National College of Ireland

A Comparative Study of Machine Learning Models for Early Corrosion Detection in Oil Pipelines

MSc Research Project
AI for Business

Aida Metenova
Student ID: 22247700

School of Computing
National College of Ireland

Supervisor: Dr. Muslim Jameel Syed

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……Aida Metenova……

Student ID: ……x22247700……..……

Programme: …MSCAIBUS…… **Year:** …2023-24..

Module: …MSc Research Project……..……

Supervisor: …Dr. Muslim Jameel Syed……..……

Submission Due Date: …12 Aug 2024……..……

Project Title: … A Comparative Study of Machine Learning Models for Early Corrosion Detection in Oil Pipelines ……..……

Word Count: …5214…… **Page Count**……18……..……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Aida Metenova……

Date: ……11 Aug 2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Comparative Study of Machine Learning Models for Early Corrosion Detection in Oil Pipelines

Aida Metenova
x22247700
MSCAIBUS
National College of Ireland

Abstract

This study seeks to explore the application of machine learning algorithms to predict corrosion rate in oil refinery pipelines in order to enhance the early detection strategies. The comparative analysis included four supervised machine learning models such as random forest, support vector regression (SVR), convolutional neural network (CNN), and the long short-term memory (LSTM). Additionally, the study focuses on using the operational data like temperature, pressure, flow rate, and pH level from the existing transmitters and flow meters to demonstrate a cost-efficient method that would not require additional investments in new equipment installation. Given, the harsh nature of hydrocarbon fluid, oil leakage caused by corrosion can significantly damage the environment leading to contamination and ecological disaster as well as substantial financial loss for the companies. The findings of the study underscore the importance of implementing effective early detection strategies to prevent such severe outcomes and show the potential of machine learning to advance corrosion management and elevate the overall safety in oil industry.

Keywords: convolutional neural network (CNN), corrosion detection, long short-term memory (LSTM), machine learning, oil and gas industry, random forest, support vector regression (SVR).

1. Introduction

The oil and gas sector is among the largest businesses across the globe positioning oil as one of the top valued resources (Statista, 2024). The main element of the industry infrastructure is pipelines. In addition to being an essential component of the production facilities, pipelines are a popular and cost-effective way to deliver energy resources to final consumers. The extensive pipeline network connecting different provinces, states, and global regions was calculated to cover 2.27 million kilometres (GlobalData, 2023). However there are many concerns, threats and risks that these pipes can meet and one of the biggest hazards is corrosion. Aggressive properties of crude oil and natural gas, coupled with intense pressure, severe temperatures, and external environmental factors like surrounding temperature, moisture, and soil profile, contribute to the formation of corrosion, Figure 1 (Choi, et al., 2015). There are many various forms that

corrosion can take but the most frequent type of it is produced by an electrochemical reaction that happens naturally when metal interacts with its surroundings to form a steady oxide.

Figure 1. Inner pipeline corrosion (Choi, et al., 2015)

Corrosion presents severe risks to the reliability of pipelines leading to the reduction of wall thickness of the pipes and the formation of corrosion pits that can cause then pipeline failures and leaks. Corrosion is responsible for 30.3% of pipeline failures making it one of the leading causes of damage in the industry as demonstrated by Figure 2 (Zakikhani, et al., 2019). Its effect has significant economic implications and presents serious operational and ecological hazards. In response to that, surveillance and management strategies become highly important. The annual cost of managing corrosion in the industry is projected to be $1.3 billion (Popoola, et al., 2013). As a result, the industry is focused on creating new techniques and strategies that can help businesses to reduce the financial losses associated with pipeline failures and unplanned shutdowns and improve the global environmental protection.

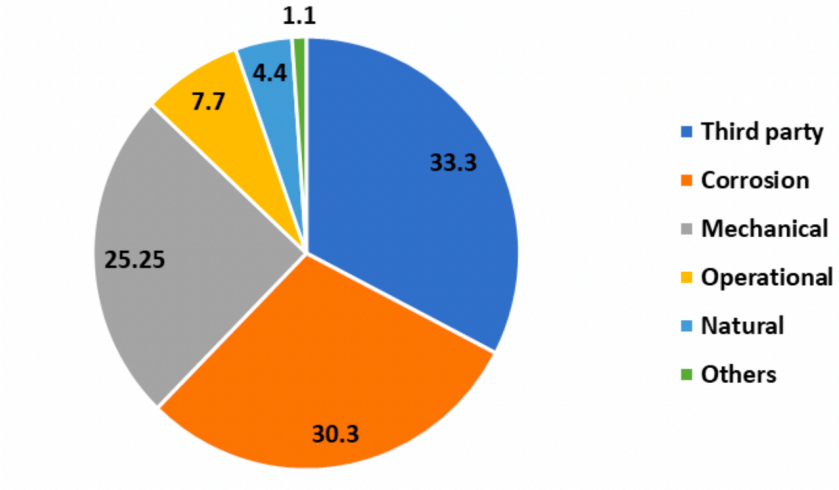

Figure 2. Percentage breakdown of damage causes in oil and gas pipelines (Zakikhani, et al., 2019)

The oil and gas industry invests significant resources to address corrosion both on the outside and inside of pipes. These efforts include both protective and detective measures. Protective actions involve technologies such as corrosion inhibitors, cathodic protection, and protective coating (Ameh, et al., 2018). Corrosion detection methods include non-destructive testing (NDT), smart inline inspection (ILI), and routine assessments. Although these methods are effective to some extent, they also have some drawbacks that range from temporary results and high expenses to ecological issues and technical difficulties in their implementation. Particularly, detection methods lack precision and consistency as they strongly rely on personnel skills for analysing outcomes, overlook certain areas like buried pipelines and other difficult to access locations, and often unavailable since they require shutting down the operation leading also to revenue loss.

Considering this, Machine Learning presents a game-changing method for early identification of corrosion without depending on scheduled maintenance, human involvement, or physical contact with pipelines. Currently, the use of Machine Learning in this context is still limited. A comprehensive review of existing literature reveals some progress in this area but there remains a significant gap for implementing these algorithms specifically for early corrosion detection through predicting its rate using the operational data from the existing sensors. This gap causes the research question: “How can Machine Learning algorithms like Random Forest, SVR, CNN, and LSTM be employed to predict corrosion rates and detect early signs of corrosion in oil refinery pipelines?”. This study aims to examine the practicality, accuracy, and methodology of implementing machine learning based corrosion detection utilizing the data from existing sensors and instruments. It focuses on supervised ML techniques – random forest, support vector regressor, convolutional neural network, and long short-term memory – and their use in forecasting corrosion rates in oilfield pipelines. Calculating the corrosion rate can play a critical role in early detection. This study intends to demonstrate how analysing temperature, pressure, flow velocity and pH level by these ML models can provide continuous real-time monitoring of pipeline conditions. Accurate predicting of corrosion rates can help to identify patterns and anomalies that indicate the beginning of corrosion before it becomes a critical issue. This approach is highly proactive as it allows to take preventive actions rather than mitigating and eliminating activities enhancing the integrity of pipelines and field facilities as safety is traditionally the first priority for oil companies. The study compares the four models in order to define the most accurate one which will best suit the industry’s needs for reliable corrosion rate prediction. The suggested method also provides significant benefits for corrosion management using current infrastructure and equipment without requiring extra investments or workforce as well as enhances the operational efficiency. The models utilize routine temperature and pressure transmitters as well as flow meters readings converting them into meaningful pieces of information and, as a result, reducing the dependence on manual checks

and minimizing the possibility of operational delays. Any advancements made in the corrosion management immediately contribute to enhanced safety and ecological protection helping oil and gas companies avoid any reputational loss or significant financial consequences.

