Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage:
https://akademiabaru.com/submit/index.php/ard
ISSN: 2289-7984

Evaluation of Machine Learning Models for Predicting Maintenance Strategies in Oil and Gas Pipelines Based on Life-cycle Cost Analysis

Adamu Abubakar Sani[1,2,*] , Mohamed Mubarak Abdul Wahab[1,3], Nasir Shafiq[1], Zafarullah Nizamani[4], Waqas Rafiq[5], Atta Ullah[6]

[1] Department of Civil and Environmental Engineering, Universiti Teknologi Petronas, Bandar Seri Iskandar 32610, Perak, Malaysia
[2] Department of Building, Abubakar Tafawa Balewa University Bauchi, PMB 0248, Bauchi Nigeria
[3] Center for Urban Resources Sustainability, Institute of Self-Sustainable Building, Universiti Teknologi Petronas, Bandar Seri Iskandar 32610, Perak, Malaysia
[4] Faculty of Engineering and Green Technology, UTAR, Kampar 31900, Malaysia
[5] King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
[6] Department of Fundamental & Applied Sciences, Universiti Teknologi Petronas, 32610, Seri Iskandar, Malaysia

ARTICLE INFO

Article history:
Received 22 March 2024
Received in revised form 6 December 2024
Accepted 15 January 2025
Available online 31 January 2025

Keywords:
Machine learning models; life cycle cost analysis; predictive maintenance strategies

ABSTRACT

The research focuses on evaluation of machine learning models in the context of predicting maintenance strategies within the oil and gas pipeline, with a primary emphasis on life-cycle cost analysis. The study underscores the crucial shift from traditional, time-based maintenance practices to data-driven, predictive maintenance strategies, which hold significant potential for enhancing safety, reliability, and cost-efficiency for pipeline operators. To address limitations associated with data availability, an innovative methodology is employed involving the generation and utilization of synthetic data. Through the simulation of diverse pipeline scenarios, the research successfully creates a comprehensive dataset for the prediction of maintenance strategies based on cost-benefit ratios. The experimental results provide valuable insights into the strengths and weaknesses of various machine learning models. Notably, Random Forest Classifier and Gradient Boosting Classifier emerge as top-performing models for classification tasks, also the predictions show that corrective maintenance has the highest frequency compared to other maintenance strategies. This study contributes significantly to the ongoing efforts to improve pipeline management within the oil and gas industry.

1. Introduction

The global economy heavily relies on the oil and gas industry, with its infrastructure being essential for ensuring a consistent energy supply [1]. Central to this infrastructure is an extensive network of pipelines responsible for transporting hydrocarbons over extended distances. However, maintaining the integrity of these pipelines while addressing safety, environmental protection, and cost-effectiveness remains a formidable challenge for operators [2]. In this context, machine

* *Corresponding author.*
E-mail address: adamu_22000793@utp.edu.my

https://doi.org/10.37934/ard.124.1.6376

learning-driven predictive maintenance strategies have emerged as a promising avenue for optimizing the management of oil and gas pipelines.

This research aims to evaluate various machine learning models, with a primary focus on life-cycle cost analysis, to determine their effectiveness in predicting maintenance strategies within the industry. Oil and gas pipelines naturally age and are continuously in operation, exposing them to

various forms of degradation, including corrosion, mechanical damage, and material fatigue [3]. These issues can lead to leaks, environmental disasters, and operational disruptions. Traditional maintenance strategies based on predefined time intervals have proven inefficient, resulting in either excessive maintenance activities or insufficient attention to critical components [4]. In contrast, predictive maintenance utilizes data-driven algorithms to anticipate potential failures before they occur, enabling proactive scheduling of maintenance interventions. This shift from reactive to proactive maintenance not only enhances the safety and reliability of pipelines but also has the potential to significantly reduce life-cycle costs.

Machine learning algorithms have demonstrated their proficiency in analysing extensive datasets, including sensor information, historical maintenance records, and relevant data, to forecast when and where maintenance is most necessary [5]. Incorporating life-cycle cost analysis into this framework empowers operators to make well-informed decisions regarding the trade-offs between immediate maintenance expenses and potential long-term savings [6]. Nevertheless, despite the significant promise of machine learning in strengthening predictive maintenance strategies for oil and gas pipelines, a notable research gap exists within the field. While some studies have explored machine learning's application in pipeline maintenance, a comprehensive evaluation of a diverse array of algorithms and their impact on life-cycle cost analysis is conspicuously lacking. This paper's primary objective is to address this gap by systematically assessing the performance of multiple machines learning models, thereby shedding light on their strengths and weaknesses. Ultimately, this study aims to provide valuable insights to industry stakeholders, facilitating the development of more efficient and cost-effective maintenance strategies that promote the safety and sustainability of oil and gas pipeline infrastructure.

2. Literature Review

2.1 Machine Learning

Is a scientific discipline that allows computers to imitate human intelligence, learning autonomously from their experiences and surroundings [7]. This capability facilitates the discovery of knowledge and supports decision-making based on data [8]. Based on specific datasets, ML employs a data-driven methodology that seeks to build computational links between dependent and independent variables [9]. To extract knowledge and information from large amounts of data, machine learning (ML) relies on effective learning algorithms, large datasets, and powerful computing power [10]. Machine learning has found applications in numerous data-intensive domains, including bioinformatics, finance, engineering, health, and medicine. Some examples of its uses in these fields are data mining, recommender systems, information retrieval, autonomous control systems, and natural language processing [11]. Figure 2 illustrates a standard process for generating machine learning models. Machine learning involves two primary stages: learning stage and prediction stage. In the learning phase, the model gains the ability to derive conclusions from the input or dataset. The predictive proficiency of the model improves gradually during the learning phase, which consists of three essential processes: preparation, instruction, and evaluation [9]. A machine learning model usually deals with input that is unorganized, contains noise, lacks consistency, and is incomplete in its raw form. Through processes such as data cleansing, integration,

reduction, transformation, extraction, and fusion, the pre-processing stage ensured that the raw data is transformed into a form suitable for use during the training phase [12]. The training phase involves selecting the learning algorithm for the model, fine-tuning the model's hyperparameters, and conducting training using pre-processed datasets. In the evaluation stage, performance metrics such as accuracy, precision, and recall are employed to evaluate the model's effectiveness. The most successful model is subsequently utilized in the prediction phase to generate predictions using new datasets.

