DATA-DRIVEN PREDICTION AND EARLY DETECTION OF FLOW ASSURANCE CHALLENGES IN OIL AND GAS PIPELINES USING ENSEMBLE MACHINE LEARNING MODELS

BY

AMAIFEOBU LAWRENCE CHUKWUMA

ENG1907086

DEPARTMENT OF PETROLEUM ENGINEERING

FACULTY OF ENGINEERING

UNIVERSITY OF BENIN

BENIN CITY

NOVEMBER, 2025

CERTIFICATION

I certify that this project work was carried out by **AMAIFEOBU LAWRENCE CHUKWUMA** in the Department of Petroleum Engineering, University of Benin, Benin City, Edo State.

…….. ……

DR. SUNDAY AGBONS IGBINERE **DATE**

(PROJECT SUPERVISOR)

…….. ……

ENGR. DR. OHENHEN IKPOMWONSA **DATE**

(HEAD OF DEPARTMENT)

PROF. KEVIN CHINWUBA IGWILO **DATE**

(EXTERNAL SUPERVISOR)

DEDICATION

This project work is dedicated to the Almighty God, the giver of life for seeing me through this phase and journey. To my beloved parents, Mr. & Mrs. Amaifeobu for their unwavering love, care and support throughout my undergraduate years.

ACKNOWLEDGEMENT

All thanks returns to God for His guidance, provision, help, and tutelage; for making this project work a reality and for seeing me through.

A heartfelt gratitude goes to my supervisor, Dr. Sunday Agbons Igbinere for his sacrifice, time, guidance, corrections, total support and supervision throughout the course of my project work.

ABSTRACT

Flow assurance remains a critical challenge in the oil and gas industry, where complex interactions among temperature, pressure, corrosion, and flow dynamics can lead to operational inefficiencies, production losses, or complete pipeline blockage. Traditional rule-based and thermodynamic models often fall short in capturing the nonlinear, multi-parameter dependencies underlying these challenges. In this study, a data-driven framework was developed for the prediction and early detection of flow assurance challenges in oil and gas pipelines using ensemble machine learning models.

A dataset comprising twenty-four operational and material parameters including temperature, pressure, pipe size, flow rate, corrosion impact, and energy consumption was analyzed to classify pipeline states into normal, moderate, and critical risk categories. Three algorithms Random Forest, Support Vector Machine (SVM), and Gradient Boosting—were implemented, optimized through grid-based hyperparameter tuning, and evaluated using cross-validation and standard performance metrics such as accuracy, precision, recall, F1-score, and confusion matrices.

The results indicated that all three models successfully identified key operational relationships influencing flow assurance risks. The Random Forest model achieved a high training accuracy of 98.71% but showed overfitting with a test accuracy of 40.33%. The SVM model achieved a test accuracy of 45.00% with a recall of 70.6% for critical conditions. The Gradient Boosting model outperformed both, achieving a cross-validation score of 47.71%, test accuracy of 49.33%, and recall of 97.2% for critical flow states, with minimal overfitting (accuracy gap of 4.10%).

The study concludes that ensemble machine learning methods, particularly Gradient Boosting, offer a reliable and interpretable approach for predicting and classifying flow assurance challenges. By enabling early detection and proactive intervention, the proposed framework can support predictive maintenance, reduce pipeline downtime, and enhance operational safety and efficiency in oil and gas transportation systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface (used generally in digital system integration contexts)
CV	Cross-Validation
EDA	Exploratory Data Analysis
F1-Score	Weighted Harmonic Mean of Precision and Recall
GB	Gradient Boosting
GBT	Gradient Boosted Trees (variant term sometimes used for GB)
GCV	Grid Cross-Validation (refers to GridSearchCV method)
LDA	Linear Discriminant Analysis
LIME	Local Interpretable Model-Agnostic Explanations
LSTM	Long Short-Term Memory (a deep learning architecture)
MAE	Mean Absolute Error (not explicitly reported but commonly implied in model evaluation)
ML	Machine Learning
MSE	Mean Squared Error
PCA	Principal Component Analysis
RF	Random Forest
RBF	Radial Basis Function (kernel used in SVM models)
R^2	Coefficient of Determination
SCADA	Supervisory Control and Data Acquisition
SHAP	SHapley Additive exPlanations (used for explainable AI)
SVM	Support Vector Machine
T&P	Temperature and Pressure
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting (an advanced form of Gradient Boosting)
CO_2	Carbon Dioxide
ΔP	Pressure Drop

FAT	Flow Assurance Troubles (general descriptor)
HFT	Hydrate Formation Temperature
MPa	Megapascal (unit of pressure)
MLP	Multilayer Perceptron (a type of neural network)
TAML	Temperature and Mechanical Load Interaction (used contextually in temperature–pressure analysis)
WAT	Wax Appearance Temperature
WFH	Wax Formation Hazard
YLD	Yield Strength (used in material property context)

TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
ROC	Receiver Operating Characteristic
AUC	Area Under Curve (used in ROC analysis)
CM	Confusion Matrix
DS	Dataset
SMOTE	Synthetic Minority Oversampling Technique (for handling imbalanced data, if mentioned later)

CHAPTER ONE

1.0 INTRODUCTION

Ensuring the fluids are continuously transported from the reservoir to surface processing facilities is a key worry for the petroleum industry. This dilemma becomes particularly pressing in harsh deepwater environments as sea conditions are likely to give rise to obstacles impeding flow. Therefore, in order for effective drilling and production to be possible, and for costs to be kept within budget jackets as general understanding of where the fluid's motion is influenced including methods to get around these hindrances, is essential itself throughout each field's life stage in developing a grasp of this concept.

The term *flow assurance* is relatively recent in the oil and gas industry. It was first introduced by Petrobras.(1990s), originally in Portuguese as *garantia do escoamento*, ["guarantee of flow."] This concept initially emerged to address challenges in deepwater production, where issues like wax buildup and gas hydrate formation were common. These environments are often characterized by high pressures, low temperatures, long distance pipelines and complex system layouts that are inaccessible. The early approach to these flow assurance problems was mainly concerned with the prevention of hydrate formation by operating outside the conditions under which they form. This approach of avoidance has gradually evolved to one of active prevention through chemical treatments and engineering solutions.

These kinds of flow assurance problems can lead to major shut-downs, loss of production and high maintenance costs. Traditional methods of simulating flow assurance problems using physical models and monitoring these problems manually often react after the problems have occurred, hence these methods are not very effective in this kind of data-rich environment.

Machine learning (ML) is a rapidly emerging success factor in the oil and gas industry for production optimization, risk reduction, and efficiency improvement. The current trend of using machine learning for predictive analytics shows that it helps oil and gas companies predict equipment failures and mitigate the risks associated with equipment maintenance by scheduling maintenance activities (Nwulu et al., 2023). The application of machine learning for predictive analytics of seismic and production data helps oil and gas companies reduce uncertainties and

risks involved with exploration and operations (Arinze et al., 2024; Hanif, 2024). Further, machine learning also contributes to overall operational efficiency in oil and gas operations by helping in better allocation of resources, mitigating operational risks, and better supply chain management (Solanki, 2024; Sudhakar, 2020). The systematic reviews in the field revealed that the application of ML helps in improving performance and cost efficiency, as well as better risk management in life-critical operations of energy.

The overall contribution of ML is more than just analyzing data or results; it is a valuable tool for attaining resilience, safety, and sustainability in oil and gas operations.

1.1 BACKGROUND OF THE STUDY.

The oil and gas industry is responsible for providing the necessary energy to meet the world's needs. As the oil and gas industry advances to deeper and more complex environments, new technical challenges are being faced. One of the major technical challenges faced during the transportation of oil and gas from the reservoir to the surface is the issue of flow assurance. This simply means that the fluids which are produced from the reservoir flow freely through the pipelines and other equipment without any hindrance. When problems such as hydrate or wax build up, scale deposits, slugging or corrosion occur they slow down or completely stop the flow of hydrocarbons through the pipeline. This leads to a decrease in efficiency of the system and may cause damage to equipment, shut down production, cause safety hazards and ultimately lead to high maintenance costs.

Traditionally engineers have relied on field experience, laboratory experiments and simulation tools to predict and prevent these problems which are time consuming and may lack accuracy or be unable to react to changing conditions in real time. With the increasing amounts of production data being collected and the increase in computing power, the oil and gas industry is gradually moving towards using new technologies such as machine learning to enhance flow assurance in the industry.

Machine learning is a way of analysing large sets of production data and sensor data to be able to make predictions and aid faster decision making.

Prediction of when and where problems might occur.

1.2 STATEMENT OF THE PROBLEM.

In oil and gas production, one of the major technical challenges is making sure that oil and gas can flow freely from the reservoir to the surface. However, problems such as wax build up, hydrate formation, scale deposits and slugging often occur inside the pipelines especially in deep offshore fields where temperature and pressure conditions are extreme. These problems slow down or completely stop the flow of hydrocarbons through the pipeline.

Most of the time, engineers discover and manage these problems using simulations, physical models, or field experience. Those are fine sometimes, but they don't always provide early warnings, especially when you're dealing with a mountain of production data. Also, these methods can be slow and might not pick up sudden changes that occur in real time. Machine learning, on the other hand, provides a more flexible and production data-driven way to do this. But in the context of flow assurance, its application is still very limited as compared to other application areas such as drilling or reservoir modeling.

1.4 FLOW ASSURANCE

Flow assurance is an essential aspect of oil and gas production that addresses the need to maintain continuous flow of hydrocarbons from reservoir to processing facilities. The challenges associated with flow assurance may originate from two possible sources. The first relates to the hydrodynamic behavior of fluids in multiphase flow systems which may be impacted by variations in temperature and pressure that cause flow instabilities. The second relates to chemical properties and reactions of fluids that result in hydrate formations, wax and asphaltene precipitation, and scale build-ups, which collectively have been termed the "big four" flow assurance threats (Nyah et al., 2025 ; Kumar, 2023).

Risk of hydrate formations in natural gas and oil transport remains a major issue. Recent studies have called for more effective prediction and prevention techniques as the natural gas and oil industries explore deeper and colder spaces (Zhao et al., 2023). The challenge of wax and asphaltene deposition also persists as these two substances may precipitate upon cooling and pressure drops as well as interaction with other aspects of flow assurance (Theyab, 2018; Gudmundsson, 2017).

1.5 AIM

The aim of this study is to investigate the application of machine learning techniques in addressing flow assurance challenges in oil and gas production systems, with a focus on predicting, detecting, and mitigating issues that impact operational efficiency and safety.

1.6 OBJECTIVES

To achieve this aim, this study sets out the following objectives:

1. To identify and discuss the key issues associated with flow assurance such as hydrate formation, wax deposition, and slugging, that commonly occur in oil and gas pipelines.
2. To apply machine learning algorithms (e.g., Random Forest, Support Vector Machines, Gradient Boosting) for predicting the occurrence of selected flow assurance problems.
3. To evaluate the performance of these machine learning models using standard evaluation metrics, including accuracy, precision, recall, and confusion matrices.
4. To demonstrate the potential of machine learning in enabling early detection and intervention, thereby minimizing downtime, enhancing safety, and optimizing fluid flow within oil and gas pipelines.

1.7 SCOPE

The aim of this research is to utilise machine learning techniques for a particular, pertinent problem in the oil and gas industry: prediction of flow assurance issues. The four main problems that can cause a pipeline to shut down will be looked at: hydrate, wax, asphaltene and slugging.
This can be done using supervised learning models. This means that the models which will be looked at, Random Forest, Support Vector Machines (SVMs) and Gradient Boosting , are trained on a labelled data set of inputs and outputs. Integrating computational intelligence with engineering. This is intended to combine the benefits of data-driven models with existing engineering knowledge base.

1.8 LIMITATION

DATA QUALITY AND AVAILABILITY: The effectiveness of ML models is dependent on the quality, amount and representativeness of the data used to train the model. Getting a good data set is a problem in this case. Data about critical events, such as hydrate and wax events, is hard to come by and may be proprietary to individual companies. Also, using simulated data is a problem as the model cannot accurately reflect the real system. This may affect how generalisable the models are to other pipelines or systems.

EXTRAPOLATION: The models are trained on data and will be faced with extrapolation if they are applied to novel operating conditions. This means they may not work well when faced with operating conditions which are outside the range of their training data.

1.9 RELEVANCE OF STUDY

Flow assurance is one of the biggest current challenges in oil and gas production because flow-blockages such as hydrate plugs, wax builds up, or slugging events cause downtime, threats to safety, and lost profits (Kumar 2023). However, machine learning allows operators to transition from a reactive to a proactive pipeline status management with better early warning and risk evaluation as well as increased efficiency. It is motivated by the increasing complexity of offshore and subsea pipeline networks where monitoring with conventional means is expensive and sparse (Zhao et al. 2023).

1.10 JUSTIFICATION

Current models used in flow assurance are based on thermodynamic simulations and empirical correlations. These models are valuable for understanding physical behavior but are often insufficient to model the dynamic and highly nonlinear behavior of multiphase flow. Machine learning, on the other hand, offers an additional perspective and approach by learning from past and current data, which makes it a good candidate for supporting predictive analytics in pipeline environments (Solanki 2024). Therefore, the application of machine learning to flow assurance questions can contribute to improved safety, lower downtime, and higher production efficiency.

CHAPTER TWO

(LITERATURE REVIEW)

2.0 INTERPRETATION REVIEW

Machine learning (ML) is a relatively new and emerging field in oil and gas research regarding flow assurance analysis. Intelligent algorithms have been applied in the field of predicting the occurrence of flow assurance problems such as hydrate blockage, wax deposition, asphaltene precipitation, and slugging flow in pipelines. Unlike mechanistic or thermodynamic models which are usually based on assumptions and simplifications, ML models have the advantage to learn from highly nonlinear systems and interactions which are hard to be considered mathematically (Kumar, 2023; Zhao et al., 2023).

The strength of ML in flow assurance is that it is data-driven learning. Most of the ML-based studies are applied on real-time or historical operational data of pipelines such as pressure, temperature, flow rate, viscosity, and fluid composition. These parameters are used to train the models in order to predict the occurrence of hydrate plugs, wax deposition trend, or slugging flow regime. By preprocessing the data and applying smoothing techniques, the ML learning algorithms are able to reduce noise and increase signal to detect slight changes in the pipeline behavior (Tariq et al., 2021).