This study is organized to deliver an in-depth analysis on how machine learning algorithms can be used for early identification of corrosion in the oil and gas business. After the introduction section, the next part provides a comprehensive review of the existing research and summarizes what has been studied so far and identifies the remaining gaps. The following part then, Research Methodology, explains the methods used in the paper, specifically the dataset, preprocessing techniques, and evaluation metrics and discusses the implementation of the selected ML models to predict corrosion rates in pipelines. The fourth chapter assesses the outcomes of the selected machine learning models, and the final section presents the research conclusions and suggests possible areas for future study.

2. Literature Review

The analysis of existing research thoroughly analyses current approaches to corrosion detection in the petroleum sector with a specific emphasis on the use of AI driven systems. The review looks into both classic methods and innovative strategies that employ machine learning models. The purpose of the background research is to provide a strong basis for understanding on what way machine learning algorithms may be implemented to elevate early corrosion detection through the prediction of corrosion rates. Considering the vital importance of pipeline safety, the current section also shows how these ML models can use basic operational parameters to predict and identify first traces of corrosion on refinery infrastructure. Also, it gives a summary of the advantages and limitations of the discussed approaches along with the gaps in the existing methods.

A. Traditional Approaches for Corrosion Detection

Traditional practices to identify corrosion are categorized into two main types based on their purpose – methods designed to examine the pipes surface area and methods intended for using within buried pipelines. Both options have their strengths and weaknesses. The current study will discuss the most commonly used ones in order to establish a foundation for comparing with ML based techniques, specifically those introduced in this paper.

A widely used traditional approach for identifying anomalies is in-line inspection (ILI) which represents a type of non-destructive testing (NDT) and involves sending an “intelligent pig” through the pipeline to assess its structural condition (Xie & Tian, 2018). The “smart pigs” or “intelligent pigs” are advanced digital devices fitted with various detectors and designed to move inside the pipelines to identify corrosion, wall

thinning, cracks, and other issues. An illustration of such device is provided on Figure 3 (Martinez , et al., 2019).

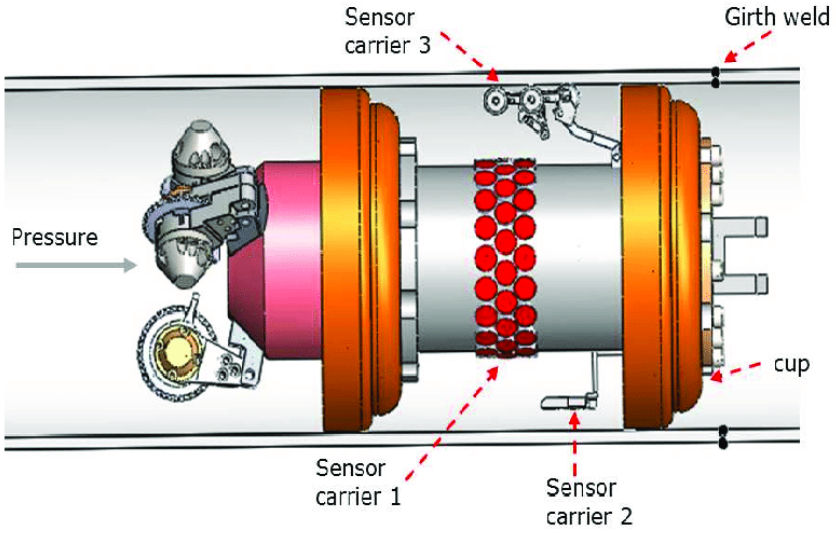

Figure 3. Pipeline inspection gauge (PIG) device (Martinez , et al., 2019)

While ILI tools became more advanced recently and demonstrated their effectiveness for identifying various forms of defects, the procedure is quite expensive. It requires shutting down either the entire oil production site or parts of it to drain the pipeline and conduct manual inspections. This not only results in financial losses but can also be difficult to perform due to the pipelines dimensions or conditions (Xie & Tian, 2018).

Another group of researchers state that the best results come from combining preventive and monitoring techniques and recommend using hydrogen corrosion monitors as part of this integrated approach (Corbin & Willson, 2007). The monitors measure the hydrogen levels that are closely related to corrosion processes within pipes and offer real-time information to the technical team. Initially, the technology used intrusive probes, but it has then evolved to non-intrusive methods and helped to prevent facilities shutdowns and decreased maintenance costs. But it still demands adding more equipment, frequent cleaning due to the hydrogen accumulation, and thorough monitoring.

A different traditional approach for detecting corrosion is coupon monitoring that involves using little samples of pipes known as “coupons” and putting them to the similar conditions as main pipeline (Reddy, et al., 2021). This method gives valuable information about what could happen inside the main tubing. Such approach is a basic and straightforward way to assess metal degradation caused by corrosion and is more cost-effective than other solutions. But it takes significant amount of time for monitoring and assessment of the coupons that sometimes can be week or months, so it is not entirely proactive.

B. Machine Learning-based Corrosion Detection Techniques

Since traditional methods for detecting and addressing corrosion present many issues and drawbacks, researchers are still seeking for more refined and efficient solutions. With the rise of Artificial Intelligence and Machine Learning exploring these technologies for detecting and predicting corrosion in the field of oil and gas production became as a natural progression. One approach involving AI focuses not just on detecting corrosion but on overall pipeline health monitoring with an emphasis on forecasting all potential faults before they happen (Chalgham, et al., 2020). This method begins with gathering data, followed by assessing risks based on the information from existing sensors, and ends with recommending corrective measures. Similar to the present paper, the technique utilizes existing data on temperature, pressure, and flow rates, but distinguishes by employing probabilistic models such as Hybrid Causal Logic (HCL) and Dynamic Bayesian Network (DBN) and is facilitated by a specialized software tool. The HCL model is used to examine any potential failure scenarios from various internal and external conditions, while DBN model forecasts corrosion progression. The goal of using the chosen systems is to estimate different pipeline malfunction situations, while the present study focuses on identifying early signs of these potential issues.
Another study examines stress corrosion cracking (SCC), a common reason of pipeline issues resulting from corrosion (Soomro, et al., 2021). Instead of focusing directly on corrosion detection, this research employs machine learning algorithms to model SCC to demonstrate a different way AI can be used to address pipeline issues caused by corrosion. The authors use deep learning techniques such as Long Short Term Memory (LSTM) and Physic Informed Deep Neural Networks (PINN) to estimate the likelihood of failure in corroded pipes. The approach enhances general understanding on how corrosion affects pipelines and underscores the capabilities of machine learning in addressing issues related to pipeline reliability and security.
One recent study of another team of researchers explored a piezoelectric-based time reversal technique integrated with a Convolutional Neural Network (CNN) to track internal corrosion in pipes (Yang, et al., 2023). The strategy requires applying piezoelectric patches to the pipeline surface to transmit and receive ultrasonic waves. The collected signals are processed then and analysed by the CNN model to evaluate the level of corrosion. This approach achieves an impressive accuracy of 99.01%, offers inexpensive pricing, and straightforward design. However, the paper does not clarify whether this method detects existing corrosion or focuses on its first traces. While the scientists acknowledge that the proposed technique can assess the degree of corrosion, an important factor for maintenance decisions, its capability to identify the onset of corrosion requires additional investigation. Also, implementing this method could lead to increased expenses because of the need to hire more personnel for installation and maintenance of the piezoelectric patches.

Corrosion and leakage identification have also been tackled using an Internet of Things (IoT) based system combined with machine learning classification techniques (Parjane, et al., 2023). In the method, IoT devices gather measurements from underwater pipeline sensors such as wall thickness and GPS coordinates at six-hour intervals enabling continuous observation. This information are processed then and analysed by a Q-learning algorithm to assess the risks of corrosion and leaks. This technique not only accurately specifies pipeline failures but also precisely locates them. However, the main disadvantages of this strategy are its operational complexity and expense along with a strong reliance on the accurate sensors readings. Moreover, the potential for either false alarms or missed leaks could lead to severe ecological damage and associated with the extensive financial consequences or just to avoidable expenses.