Based on the type of input that is provided to the learning system, the area of machine learning can be classified into three subdomains: (i) supervised learning, (ii) unsupervised learning, and (iii) reinforcement learning [13]. Below is a quick explanation of each ML subdomain.

2.1.1 Supervised learning

By combining independent variables with the identified dependent variable, supervised learning can deduce the connections between them. Common techniques employed in supervised learning encompass decision trees, neural networks, support vector machines, and k-nearest neighbours. Supervised learning can be categorized into two subtypes: classification and regression. When the dependent variable consists of a limited set of distinct values, the problem is categorized as a classification problem. In the field of PIM (Pipeline Integrity Management), classification is commonly employed for tasks such as identifying leaks, determining defect types, and predicting risk levels. On the other hand, regression is the classification for problems where the dependent variable has a continuous value. In PIM, regression is applied to forecast degradation rates and estimate the sizes of defects.

2.1.2 Unsupervised learning

Unsupervised learning focuses on discovering patterns or concealed structures within datasets that consist of diverse input variables and unknown output variables. Clustering, a specific unsupervised learning task, aims to categorize items into distinct clusters [14]. This grouping ensures that items within the same cluster are connected to each other while being separate from those in other clusters, determined by predefined criteria [15]. In the process of clustering, various techniques such as K-means clustering, hierarchical clustering, and the Gaussian mixture model are commonly employed. Applying clustering in PIM can streamline risk assessment by grouping together pipeline segments that share similar operating conditions, construction materials, and degradation mechanisms.

2.1.3 Reinforcement learning

The learning process in reinforcement learning is facilitated by the feedback received through rewards and punishments linked to specific actions [16]. A reinforcement learning system, like unsupervised learning, does not receive datasets containing sets of predetermined input-output pairs. The diagram in Figures 1 and 2 illustrates a classification of machine learning methods and Progression stages of a Machine Learning model’s development respectively.

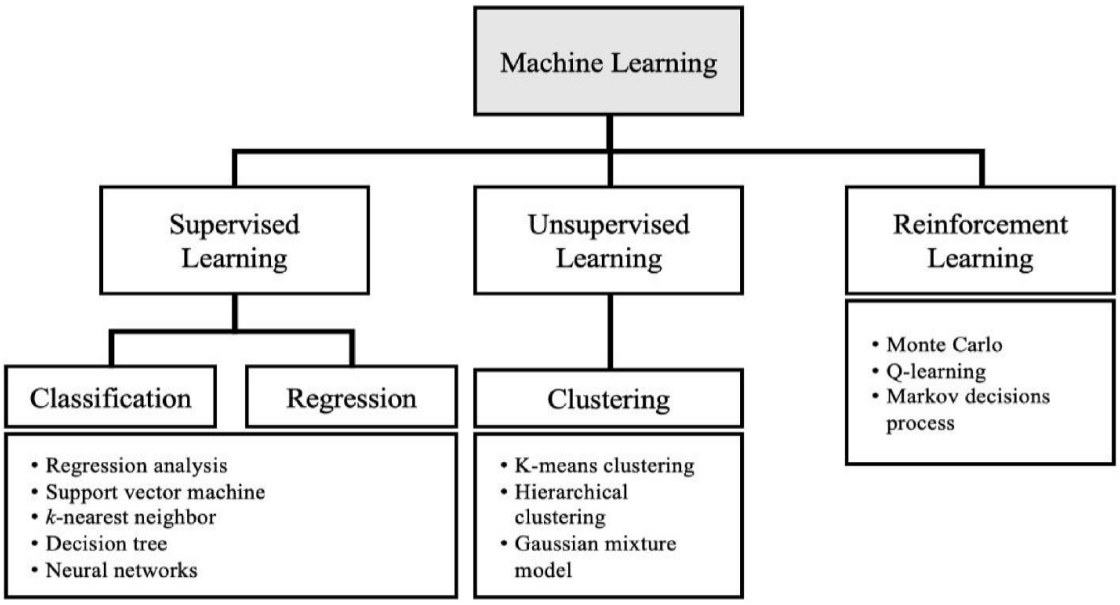

Fig. 1. Classification of machine learning methods [10]

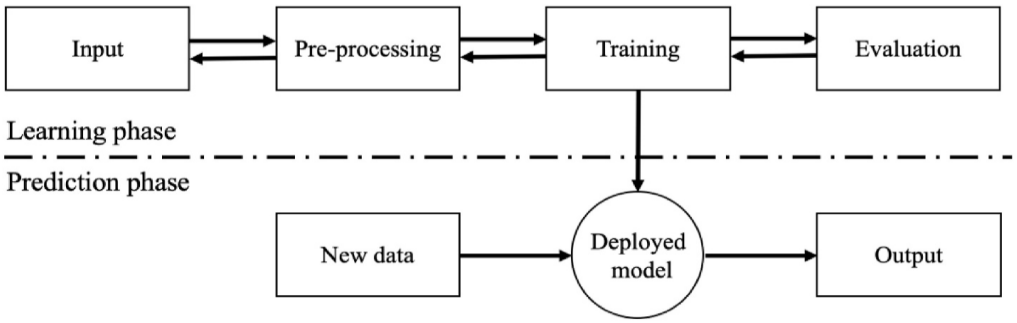

Fig. 2. Progression stages of a machine learning model's development [10]

2.2 Machine Learning Applications in Pipeline Integrity Management

ML significantly contributes to PIM by identifying irregularities, forecasting maintenance requirements, evaluating risks, optimizing inspection timelines, providing decision support, improving leak detection, monitoring corrosion, integrating various data sources, and supporting continuous enhancement [7]. These applications collectively enhance the reliability, safety, and efficiency of pipeline operations.

2.3 The Application of Life-cycle Costing (LCC) in the Oil and Gas Sector

Examining all expenses accrued over the entire lifespan of an asset, encompassing initial investment, continuous maintenance, and operational costs, as well as salvage and resale value, life-cycle costing (LCC) stands as a vital economic assessment to ascertain the comprehensive expenses associated with owning an asset. In the early 1960s, the concept of LCC was established by the US Department of Defence (DoD) with the aim of enhancing the effectiveness of federal procurement processes [17,18]. Subsequently, it has been utilized in a range of projects spanning different fields such as transportation, energy, manufacturing, and healthcare. This section provides a concise summary of pertinent LCC investigations related to the offshore oil and gas sector. Smith and Celant [19] employed the LCC approach to evaluate and compare the net present value (NPV) of various materials designed for downhole tubing. Consequently, the most advantageous option was chosen

in the initial stages of the Ekofisk redevelopment project in the Norwegian North Sea. Winkel [20] used LCC analysis to arrive at an ideal blend of materials and equipment. Paula *et al.,* [21] Utilized the LCC methodology to assess different iterations of this equipment and explored viable strategies enhancing the configuration of underwater manifold systems. The objective was to enhance the decision-making procedures in facility management for the refinery sector, including activities like inspections and the implementation of maintenance tasks. Iwawaki *et al.,* [22] created a methodology focused on activities for conducting Life Cycle Cost (LCC) analysis. This strategy is crafted to assess and compare expenses related to the introduction of new offshore structures.