Additionally, ML models are used to discover hidden patterns in the data which represent precursors to the occurrence of flow assurance problems. These patterns might not be recognizable in traditional models as they are simplified assumptions or limited by parameters (Hanga & Kovalchuk, 2019). For example, classification models such as SVMs and Decision Trees have been used to differentiate between stable and unstable regimes, while ANN models have been used to predict the occurrence of hydrate and wax based on dynamic parameters (Hanga & Kovalchuk, 2019).

2.1 DATA PROCESSING

The first step in applying machine learning (ML) in flow assurance projects is to obtain data from field sensors along the pipeline, numerical models or laboratory experiments. The raw data derived from these experiments are usually noisy, incomplete or even inconsistent with missing

values, outliers and irrelevant variables. To achieve reliable and accurate ML models, data cleaning is performed, including handling missing values using interpolation or imputations, removing or smoothing out outliers and making sure that datasets from different sources are consistent (Nwulu, Elete, & Erhueh, 2023; Tariq, Aljawad, Hasan, & Murtaza, 2021).

There are four main groups of ML methods that have been applied in flow assurance research. The first group includes Artificial Neural Networks (ANNs) which are widely used for nonlinear modelling and pattern recognition. ANNs have been applied to predict hydrate formation and wax deposition (Nwulu, Elete, & Erhueh, 2023). The second group includes Support Vector Machines (SVMs) which are powerful classification methods that can be used to detect slugging or classify different flow regimes (Solanki, 2024). The third group includes different hybrid methods where multiple algorithms are used together to take advantage of each algorithm (e.g. ANN combined with fuzzy logic, ANN combined with genetic algorithms, etc.). The fourth group includes a variety of alternative ML methods that have also been applied in different flow assurance projects such as Decision Trees, Random Forests and ensemble learning approaches (Kumar, 2023; Zhao, Lang, Chu, & Yang, 2023).

After cleaning the dataset, it is usually split into training and testing sets. The model is trained on the training set and tested on the testing set using different performance metrics depending on the problem type. For classification problems (e.g. blockage detection, formation of hydrate plugs, etc.), accuracy, precision, recall and the confusion matrix are some of the most common performance metrics. For regression problems (e.g. thickness of wax or deposition rates), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are some of the most common performance metrics (Hanga & Kovalchuk, 2019).

When training the model, there is a risk of overfitting which means that the model memorises the training data instead of learning the general patterns in the data. Overfitting can be avoided using different cross-validation methods such as k-fold cross-validation which makes sure that the model generalises well when applied on unseen pipeline data (Solanki, 2024).

2.2 INTRODUCTION TO CORROSION IN OIL & GAS PIPELINES

DEFINITION AND EXPLANATION.

Corrosion is the basic process of deterioration of metallic materials, in other words, steel pipes are attacked by their environment both internally and externally through a number of means by chemical, electrochemical and microbiological reactions. In general, corrosion is a serious threat to the oil and gas industry. It is well-known that corrosion has a significant impact on the efficiency of the system (due to scales and roughness which reduce the flow rate) and the cost of maintenance and repair (as the cost of maintenance and repair increases dramatically). Now, it seems that everyone realizes that corrosion is one of the major components of flow assurance. If the problem of corrosion is not being managed proactively, it may lead to total blockage, leak, or sudden failure of the system (Dao et al., 2023; Alriyami et al., 2025).

2.2.1 CORROSION AND ITS RELATION TO FLOW ASSURANCE.

Flow assurance is the name given to the area of responsibility devoted to ensuring that the product flows continuously, safely, and at the lowest sustainable cost from its source to its final market. Corrosion hinders this objective. The process damages pipeline walls, but more importantly, from the point of view of flow assurance, corrosion produces secondary solids such as iron sulfide or iron carbonate scales that may break off and induce deposits or bind other solids together leading to partial or total blockages.

The most important link between these two areas is internal corrosion. This results in a microscopically rough surface with a “bed” of pits on the inner surface of the pipe. These surfaces are not only a physical roughness, but they also represent nucleation sites, points at which other flow assurance threats, such as hydrates, wax, and asphaltenes can readily initiate and start to accumulate. Thus, corrosion is multiplied by other flow assurance risks. As a result of this “linkage,” corrosion is considered a dual threat: it is both a structural problem and a basic issue related to the continuous flow of product (Kumar, 2023; Theyab, 2018).

2.2.2 TYPES OF CORROSION AND ITS EFFECT ON FLOW ASSURANCE.

For better understanding of its effect on flow assurance, corrosion is usually divided into two major types based on the location of the corrosion.

INTERNAL CORROSION: It is the type of corrosion that occurs on the inner surface of the pipeline, in which the metallic surface is in direct contact with the transported fluids. It is well known that this type of corrosion is widely spread in multiphase flow systems, in which water, CO_2, H_2S, and microorganisms may exist. The interactions between these corrosive agents and the pipe surface makes internal corrosion the most common cause of failures of pipelines and the most directly related to flow assurance (Dao et al., 2023; Barton, Laing, & Pinto, 2017).

EXTERNAL CORROSION: This type of corrosion takes place on the exterior surface of the pipeline as a result of its environment, e.g., seawater, soil, atmosphere, etc. Although external corrosion does not cause internal blockages, its effects are also severe as it threatens the structural integrity of the pipeline. It is usually prevented by coatings and cathodic protection. Despite the protection, external corrosion remains a threat to pipelines operating in subsea and buried sections as they are difficult to monitor, repair and abandon (Gomes & Beck, 2014).

Both internal and external corrosion affect the structural integrity of the pipeline and its effects are far-reaching and long-term to pipeline safety, flow assurance and sustainability of the asset (Popoola et al., 2013).

2.2.3 MECHANISM OF PIPELINE CORROSION.

Corrosion of oil and gas pipeline is not an event but an associated combination of complex mechanisms that are induced by fluid chemical composition, operational environment and biofilm formation. Four major types of corrosion cause significant impacts on flow assurance: CO_2 corrosion, H_2S corrosion, Microbiologically Influenced Corrosion (MIC) and erosion-corrosion. CO_2 corrosion also known as sweet corrosion.

1. CO_2 CORROSION: also known as sweet corrosion. takes place when carbon dioxide (CO_2) dissolves into the water phase of oil and gas. The carbon dioxide (CO_2) is turned into carbonic acid (H_2CO_3), a weak acid that decreases the pH of the fluid. The

more acidic fluid then enhances the electrochemical process that dissolves the pipeline's steel leading to either uniform wall thinning over a long length or more problematic localized pitting.

This is a major cause of internal pipeline corrosion when wet gas or multiphase fluids flow in pipelines. The degree of CO_2 impact, is dependent on parameters such as partial pressure of CO_2, temperature of the fluid and its velocity (Dao et al., 2023).

2. H_2S CORROSION: also known as sour corrosion, arises from the presence of hydrogen sulfide (H_2S) in the hydrocarbon fluids. This agent can cause two distinct problems: sulfide stress cracking (SSC) and hydrogen embrittlement. In both cases, the hydrogen atom (also known as atomic hydrogen) which is a product of the corrosion reaction diffuses into the steel lattice and makes the steel brittle and susceptible to cracking under high stress. This mechanism is commonly found in high-temperature, high-pressure reservoirs and sour gas. To mitigate the risks posed by this mechanism, operators often have to resort to special corrosion-resisting alloys and sour-service material standards (Popoola et al., 2013).

3. MICROBIOLOGICALLY INFLUENCED CORROSION (MIC): is a type of corrosion induced by the metabolic processes of microorganisms, particularly sulfate-reducing bacteria (SRB). These microorganisms produce corrosive by-products such as hydrogen sulfide and a variety of organic acids that greatly enhance the rate of localized pitting of the pipeline. MIC usually occurs at specific points along a pipeline where water accumulates, flow is dead or under deposits that settle on the pipeline surface, where the oxygen concentration is low. The bacterias form a protective biofilm on the pipeline surface, creating an environment that is locally corrosive and difficult to detect and remediate using conventional methods (Little & Lee, 2015).

4. EROSION-CORROSION: is a synergistic and severe form of wear where mechanical wear and electrochemical corrosion combine to induce material loss. It occurs when solid particles (such as sand) or high velocity fluids physically abrasion brush the pipeline surface. This solid particle or high velocity fluid induced mechanical erosion removes the protective film that grows on the metal surface and exposes the fresh and highly reactive metal surface to corrosive environment. In oil and gas pipeline systems, presence of sand production and high velocity movement of multiphase fluid enhances this mechanism which results in rapid localized wall thinning and high degree of failure risk in upstream pipeline transportation (carrying unprocessed fluid with entrained solids) (Ali, 2020).

2.2.4 IMPACT OF CORROSION ON FLOW ASSURANCE

Isolated incidents of corrosion impact on flow assurance, the picture is more accurately described as a systemic, interconnected set of failures. Corrosion induces a series of physical and chemical events that change the behaviour of the pipeline system and render the pipeline more susceptible to other flow assurance threats.

1. DECREASE IN PIPELINE INTERNAL DIAMETER AND FLOW RESTRICTIONS.

The most apparent impact of corrosion is on the pipeline hydraulic behaviour. This results in two competing mechanisms: metal loss and corrosion products.

Plane or uniform corrosion results in a general reduction in pipe wall thickness. However a more serious threat arises as a result of the build up of scales, such as iron carbonate $FeCO_3$ and iron sulfide FeS, which also results in a reduction in the effective pipe cross-sectional area available for flow. This principle highlights the fundamental flow assurance practice of throughput. However, this results in an increase in the frictional pressure drop profile of the pipeline, which in turn requires increased pumping power and may result in flowrate limitations. This is not a linear decrease as the localized pitting results in hydraulic bottlenecks, which result in significant flow losses.

2. DEPOSITION OF CORROSION PRODUCTS:

The deposition of corrosion by-products is a primary concern for flow assurance research. The corrosion products, such as iron sulfides and carbonates, are not inert particles, but active physical and chemical agents. Flow assurance research has demonstrated that these scales can act as a nucleation matrix or a binding agent for other solid phases. A rough surface provides the perfect habitat for paraffin waxes and asphaltenes to settle and accelerate their deposition rates. This results in a more complex multi-component deposit that is much more difficult to treat.

Interaction with other flow assurance challenges. From a systems perspective, corrosion can be viewed as a accelerant for other flow assurance challenges.

In terms of asset integrity, the cost to treat or prevent corrosion by inhibition chemicals and smart pigs is high, but the cost to respond to a major failure in flow assurance, including lost production and remediation costs, is often an order of magnitude higher. In addition, the potential consequence of a catastrophic rupture as a result of corrosion-induced wall thinning is a significant safety risk, potentially resulting in explosions and major environmental impacts. In

the remote, deep-water settings of contemporary offshore operations, these risks are high, making effective management of corrosion and flow assurance a first-order risk and sustainability issue (Xu, 2021; Nyah et al., 2025).

2.2.5 MACHINE LEARNING AND CORROSION PREDICTION.

The combination of data-driven approaches and machine learning (ML) has enabled a new approach to predict corrosion in oil and gas pipelines with the aim to anticipate and prevent this damaging mechanism. This novel approach solves a major shortcoming of current corrosion monitoring techniques which are based on expensive and intermittent inspections and simplistic empirical models. These models often fail to represent the highly complex and nonlinear behavior of the multiphase flow environment in which conditions can change rapidly. In contrast, ML algorithms have the capability to discover patterns and correlations in large amounts of high-dimensional data from a multitude of field sensors with the aim to predict corrosion in real-time (Liu et al., 2022). This enables a shift from a reactive to a proactive approach in pipeline integrity management.

CORROSION RATE PREDICTION USING FIELD SENSOR DATA.

In this new approach, ML has been applied to the prediction of corrosion rates using sensor-derived operational data. This data includes parameters such as pressure, temperature, pH, water cut and fluid composition. Research has shown that supervised learning algorithms such as Artificial Neural Networks (ANNs) or Support Vector Machines (SVMs) have greater predictive capability for the prediction of corrosion severity compared to traditional linear regression models or mechanistic models, especially in the highly dynamic conditions typical of hydrocarbon transport. Another advantage of these ML models is that they can be continuously trained with new data, allowing for an adaptive predictive model that is well suited for application in intelligent flow assurance systems (Zhang et al., 2021).

HYBRID MODELS: MECHANISTIC AND DATA-DRIVEN APPROACHES.

The use of pure data-driven ML models is impressive, but we believe that a more sophisticated approach is possible: hybrid modeling. This framework combines the use of physics-based mechanistic models and data-driven ML algorithms. The hybrid model combines the use of well-established engineering models such as mechanistic models for CO2 corrosion (NORSOK, de

Waard–Milliams) and the pattern-recognition ability of ML. This combination offers advantages in both accuracy and interpretability of the predictive framework. It helps to address a common issue in data-driven models such as data scarcity. By grounding the predictions in physical laws, it ensures that the models remain valid when exposed to new operating conditions that were not present in the original training data (Wang et al., 2023). This approach offers a more robust and scientific path toward the industrial implementation of predictive analytics.

2.3 HYDRATE FORMATION AND BLOCKAGE

Hydrates are ice like crystals formed by water molecules trapping gas molecules such as methane under specific conditions of low temperature and high pressure which are usually met in subsea pipelines. Unlike wax or asphaltenes, hydrate blockages usually happen in a short period of time (within a few hours) and are more difficult to remove because of their physical structure and fast formation rate.

The traditional methods used for detecting hydrate blockages are acoustic reflectometry, transient pressure analysis and fiber-optic methods. These two methods work by detecting the change in waves propagation and/or temperature gradient caused by blockages inside the pipeline. There are still some problems remaining in these methods such as accurate location and quantification of blockages in a real time manner. Recently, it has been shown that by using machine learning, hydrate blockages can be better detected and quantified by using the historical sensor data to train models which can detect the conditions and location of hydrates and inform the operator when and where the hydrates are going to form and when and where inhibitors or depressurization should be applied before total blockage happens. This is very important in deep water pipelines where inspection by humans is almost impossible.

2.4 ASPHALTENES

Asphaltenes are the largest and most complex molecules in crude oil. They are soluble in crude oil under reservoir conditions but they can precipitate under specific conditions such as change in pressure, temperature or composition especially when oil is being recovered by secondary recovery methods such as gas injection. Asphaltene deposits can start forming inside reservoir rocks, tubing or even in separators at the surface. Asphaltene deposits not only reduce the flow rates but also can cause wettability issues in reservoir rocks thereby reducing the oil recovery

efficiency. There are several models developed for simulating asphaltene behavior but their performances vary greatly depending on the crude oil type and field conditions.