C. Application of Machine Learning Algorithms in Corrosion Prediction

While detecting corrosion is crucial for timely maintenance, it involves expensive and complex implementation, requires extensive personnel, and demands additional equipment. In contrast, predicting the corrosion rate offers a proactive and cost-effective approach for managing pipeline integrity. Recognising the potential of machine learning for such predictive tasks, a group of researchers demonstrated the use of machine learning methods, specifically random forest, to predict the marine atmospheric corrosion behaviour of low-alloy steels (Yan, et al., 2020). They used a database of corrosion data of steels exposed to marine environments and found that random forest algorithm can predict corrosion rates with high accuracy. The important factors identified included temperature, alloying elements, and humidity. However, while they show the models potential in corrosion prediction, they focus on the marine environments and low-alloy steels whereas the broader applicability to different environments and materials remains unexplored.

Another study explored support vector regression (SVR) combined with meta-heuristic algorithms to predict the maximum depth of pitting corrosion in oil and gas pipelines (Ben Seghier, et al., 2020). The researchers used SVR with genetic algorithm (GA), particle swarm optimization (PSO), and firefly algorithm (FFA) to optimize the models parameters. Applying these hybrid models they found that the SVR-FFA provided the most accurate predictions. They also emphasized the importance of soil properties and environmental factors in corrosion depth and showed that these hybrid models can outperform traditional empirical methods. But their study mainly targets the depth of pitting corrosion rather than the overall corrosion rate leaving a gap for approaches that address comprehensive corrosion rate predictions.

A study on applying deep learning techniques for automatic corrosion detection was presented by authors who developed a deep learning model capable of pixel-level segmentation of corrosion (Nash, et al., 2022). They introduced three Bayesian variants to provide uncertainty estimates at each pixel using a dataset of 225 images and achieved a significant accuracy with the Mask R-CNN model. Despite the good results the

reliance on a limited dataset presents challenges and leads to potential false positives and negatives in real-world applications. While the paper shows promise its focus is on detecting corrosion through image analysis, whereas the current study uses sensor data instead of images and focuses on predicting the corrosion rate. What can be beneficial in further studies is integrating real-time detection with future corrosion rate predictions for greater accuracy and efficient corrosion management.

Following the focus on deep learning for corrosion detection, a study by Oyedeji et al developed a multiple phase convolutional neural network (CNN) model to detect corrosion on metallic materials (Oyedeji, et al., 2023). Their model not only identifies the presence of corrosion but also determines its severity, stage, and exact location with high accuracy. The study uses binary classification, multi-label classification, and patch distribution algorithms on a dataset of 600 images and provides a detailed picture on corrosion characteristics. Integrating these detection capabilities with real-time predictive techniques for corrosion rates as proposed in the current study could solve both immediate detection with ongoing maintenance needs.

Expanding on that, a study of other researchers focuses on predicting pipeline corrosion using a long short-term memory (LSTM) neural network model (Sow & Ghazzali, 2024). It was designed specifically to handle temporal sequences in order to predict the evolution of pipeline wall thickness over time. The authors used a dataset from Quebec Metallurgy Centre to train and test the model and achieved an 80% accuracy in predicting the changes of thickness over a hundred days. Their approach differs from the current study which uses sensors data for corrosion prediction. This highlights the importance of exploring diverse and alternative data sources to enhance the use of the model.

Other group of authors focus on the application of ML in corrosion detection using various datasets (Daoudi, et al., 2024). They examined datasets like X-ray computed tomography (XCT) images, the corrosion and materials collection database (CAMCD), and a specialized dataset for cross-country pipeline inspection data analysis. Integrating different datasets, the review shows how machine learning models can achieve greater accuracy and reliability in detecting and predicting corrosion. This aligns with the current study's focus to use sensor data for corrosion prediction and contributes to better understanding of effective corrosion management techniques.

Building on this, a study on applying multi-sensor data fusion technology and wireless sensor network (WSN) in rebar corrosion adds to the point of using diverse sensor data for corrosion prediction (Yu, et al., 2023). The researchers selected five parameters such as chloride ion concentration, pH level, rebar corrosion potential, and internal temperature and humidity of concrete to increase the accuracy. Based on the measurements, the system provides a detailed analysis of the corrosion potential index and reinforcement corrosion rate. Similarly to this approach, the current study focuses on the temperature,

pressure, flow velocity, and pH level as they crucial in affecting the corrosion processes. Unlike the multifunctional sensor comprising various probes, this study proposes utilizing existing sensors to simplify the implementation process and reduce costs.

Prediction corrosion in the oil and gas industry is crucial for ensuring the safety of environment and protecting the expensive infrastructure that includes pipelines, storage tanks, and offshore platforms as emphasised in one of the recent studies (Odili, et al., 2024). Advanced technologies including AI-driven predictive models offer important benefits to proactive maintenance. Major oil companies can significantly improve asset reliability, reduce downtime, and lower maintenance costs. While the existing AI-powered solutions provide a good foundation, the current study aims to enhance them further using existing temperature transmitters, pressure transmitters and flow meters data for improved accuracy and practical application.

This research analysis revealed notable gaps in the deployment of machine learning strategies for early detection of corrosion specifically in using supervised learning algorithm with existing sensors data within the oil and gas sector. While traditional mechanisms offer adequate outcomes they also come with significant challenges, such as substantial expenses, production halts that lead to large financial damages, and measurement inconsistencies. On the other hand, more progressive ML-powered techniques offer compelling options but overlook the crucial need for early detection which is essential for preventing critical pipeline harm. This literature overview highlights the gap that the paper aims to address by deploying supervised learning algorithms like random forest, SVR, CNN, and LSTM for predicting early signs of corrosion based on the current sensors readings. Upcoming chapters will delve deeper into the methodology, implementation, and evaluation of the startegy.

3. Research Methodology

The research involved several steps starting from preliminary research and literature review followed by problem validation and leading to design, implementation, and evaluation of machine learning models to assess their effectiveness in predicting corrosion rate as shown on Figures 4.

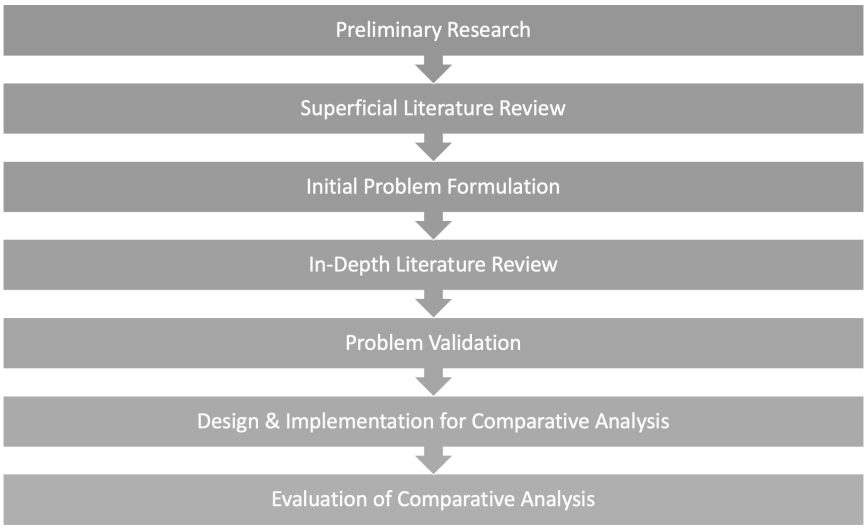

Figure 4. Research Methodology

The implementation of machine learning models included data collection, processing, features extraction, models training, evaluation of the results and comparison as shown on Figure 5.