According to the LCC analysis, choosing in-situ HDPE linings emerged as the most cost-effective choice, whereas opting for carbon steel with inhibitors or corrosion-resistant alloys (CRAs) was found to be less economically appealing. Creating plans for operations, maintenance, and support services is essential in devising strategies to minimize risks, Kayrbekova and Markeset [23] employing the LCC concept, the study tackled the difficulties related to utilizing LCC analysis for planning the operation and maintenance of intricate offshore oil and gas production facilities in challenging environments. In 2010, Li *et al.,* [24] proposed a cost-efficient optimal design idea aimed at minimizing the LCC of ice-resistant platforms within the Bohai Bay oil field in China. In research published in 2010, Kayrbekova and Markeset [23] reviewed the existing methodologies utilized on the Norwegian Continental Shelf (NCS) to apply the LCC concept. The document outlines the outcomes derived from the examination of diverse LCC regulations and discussions with industry experts. This presents a different perspective on traditional LCC in the field of engineering design, Kayrbekova and Markeset [23] developed a LCC model cantered around activities and applied it in a real-world setting in an offshore Arctic environment. Results from the study revealed that the activity based LCC, as opposed to the conventional approach concentrating solely on cash flows, adeptly oversees both costs and cash flows.

Evaluate possible offshore process options in the conceptual design stage, taking into account the related expenses and potential risks, Nam *et al.,* [25] proposed a new approach to LCC. According to the research results, the incorporation of LCC is crucial in determining the optimal liquefaction method for floating LNG production facilities. An LCC approach was discovered by Ortiz *et al.,* [26] on guide in choosing suitable production technologies for the development of heavy oil well construction projects. LCC analysis was utilised by Burlini and Araruna [27] Examining waste management efforts during the exploration phase of offshore oil and gas projects was the focus of this study. With the goal of aligning with existing regulatory standards, the objective was to support Brazilian businesses involved in oil and gas exploration by integrating the LCC approach into the decision-making processes related to waste management. A novel LCC approach was also discovered by Wang and Weng [28], assessing the potential cost savings of implementing base isolation for large LNG tanks requires an examination of how a reduction in seismic force during an earthquake might influence the overall expenditures. An integrated bottom-up LCC method was developed by Marten and Gatzen [29] in order to foster impartial decision-making, the method includes advocating for openness in the disclosure of costs among oilfield service providers. This approach was employed to suggest a viable plan for introducing a closed-loop rotary steering service for a company within the oilfield sector.

3. Methodology

The work specifically focuses on the evaluation of a machine learning models. It is essential to have adequate data to train a machine learning model when developing it. However, there may be situations when there is insufficient real data available or when the existing data has privacy or

confidentiality considerations that prevent it from being used. Synthetic data can be generated in such instances to simulate real data. Synthetic data is data that has been deliberately generated to replicate the patterns and relationships seen in real data. It can be used as a substitute for actual data when there is insufficient data or when real data is unavailable. Machine learning algorithms that learn from real data to create new data with similar relationships and trends can be used to generate synthetic data. Using synthetic data during the model's development could help cover data gaps and generate larger datasets for training the machine learning model.

A random sample of real data was obtained. The data was statistically analysed to identify relationships and trends between the attributes. Following that, synthetic data was created using machine learning algorithms that mimic the patterns and relationships identified in the real data. To ensure the synthetic data appropriately represents the real data, it must be assessed. Finally, the machine learning model was trained on synthetic data, and its performance was compared to that of a model trained on real data. While synthetic data can be a useful tool for developing models, it should always be checked against real-world data to ensure its accuracy and usefulness.

3.1 Experimental Setup

This approach for generating synthetic data within the experiment involved the consideration of a wide range of parameters associated with oil and gas pipelines, including pipeline characteristics, costs, and benefits. The experimental approach included the random selection of values within predefined ranges for each parameter, ensuring diversity within the synthetic dataset. Utilizing a random number generator, data points was generated that adhered to these specified limits, enabling us to simulate various scenarios related to pipelines. Subsequently, the total LCC was calculated and total life cycle benefit (LCB) for each synthetic case, maintaining adherence to the provided criteria for computing the cost-benefit ratio (CBR). This method helps to establish a comprehensive dataset, from which we can derive predictions concerning maintenance strategies based on CBR thresholds. Specifically, CBR values below 0.35 suggest routine maintenance, those ranging from 0.36 to 0.7 imply preventive maintenance, values from 0.71 to 1.0 indicate corrective maintenance, and values exceeding 1.0 signal the need for pipeline replacement. Table 1 below, shows the variable and the range of their corresponding values.

Table 1
List of variables

SN.	Variables/Features	Range
1	Nominal diameter	4 inches to 48 inches
2	Wall thickness	1.8mm to 24mm
3	Pipe grade	P5 to P91
4	Operating pressure (kg/mm^2)	200psi to 1500psi
5	Age (years)	50 to 100
6	Age of coating (years)	20 to 30
7	Type of pipeline	Gathering, transmission, distribution, flowlines, and feeder
8	Coating type	FBE, Polyolefin, Galvanizing
9	Discharge temperature	121°C to 135°C
10	Corrosion rate	0.4mm/year to 0.6mm/year
11	History of leaks due to intimal corrosion	250/year to 300/year
12	Number of employees	190 to 230
13	Design and construction cost	9 million to 13 million
14	Operational cost	7 million to 8 million
15	Maintenance cost	4 million to 7 million
16	Disposal cost	2 million to 4 million

17	Total life cycle cost (LCC)	22 million to 32 million
18	Social benefit	2 million to 14 million
19	Operational Benefit	2 million to 16 million
21	Environmental Benefit	2 million to 13 million
21	Total life cycle benefit (LCB)	6 million to 43 million
22	Cost Benefit Ratio (CBR)	< 1 or > 1

3.2 Performance Evaluation

Accuracy is a measure of how successfully a prediction model detects the true class labels of a set of test data. The accuracy score is determined as a percentage of the number of correct predictions divided by the total number of forecasts made.