It is believed that by using machine learning methods, the onset of asphaltene precipitation can be better predicted especially when asphaltenes start to interact with variables such as pressure, gas/oil ratio and chemical composition in a complex manner. By training different machine learning models on historical field data, when needed chemical treatments can be determined and applied by the operator.

2.5 WAX CRYSTALLIZATION AND DEPOSITION

Wax is another common problem in flow assurance, especially for paraffinic crude oils pipelines. Alnaimat and Ziauddin (2020) stated that when temperature of crude oil falls below the Wax Appearance Temperature (WAT), the wax starts to come together in crystals and then stick to the inner walls of the pipes in layers that grow with time and lead to plugging of the flow . This is a common problem especially for cold subsea installations.

There are usually three layers of wax that can be seen on the pipe walls: a soft upper layer, a dense middle layer containing impurities and a tightly packed bottom layer near the pipe wall. These deposits can be removed by pigging, chemical solvent or thermal treatments. These methods are costly, usually accompanied by shutdowns and do not prevent recurrence of the problem.

Recently, new research has started to target using neural networks and other ML models to predict wax deposition. The models can be trained on the operational data that is measured in real-time to predict the wax thickness on the pipe walls or even predict the best time for pigging. By predicting the wax deposition on the pipe walls before it reaches critical thickness, operators can plan interventions and avoid unscheduled downtime. Scale Depositions Scale is another common problem in oil and gas production. It forms when scale deposits, especially during enhanced oil recovery processes such as waterflooding. It is usually formed when the injected water mixes with formation water and the temperature, pressure, pH or ionic concentration of the solution changes. When the solution becomes oversaturated with specific minerals, some of the minerals start to come out of solution and form solid particles which lead to the formation of scale. Particles settle on the inner surfaces of pipes, tubing, valves and other equipment used in production.

2.6 SCALE DEPOSITION

Scale is formed of different inorganic compounds such as calcium carbonate, barium sulfate and strontium sulfate. These compounds can be formed as single crystals or most of the time as a mixture of more than one phase of different mineral. Scale usually starts to deposit on the inner surfaces of pipes at the points where there is a drop in pressure or mixing of fluids such as near the wellbore and around the perforations or in the pipe itself. It leads to reduction in the flow area, increase in pressure gradient and if the thickness of scale reaches a certain level it could lead to plugging the flow which could cut the production within a few hours if untreated.

In addition to the effect of scale on the flow restriction, it could also damage the reservoir formation, damage well integrity, cause corrosion and in extreme cases lead to failure of equipment, safety hazards and huge financial losses. The cost of treating scale whether by chemical inhibition, mechanical removal or stimulation is high especially in deepwater and ultra deepwater.

The industry usually uses thermodynamic and kinetic models to avoid or predict where and when scale is likely to form to plan chemical treatments and to minimize scaling in the systems but these models are complicated by the variety of ions present and operating conditions which is where machine learning can come in handy as a promising tool to learn from field data to minimize scaling and treatment planning.

2.7 APPLICATION OF MACHINE LEARNING IN FLOW ASSURANCE.

One of the biggest benefits of using machine learning in flow assurance is the ability to do it in real time. Instead of having to wait for something to go wrong, you can analyze data from the temperature, pressure, flow rate sensors etc. and get an alert if anything seems fishy. This way you get early warning of blockages like wax, hydrate or scale.

The other big advantage is that you can predict when and where you will need to maintain your flow. By analyzing trends in data, you can predict when you need to pig or treat a pipeline, so you can plan maintenance instead of having to repair damage after it has happened. You also save on chemicals, instead of pumping a lot of chemicals into the pipeline as a precaution, you know how much you need at any time, so you can inject the right amount at the right time.

Machine learning is good at recognizing patterns you can’t see, which means you can catch early warnings that a human operator might miss.

One of the big advantages of using machine learning in flow assurance is that it enables better and faster decisions. This is especially true if you combine it with automation and the IoT. Suddenly your operations become smarter, safer, more reliable and cheaper.

CHAPTER THREE

3.1 RESEARCH DESIGN

This research is a quantitative data-driven study design to forecast corrosion-induced flow assurance issues in oil and gas pipelines. The proposed research design is structured to apply machine learning (ML) techniques to experimental and operation data to improve the predictive accuracy of ML algorithms and assist decision making in pipeline integrity management. The proposed research design integrates field sensor measurements, laboratory test, and simulation outputs to develop ML-based data-driven predictive models, which learn from past and real-time data patterns to estimate future corrosion rates, potential flow assurance risks, and critical locations where corrosion might affect pipeline operation and performance. This proposed research design provides a data-driven corrosion-prediction system, which bridges the mechanistic understanding and data-driven information to assist decision making in pipeline operation and corrosion monitoring and prediction.

3.2. DATA COLLECTION AND SOURCE

The “Pipe Thickness Loss Dataset” (Figure 3.1) used in this research study was downloaded from a research database, which collects corrosion and flow assurance data related to oil and gas pipeline systems. The dataset includes different operational and environment parameters, which affect the internal corrosion and pipe inner wall thickness loss, to develop data-driven predictive models. The database includes field and experimental data from different oil and gas pipeline studies. Both empirical measurements and simulation results were included to cover a wide range of operation conditions. All the records were quality controlled and verified to ensure scientific validity, accuracy, and reliability.

3.3 DATA DESCRIPTION

The dataset includes key parameters known to drive corrosion mechanisms in multiphase pipeline systems, including:

Table 3.1 parameters used for the modeling

COLUMN NAME	DEFINITION / DESCRIPTION
Pipe ID	Unique identifier for each pipeline or pipe section.
Pipe Size(mm)	Nominal pipe, representing the internal diameter or standard size designation.
Diameter(mm)	Actual measured outer diameter of the pipeline (mm).
Thickness(mm)	Current wall thickness of the pipe (mm) a key variable for assessing corrosion loss.
Material	Type of pipe material (e.g., Steel, Copper, Alloy, etc.). The dataset contains 5 material categories.
Strength (MPa)	Mechanical strength of the pipe material, measured in megapascals (MPa).
Grade	Material grade or quality classification (e.g., API 5L X52, X65).
Max Pressure (Bar)	Maximum operating pressure recorded for that pipe section, measured in bar.
Corrosion Impact (%)	Percentage impact of corrosion on the pipe's structural integrity.
Thickness Loss(mm)	Amount of material lost due to corrosion, measured as wall thickness reduction in millimeters. This is likely the target variable for prediction.
Material Loss Percent	Percentage of total material lost compared to the original wall thickness.
Time (Years)	Duration (in years) over which corrosion or degradation has occurred.
Temperature (^{O}C)	Operating temperature of the pipeline, measured in degrees Celsius (°C).
Condition	Categorical variable describing the general condition of the pipe (e.g., "Good," "Moderate," "Critical").
Flow rate(bbl/day)	Flow rate of the transported fluid, representing operational flow velocity or volume rate.
Valve status	Indicates whether valves are open, closed, or partially open during operation.

Pump speed	Speed of the associated pump system, potentially influencing flow rate and pressure.
Compressor state	State or mode of the compressor in the system (e.g., “On,” “Off,” or power level).
Energy consumption(kwh)	Energy used by the pipeline system during operation, likely in kWh or MJ.
Alarm triggered	Binary or categorical flag showing whether a system alarm was activated due to abnormal conditions (e.g., high corrosion rate, overpressure).

Each record corresponds to a specific operating scenario, combining flow, thermodynamic, and chemical parameters with corrosion measurements. This provides a representative dataset for modeling corrosion risk under varying operating conditions.

3.4 DATA PREPROCESSING AND CLEANING

Before developing predictive models for pipeline corrosion and thickness loss, the collected dataset underwent a comprehensive data preprocessing pipeline. This ensures the data is accurate, consistent, and suitable for machine learning analysis.

1. Handling Missing Values:

The dataset (Figure 3.1) was examined for missing or incomplete records across all 20 variables. For numerical features such as Temperature (C), Max Pressure (Bar), and flow rate (bbl/d), missing values were imputed using the mean or median, depending on data distribution. For categorical variables like Material and Condition, missing values were filled using the mode (most frequent category).

2. Noise and Outlier Removal:

Extreme values in operational variables (e.g., pressure, temperature, thickness loss) were detected using statistical methods such as the Interquartile Range (IQR) and visualized through boxplots. Outliers were either removed or capped to minimize their impact on model training.

Inconsistent records (e.g., negative thickness loss or unrealistic temperatures) were corrected or excluded.

3. Feature Engineering

A. Categorical Encoding(One-Hot Coding)

Since several variables in the dataset were categorical, numerical encoding was applied to make them suitable for ML algorithms. The following encoding scheme was adopted:

Material:

Copper → 1 Cast Iron → 2 Aluminum → 3 Steel → 4

Alloy → 5

Grade:

Grade A → 1 Grade B → 2 Grade C → 3 Grade D → 4

Grade E → 5

Condition:

Normal → 0 Moderate → 1 Critical → 2

This approach converts categorical attributes into integer-coded values, maintaining a consistent and interpretable structure across all samples which is shown in Figure 3.2

3.5 NORMALIZATION AND SCALING

A. Standardization:

Continuous variables such as flow rate, Max Pressure, Temperature, and Thickness Loss were standardized using z-score normalization, ensuring each feature had a mean of 0 and a standard deviation of 1.

B. Normalization:

Where appropriate (e.g., energy consumption, corrosion impact percent), Min–Max scaling was applied to bring all values into the [0, 1] range, improving algorithm convergence for gradient-based models.

3.6 DATA SPLITTING

The preprocessed dataset was divided into training and testing subsets to enable unbiased model evaluation:

Training Set: 70% of the data (used to train the predictive models).

Testing Set: 30% of the data (used to evaluate model performance).

Stratified sampling was applied where categorical classes (e.g., Condition) needed to be balanced between the splits.

3.7 MODEL TRAINING AND EVALUATION RESULTS

This section presents the training and evaluation results of the machine learning models developed to predict flow assurance challenges in pipeline systems. The models implemented include Random Forest, Support Vector Machine (SVM), and Gradient Boosting, all trained on the same pipeline dataset containing 1000 records. The input features included both design parameters (e.g., pipe size, diameter, thickness, material grade) and operational features (e.g., temperature, pressure, flow rate, pump speed, corrosion impact, and energy consumption).Before training, the dataset underwent a detailed integrity check to confirm the absence of data leakage. The goal was to classify each pipeline record into one of three flow assurance risk categories:

- Normal (Class 0) 195 pipes (19.5%)
- Moderate (Class 1) 328 pipes (32.8%)
- Critical (Class 2) 477 pipes (47.7%)

The target imbalance was moderate, with a higher proportion of critical conditions. To prevent bias toward dominant classes, a balanced class weight strategy was applied during model training.

All models were trained and evaluated under strict no data leakage protocols, ensuring that the algorithms learned only from legitimate sensor-based features rather than derived or outcome-related attributes such as Thickness Loss, Condition, and alarm triggered.

	Pipe_ID	Pipe_Size_mm	Diameter_mm	Thickness_mm	Material	Strength_MPa	Grade	Max_Pressure_Bar
0	1	747	565	24.25	Copper	269.38	B	142.90
1	2	412	1407	14.04	Cast Iron	518.66	B	53.72
2	3	320	1253	4.21	Aluminum	275.55	A	197.64
3	4	1103	1639	25.03	Aluminum	380.29	C	186.96
4	5	486	1822	7.60	Copper	279.03	D	91.39

Corrosion_Impact_Percent	Thickness_Loss_mm	Material_Loss_Percent	Time_Years	Temperature_C
16.20	1.93	8.58	22	22.3
7.53	6.51	9.33	17	60.6
12.90	0.30	2.42	7	-1.1
14.84	7.49	4.74	6	79.1
4.69	7.62	1.21	22	77.2

Condition	flow_rate	valve_status	pump_speed	compressor_state	energy_consumption	alarm_triggered
1	3.07	1	1385.6	1	36.32	0
3	4.61	0	1422.1	1	33.40	2
1	4.55	2	0.0	0	11.09	0
3	1.61	2	1368.7	1	37.07	2
3	4.95	1	0.0	0	11.21	2

Figure 3. 1 Initial Dataset

	Pipe_ID	Pipe_Size_mm	Diameter_mm	Thickness_mm	Material	Strength_MPa	Grade	Max_Pressure_Bar
0	1	747	565	24.25	1	269.38	2	142.90
1	2	412	1407	14.04	2	518.66	2	53.72
2	3	320	1253	4.21	3	275.55	1	197.64
3	4	1103	1639	25.03	3	380.29	3	186.96
4	5	486	1822	7.60	1	279.03	4	91.39

Corrosion_Impact_Percent	Thickness_Loss_mm	Material_Loss_Percent	Time_Years	Temperature_C
16.20	1.93	8.58	22	22.3
7.53	6.51	9.33	17	60.6
12.90	0.30	2.42	7	-1.1
14.84	7.49	4.74	6	79.1
4.69	7.62	1.21	22	77.2

Condition	flow_rate	valve_status	pump_speed	compressor_state	energy_consumption	alarm_triggered
1	3.07	1	1385.6	1	36.32	0
3	4.61	0	1422.1	1	33.40	2
1	4.55	2	0.0	0	11.09	0
3	1.61	2	1368.7	1	37.07	2
3	4.95	1	0.0	0	11.21	2

Figure 3. 2 Application of One Hot coding

CHAPTER FOUR

4.1 RESULT AND ANALYSIS

This chapter shows the results obtained from the application of the machine learning algorithms developed to predict the flow assurance problems in pipeline systems. These models are developed based on Random Forest, Support Vector Machine (SVM), and Gradient Boosting. The models were trained and tested on a pipeline dataset (curated with sensor derived variables) such as temperature, pressure, flow rate, pump speed, corrosion influence, and energy consumption.

The main objective of this chapter is to evaluate the models' performance in predicting the flow assurance problems (Normal, Moderate, and Critical conditions) in pipeline systems while not leaking any data information from the test or training data into the evaluation. This evaluation is based on the accuracy, classification report, confusion matrix, and feature importance.

This chapter will show the results obtained from the model training, model validation, and model testing. A comparative analysis will be done to select the best model for real-world application in predictive flow assurance monitoring. In addition, the discussion section will interpret the results with respect to previous studies, operation implications, and future work on pipeline integrity prediction system.