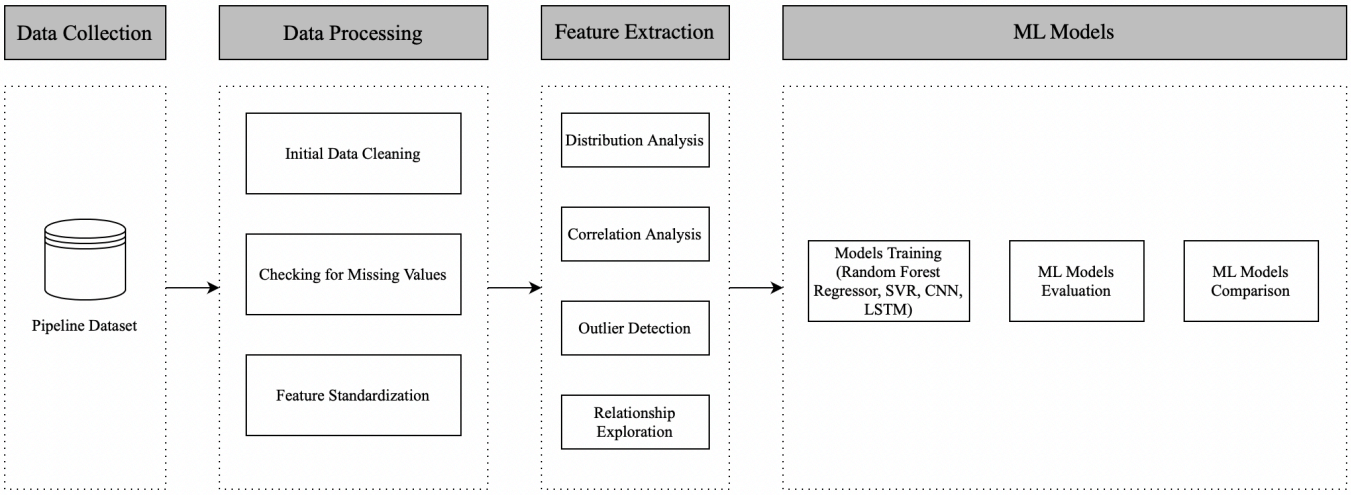

Figure 5. ML Experiment Steps

In the first step the dataset used for the study consists of various operational features of oil pipelines, such as temperature, flow velocity, internal pressure, and pH level (Khakzad & Khakzad, 2021).

Next step starts with cleaning data to remove irrelevant metadata columns and rows containing headers to ensure the data is clean and ready for analysis. Then, the preprocessing includes checking for missing values and since no missing values found the data is standardized using StandardScaler in order to ensure that all features contribute equally to the analysis and model training. After that, the preprocessing continues with analysing the distribution of each feature and examining the relationships between features. The analysis

of the dataset helps to understand the range and variability of each parameter. The main features like temperature, flow velocity, CO2 pressure and internal pressure showed consistent values with minimal variability which means that the dataset is relatively stable and does not have extreme outliers that could skew the analysis. The correlation analysis was conducted to examine the relationships between parameters indicating strong correlation between pH and CO2 pressure meaning that these variables potentially significantly influence each other (Figure 6). For instance, the changes in CO2 pressure can directly affect the pH level which in its turn can influence the corrosion process.

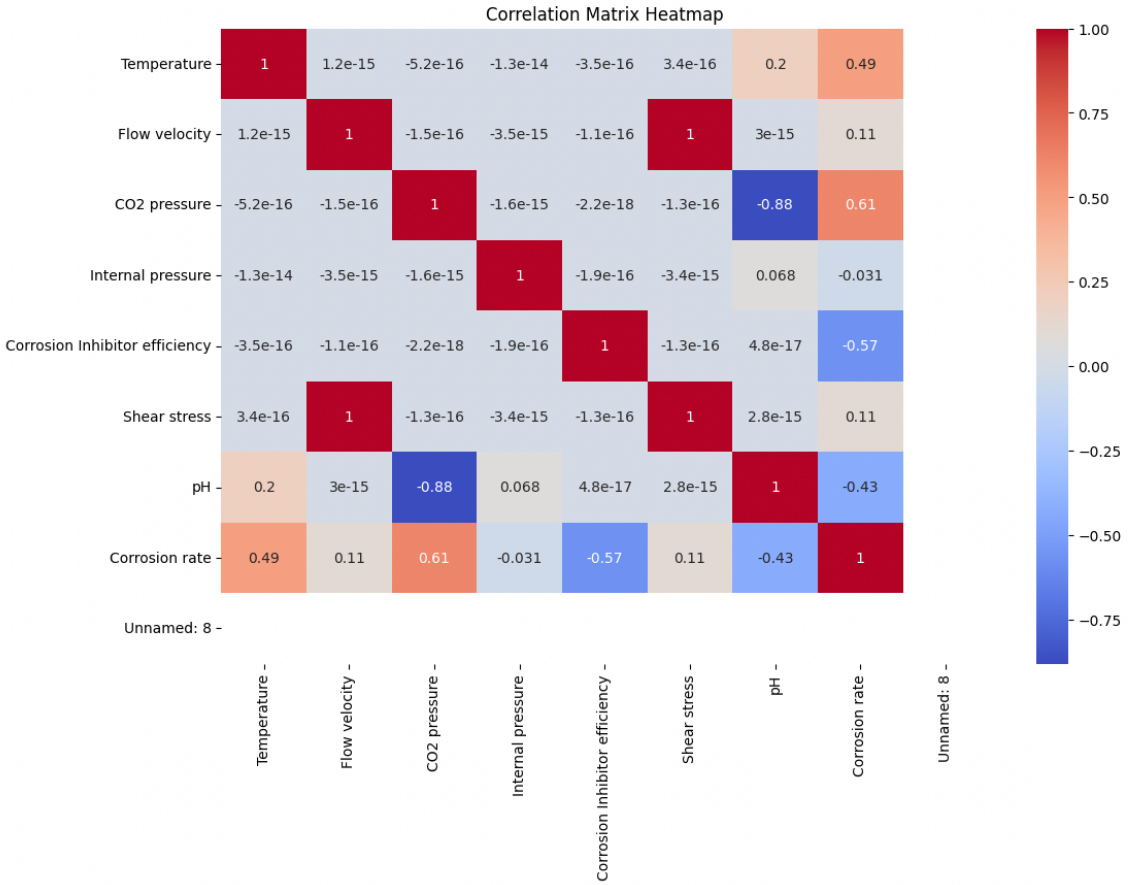

Figure 6. Correlation Matrix

Histograms plotted for better understanding of the distribution of each individual parameter illustrated the frequency of observations and the data central tendency and dispersion on Figure 7. For instance, the histogram of temperature showed an even distribution while the histogram of flow velocity demonstrated a greater frequency of certain flow velocity values and slightly skewed distribution.

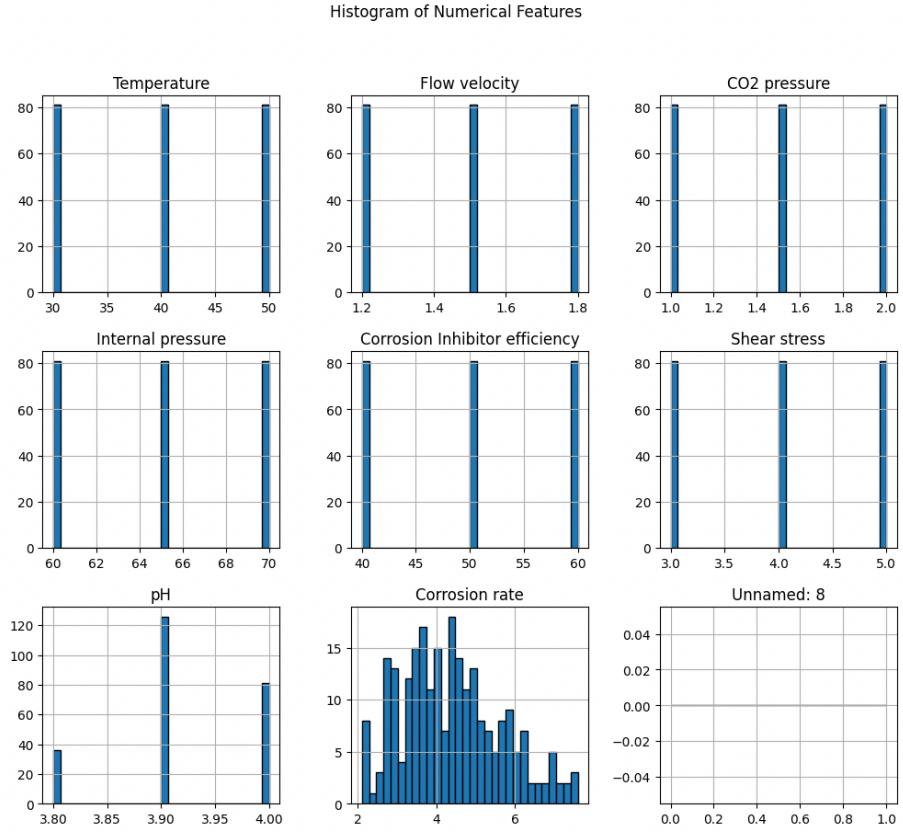

Figure 7. Histogram

Afterwards, the box plots were used to identify the outliers that are crucial to understand the impact on mean and standard deviation (Figure 8). Identifying and addressing outliers makes the analysis more robust and provide reliable results.