Recall is a metric used to assess the efficacy of a classification model, especially when the class distribution is skewed. It calculates the percentage of real positive samples. Recall is calculated as shown in Eq. (1):

$$recall = \frac{true\ positives}{(true\ positives + false\ negatives)} \quad (1)$$

Precision is a metric used to assess the performance of a classification model, especially when the class distribution is skewed. It calculates the percentage of expected positive samples. F1-score (also known as F-score or F-measure) is a metric often used to assess classification model performance. It's a weighted average of precision and recall that maintains a balance between the two. F1-score is computed as shown in Eq. (2):

$$F1 - score = 2 \times \frac{(precision \times recall)}{(precision + recall)} \quad (2)$$

4. Results

Twenty thousand (20,000) samples of dataset were generated in four categories: routine maintenance, preventative maintenance, corrective maintenance, and replacement. The predictions are based on the assumptions that the CBR accurately reflects the asset's condition and are meant to assist maintenance personnel in making well-informed choices regarding the proper maintenance actions to conduct based on the asset's current condition.

4.1 Data Exploration or Visualization

Figure 3(a) to 3(g) below, show the data pattern of some features (variables). Each plot depicts the characteristics of a specific data pattern within a defined range of data points, typically the initial 1000 points. It likely serves as a component of a data exploration or visualization process, enabling an examination of how different data operate in the context of cost-benefit analysis. Frequency of each maintenance strategy also shown in Figure 4.