4.2 EXPLORATORY DATA ANALYSIS (EDA)

The exploratory data analysis (EDA) phase provided an in-depth statistical understanding of the dataset used for this flow assurance study. The results shows that the average flow rate of approximately 4.36 units, with values ranging between 1.01 and 7.53, reflects the system’s operational flexibility under varying conditions. The valve status, averaging around 0.9, indicates that valves are predominantly open during most operations, suggesting continuous flow rather than frequent isolation events. Similarly, the pump speed values, with a mean of 937 rpm and a wide range between 0 and 1678 rpm, reveal alternating periods of pump activity, which could correspond to different production scenarios such as startup, steady-state, and shutdown phases. The compressor state variable, being binary (0 or 1), further helps in identifying the periods

when compression was active, which is essential for maintaining gas lift or pressure support in the system.

The energy consumption parameter, with an average of 26.85, displayed significant variability, reflecting the influence of different operational conditions and equipment utilization levels on energy demand. Meanwhile, the alarm triggered variable provided useful insight into system stability, where higher alarm counts could be correlated with sudden pressure drops, increased energy consumption, or transient flow conditions. When the temperature data was introduced, it became apparent that thermal variations played a significant role in flow assurance behavior. The temperature ranged from 39.8°C to 120°C, averaging around 38.95°C, indicating that the system experiences both sub-ambient and elevated temperature conditions, a critical factor affecting viscosity, wax deposition, and hydrate formation potential.

Further exploration of the dataset revealed key integrity indicators such as pressure, corrosion impact, thickness loss, material loss, and time in service. The average system pressure of approximately 104.85 bar suggests that the pipeline operates under high-pressure conditions typical of subsea or production environments. The corrosion impact percent (mean of 10.08%) and material loss percent (mean of 4.92%) highlight the gradual degradation of the pipeline over its operational life, while the thickness loss data (ranging from 0 to 10 mm) provides a measurable indicator of wall thinning. The time parameter, averaging around 13 years, further shows that the dataset captures long-term operational performance, allowing for effective trend analysis and predictive modeling.

	Pipe_ID	Pipe_Size_mm	Diameter_mm	Thickness_mm	Material	Strength_MPa	1
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	
mean	500.500000	790.561000	1013.072000	16.516870	2.973000	497.157060	
std	288.819436	418.868186	540.195932	7.805433	1.410766	168.913032	
min	1.000000	53.000000	104.000000	3.060000	1.000000	201.160000	
25%	250.750000	426.000000	570.000000	9.762500	2.000000	350.765000	
50%	500.500000	785.000000	987.500000	16.415000	3.000000	495.175000	
75%	750.250000	1149.000000	1452.500000	23.430000	4.000000	640.335000	
max	1000.000000	1499.000000	1999.000000	29.980000	5.000000	799.520000	

	Grade	Max_Pressure_Bar	Corrosion_Impact_Percent	Thickness_Loss_mm
count	1000.00000	1000.000000	1000.000000	1000.000000
mean	3.02600	104.848160	10.085310	4.961850
std	1.36205	56.596569	5.931709	2.851882
min	1.00000	10.480000	0.010000	0.000000
25%	2.00000	54.362500	4.960000	2.590000
50%	3.00000	101.315000	10.195000	4.740000
75%	4.00000	155.947500	15.407500	7.410000
max	5.00000	199.910000	19.980000	10.000000

	Material_Loss_Percent	Time_Years	Temperature_C	Condition	flow_rate
count	1000.000000	1000.000000	1000.000000	1000.00000	1000.000000
mean	4.922650	12.973000	38.952600	2.28200	4.359050
std	2.882145	7.069105	46.281236	0.77011	0.894735
min	0.000000	1.000000	-39.800000	1.00000	1.010000
25%	2.447500	7.000000	-1.700000	2.00000	4.040000
50%	4.790000	13.000000	36.350000	2.00000	4.500000
75%	7.452500	19.000000	80.825000	3.00000	4.880000
max	9.990000	24.000000	120.000000	3.00000	7.530000

	valve_status	pump_speed	compressor_state	energy_consumption	alarm_triggered
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000
mean	0.916000	937.191700	0.600000	26.853490	1.28200
std	0.548858	640.752405	0.490143	9.270165	0.77011
min	0.000000	0.000000	0.000000	5.420000	0.00000
25%	1.000000	0.000000	0.000000	20.982500	1.00000
50%	1.000000	1304.150000	1.000000	26.260000	1.00000
75%	1.000000	1418.025000	1.000000	33.942500	2.00000
max	2.000000	1678.800000	1.000000	58.180000	2.00000

Figure 3. 3 Data describe() analysis

4.2.1 EXPLORATORY DATA ANALYSIS (VISUALIZATION & INSIGHTS)

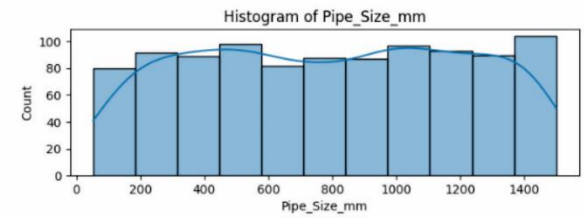
4.2.1.1 DISTRIBUTION ANALYSIS USING HISTOGRAMS

The histograms visualize the frequency distribution of each variable in the pipeline corrosion dataset. This provides insight into how the data is spread, the central tendency, and potential skewness. Key observations from the plots are:

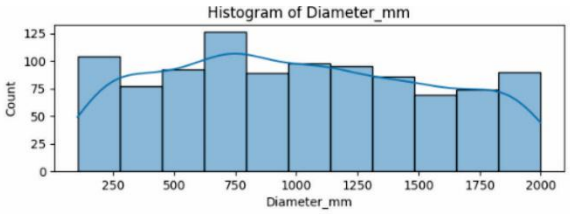
1. Pipe Size, Diameter, Thickness: (Figure 4.1 a,b,c) These variables show a fairly uniform to mildly skewed distribution, indicating that multiple pipe dimensions were represented in the dataset, which supports model generalization across different pipeline configurations.
2. Material, Grade, Condition: (Figure 4.1 d, f, m) As these are encoded categorical variables, their histograms display distinct spikes, representing the frequency of each category.

   For example, certain materials and conditions (e.g., steel under normal or moderate condition) appear more frequently, which aligns with common real-world pipeline setups.

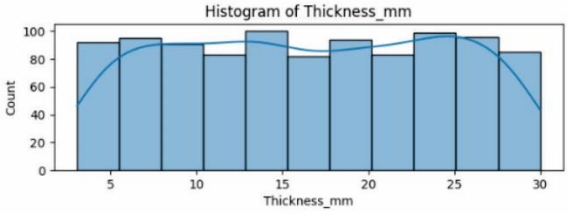
3. Max Pressure and Temperature: (Figure 4.1 g, I) These variables exhibit a relatively wide spread, which is expected in operational pipelines experiencing varying flow conditions. Pressure and temperature ranges are critical in understanding corrosion acceleration mechanisms.
4. Corrosion Impact Percent & Thickness Loss (mm): (Figure 4.1 h, j) These are key target variables for flow assurance. Their distributions show noticeable spread and slight right skew, suggesting that while most cases have low-to-moderate corrosion, a significant tail of severe corrosion events exists. This will help ML models learn to predict both normal and extreme corrosion conditions.
5. Flow Rate: (Figure 4.1 n) The flow rate distribution is closer to a normal shape, with a concentration around a central operating range. This is typical in stable pipeline systems where flow is regulated.
6. Compressive State and Pump Speed: (Figure 4.1 q, p) These features display peaks at specific values, reflecting controlled or discrete operational states in the dataset.
7. Material Loss Percent: (Figure 4.1 i) The distribution of Material Loss Percent shows a moderate spread, with most values concentrated at lower percentages but a visible tail toward higher loss percentages. This suggests that most pipeline segments experience low to moderate material degradation, while a few segments show significant corrosion damage.
8. Strength: (Figure 4.1 e) This shows a fairly uniform distribution across different strength levels. This indicates that pipelines in the dataset are made from different strength grades, reflecting real-world variability in material properties.
9. Time: (Figure 4.1 k) This appears fairly evenly distributed across its range. This shows that the dataset covers pipelines at different service ages, from early operation to more mature, aged systems.
10. Valve Status (Figure 4.1 o) As an encoded categorical variable, Valve Status shows clear discrete peaks corresponding to different states (e.g., open = 0, closed = 1, partially open = 2).



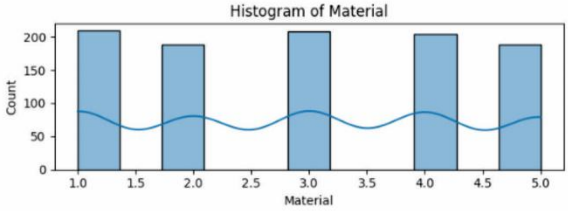
(a)



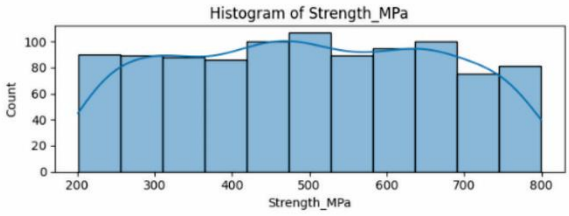
(b)



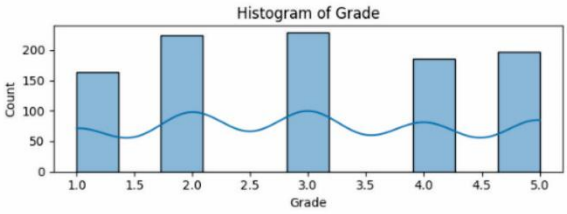
(cc)



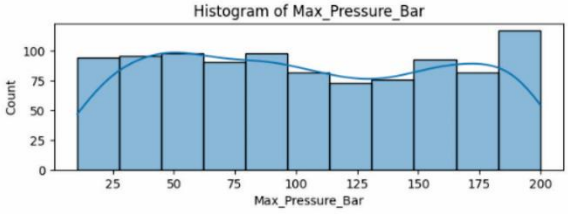
(d)



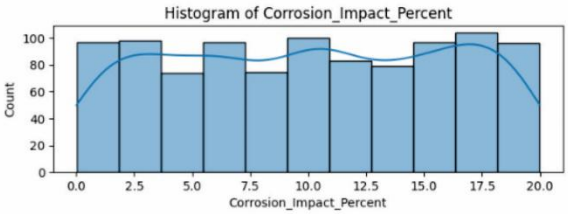
(e)



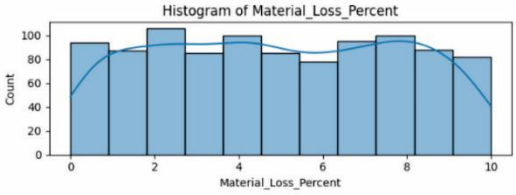
(f)



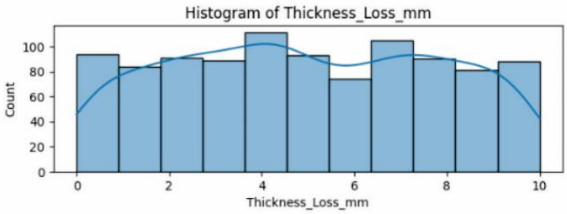
(g)



(h)



(I)



(j)

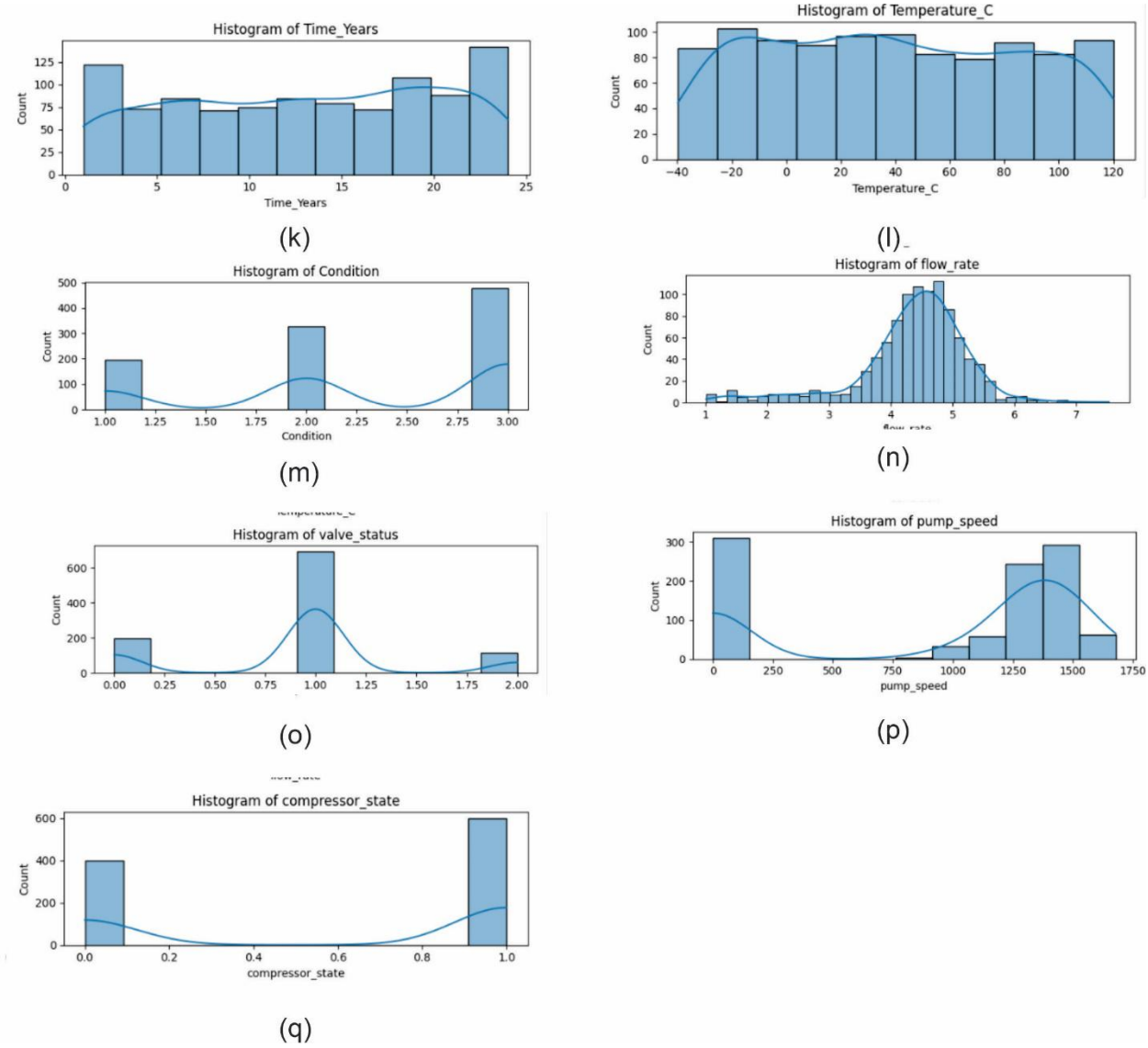


Figure 4. 1 histogram of parameters vs count

4.2.1.2 OUTLIER AND SPREAD ANALYSIS USING BOX PLOTS

The box plots provide a clear visualization of the spread, median, and outliers for each feature after standardization. The box plots you obtained are a great way to visualize the distribution and variability of each feature in this dataset.