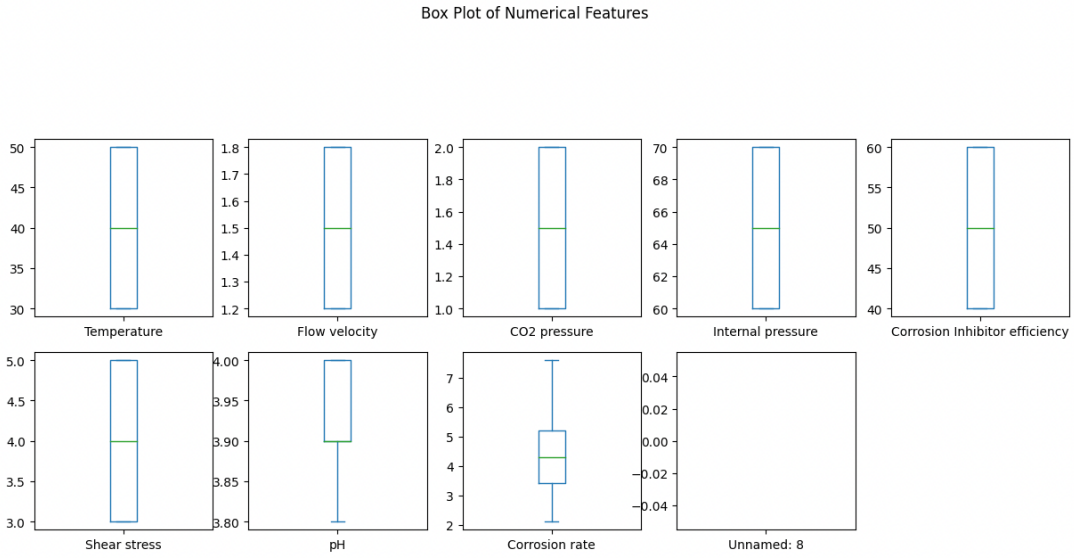

Figure 8. Box Plot

And finally, a pair plot was generated to explore the relationships between the parameters (Figure 9). It illustrated the scatter plots for each pair of features and the potential correlations and patterns. This visualization helps to identify any linear or non-linear relationships between the variables in order to ensure that all possible interactions are considered.

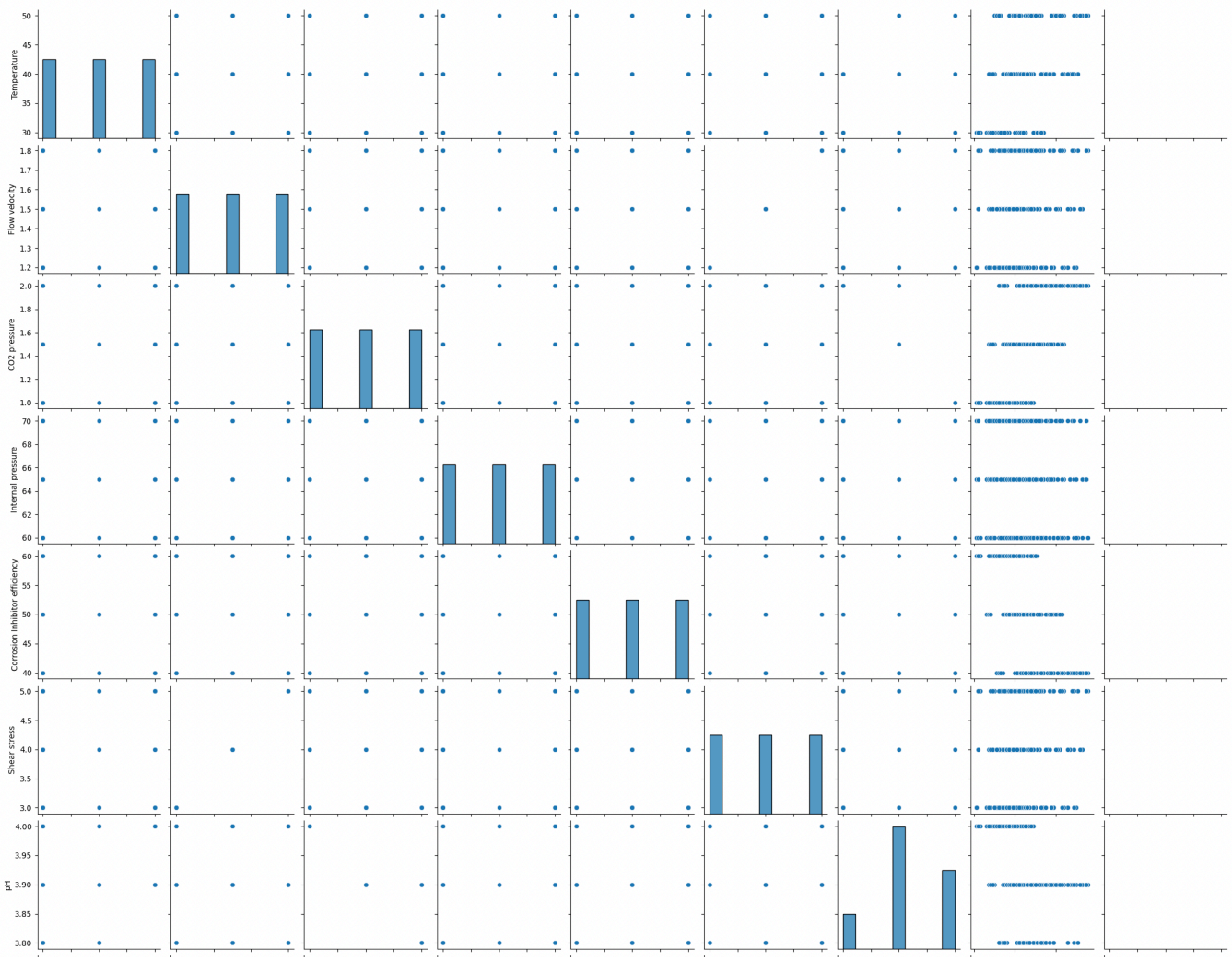

Figure 9. Pair Plot

After preprocessing, the dataset is ready for model training. The data was split into training and testing sets at the ratio 80-20 to evaluate the models performance on unseen data (Chaising, et al., 2023). This step ensures not only the effective training but also robust application to new data. The models used to train and evaluate the outcomes include random forest regression, support vector regressor (SVR), convolutional neural network (CNN), and long short-term memory network (LSTM).

Random forest regression is used to handle non-linear relationships and capture complex feature interactions (Sutaria & Jain, 2023). This method involves creating several decision trees during the training

phase and then combines their predictions to improve accuracy and prevent overfitting. It is suitable for this study due to its effectiveness in dealing with noisy data and its ability to handle complex relationships between features like temperature, flow velocity, and CO2 pressure. The model was trained on the preprocessed dataset and default parameters. The training involved fitting the model to the training data and using it to predict the target variable using the RandomForestRegressor class from the Scikit-learn library.