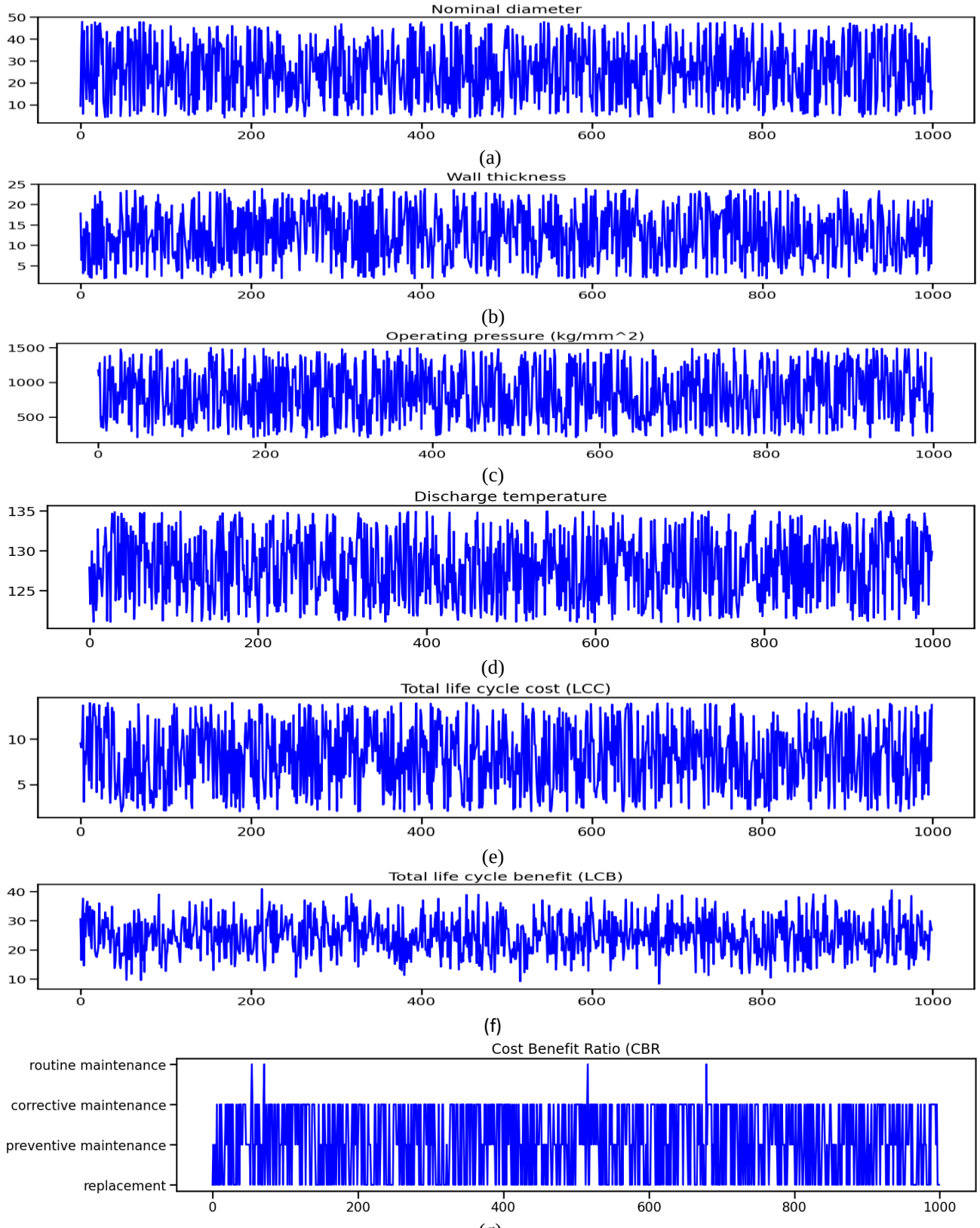

Fig. 3. Data visualization

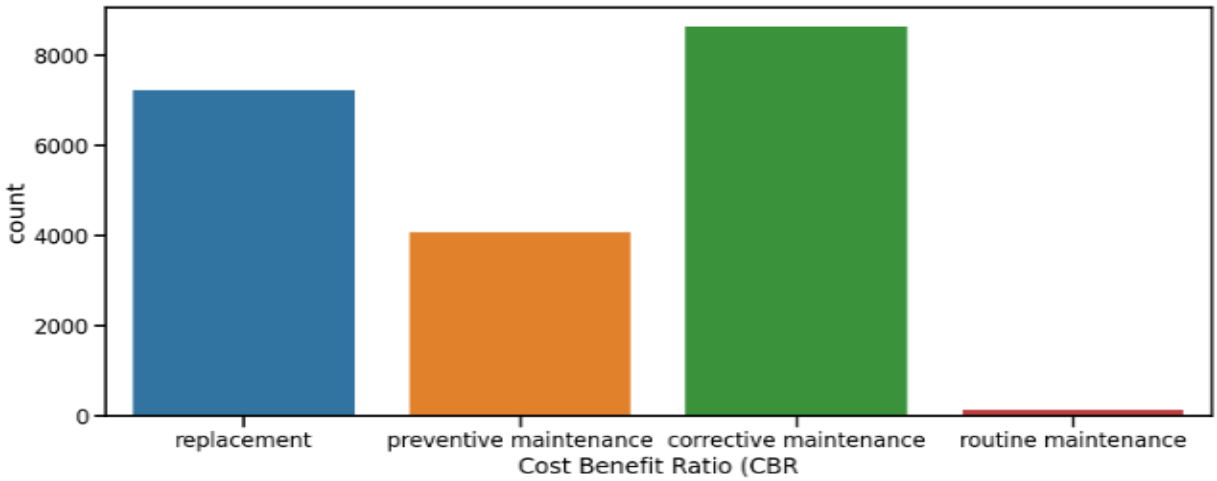

Fig. 4. Frequency of each maintenance strategies

4.2 Model Development

PyCaret was used to develop the machine learning model because it is easier to deploy and integrate machine learning models into real-world applications and can handle large datasets and parallel processing, making it faster than other machine learning models. Several classifiers can be used on the dataset, including Linear Discriminant Analysis, Decision Tree Classifier, Extra Trees Classifier, Random Forest Classifier, Light Gradient Boosting Machine, Extreme Gradient Boosting, Gradient Boosting Classifier, and Logistic Regression as shown in Table 2.

Table 2
Performance comparison of the models

Acronym	Model	Acc.	AUC	Recall	Prec.	F1	Kappa	MCC	TT (sec)
rf	Random Forest classifier	0.8872	0.9709	0.8872	0.8880	0.8873	0.8245	0.8247	1.4060
gbc	Gradient Boosting Classifier	0.8871	0.9750	0.8871	0.8878	0.8871	0.8244	0.8246	7.1250
qda	Quadratic Disc. Analysis	0.8850	0.9752	0.8850	0.8865	0.8852	0.8210	0.8214	0.0770
lgbm	Light GB Machine	0.8814	0.9732	0.8814	0.8820	0.8814	0.8156	0.8158	0.6880
et	Extra Tress Classifier	0.8805	0.9714	0.8805	0.8818	0.8801	0.81360	0.8140	0.4720
lr	Logistic Regression	0.8746	0.9699	0.8746	0.8752	0.8746	0.8052	0.8054	0.9140
nb	Naives Bayes	0.8712	0.9675	0.8712	0.8726	0.8716	0.8001	0.8003	0.0270
dt	Decision Tress Classifier	0.8404	0.8731	0.8404	0.8410	0.8404	0.7524	0.7525	0.0620
ridge	Ridge Classifier	0.8108	0.0000	0.8108	0.8186	0.8033	0.7004	0.7082	0.0670
ada	Ada Boost Classifier	0.7784	0.9359	0.7784	0.8237	0.7603	0.6643	0.7005	0.4670
svm	SVM – Linear Kernel	0.7551	0.0000	0.7551	0.8073	0.7373	0.6126	0.6509	0.1470
knn	K Near Neighbours Classifier	0.6774	0.8084	0.6774	0.6943	0.6720	0.4814	0.4930	0.0800
dummy	Dummy Classifier	0.4248	0.5000	0.4248	0.1805	0.2533	0.0000	0.0000	0.0220

4.3 Receiver Operating Characteristic (ROC) Curve

Figure 5 illustrate the ROC curve. The ROC curve visually depicts the trade-off between sensitivity and specificity at different classification thresholds, with the area under the curve (AUC) plot summarizing overall model performance. The x-axis represents the false positive rate, indicating the proportion of negative instances incorrectly classified as positive, while the y-axis represents the true positive rate or sensitivity. The curve is formed by connecting points corresponding to various

threshold values, illustrating the trade-off between sensitivity and specificity. The AUC value, a single metric derived from the curve, indicates the model's discriminatory power—higher values suggest better performance, while around 0.5 implies performance no better than random chance. In essence, the AUC plot offers a visual understanding of the model's sensitivity-specificity trade-off, and the AUC value provides a concise assessment of its overall effectiveness.

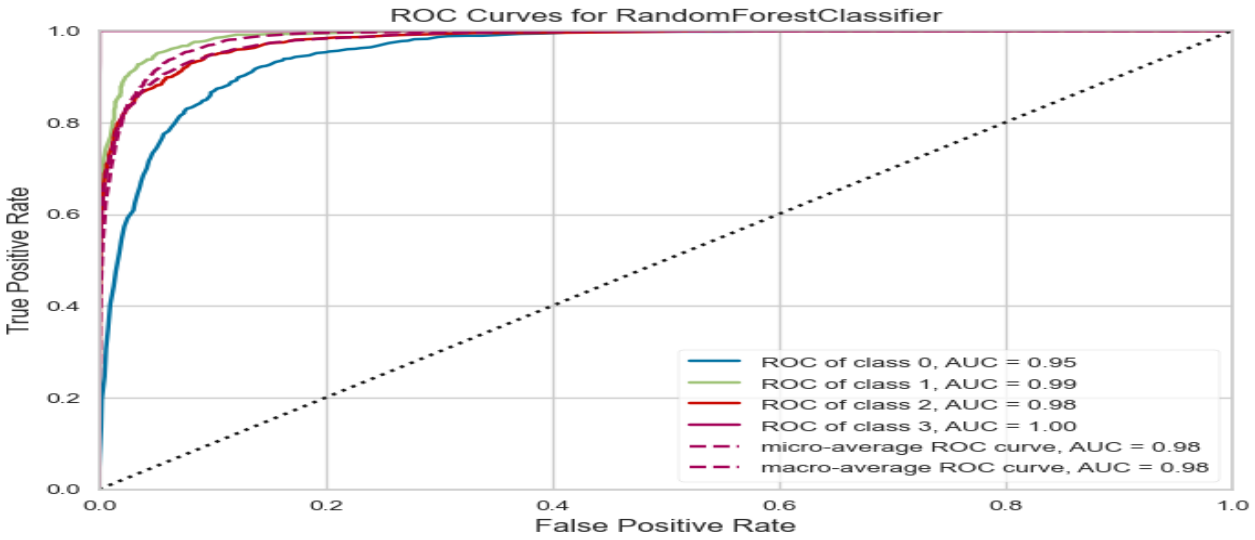

Fig. 5. ROC curve for random forest classifier

4.4 Validation Curve

A validation curve is a graphical representation illustrating how the model's performance metric, such as accuracy or AUC, changes with different values of a hyperparameter. The x-axis likely depicts varying values of the hyperparameter, while the y-axis shows the model's performance metric as shown in Figure 6. The curve typically consists of both a training curve, indicating performance on the training set, and a validation curve, indicating performance on a separate validation set. Analysing the validation curve aids in identifying optimal hyperparameter values that achieve a balance between good performance on both training and validation data, guiding the process of hyperparameter tuning.