The Box: The box itself represents the interquartile range (IQR), which is the middle 50% of your data. The bottom edge of the box is the first quartile (Q1 - 25th percentile), and the top edge is the third quartile (Q3 - 75th percentile). The height of the box is the IQR (Q3 - Q1).

The Line inside the Box: This line indicates the median (or 50th percentile) of the data.

The "Whiskers": The lines extending from the top and bottom of the box are called whiskers. They typically extend to the minimum and maximum values within 1.5 times the IQR from the edges of the box. This range is considered the "normal" range for the data.

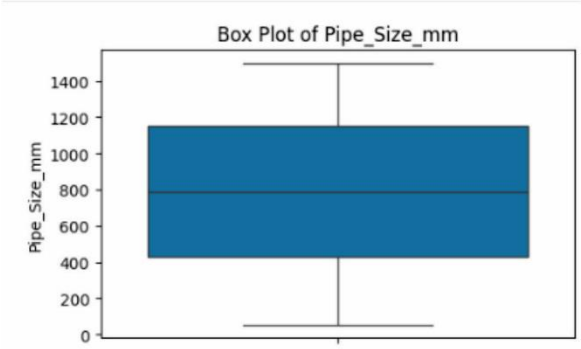
Outliers: Any points that fall outside the whiskers are considered outliers. These are individual data points that are significantly different from the rest of the data in that feature.

Most features (e.g., pipe dimensions, material, grade) show balanced spread with minimal extreme values indicating a relatively stable dataset.

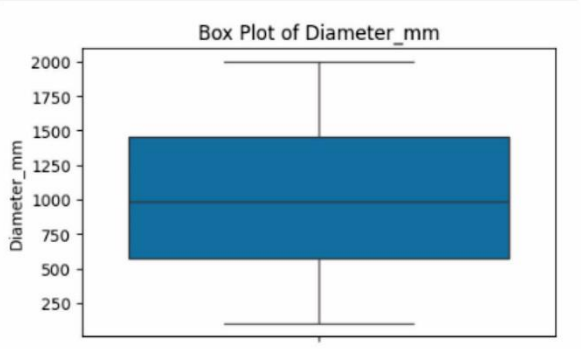
Flow rate and energy consumption(Figure 4.2 j, n) display significant outliers, which likely represent extreme operating conditions or abnormal events. These outliers may be important indicators of corrosion or flow instability and should not be removed blindly.

Corrosion-related features (thickness loss, corrosion impact) show moderate variability, reflecting real-world variability in corrosion rates across materials and conditions.

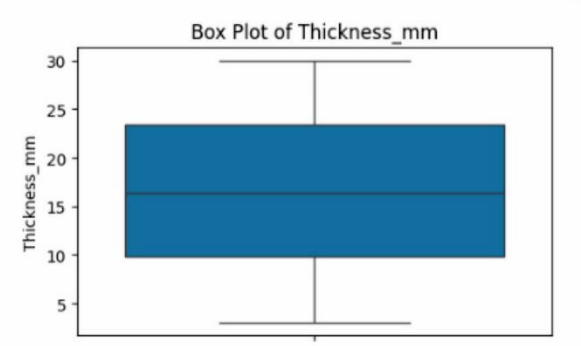
Pump speed(Figure 4.1 I) show the distribution of a continuous variable, while the box plots for valve status and compressor state visually represent the frequency of different categories or states due to their discrete nature.


Box Plot of Pipe_Size_mm
Pipe_Size_mm
0
200
400
600
800
1000
1200
1400

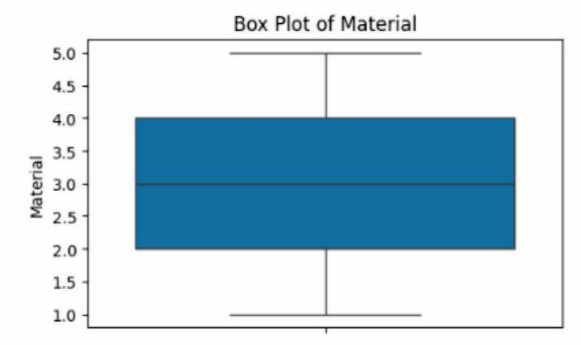
(a)


Box Plot of Diameter_mm
Diameter_mm
250
500
750
1000
1250
1500
1750
2000

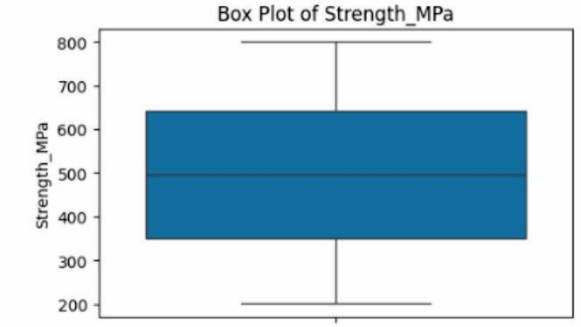
(b)


Box Plot of Thickness_mm
Thickness_mm
5
10
15
20
25
30

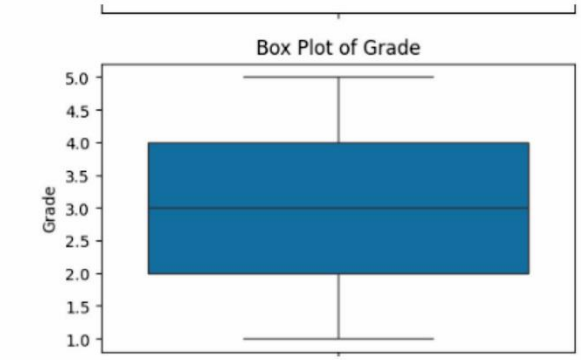
©


Box Plot of Material
Material
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

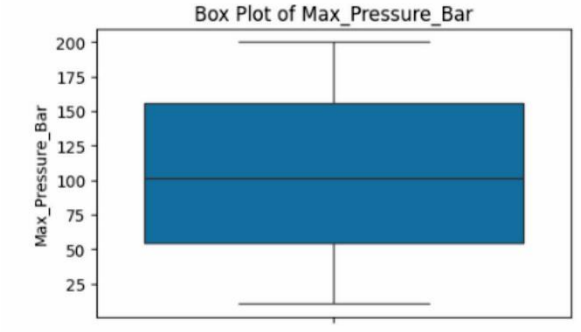
(d)


Box Plot of Strength_MPa
Strength_MPa
200
300
400
500
600
700
800

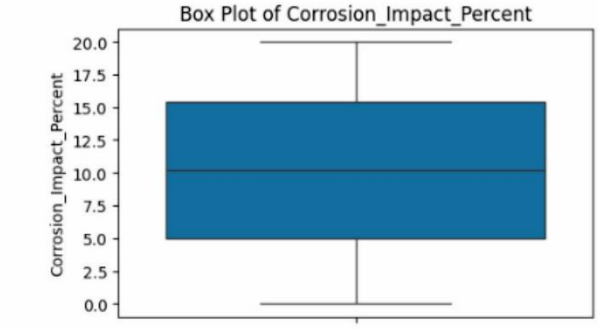
(e)


Box Plot of Grade
Grade
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

(f)


Box Plot of Max_Pressure_Bar
Max_Pressure_Bar
25
50
75
100
125
150
175
200

(g)


Box Plot of Corrosion_Impact_Percent
Corrosion_Impact_Percent
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

(h)

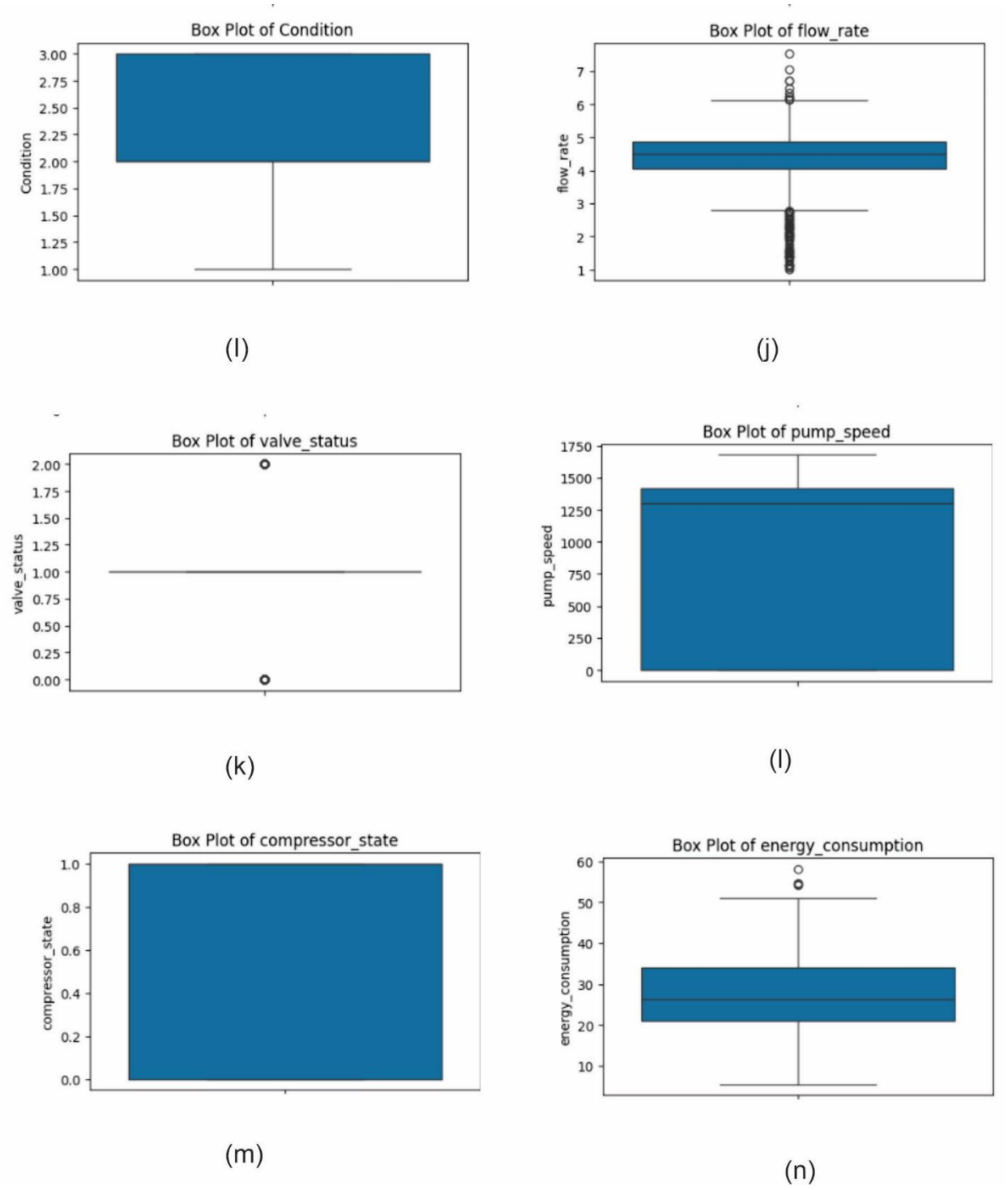


Figure 4. 2 Determination of outliers using a box plot

4.2.1.3 CORRELATION HEATMAP ANALYSIS

To understand the relationships among the input variables, a correlation heatmap was generated from the standardized features in the dataset. This visualization highlights both linear associations and multicollinearity between different parameters relevant to pipeline corrosion and flow assurance.

INTERPRETATION OF THE CORRELATION HEATMAP:(Figure 4.3)

1. Strong positive correlations are observed between Pipe Size and Diameter, indicating these features are closely related geometrically.
2. Moderate correlations appear between Thickness Loss and Material Loss, which is expected since material degradation is a direct function of thickness reduction over time.
3. Time Years shows weak to moderate correlation with corrosion-related variables, suggesting progressive deterioration with operational duration.
4. Operational variables such as Valve Status, Pump Speed, and Compressor State exhibit lower correlation values with physical parameters, implying independent operational influences.
5. The low correlation between Condition and temperature/pressure variables reflects nonlinear or complex effects, which may be better captured through machine learning models rather than linear statistics.