Support vector regressor predicts the target variable by fitting the data within a specified margin of tolerance. The SVR works well in high-dimensional spaces and suitable for regression tasks (Yang, et al., 2023). That is crucial in understanding how changes in conditions like pH and internal pressure impact the corrosion rate. The model by default uses a radial basis function (RBF) kernel to capture non-linear relationships. The implementation involved training the model on the preprocessed data, fitting it to the training data, and using to predict the target variable.

A simple convolutional neural network with one layer was implemented to detect patterns in the data (Patil & Rane, 2020). While CNN is typically used for image data it can also be applied to tabular data to extract local patterns and features which helps to identify how local changes in parameters like temperature and flow velocity affect corrosion rate. The model included a 1D convolutional layer with 64 filters and a kernel size of 2, then a flattening layer and a dense layer with one output neuron. It utilized the Adam optimizer for training and measured errors with the mean squared error (MSE) loss function. The process involved adjusting the input data to fit the CNN architecture and training the model for 10 epochs with a validation split of 20%.

Long short-term memory is a type of recurrent neural network that can identify patterns over time in sequential data (Kumar, 2023). In this study an LSTM model with two layers and an additional dense layer was used to analyze temporal patterns in the pipeline condition data. The architecture included two layers each with 50 units and a final dense layer with a single output neuron. The model employed the Adam optimizer along with the MSE loss function. The input data was reshaped to fit the models requirements and the training was done over 10 epochs with a validation split of 20%.

4. Experimental Results and Discussion

To evaluate the models performance a few key metrics were used such as mean absolute error (MAE), root mean squared error (RMSE), and R2 score (Wang, et al., 2023). MAE measures the average difference between predictions and the actual values. RMSE also measures the differences between predicted and actual values but squares them before averaging. This makes RMSE more sensitive to big errors. R2 score

shows how well the models predictions match the actual data indicating the proportion of variance in the corrosion rates. The results of the experiment are summarized on Figure 10.

Model	Train MAE	Validation MAE	Train RMSE	Validation RMSE	Train R2 Score	Validation R2 Score
Random Forest	0.230	0.313	0.326	0.451	0.891	0.819
SVM	0.294	0.354	0.395	0.484	0.841	0.802
CNN	0.099	0.318	0.141	0.419	0.977	0.835
LSTM	0.168	0.286	0.236	0.384	0.938	0.854

Figure 10. Results

The random forest regressor achieved a validation R2 score of 0.819 which means that the model can explain 81.9% of the variance in the data. It had reasonable MAE and RMSE values making the model a robust choice for the regression task. SVR had a slightly lower performance with a validation R2 score of 0.802 and higher MAE and RMSE values which suggests that SVR has a slightly higher prediction error. The CNN model showed promising results with validation R2 score of 0.835. Despite having the lowest train MAE it performed well on the validation set with reasonable MAE and RMSE values. The LSTM model outperformed all other models with the highest validation R2 score of 0.854 which indicates the best predictive capability. It also had the second lowest train MAE and validation MAE showing a strong ability to perform well on unseen data.

Additionally, three visuals were plotted to illustrate the models performance: actual vs. predicted values for all models, residual plots to show the distribution of errors, and training and validation loss for LSTM and CNN models.

The actual vs. predicted plots on Figure 11 reveal how much closely the models performance align with the true values. The random forest and SVR showed a strong correlation meaning high accuracy, while CNN and LSTM models were less consistent and provided more scattered points.

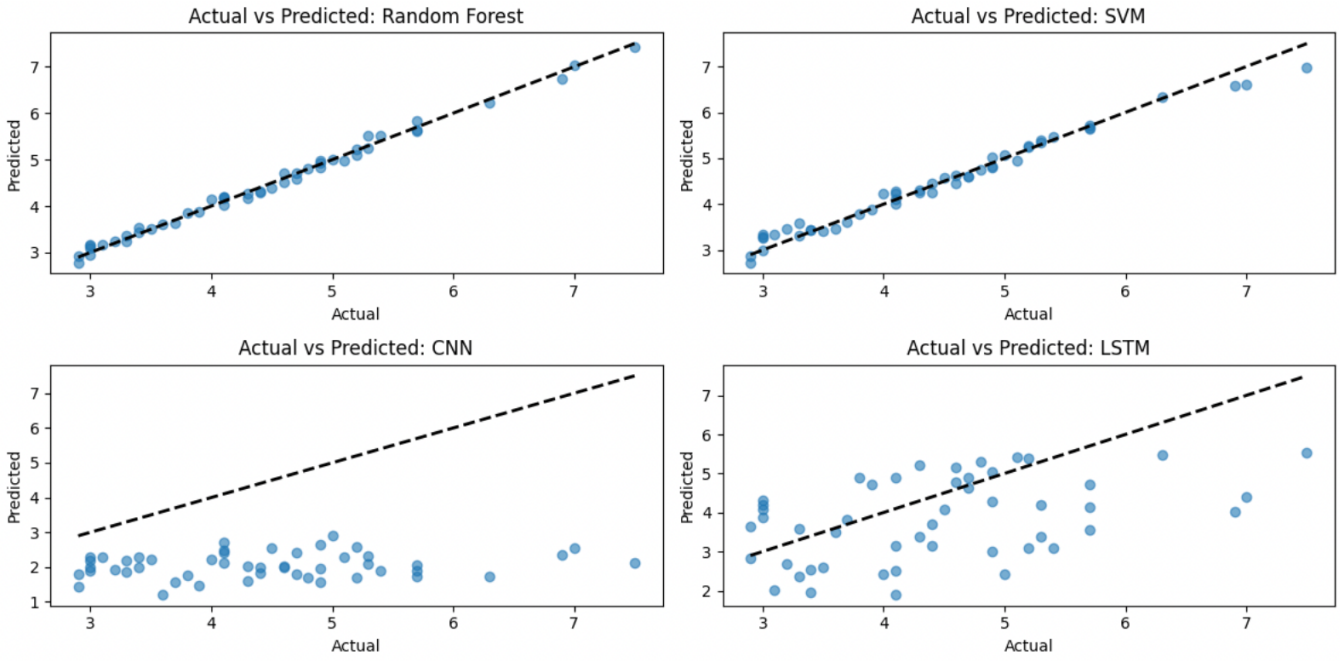

Figure 11. Actual vs. predicted values for all models

The residual plots on Figure 12 show the distribution of prediction errors. Random forest and SVR again indicated low prediction error, whereas the CNN and LSTM showed wider residual distributions. The CNN model had more errors at higher values which can mean underfitting and the LSTM indicated variability in its predictions and potentially overfitting.

