International Journal of Chemical and Biochemical Sciences
(ISSN 2226-9614)

Journal Home page: www.iscientific.org/Journal.html

© International Scientific Organization

Employing Ensemble Learning for Predicting External Corrosion Rates on Oil and Gas Platforms

Shweta Singh[1], Rupal Gupta[2], Hiral Gaud[3], Apurva Kumar R Joshi[4]

[1]*Maharishi School of Engineering & Technology, Maharishi University of Information Technology, Uttar Pradesh, India*

[2]*College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India*

[3]*Department of Biotechnology, Parul University, PO Limda, Vadodara, Gujarat, India*

[4]*Department of Biochemistry, School of Sciences, JAIN (Deemed-to-be University), Bangalore, India*

Abstract

The structural strength of gas and oil installations is threatened by corrosion, which creates maintenance problems and possible safety hazards. Properly estimating the rates of exterior corrosion is essential for preventative upkeep plan implementation and long-term structural reliability. This study investigates the use of ensemble learning techniques to improve corrosion rate prediction accuracy. This paper proposes a novel corrosion rate prognosis approach called Penguins Search Optimized Random Forest (PeSO-RF). Data samples were gathered from multiple onshore pipeline locations to evaluate the suggested PeSO-RF approach. The suggested approach has been implemented via the Python language and trained on the available data. Additionally, the suggested method's effectiveness is evaluated against other methods that are currently in use and examined in terms of multiple metrics like MAE, R2, RMSE and MSE. These results show that, when it comes to corrosion rate prediction, the suggested strategy performs better than the others. The suggested methodology provides property managers to upkeep specialists with a useful and precise instrument to evaluate as well as reduce corrosion-related hazards on oil coupled with gas sites.

Keywords: Oil and gas stations, corrosion, safety hazards, ensemble learning, Penguins Search Optimized Random Forest (PeSO-RF)

Full-length article **Corresponding Author*, e-mail: shwetasingh580@gmail.com

1. Introduction

The Metallurgical structures and machinery in oil, gas and refinery facilities are exposed to natural gas, crude oil, petroleum-based products along with sources of energy, water, chemicals, environment and soil. The oil, gas and refinery sectors are classified as high-risk industries because of the presence of substances that are flammable, explosive, poisonous to human health, or harmful to the environment [1]. An oil well is a bore that is drilled into the Earth to extract petroleum products and hydrocarbons from underground. In oil wells, natural gas, as well as water, is found in conjunction with other hydrocarbons, resulting in the production of natural gas. Gas wells are specifically constructed for the sole purpose of extracting natural gas [2]. Microbially Induced Corrosion (MIC) is the outcome of cooperative interactions among the metal surface, non-living products of corrosion as well as microorganisms and their byproducts. MIC, or Microbiologically Influenced Corrosion, is a complex phenomenon that is misunderstood by corrosion experts. Microorganisms, when present in specific concentrations and types, have been observed to increase the rate of corrosion in offshore systems [3]. The gas and oil sector has been the most corrosive source of energy since its inception. The power sector bears the largest portion of the overall cost associated with corrosion issues [4]. The materials utilized in the production of oil and gases are subjected to highly corrosive industrial conditions. While the overall rate of significant events in the oil and gas business is not too high, especially in the offshore field, the degradation of materials has the potential to cause expensive catastrophic failures that might have severe impacts on human life as well as the environment [5]. Corrosion affected by the presence or activity is referred as MIC. Multiple species are involved in MIC, thriving in colonies and constructing bio-films. They are capable of enduring and flourishing in harsh environments characterized by oxygen deprivation, absence of light, elevated salinity, extreme pH levels ranging from acidic to highly alkaline and varying temperatures [6]. “Corrosion-resistant alloys

(CRAs)" are employed in such situations due to their minimal susceptibility to general corrosion when exposed to elevated levels of "Carbon dioxide (CO2)" and "hydrogen sulfide (H2S)" under high pressures and temperatures. CRA is ambiguous because it refers to an element's response to oil field conditions rather than any inherent characteristics of the material, as opposed to carbon steel [7].