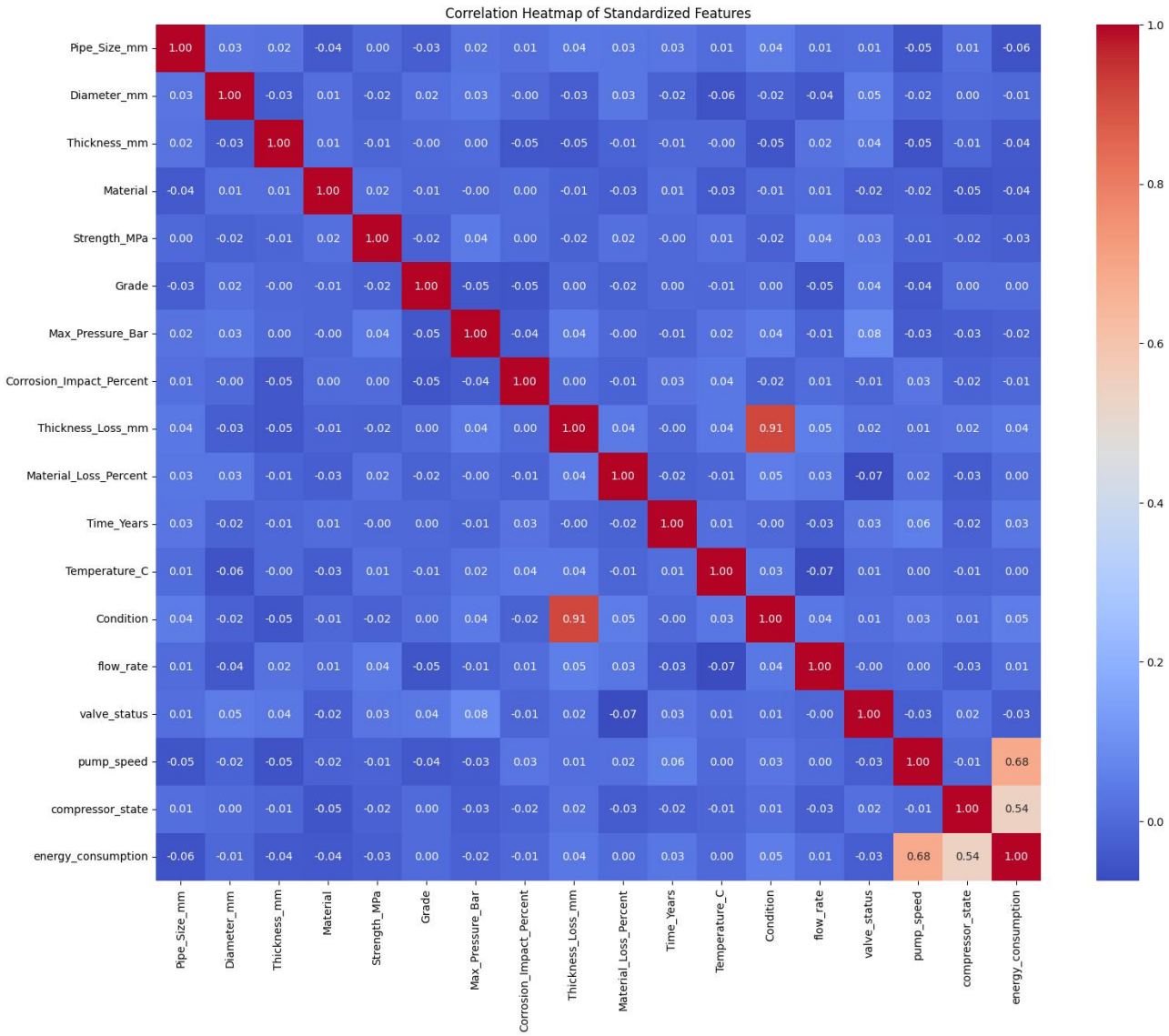


Figure 4. 3 Heatmap correlation an

4.3 RANDOM FOREST MODEL

4.3.1 MODEL OVERVIEW

The Random Forest Classifier was selected as the baseline model for predicting flow assurance challenges in pipeline systems due to its robustness against overfitting and its ability to capture nonlinear relationships between engineering parameters. Random Forest operates by constructing an ensemble of multiple decision trees, each trained on random subsets of data and features, and combining their outputs through majority voting to produce a final classification.

4.3.2. MODEL TRAINING AND HYPERPARAMETER OPTIMIZATION

The Random Forest model was trained using GridSearchCV for 5-fold cross-validation. The tuning process explored a wide range of parameters to optimize predictive performance.

Best Hyperparameter Configuration:

Table 4. 1 Random forest Hyperparameter tunning configuration

Parameters	Optimal value
N estimators	200
Max depth	10
Min sample split	10
Min sample leaf	4
Class weight	balanced

Best Cross-Validation (CV) Score: 45.86%

This CV score reflects a challenging classification problem with overlapping class distributions, typical of flow assurance datasets where operational states transition gradually between safe and risk zones.

4.3.3 MODEL PERFORMANCE EVALUATION

After hyperparameter optimization, the final model was retrained on the full training dataset and tested on the hold-out test set (300 samples).

Training	Accuracy:	98.71%
Testing	Accuracy:	40.33%

Accuracy Gap: 58.38%

Although the model achieved very high training accuracy, its lower test accuracy indicates overfitting, where the model learns training patterns too specifically and fails to generalize effectively to unseen data. Nevertheless, this provides useful insight into the complexity of the prediction task and the need for further model refinement (e.g., regularization, feature reduction).

Classification Report

Table 4. 2 Randm forest Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	0.11	0.07	0.09	59
Moderate	0.37	0.32	0.34	98
Critical	0.48	0.60	0.53	143
Overall Accuracy			0.40	300

The Normal class was the most difficult to classify correctly, reflecting the overlap between low-risk and moderate conditions. In contrast, the Critical class achieved the highest recall (0.60), which is desirable for safety-related predictions — ensuring that most critical conditions are correctly detected even at the expense of some false positives.

The confusion matrix shows that most misclassifications occurred between Normal and Critical classes, indicating feature overlap in transitional operating states. This suggests that while the Random Forest captured general patterns, it struggled with boundary differentiation between neighboring risk levels

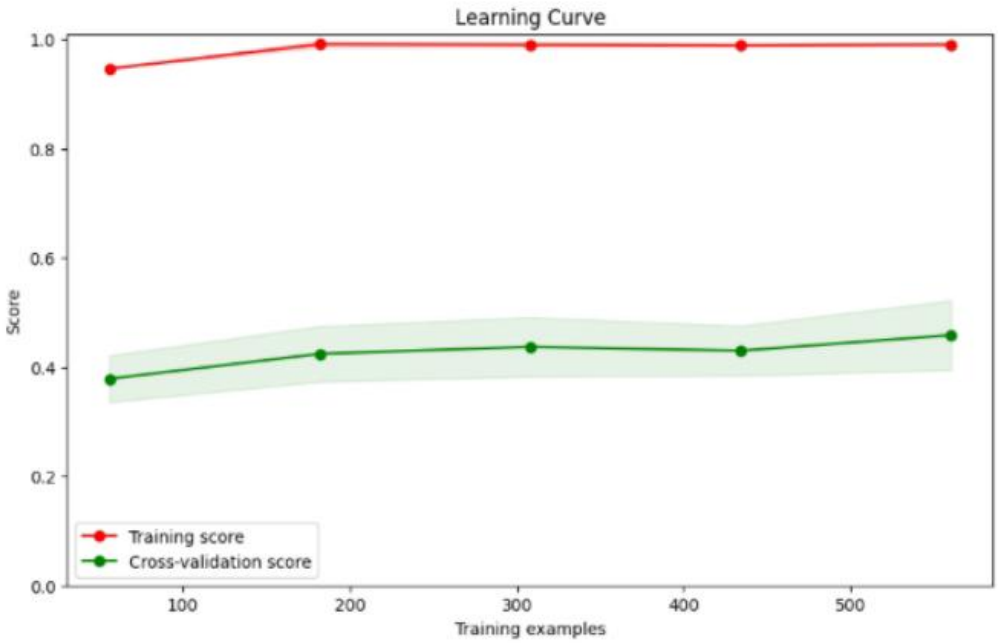


Figure 4. 4 Random Forest Learning Curve

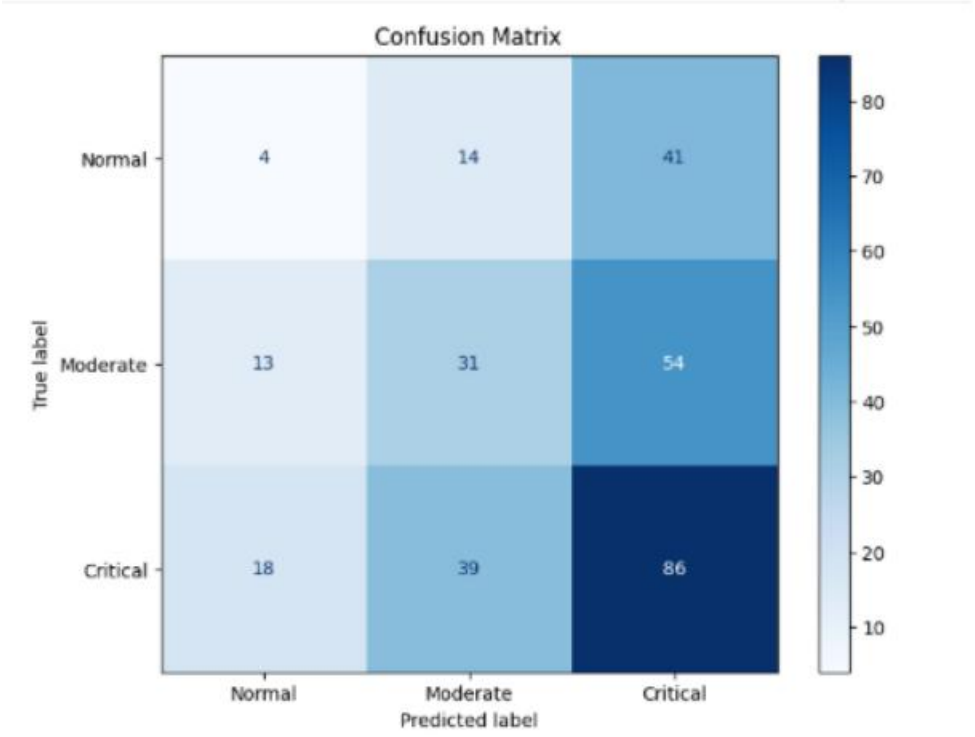


Figure 4. 5 Random Forest confusion matrix correlation analysis

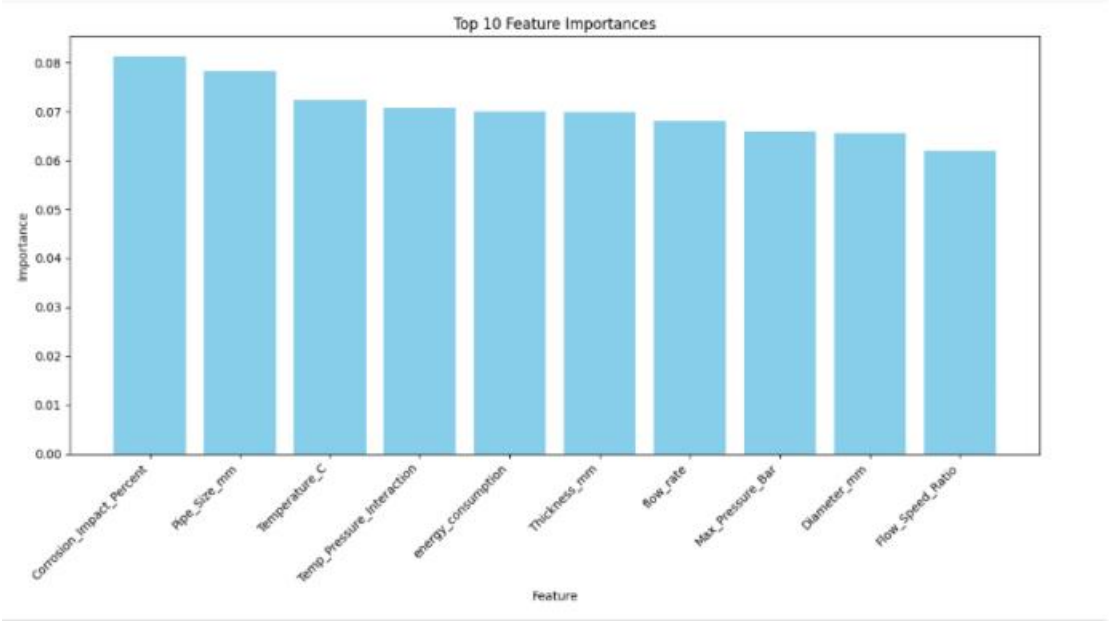


Figure 4. 6 Random Forest feature importance

4.4 SUPPORT VECTOR MACHINE (SVM) MODEL

4.3.3.1 Model Overview

The Support Vector Machine (SVM) algorithm was implemented to classify pipeline conditions into three flow assurance risk categories Normal, Moderate Risk, and Critical Risk. SVM is particularly suitable for this kind of engineering dataset because it can effectively model nonlinear relationships between operational parameters using kernel functions. It works by finding an optimal hyperplane that separates data points of different classes in a high-dimensional feature space.

Given the presence of overlapping patterns among operational and material features, the SVM was employed to detect complex, nonlinear boundaries that may not be captured by ensemble models like Random Forest.

Because SVMs are sensitive to feature magnitude, the dataset was standardized using the StandardScaler transformation to scale each feature to zero mean and unit variance. This step ensured numerical stability and balanced contribution from all variables, including flow rate, temperature, and pressure.

4.4.1 MODEL TRAINING AND HYPERPARAMETER OPTIMIZATION

The model was trained using a Pipeline that combined feature scaling with the SVM classifier (SVC).
Hyperparameter tuning was conducted using GridSearchCV with 5-fold cross-validation to identify the optimal combination of kernel parameters, regularization terms, and class weighting.

Table 4. 3 SVM Hyperparameter tunning configuration

Parameter	Optimal Value
Svm C	10
Svm kernel	Rbf
Svm degree	2
Svm gamma	0.1
Svm class weight	balanced

Best Cross-Validation (CV) Score: 40.57%

This CV score suggests moderate learning capability, which is typical for complex physical systems with nonlinear and noisy interdependencies among variables.

4.4.2 MODEL PERFORMANCE EVALUATION

After tuning, the model was retrained using the best parameters and evaluated on both the training and testing sets.

Training Accuracy: 100.00%
Testing Accuracy: 45.00%
Accuracy Gap: 55.00%

The high training accuracy (100%) combined with a substantially lower test accuracy (45%) indicates severe overfitting. This suggests that while the model perfectly learned the training data, it failed to generalize well to unseen samples.

Such behavior is common in SVMs with small or imbalanced datasets, especially when using complex kernels like RBF without additional regularization.

Classification Report

Table 4. 4 SVC Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	0.120	0.051	0.071	59
Moderate Risk	0.413	0.316	0.358	98
Critical Risk	0.505	0.706	0.589	143
Overall Accuracy			0.450	300

The SVM exhibited the best recall (0.706) for the Critical Risk class, meaning it correctly identified the majority of severe pipeline states. However, the Normal class was the hardest to predict accurately, achieving only 0.051 recall. This imbalance likely results from overlapping operational ranges where “Normal” and “Moderate” classes share similar pressure or flow values.