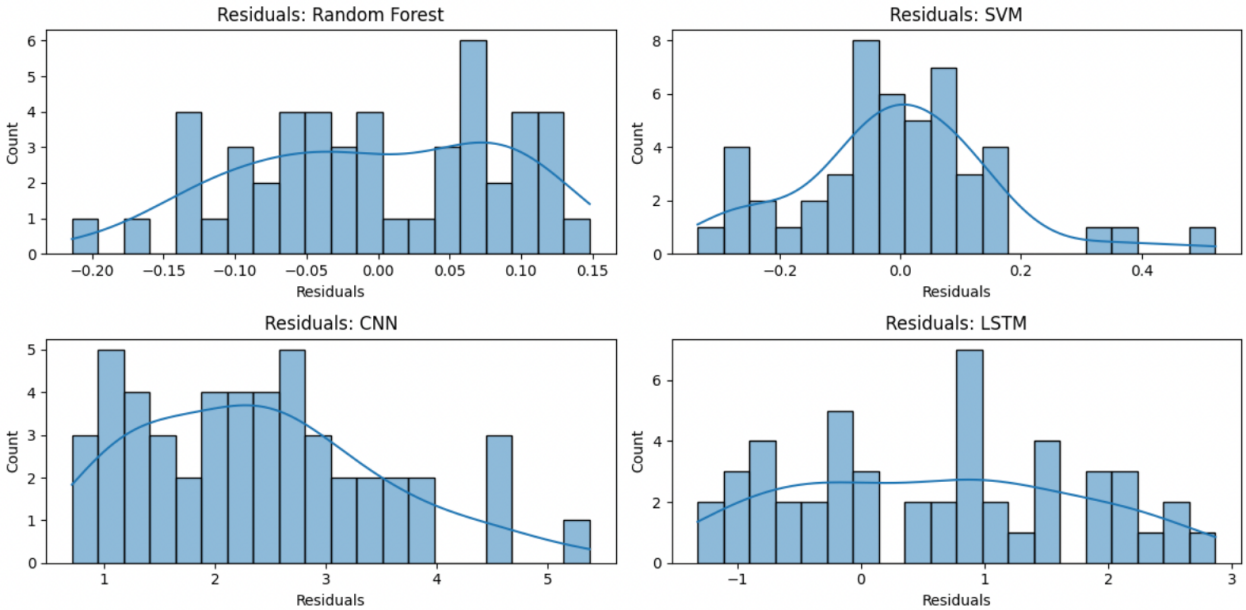

Figure 12. Residuals

The training and validation loss plots on Figures 13 track the learning processes of the CNN and LSTM models. For CNN both training and validation loss decreased meaning effective learning and overall good generalization. But the gap between the two suggests some overfitting. LSTM model showed a sharp decline for both training and validation and then stabilized indicating more robust training and better generalization compared to CNN.

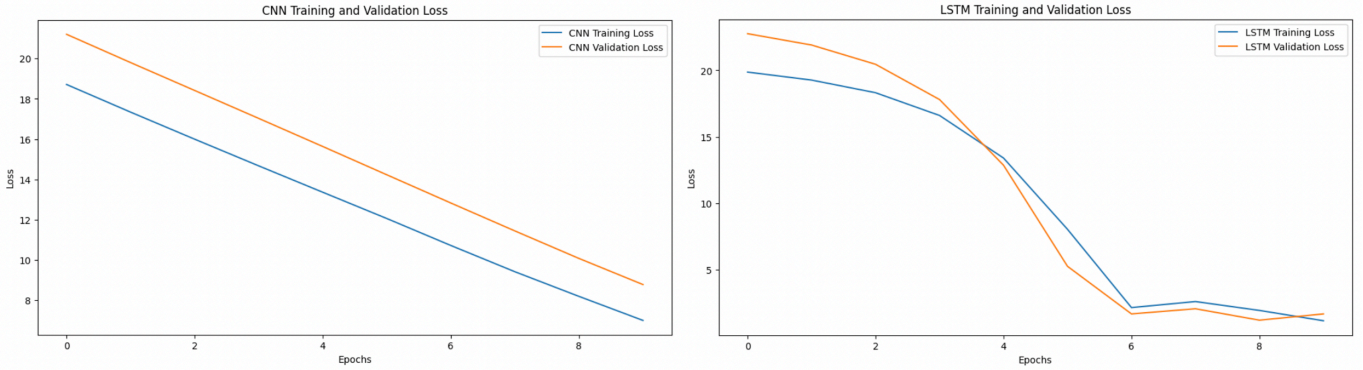

Figure 13. Training and Validation Loss

In this experiment, the LSTM model outperformed all other models in predictive performance with the highest validation R2 score. The CNN model also performed well closely following LSTM model. The random forest and SVR models showed reasonable performance but were slightly less effective than the neural network models. While the random forest model showed strong interpretability and consistency in residuals, the LSTM model accuracy and robust training and validation loss patterns make it the best choice for early corrosion detection. Its high predictive accuracy indicates that the model can effectively identify early signs of corrosion helping to take timely actions and avoid damages and operational disruptions.

5. Conclusion and Future Work

The aim of this study was to explore the feasibility, accuracy, and methodology for implementing machine learning algorithms to predict corrosion rates in oil refinery pipelines in order to contribute to the early corrosion detection. Through a comparative analysis of four supervised machine learning models that include random forest, support vector machine (SVM), convolutional neural network (CNN), and long short-term memory (LSTM) the study provided some understanding on their performance and applicability. The LSTM model demonstrated the highest predictive accuracy from all the models and confirmed its potential in identifying early signs of corrosion. The CNN model also showed strong performance but slightly less effective than LSTM. Random forest and SVR were good enough but were outperformed by

the neural network models in terms of predictive accuracy. The research also highlighted the effectiveness of using existing sensors data such as temperature, pressure, flow velocity, and pH level, for accurate and timely corrosion rate predictions. Using these parameters allows to perform real-time monitoring and avoid the high expenses on additional equipment, installation, and manpower. The findings underscored the importance of integrating advanced machine learning techniques into the broader corrosion management as their implementing can significantly improve the operations efficiency, reduce downtime, and enhance the safety.

Future works can focus on refining these models even further using more diverse data sources and exploring different environmental conditions and pipeline materials. Additionally, the real-world implementation and testing could be another important step in improving the models effectiveness.

Overall, the study contributes to the knowledge of applying the machine learning in industrial maintenance and provides a foundation for future advancements in early corrosion detection and management.

REFERENCES

Aljameel, S. S. et al., 2022. *An Anomaly Detection Model for Oil and Gas Pipelines Using Machine Learning.* [Online]
Available at: https://www.mdpi.com/2079-3197/10/8/138
[Accessed 9 April 2024].

Ameh, E. S., Ikpeseni, S. C. & Lawal, L. S., 2018. *A Review of Field Corrosion Control and Monitoring Techniques of the Upstream Oil and Gas Pipelines.* [Online]
Available at:
https://www.researchgate.net/publication/322935097_A_Review_of_Field_Corrosion_Control_and_Monitoring_Techniques_of_the_Upstream_Oil_and_Gas_Pipelines
[Accessed 3 April 2024].

Ben Seghier, M. E. A. et al., 2020. *Prediction of maximum pitting corrosion depth in oil and gas pipelines.* [Online]
Available at:
https://www.sciencedirect.com/science/article/pii/S1350630719318746?casa_token=LoznGsjCRB8AAAAA:xY0sf0jWgWWpDOGyV-MwTpc5jt7f_ieixeOu1kDkr31njq-379J0UHXjRxlCIUtzoTFybdeUg5r7IA
[Accessed 3 July 2024].

Chaising, S., Syukur, M. & Nithibandanseree, P., 2023. *Comparison of Machine Learning Algorithms for Prediction of Total Assets.* [Online]
Available at: https://ieeexplore.ieee.org/document/10051085/authors#authors
[Accessed 17 June 2024].

Chalgham, W., Wu, K.-Y. & Mosleh, A., 2020. *System-level prognosis and health monitoring modeling framework and software implementation for gas pipeline system integrity management.* [Online]
Available at: https://www.sciencedirect.com/science/article/pii/S1875510020305254?via=ihub#bib11
[Accessed 7 April 2024].

Choi, K.-H.et al., 2015. *Comparison of computational and analytical methods for evaluation of failure pressure of subsea pipelines containing internal and external corrosions.* [Online]
Available at: https://link.springer.com/article/10.1007/s00773-015-0359-5
[Accessed 3 April 2024].

Corbin, D. & Willson, E., 2007. *New Technology for Real-Time Corrosion Detection.* [Online]
Available at: https://www.dau.edu/sites/default/files/Migrated/CopDocuments/New%20Technology%20for%20Real-Time%20Corrosion%20Detection.pdf
[Accessed 7 April 2024].

Daoudi, N. et al., 2024. *Applications of Machine Learning in Corrosion Detection.* [Online]
Available at: https://ieeexplore.ieee.org/abstract/document/10541125/authors#authors
[Accessed 17 July 2024].