Study [8] developed an analytical approach for analyzing the time variation failure probability caused by the advancement of corrosion in pipelines for oil and gas. It was projected how long the pipes would last before failing and required to be replaced or repaired. Study [9] developed a hybrid intelligence algorithm technique for predicting corrosion rates of multiple-phase flow pipelines. PCA-CPSO-SVR, the suggested model, included "principal component analysis (PCA)", "chaos particle swarm optimization (CPSO)" and "support vector regression (SVR)". Research [10] created failure prediction models for exterior corrosion in under-the-ground gas transmission networks by considering conventional and environmental/geographical characteristics. Study [11] offered a platform risk assessment by forecasting the elimination age of existing frameworks in the "Gulf of Mexico (GoM)" Employing multiple ML methods, namely "gradient boosted regression tree (GBRT)" & "artificial neural network (ANN)." The study [12] introduced a probabilistic method for evaluating the seismic risk of pipeline infrastructure in Canada. Research [13] proposed a novel approach utilizing a "hybrid Bayesian network (BN)" and "Markov process" for accurately determining the "MIC" score, faults possibility and catastrophic failure duration of a subsea pipeline with internal corruption. "Bayesian Network (BN)" model was created to calculate an amount of MIC using a probabilistic approach, considering the interaction and dynamic non-linearity of important parameter values. Study [14] introduced an in-depth "Prognosis and Health Monitoring (PHM)" modeling structure for managing the integrity of gas pipeline systems. Its purpose was to prevent or minimize the occurrence of failures. The proposed PHM approach accounted for every possible mode of pipeline failure.

Research [15] provided a semi-supervised region generalization diagnosis technique for rusting leakage risk. Optical sensing equipment was successfully employed to detect and assess existing and potential leaks in pipelines. Study [16] created a functional ANN model to forecast the rate of air corrosion on carbon steel. Research [17] presented a methodology for forecasting corrosion rates based on a limited sample of laboratory-based metal corrosion data. The framework was devised to offer a novel approach for addressing the issue of pipeline corrosion in situations when there was a lack of enough genuine samples. The study [18] provided a novel evaluation system that utilized a blend of "fuzzy logic inference & machine learning" approaches. The factors that influence the criticality of pipeline failures in the framework encompass the impact of "transportation disruptions, safety as well as health considerations, environmental along with ecological effects and equipment maintenance". Study [19] presented an innovative experimental framework, which employed multiple learning algorithms to anticipate the decommissioning choice based on a newly acquired dataset. Study [20] introduced a novel method for identifying degradation characteristics of a "GE MS 5002B" air turbine that served in the Hassi R'Mel gases field located in southern Algeria. The suggested methodology mainly depends on "Long Short-Term Memory (LSTM)" systems, employing deep learning techniques to analyze operational data.

The primary objective of the suggested approach (PeSO-RF) serves to improve the precision of forecasting exterior corrosion rates in oil and gas systems. This improvement will facilitate efficient planning of preventive maintenance and guarantee long-term structural dependability and safety. The remaining research can be classified into the following categories: Section 2 is dedicated to the presentation of our proposed method. Section 3 outlines the experimental results of this study. Section 4 presents the results of this research.

2.2 Penguins Search Optimization Algorithm

The Penguins Search Optimization Algorithm is a method that draws inspiration from the hunting behavior of penguins. It involves collaborative efforts among penguins to optimize the overall energy expenditure and locate abundant food sources. The division of penguins into groups allows for an extensive exploration of the ocean to find the most abundant fishing spots. This objective is achieved by multiple dives and communication among the groups. The hunting procedure conducted on the penguins involves the modification of an algorithm called PeSO, which is designed to address problems related to combinatorial optimization. The PeSO process consists of five distinct steps:

Step 1: Initializing the algorithm parameters.