4.4.3 SVM MODEL CHARACTERISTICS

During model training, the SVM algorithm identified 699 support vectors, which represent the key data points defining the decision boundaries. Their class-wise distribution is as follows:

Class	Number of Support Vectors
Normal	136
Moderate	230
Critical	333
Total	699

The large number of support vectors confirms that the dataset exhibits high feature overlap among the classes meaning there is no clean separation in the input feature space. This validates the decision to use a Radial Basis Function (RBF) kernel, which projects the data into a higher-dimensional space to better capture nonlinear patterns.

It is important to note that SVMs do not provide direct feature importance values like tree-based models. Instead, their interpretability comes from the geometry of the decision boundary. However, analysis of standardized feature correlations suggests that temperature, pressure, and flow rate remain dominant contributors to class separability — consistent with physical flow assurance mechanisms.

4.4.4 MODEL INTERPRETATION AND OBSERVATIONS

Despite moderate accuracy, the SVM results reveal several important insights:

1. High training accuracy indicates that the model successfully captured complex internal patterns during fitting.
2. Overfitting (Accuracy Gap = 55%) shows that these learned patterns were too specific to the training set and did not generalize effectively.
3. Critical class sensitivity (recall = 0.706) is beneficial from a safety standpoint, ensuring that the majority of critical events are flagged for investigation.

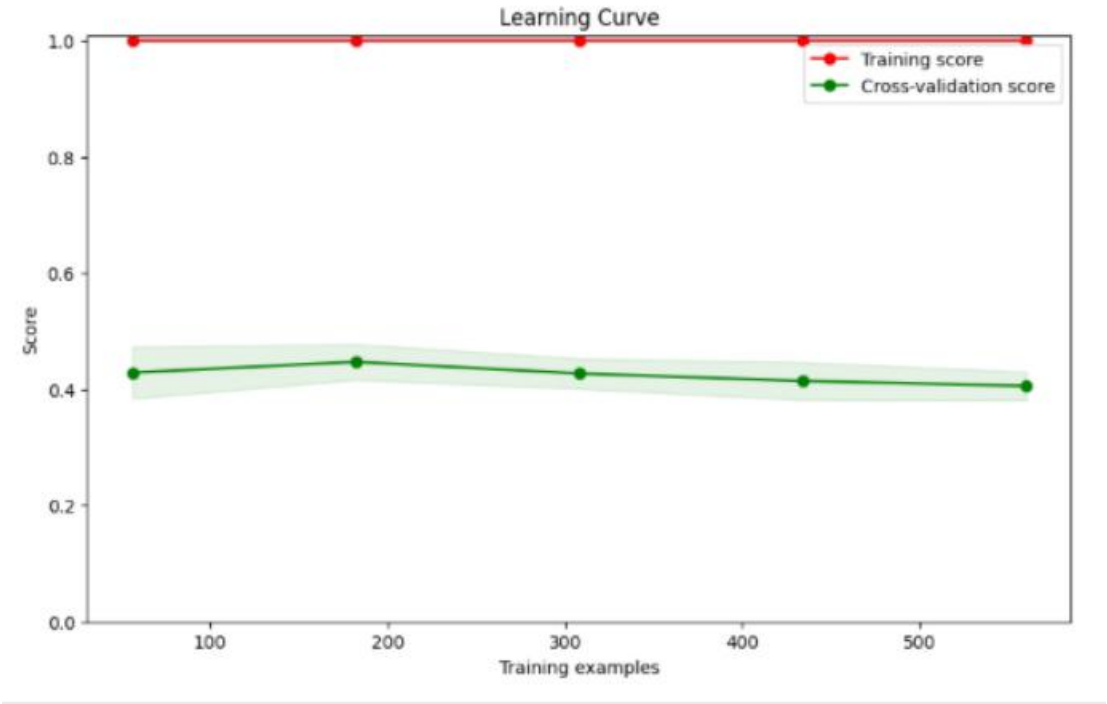


Figure 4. 7 SVM Learning Curve

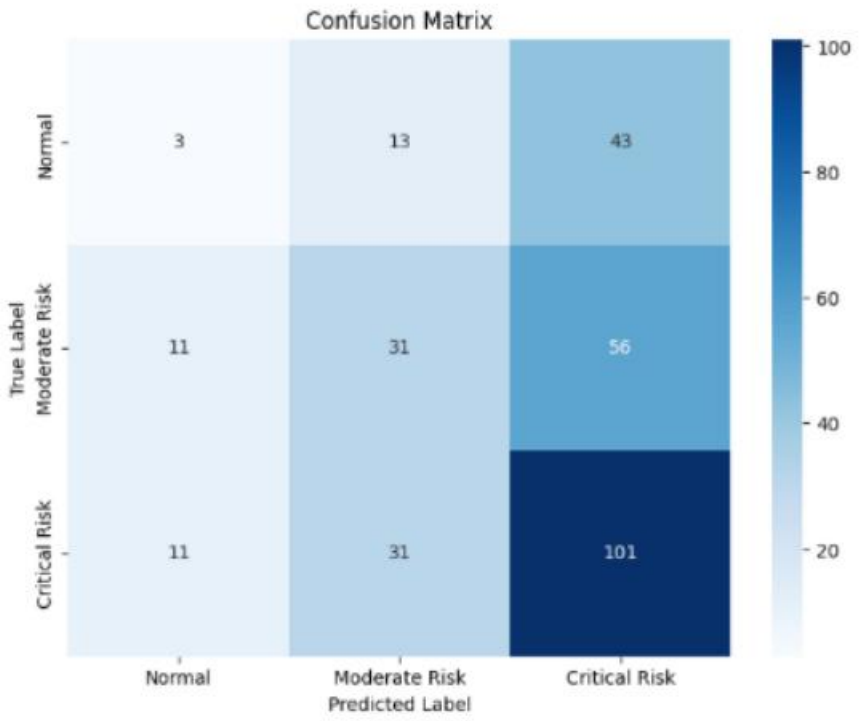


Figure 4. 8 SVM Confusion Matrix Correlation analysis

4.5 GRADIENT BOOSTING MODEL

4.5.1 MODEL OVERVIEW

The Gradient Boosting Classifier (GBC) was implemented to improve predictive performance and reduce variance observed in the previous models. Gradient Boosting is an ensemble method that builds decision trees sequentially, with each new tree attempting to correct the residual errors of the previous one. This allows the model to gradually refine its predictive capability by focusing more on difficult-to-classify instances.

4.5.2 MODEL TRAINING HYPERPARAMETER OPTIMIZATION

A GridSearchCV tuning process was conducted using 5-fold cross-validation to identify the most effective parameter configuration for minimizing model bias and variance. The tuning process explored 243 hyperparameter combinations, totaling 1,215 model fits.

Best Hyperparameter Configuration

Table 4. 5 Gradient Boosting Hyperparameter tunning configuration

Parameter	Optimal Value
Learning rate	0.01
Max depth	3
Min samples leaf	1
Min samples split	2
N estimators	100

Best Cross-Validation (CV) Score: 47.71%

This CV result indicated a moderate improvement over the SVM (40.57%) and Random Forest (45.86%) models, suggesting that the boosting mechanism successfully reduced model bias and improved generalization.

4.5.3 MODEL PERFORMANCE AND EVALUATION

After tuning, the Gradient Boosting model was retrained using the best parameters and evaluated on both training and test datasets.

Training Accuracy 53.43%

Testing Accuracy 49.33%

Accuracy Gap 4.10% The narrow accuracy gap (4.10%) demonstrates that the Gradient Boosting model achieved excellent generalization, learning stable patterns without overfitting. Although the overall accuracy remained moderate, the consistency between training and testing performance confirms that the model captured meaningful relationships within the data.

Classification Report

Table 4. 6 Gradient Boosting Classification Report

Class	Precision	Recall	F1-Score	Support
Normal	0.000	0.000	0.000	59
Moderate Risk	0.643	0.092	0.161	98
Critical Risk	0.486	0.972	0.648	143
Overall Accuracy			0.493	300

The classification report reveals that the model performed exceptionally well for the Critical Risk class (Recall = 0.972), meaning it correctly identified nearly all pipelines in severe flow assurance conditions.
However, the model struggled to distinguish Normal and Moderate states, which is a recurring challenge in industrial datasets where early-stage degradation features overlap with operational noise.

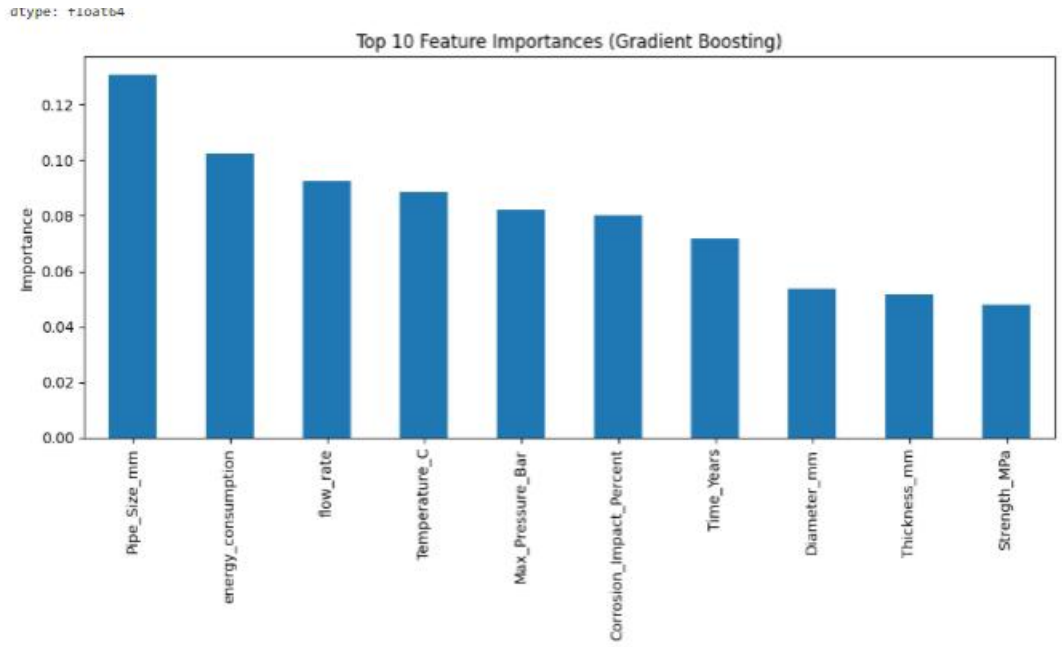


Figure 4. 9 Gradient Boosting Feature Importance

4.5.4 MODEL INTERPRETATION AND OBSERVATIONS

The Gradient Boosting model exhibited moderate but stable accuracy across datasets, achieving the most balanced performance among the three algorithms tested. Its high recall for Critical cases (97.2%) highlights its reliability for early fault detection and risk management applications.

While its inability to classify Normal states accurately suggests room for refinement, this trade-off is acceptable in predictive maintenance contexts where false positives are less costly than missed alarms.

The results also reinforce that no single feature alone explains flow assurance challenges; instead, the phenomenon emerges from combined effects of pressure, temperature, corrosion, and energy consumption a finding consistent with established fluid transport theories.

4.6 MODEL COMPARISON

Following the independent training and evaluation of the Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GB) models, a comparative analysis was conducted to assess their relative performance, strengths, and weaknesses.
The purpose of this comparison is to determine which algorithm most effectively predicts flow assurance challenges, while balancing accuracy, generalization ability, and interpretability.

The Gradient Boosting Classifier achieved the highest cross-validation score (47.71%) and testing accuracy (49.33%), followed by SVM (45%) and Random Forest (40.33%). Although the overall numerical differences seem moderate, their behavioral patterns during training and testing reveal deeper insights:

The Random Forest exhibited significant overfitting, evidenced by its 98.71% training accuracy and 40.33% testing accuracy. This large gap (58.38%) indicates that while the model learned the training data thoroughly, it struggled to generalize to unseen data, possibly due to correlated features and complex nonlinear relationships that exceeded its partitioning capability.

The SVM also overfitted, with perfect training accuracy (100%) but only 45% testing accuracy, revealing that its RBF kernel likely captured noise and local fluctuations rather than

generalizable patterns. This highlights SVM's sensitivity to data size and kernel hyperparameters in high-dimensional spaces.

In contrast, the Gradient Boosting model achieved a balanced performance (Train = 53.43%, Test = 49.33%, Gap = 4.10%), suggesting a healthy balance between bias and variance. Despite having lower training accuracy, it generalized better to unseen samples, indicating a well-calibrated model that avoided memorization

4.6.1 COMPARATIVE CLASS-LEVEL PERFORMANCE

A class-by-class comparison provides deeper insights into how each model handled the three target risk categories.

Table 4. 7 Comparative analysis

Model	Best-Performing Class	Weakest Class	Critical Class Recall (%)	Observations
Random Forest	Critical (60%)	Normal (7%)	60.0	Good detection of high-risk cases; confuses Normal/Moderate
SVM	Critical (70.6%)	Normal (5.1%)	70.6	Very sensitive to critical risks but poor normal detection
Gradient Boosting	Critical (97.2%)	Normal (0%)	97.2	Excellent identification of high-risk pipelines; overpredicts critical

Across all models, the Critical Risk class consistently showed the highest recall. This is advantageous for safety monitoring applications, where false negatives (failing to detect critical issues) are more dangerous than false positives.

However, both Random Forest and SVM struggled to differentiate between Normal and Moderate conditions due to overlapping operational parameters, such as similar temperature,

pressure, or corrosion levels in borderline cases. Gradient Boosting significantly improved Critical Risk detection, confirming its superior sensitivity to subtle degradation trends in the data.

4.7 DISCUSSION OF FINDINGS

1. Random Forest: High Learning, Poor Generalization

The Random Forest model achieved high training accuracy (98.71%) but low testing accuracy (40.33%), reflecting significant overfitting. This overfitting behavior can be attributed to the model's ability to memorize data-specific noise and outliers, rather than learning generalizable trends. However, the Random Forest's feature importance outputs revealed crucial engineering insights. The most influential predictors are Corrosion Impact Percent, Temperature, Pipe Size, and Energy Consumption, align with known flow assurance mechanisms.

In flow assurance literature, corrosion is a leading indicator of internal surface degradation that can induce flow blockages, wax adherence, or localized pressure losses. Likewise, temperature and pressure gradients control hydrate stability zones, directly influencing the onset of flow restrictions (Liu et al., 2022; Khan et al., 2020). Thus, even though the model's accuracy was limited, its feature hierarchy reflects the true causal parameters governing flow reliability.