Elmrabit, N., Zhou, F., Li, F. & Zhou, H., 2020. *Evaluation of Machine Learning Algorithms for Anomaly Detection.* [Online]
Available at: https://ieeexplore.ieee.org/abstract/document/9138871/authors#authors
[Accessed 10 April 2023].

Feng, G. & Buyya, R., 2016. Maximum revenue-oriented resource allocation in cloud. *International Journal of Grid and Utility Computing,* 7(1), pp. 12-21.

GlobalData, 2023. *GlobalData.* [Online]
Available at: https://www.globaldata.com/store/report/oil-and-gas-pipelines-market-analysis/#:~:text=The%20total%20length%20of%20the,total%20length%20of%202%2C113%2C065%20km.
[Accessed 3 April 2024].

Khakzad, S. & Khakzad, N., 2021. *Simulation data for CO2 corrosion rate of oil pipeline.* [Online]
Available at: https://data.mendeley.com/datasets/4nydhxjymw/1
[Accessed 15 March 2024].

Kumar, A., 2023. *Stock Market Prediction using LSTM and Markov Chain Models: A Case Study of Royal Bank of Canada Stock.* [Online]
Available at: https://dspace.library.uvic.ca/items/322b5f0a-5903-48e5-ae97-2eae5b19269d/full
[Accessed 7 July 2024].

Kune, R. et al., 2016. The anatomy of big data computing. *Software—Practice & Experience,* 46(1), pp. 79-105.

Martinez , A. R. et al., 2019. *Design and Validation of an Articulated Sensor Carrier to Improve the Automatic Pipeline Inspection.* [Online]
Available at: https://www.researchgate.net/publication/331945363_Design_and_Validation_of_an_Articulated_Sensor_Carrier_to_Improve_the_Automatic_Pipeline_Inspection
[Accessed 5 April 2024].

Nash, W., Zheng, L. & Birbilis, N., 2022. *Deep learning corrosion detection with confidence.* [Online]
Available at: https://www.nature.com/articles/s41529-022-00232-6
[Accessed 5 July 2024].

Nassif, A. B., Talib, M. A., Nasir, Q. & Dakalbab, F. M., 2021. *Machine Learning for Anomaly Detection: A Systematic Review.* [Online]
Available at: https://ieeexplore.ieee.org/abstract/document/9439459/authors#authors
[Accessed 8 April 2024].

Odili, P. O. et al., 2024. *INTEGRATING ADVANCED TECHNOLOGIES IN CORROSION AND INSPECTION MANAGEMENT FOR OIL AND GAS OPERATIONS.* [Online]
Available at: https://fepbl.com/index.php/estj/article/view/835
[Accessed 13 July 2024].

Oyedeji, O. A., Khan, S. & Erkoyuncu, J. A., 2023. *Application of CNN for multiple phase corrosion identification and region detection.* [Online]
Available at: https://www.sciencedirect.com/science/article/pii/S1568494624007828?casa_token=2fipUJ8uV44AAAAA:pTYQYr0BC4DiqYXSOHkbFF0sPD47FFUzEoTqf-j2XZdTD6j_NyCvjPgoqjjwa_qeGJsmrUdCoNK9
[Accessed 5 July 2024].

Parjane, V. A., Arjariya, T. & Gangwar, M., 2023. *Corrosion Detection and Prediction for Underwater pipelines using IoT and Machine Learning Techniques.* [Online]
Available at: https://ijisae.org/index.php/IJISAE/article/view/2626
[Accessed 6 April 2024].

Patil, A. & Rane, M., 2020. *Convolutional Neural Networks: An Overview and Its Applications in Pattern Recognition.* [Online]
Available at: https://link.springer.com/chapter/10.1007/978-981-15-7078-0_3
[Accessed 3 July 2024].

Popoola, L. T. et al., 2013. *Corrosion problems during oil and gas production and its mitigation.* [Online]
Available at: https://link.springer.com/article/10.1186/2228-5547-4-35
[Accessed 3 April 2024].

Reddy, M. S. B. et al., 2021. *Sensors in advancing the capabilities of corrosion detection: A review.* [Online]
Available at: https://www.sciencedirect.com/science/article/pii/S0924424721005513
[Accessed 7 April 2024].

Soomro, A. A., Mokhtar, A. A., Kurnia, J. C. & Lu, H., 2021. *Deep Learning-Based Reliability Model for Oil and Gas Pipeline Subjected to Stress Corrosion Cracking: A Review and Concept.* [Online]
Available at: https://www.researchgate.net/publication/352165086_Deep_Learning-Based_Reliability_Model_for_Oil_and_Gas_Pipeline_Subjected_to_Stress_Corrosion_Cracking_A_Review_and_Concept
[Accessed 7 April 2024].

Sow, K. M. & Ghazzali, N., 2024. *Developing a predictive model using multivariate analysis and Long Short-Term Memory (LSTM) to assess corrosion degradation in mining pipeline thickness.* [Online]
Available at: file:///Users/mac/Downloads/FLAIRS_37_74.pdf
[Accessed 11 July 2024].

Statista, 2024. *Statista.* [Online]
Available at: https://www.statista.com/topics/1783/global-oil-industry-and-market/#topicOverview
[Accessed 3 April 2024].

Sutaria, R. & Jain, R., 2023. *Auto-Price Forecast: An Analysis of Car Value Trends.* [Online]
Available at: https://ieeexplore.ieee.org/document/10170263
[Accessed 3 July 2024].

Wang, Q. et al., 2023. *Evolution of corrosion prediction models for oil and gas pipelines: From empirical-driven to data-driven.* [Online]
Available at: https://www.sciencedirect.com/science/article/pii/S1350630723000511?casa_token=V9dK4QMF-v4AAAAA:CVvJnDLS63O5o32VIpJgvqE-2fb9bx1Hjwewv8LSoD--rIJsp5GeDb0q7M_WEg4OKmAesKhcv2HE
[Accessed 9 July 2024].

Xie, M. & Tian, Z., 2018. *A review on pipeline integrity management utilizing in-line inspection data.* [Online]
Available at: https://pdf.sciencedirectassets.com/271094/1-s2.0-S1350630718X00077/1-s2.0-S1350630717313067/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEAsaCXVzLWVhc3QtMSJHMEUCIFtE5i7GWpgaer14CtbqNYSYx0Yjsw19hZ8onf6G4MMDAiEAy7RwEJfQ%2BcjJJ3F7iudp9xjGuLud94AshkG2WJltIU
[Accessed 5 April 2024].

Yang, D. et al., 2023. *A Novel Pipeline Corrosion Monitoring Method Based on Piezoelectric Active Sensing and CNN.* [Online]
Available at: https://www.mdpi.com/1424-8220/23/2/855
[Accessed 9 April 2024].

Yang, L., Ma, H., Zhang, Y. & Li, S., 2023. *Research on energy consumption prediction of public buildings based on improved support vector machine.* [Online]
Available at: https://ieeexplore.ieee.org/document/10327420/authors#authors
[Accessed 7 July 2024].
Yan, L., Diao, Y., Lang, Z. & Gao, K., 2020. *Corrosion rate prediction and influencing factors evaluation of low-alloy steels in marine atmosphere using machine learning approach.* [Online]
Available at: https://www.tandfonline.com/doi/full/10.1080/14686996.2020.1746196#abstract
[Accessed 3 July 2024].
Yu, A., Shang, Z., Sun, H. & Kuang, H., 2023. *Research on reinforced corrosion monitoring system based on Multi-sensor data fusion and wireless sensor network.* [Online]
Available at: https://ieeexplore.ieee.org/document/10212580/authors#authors
[Accessed 17 July 2024].
Zakikhani, K., Zayed, T., Abdrabou, B. & Senouci, A., 2019. *Modeling Failure of Oil Pipelines. Journal of Performance of Constructed Facilities.* [Online]
Available at: https://doi.org/10.1061/(ASCE)CF.1943-5509.0001368
[Accessed 3 April 2024].