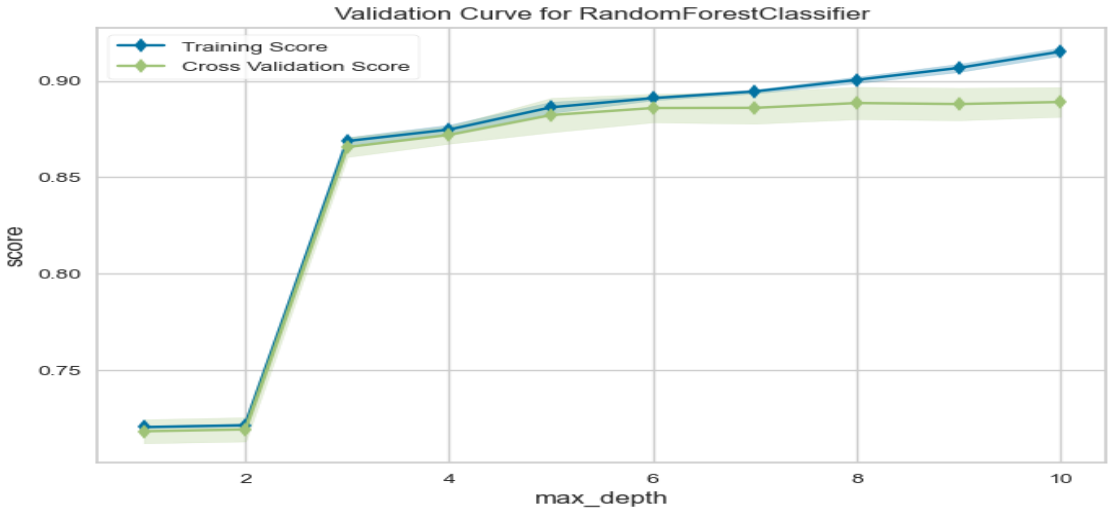

Fig. 6. Validation curve for random forest classifier

4.5 Confusion Matrix

The confusion matrix serves as a crucial tool in assessing the effectiveness of classification models. It offers a comprehensive summary of a model's performance by presenting the counts of various prediction outcomes in comparison to the actual class labels. The key components of a confusion matrix encompass Figure 7 illustrate confusion Matrix for random forest classifier.

I. True Positives (TP): Instances in which the model accurately predicted the positive class.
II. True Negatives (TN): Instances where the model correctly predicted the negative class.
III. False Positives (FP): Instances in which the model erroneously predicted the positive class when the actual class was negative.
IV. False Negatives (FN): Instances where the model incorrectly predicted the negative class when the actual class was positive.

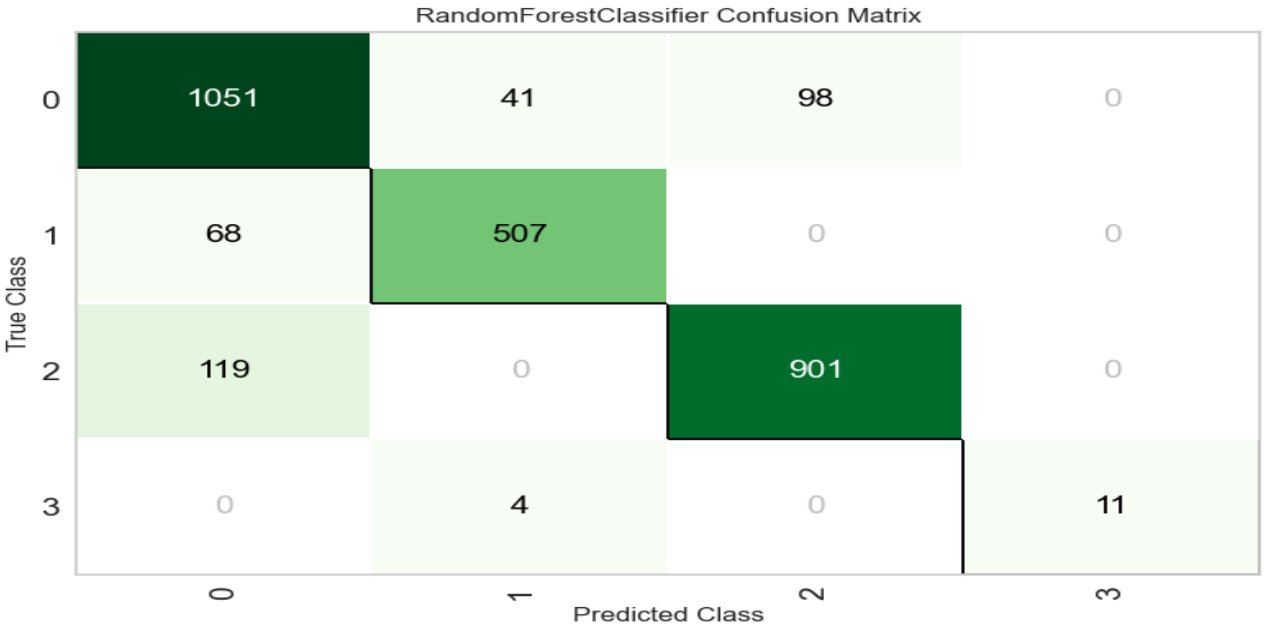

Fig. 7. Random forest classifier confusion matrix

4.6 Classification Report

Figure 8 below, represents a classification report, serving as a tool for assessing performance. It offers a detailed breakdown of metrics related to a model's classification performance across multiple classes. This report commonly encompasses precision, recall, F1-score, and support values for each class. The graphical depiction illustrates these metrics for each class within the random forest classification model. This visualization helps in comprehending the model's efficacy across different classes, enabling a more nuanced evaluation beyond a singular accuracy metric.

