193*, 106592.
91. Thomas, M. (2000). *Complementarity of TPM and TQM-The Indian Experience*. Sevilla First World Conference on Production and Opeations Management.
92. Tiddens, W. W., Braaksma, A. J. J., & Tinga, T. (2015). The adoption of prognostic technologies in maintenance decision making: a multiple case study. *Procedia CIRP, 38*, 171-176.
93. Tiwari, K., Mehta, K., Jain, N., Tiwari, R., & Kanda, G. (2007). Selecting the appropriate outlier treatment for common industry applications. *NESUG Conference Proceedings on Statistics and Data Analysis, Baltimore, Maryland, USA*.
94. Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. P*rocedia-Social and Behavioral Sciences, 106*, 234-240.
95. Wang, F.-K., & Mamo, T. (2020). Gradient boosted regression model for the degradation analysis of prismatic cells. *Computers & Industrial Engineering, 144*, 106494.
96. White, G., Zink, A., Codecá, L., & Clarke, S. (2021). A digital twin smart city for citizen feedback. C*ities, 110*, 103064.
97. Wintle, J. (2010). *Management of Ageing and Life Extension Issues*. Aberdeen. TWI Ltd.
98. Wintle, J., Johnston, C., Miles, J., & McGrath, B. (2012). Management of ageing: A framework for nuclear chemical facilities. *HSE Research Reports. Prepared by TWI for the Health and Safety Executive, UK*.
99. World Oil (2016). [Medium] Place Published: Updated Last Update Date. http://www.worldoil.com/news/2016/11/29/over-600-offshore-projects-to-be-decommissioned-over-the-next-five-years
100. Zakowski, K., Narozny, M., Szocinski, M., & Darowicki, K. (2014). Influence of water salinity on corrosion risk—the case of the southern Baltic Sea coast. *Environmental monitoring and assessment, 186*(8), 4871-4879.
101. Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. Applied artificial intelligence, 17(5-6), 375-381.
102. Zhang, X., Yan, C., Gao, C., Malin, B. A., & Chen, Y. (2020) Predicting missing values in medical data via XGBoost regression. *Journal of Healthcare Informatics Research, 4*(4), 383-394.