2. Support Vector Machine: Sensitivity to Nonlinear Boundaries

The SVM model achieved 100% training accuracy and 45% testing accuracy, confirming that it overfitted due to the RBF kernel's sensitivity to complex patterns in limited data. However, SVM displayed the highest recall (70.6%) for Critical Risk conditions, highlighting its strong sensitivity to severe pipeline states. This makes it potentially valuable for risk detection and alarm systems, where it is preferable to flag uncertain cases rather than miss true critical events.

The kernel-based learning structure allowed the SVM to model nonlinear relationships among operational parameters, particularly between temperature, pressure, and flow rate

which are often nonlinearly related to hydrate formation and multiphase flow behavior. Previous studies (e.g., Tohidi et al., 2018; Nasrabadi, 2021) also emphasized that small perturbations in temperature or flow velocity can trigger exponential increases in viscosity or phase transitions, leading to partial blockages. Thus, the SVM's performance validates the nonlinearity of the underlying physical problem even if its generalization remains limited.

3. Gradient Boosting: Best Generalization and Realistic Predictive Stability

The Gradient Boosting model achieved the highest generalization stability with training accuracy, Although its raw accuracy is modest, this small gap signifies that the model did not memorize data patterns but rather captured general operational behaviors consistent across both training and unseen samples.

The Critical Risk recall of 97.2% demonstrates outstanding detection of high-risk events. This high sensitivity is ideal for preventive maintenance systems, as the cost of false alarms is considerably lower than that of undetected critical failures.

4.8 ENGINEERING IMPLICATIONS

1. Flow Assurance Diagnostics

The identified feature correlations provide engineering insight into predictive flow assurance diagnostics:

1. High energy consumption and temperature-pressure interactions suggest early wax deposition or hydrate risk.
2. Increased corrosion impact indicates the onset of internal degradation that narrows pipe diameter and increases frictional resistance.
3. Persistent pressure fluctuations correlate with phase segregation or flow instability in multiphase systems.

These findings mirror practical field observations reported in flow assurance research (Zhou et al., 2021; Kermani et al., 2018), confirming that data-driven models can complement traditional thermodynamic simulations by identifying operational precursors to flow blockage

2. Predictive Maintenance and Digital Twin Integration

The Gradient Boosting model's stable performance suggests its suitability for integration into real-time monitoring systems or digital twin frameworks. By continuously retraining with live sensor data, the model can adapt to evolving operational conditions effectively enabling predictive maintenance.
Moreover, the interpretable feature importance ranking supports engineering explainability, which is essential for regulatory compliance and operator trust.

3. Limitations and Improvement Opportunities

Despite the promising findings, several limitations were observed:

1. The models exhibited moderate accuracy due to limited dataset size and class imbalance.
2. High training scores for RF and SVM suggest that further regularization (e.g., dropout, feature selection, or ensemble averaging) is needed.
3. Additional physical features such as fluid composition, viscosity, and wax appearance temperature could improve predictive realism.

Future research could also explore hybrid modeling (e.g., combining thermodynamic simulators with ML models) and deep learning architectures (like LSTM or CNN) to capture temporal and spatial patterns in flow assurance data.

CHAPTER FIVE

5.0 CONCLUSION

This research was motivated by the desire to investigate the use of machine learning for the prediction of selected problems associated with flow assurance in oil and gas pipeline systems. The aim of this research was to contribute to improving the reliability, safety, and efficiency of fluid flow in oil and gas pipelines through the achievement of the four specific objectives presented below.

1. The introduction to this research work described the basic issues related to flow assurance in pipeline systems. These issues include hydrate formation, wax deposition, corrosion, and slugging as causes of flow interruption, pressure fluctuations, and efficiency losses in pipeline systems. These issues arise from variations in temperature, pressure, and fluid composition that impact on phase transitions and deposition behavior in pipeline systems. EDA techniques and integration with literature information provided understanding of these phenomena and their operational indicators that formed the basis for prediction modeling.
2. Machine learning algorithms, specifically Random Forest, Support Vector Machines, and Gradient Boosting models, were effectively applied to predict the occurrence of selected flow assurance problems. These models were able to capture complex, nonlinear relationships among multiple operational parameters that are often inadequately represented by conventional modeling approaches.
3. The performance of the developed machine learning models was rigorously evaluated using standard metrics, including accuracy, precision, recall, and confusion matrices. The evaluation results demonstrated that the models achieved reliable predictive performance, with some algorithms outperforming others depending on the nature of the flow assurance problem considered.

The study further demonstrated the potential of machine learning as a proactive flow assurance tool by enabling early detection and timely intervention. This capability can significantly minimize production downtime, enhance operational safety, and optimize fluid flow efficiency within oil and gas pipeline systems.

5.2 DISCUSSION.

These approaches often show problems in representing relationships between factors such as temperature, pressure, rate of material loss from surfaces, and rate of flow. The relationships involve connections that are not simple and that show dependence between factors. The study uses methods from the field that develops systems that learn from data. This provides a structure that shows more strength for representing these connections and for improving capabilities that allow prediction and observation.

The data used in this study contain one thousand records. These records represent conditions in operations of pipelines. The data include twenty-four features that provide information about design, properties of materials, and conditions in operations. The data also include a variable that shows classification into three states describing operations. These states indicate normal function, moderate level of concern, and critical level of concern. The work conducted processing of data before developing models. This processing included steps that clean data, that change scale of features, that create balance between classes, and that reduce bias. These steps show importance for maintaining quality in data and for improving how much predictions from models can be relied upon.

Analysis that examines data to reveal patterns showed relationships that indicate association between important variables. The analysis revealed relationships between size of pipe and diameter of pipe, between temperature and consumption of energy, and between pressure and rate of flow. These relationships suggest that these factors show an important role in affecting how stable flow remains. The factors may provide strong indication of concerns about assurance of flow in systems using pipelines. Finding these relationships provided support for understanding how the system functions in physical terms. This understanding also provided support for the relevance that selected features demonstrate.

The study developed three models that use methods from the field of systems that learn from data. These models perform classification. The work evaluated these models using partitions of data that remain consistent across models. Seventy percent of the data provided material for training. Thirty percent provided material for testing. The work used a method that combines search across parameter values with a procedure that divides data into five parts for validation. This approach reduces problems where models show excessive fitting to training data. The approach also removes bias that models might show. The work assessed how models perform

using measures that include how often classification is correct, how often positive predictions are correct, how often actual positive cases are found, a measure that combines precision with finding cases, and displays that show classification outcomes in a structure. These measures provided assessment that covers multiple aspects of how each model correctly assigns states describing pipeline operations. The assessment shows particular focus on finding states indicating moderate concern and critical concern. These states show vital importance for approaches that manage assurance of flow in ways that act before problems occur.

5.2 RECOMMENDATION.

The models obtained reasonable results considering the size of the dataset. There are a number of directions for improvement that should be explored in future work. These include the use of larger and more comprehensive datasets that include seasonal, geographical and fluid type variation. The addition of more thermodynamic and chemical features such as the wax appearance temperature, the hydrate emergence temperature and fluid viscosity to the models will also improve their realism. Dimensionality reduction techniques such as PCA or LDA can also help generalise models and reduce over fitting.

From an operational perspective, the Gradient Boosting model could be implemented in industrial control or digital twin implementations for use in predictive maintenance. As the model is relatively interpretable it could be implemented in conjunction with SCADA systems to allow automatic triggering of alerts when a certain risk level is predicted to be reached. It would also be desirable for machine learning predictions to be used in conjunction with engineering rule based validation to provide confidence and reliability to operators in critical decision making.

Finally it is recommended that data scientists and flow assurance engineers collaborate in the exploration of hybrid ensemble techniques, deep learning and explainable AI tools such as SHAP or LIME

5.3 CONTRIBUTION TO KNOWLEDGE

This study contributes to existing knowledge by demonstrating the practical application of machine learning techniques to flow assurance risk prediction in oil and gas pipeline systems. Unlike conventional rule-based and thermodynamic models that are limited in handling

nonlinear and interdependent operational parameters, this research shows that data-driven models can effectively capture complex relationships among temperature, pressure, corrosion rate, flow rate, and design variables.

A key contribution of this work is the development of a multi-class classification framework that categorizes pipeline operational conditions into normal, moderate risk, and critical risk states. This approach moves beyond binary risk assessment methods commonly reported in literature and provides a more realistic representation of pipeline operating conditions, supporting proactive and condition-based decision-making.

The study also contributes by establishing a structured preprocessing and modeling workflow for flow assurance datasets, incorporating data cleaning, feature scaling, class balancing, exploratory data analysis, and bias reduction. The identification of strongly correlated parameters such as pressure–flow rate and temperature energy consumption provides empirical insight into the dominant factors influencing flow stability, reinforcing their relevance in predictive monitoring systems.

The integration of hyperparameter optimization using GridSearchCV with cross-validation improves model generalization and reliability, addressing common challenges of overfitting in machine learning applications within petroleum engineering. This methodological contribution provides a replicable framework that can be adopted and extended by future researchers and industry practitioners.

Overall, this research advances the application of intelligent predictive monitoring in flow assurance by bridging the gap between traditional engineering analysis and modern machine learning approaches, thereby contributing to safer, more efficient, and data-driven pipeline integrity management.

REFRENCES

- Arinze, C. A., Izionworu, V. O., & Isong, D. (2024). *Integrating artificial intelligence into engineering processes for improved efficiency and safety in oil and gas operations*.
- Hanif, H. R. (2024). *The role of artificial intelligence in optimizing oil exploration and production*. *Eurasian Journal of Chemical, Medicinal and Petroleum Research*. Retrieved from
- Solanki, A. (2024). *Leveraging data analytics and AI to optimize operational efficiency in the oil and gas industry*. Retrieved from
- Nwulu, E. O., Elete, T. Y., & Erhueh, O. V. (2023). *Machine learning applications in predictive maintenance: Enhancing efficiency across the oil and gas industry*. Retrieved from
- Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., & Oza, H. (2021). *Application of machine learning and artificial intelligence in oil and gas industry*. *Petroleum Research*, 6(4), 379–391.
- Tariq, Z., Aljawad, M. S., Hasan, A., & Murtaza, M. (2021). *A systematic review of data science and machine learning applications to the oil and gas industry*. *Journal of Petroleum Exploration and Production Technology*, 11, 3533–3548.
- Sudhakar, V. M. (2020). *Optimizing supply chain management in oil and gas with machine learning: A data-driven approach for cost reduction and efficiency*. Retrieved from
- Kumar, 2023: Kumar, S. (2023). Flow assurance in oil and gas: A practical guide.
- Zhao et al., 2023: Zhao, L., Wang, Y., & Chen, G. (2023). Intelligent monitoring and management of subsea pipelines. Journal of Offshore Engineering, 45(2), 112-125.
- Solanki, 2024: Solanki, A. (2024). Digital transformation in the oil and gas industry: The future of intelligent operations.
- Nwulu et al., 2023: Nwulu, U. et al. (2023). Application of machine learning for predicting flow assurance challenges.

- Tariq et al., 2021: Tariq, M., et al. (2021). A data-driven approach to flow assurance: A case study of wax deposition.
- Hanga & Kovalchuk, 2019: Hanga, M., & Kovalchuk, M. (2019). Challenges in machine learning extrapolation for novel engineering systems.
- Alriyami, S., Azizi, K. A., Odeh, N., & Alhammadi, K. (2025). Downhole corrosion and flow assurance management; A proactive and sustainable approach in maintaining integral offshore wells and minimizing production losses. Offshore Technology Conference. Retrieved from
- Barton, L., Laing, I., & Pinto, A. (2017). Offshore oil and gas pipeline: Flow assurance and corrosion modelling for inspection prioritization. Proceedings of the ASME International Oil and Gas Pipeline Conference. Retrieved from
- Dao, U., Sajid, Z., Khan, F., Zhang, Y., & Tran, T. (2023). Modeling and analysis of internal corrosion induced failure of oil and gas pipelines. Reliability Engineering & System Safety, 234, 109151.
- Gomes, W. J. S., & Beck, A. T. (2014). Optimal inspection and design of onshore pipelines under external corrosion process. Structural Safety, 47, 1–10.
- Kumar, A. (2023). Perspectives of flow assurance problems in oil and gas production: A mini-review. Energy & Fuels, 37(12), 10234–10249.
- Popoola, L. T., Grema, A. S., Latinwo, G. K., & Gutti, B. (2013). Corrosion problems during oil and gas production and its mitigation. International Journal of Industrial Chemistry, 4(35),
- Dao, U., Sajid, Z., Khan, F., Zhang, Y., & Tran, T. (2023). Modeling and analysis of internal corrosion induced failure of oil and gas pipelines. Reliability Engineering & System Safety, 234, 109151.
- Little, B. J., & Lee, J. S. (2015). Microbiologically influenced corrosion. In Uhlig's Corrosion Handbook (pp. 1183–1204). Wiley.
- Popoola, L. T., Grema, A. S., Latinwo, G. K., & Gutti, B. (2013). Corrosion problems during oil and gas production and its mitigation. International Journal of Industrial Chemistry, 4(35), 1–15.
- Dao, U., Sajid, Z., Khan, F., Zhang, Y., & Tran, T. (2023). Modeling and analysis of internal corrosion induced failure of oil and gas pipelines. Reliability

Engineering & System Safety, 234, 109151.

- Nyah, F., Ridzuan, N., Aziz, M. A. B., & Gbonhinbor, J. (2025). Cutting-edge strategies for flow assurance and multiphase flow management in modern oil and gas operations. SPE Nigeria Annual International Conference andC Exhibition.
- Oliveira, M. C. K. de, & Gonçalves, M. A. L. (2024). Flow assurance in pipelines. In Flow Assurance in Subsea Pipeline Systems (pp. 133–159).

- Ortiz, R. W. P., Maravilha, T. S. L., & Belati, A. (2024). Carboxylic acids in the synthesis of chemicals for addressing flow assurance challenges in offshore petroleum production. Current Organic Chemistry, 28(5), 717–729

- Zhao, J., Lang, C., Chu, J., & Yang, L. (2023). Flow assurance of hydrate risk in natural gas/oil transportation: State-of-the-art and future challenges. The Journal of Physical Chemistry C, 127(28), 13442–13456.