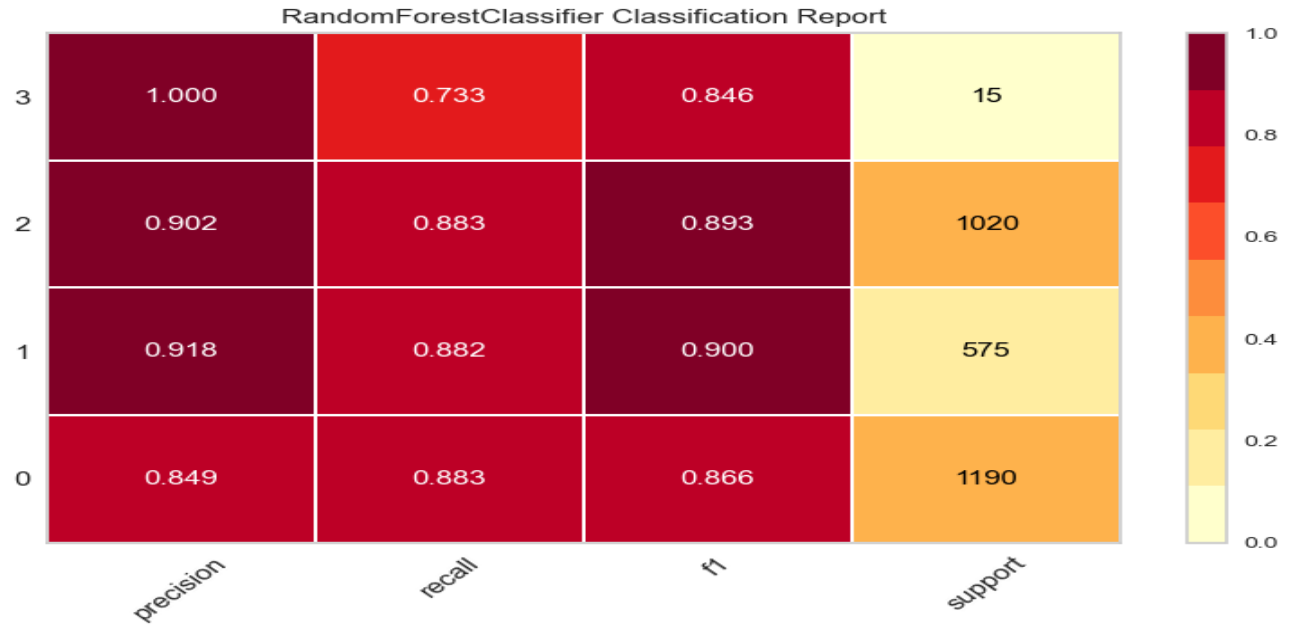

Fig. 8. Random forest classifier classification report

5. Discussions

Table 2 above shows how various machine learning models performed on a classification task. The accuracy, area under the receiver operating characteristic curve, AUC, recall, precision, F1 score, kappa, Matthew's correlation coefficient (MCC), and training time (TT) of the models are all examined. Random Forest Classifier and Gradient Boosting Classifier, linear discriminant analysis, decision tree classifier, extra trees classifier, linear regression, naive Bayes, quadratic discriminant analysis, AdaBoost classifier, ridge classifier, support vector machine with linear kernel, k-nearest neighbours' classifier, and dummy classifier are among the models available. The outcomes reveal that Random Forest Classifier has the highest accuracy and recall, although the AUC values for most of the models are similar. It is important to highlight, however, that AUC values are unaffected by class imbalance, which may exist in this dataset. Most models have excellent precision and F1 scores, indicating good performance in properly detecting positive samples. For most models, the kappa and MCC values are high, showing significant agreement between projected and actual labels. Different models have different training times, with some being faster than others.

AUC is a statistic often used to assess the performance of binary classification algorithms. The curve in question is the ROC curve, which is a plot of the true positive rate (TPR) vs. the false positive rate (FPR) as the classification threshold changes. The AUC score is a helpful indicator since it represents a model's overall performance over multiple threshold settings rather than just one. Lastly, the results indicate that, based on the provided assessment metrics, the top-performing models for this classification task are Random Forest Classifier, and Gradient Boosting Classifier. When choosing a model for deployment, it's crucial to take additional aspects into account, including interpretability, scalability, and computational resources.

6 Conclusions

This research paper emphasizes the significance of implementing machine learning-driven predictive maintenance strategies within the oil and gas pipeline sector. The assessment of various machine learning Models, with a specific focus on life-cycle cost analysis, provides valuable insights

into their potential for optimizing maintenance approaches. The transition from traditional, schedule-based maintenance practices to data-driven, predictive maintenance holds the promise of enhancing safety, reliability, and cost-efficiency for pipeline operators. The methodology applied in this study, including the generation and application of synthetic data to address data limitations, demonstrates an innovative approach to model development. By simulating a wide range of pipeline scenarios, the research effectively establishes a comprehensive dataset for predicting maintenance strategies based on cost-benefit ratios (CBR). The experimental findings illuminate the strengths and limitations of various machine learning models, with Random Forest Classifier and Gradient Boosting Classifier emerging as the most effective options for the binary classification task. However, it is crucial to consider other factors such as interpretability, scalability, and computational resources when selecting a model for real-world deployment.

In essence, this research contributes to ongoing efforts to enhance the management of oil and gas pipelines, ultimately promoting safety, sustainability, and cost-effectiveness within the industry. By addressing the research gap and providing a systematic evaluation of machine learning algorithms, this study empowers industry stakeholders to make informed decisions that ensure the integrity and reliability of pipeline infrastructure while minimizing life-cycle costs. As the oil and gas industry continues to evolve, the integration of machine learning into predictive maintenance strategies holds great promise for the future.

Acknowledgment

The authors gratefully acknowledge the support provided by Yayasan Universiti Teknologi PETRONAS (YUTP) Malysia, with the grant cost center 015LCO-421.

References

[1] Razi, Faran, and Ibrahim Dincer. "Renewable energy development and hydrogen economy in MENA region: A review." *Renewable and Sustainable Energy Reviews* 168 (2022): 112763. https://doi.org/10.1016/j.rser.2022.112763

[2] Chandima Ratnayake, R. M., and T. Markeset. "Asset integrity management for sustainable industrial operations: measuring the performance." *International journal of sustainable engineering* 5, no. 2 (2012): 145-158. https://doi.org/10.1080/19397038.2011.581391

[3] Wasim, Muhammad, and Milos B. Djukic. "External corrosion of oil and gas pipelines: A review of failure mechanisms and predictive preventions." *Journal of Natural Gas Science and Engineering* 100 (2022): 104467. https://doi.org/10.1016/j.jngse.2022.104467

[4] Juselius, Juha. "Advanced condition monitoring methods in thermal power plants." (2018).

[5] D. A. Data. "Machine learning." *Map Reduce programing, Demand Supply Management and Time series*. 2017.

[6] Schoenmaker, Dirk, and Willem Schramade. *Principles of sustainable finance*. Oxford University Press, 2018.

[7] Ali, Tasnuva, Azni Haslizan Ab Halim, and Nur Hafiza Zakaria. "3D Lightweight Cryptosystem Design for IoT Applications Based on Composite S-Box." *International Journal of Computational Thinking and Data Science* 3, no. 1 (2024): 40-54. https://doi.org/10.37934/ctds.3.1.4054

[8] Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science* 349, no. 6245 (2015): 255-260. https://doi.org/10.1126/science.aaa8415

[9] Worden, Keith, and Graeme Manson. "The application of machine learning to structural health monitoring." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365, no. 1851 (2007): 515-537. https://doi.org/10.1098/rsta.2006.1938

[10] Zhou, Lina, Shimei Pan, Jianwu Wang, and Athanasios V. Vasilakos. "Machine learning on big data: Opportunities and challenges." *Neurocomputing* 237 (2017): 350-361. https://doi.org/10.1016/j.neucom.2017.01.026

[11] Qiu, Junfei, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. "A survey of machine learning for big data processing." *EURASIP Journal on Advances in Signal Processing* 2016 (2016): 1-16. https://doi.org/10.1186/s13634-016-0355-x.

[12] Hu, Xiaohua. "DB-HReduction: A data preprocessing algorithm for data mining applications." *Applied Mathematics Letters* 16, no. 6 (2003): 889-895. https://doi.org/10.1016/S0893-9659(03)90013-9

[13] Russell, Stuart J., and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

[14] Suhaili, Shamsiah, Joyce Shing Yii Huong, Asrani Lit, Kuryati Kipli, Maimun Huja Husin, Mohamad Faizrizwan Mohd Sabri, and Norhuzaimin Julai. "Development of digital image processing algorithms via fpga implementation." *Semarak International Journal of Electronic System Engineering* 3, no. 1 (2024): 28-45. https://doi.org/10.37934/sijese.3.1.2845

[15] Prajapat, Rajendra, Ram Narayan Yadav, and Rajiv Misra. "Energy-efficient k-hop clustering in cognitive radio sensor network for internet of things." *IEEE Internet of Things Journal* 8, no. 17 (2021): 13593-1360. https://doi.org/10.1109/JIOT.2021.3065691

[16] Sutton, Richard S. "Introduction: The challenge of reinforcement learning." In *Reinforcement learning*, pp. 1-3. Boston, MA: Springer US, 1992. https://doi.org/10.1007/978-1-4615-3618-5_1

[17] Shields, Michael D., and S. Mark Young. "Managing product life cycle costs: an organizational model." *Journal of cost management* 5, no. 3 (1991): 39-52.

[18] Ghosh, Chanchal, J. Maiti, Mahmood Shafiee, and K. G. Kumaraswamy. "Reduction of life cycle costs for a contemporary helicopter through improvement of reliability and maintainability parameters." *International Journal of Quality & Reliability Management* 35, no. 2 (2018): 545-567. https://doi.org/10.1108/IJQRM-11-2016-0199

[19] Smith, L. M., and M. Celant. "Life cycle costing-are duplex stainless steel pipelines the cost-effective choice?." In *Offshore Technology Conference*, pp. OTC-7789. OTC, 1995. https://doi.org/10.4043/7789-MS

[20] Winkel, J. D. "Use of life cycle costing in new and mature applications." In *SPE European Production Operations Conference and Exhibition*, pp. SPE-35565. SPE, 1996. https://doi.org/10.2118/35565-MS

[21] Paula, M. T. R., E. L. Labanca, and Paulo Childs. "Subsea manifolds design based on life cycle cost." In *Offshore Technology Conference*, pp. OTC-12942. OTC, 2001. https://doi.org/10.4043/12942-MS

[22] Iwawaki, Hirohito, Yoshio Kawauchi, Masaaki Muraki, Duane Evans, and Shinobu Matsuoka. "Life Cycle Costing (LCC) Based Decision Making for Reactor Effluent Air Coolers in Refineries." In *NACE CORROSION*, pp. NACE-02483. NACE, 2002.

[23] Kayrbekova, D., and T. Markeset. "Economic decision support for offshore oil and gas production in arctic conditions: identifying the needs." In *Proceedings of the European Safety and Reliability Conference*, pp. 5-9. 2010.

[24] Li, Gang, Dayong Zhang, and Qianjin Yue. "Life-cycle cost-effective optimum design of ice-resistant offshore platforms." (2009): 031501. https://doi.org/10.1115/1.3124138

[25] Nam, Kiil, Daejun Chang, Kwangpil Chang, Taejin Rhee, and In-Beum Lee. "Methodology of life cycle cost with risk expenditure for offshore process at conceptual design stage." *Energy* 36, no. 3 (2011): 1554-1563. https://doi.org/10.1016/j.energy.2011.01.005

[26] Ortiz Volcan, Jose Luis, and Ramez Antonio Iskandar. "A life cycle approach for assessing production technologies in heavy oil well construction projects." In *SPE International Heavy Oil Conference and Exhibition*, pp. SPE-150709. SPE, 2011. https://doi.org/10.2118/150709-MS

[27] Burlini, Patricia Soares, and José Tavares Araruna. "Life Cycle Concept (LCC) in Waste Management in the O&G Offshore Exploration." In *SPE North Africa Technical Conference and Exhibition*, pp. SPE-164787. SPE, 2013. https://doi.org/10.2118/164787-MS

[28] Wang, Hao, and Dagen Weng. "Life-cycle cost assessment of seismically base-isolated large tanks in liquefied natural gas plants." *Journal of Pressure Vessel Technology* 137, no. 1 (2015): 011801. https://doi.org/10.1115/1.4027461

[29] Marten, Christian, and Matthias M. Gatzen. "Decreasing operational cost of high performance oilfield services by lifecycle driven trade-offs in development." *CIRP Annals* 63, no. 1 (2014): 29-32. https://doi.org/10.1016/j.cirp.2014.03.062