Step 2: Generate a population: This stage involves the random creation of a population of solutions.

Step 3: Construction of a new position: Once population p is generated, consisting of a certain number of groups, each group explores a place. After each exploration, the groups exchange information about the snow to enhance the research position for the next exploration, utilizing equation 1:

$$C\_new=C\_last+rand\ (\ )|W\_best-W\_id\ | \quad (1)$$

The rand () function produces a random integer based on a specified distribution. It is employed to ascertain the present solution (D_last), the optimal localized solution (X_best), the ultimate solution (X_id) and the fresh solution (D_new).

Step 4: Update Best Groups/Optimal solution: Following each dive and exchange of information, penguins ascertain the optimal group, namely the group that consumed the most quantity of fish.

Step 5: Verification of the stopping condition: The algorithm checks whether the stop condition is satisfied.

2.3. The Random Forest Method

RF algorithm is an ensemble approach that combines decision trees also trains several models using sampling from statistical data. When a new sample needs to be predicted, these models are employed to predict the new data set independently. The category of the new samples can be selected based on the premise that the minority is subordinate to the majority. Individual models exhibit sensitivity to data noise, leading to elevated variance. RF is a Bagging technique that relies on the bootstrap sampling approach. It has the capacity to reduce vulnerability to data disturbance and improve the accuracy and reliability of the

model. This method can be applied by training multiple algorithms on the same dataset and doesn't require any additional input. The RF method is trained by selecting a subset of features at random, rather than using all features. The model has two essential parameters: the total count of decision trees and the number of features employed to construct each distinct decision tree. The construction procedures of the RF system are as follows:

Step 1: A new data set containing m samples can be created by running m iterations of the samples using replacements from the previous data set. Moreover, a total of "f features" are chosen from the "n" available features to act as input features by applying the sample without replacement concept.

Step 2: Determine the coefficient of Gini impurities of the subgroup specimens belonging to "classC_k", represented as "C_k", in the new "sample-set D".

$$\text{Gini(D)} = 1 - \sum_{K=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2 \quad (2)$$

Compute the Gini index, denoted asGini(D, A), for each "feature A" & its "corresponding value a".

$$\text{Gini(D, A)} = \frac{|D_1|}{|D|}\text{Gini}(\text{D}_1) + \frac{|D_2|}{|D|}\text{Gini}(\text{D}_2)$$
$$D_1 = \left\{(\vec{x}, y) \in D \middle| \vec{x}^{(A)} = a\right\} \quad (3)$$
$$D_1 = \left\{(\vec{x}, y) \in D \middle| \vec{x}^{(A)} \neq a\right\} = D - D_1$$

Step 3: Choosing the most suitable feature and the most effective segmentation point: The "features A" and "corresponding value a", the features and segmentation points that minimize Gini impurity are considered ideal. Regarding their assertion, their training set is divided into two sub-nodes.

Step 4: Utilize a "recursive function" to carry out step 2 and 3 for both sub-nodes (m, f). Ultimately, a "decision tree" is generated.

Step 5: Repeat the specified actions in a sequential manner. To construct a Random Forest model, t decision trees will be generated over a span of one to four t periods.

2.4. Penguins Search Optimized Random Forest (PeSO-RF)

The Penguins Search Optimized random forest (PeSO-RF) is an innovative and sophisticated method for forecasting corrosion in oil and gas sites. This advanced prediction model combines a random forest algorithm with a Penguin Search Optimization (PeSO) approach, resulting in a strong and effective system for predicting corrosion. Random forests are highly effective in managing intricate data sets, but PeSO fine-tunes the hyper parameters of the model to improve accuracy. PeSO-RF utilizes a synergistic combination to investigate several aspects that affect corrosion in gas and oil platforms, including ambient conditions, material qualities and operational parameters. By incorporating Penguins Search Optimization, the random forest model is optimized to achieve optimal performance, resulting in more accurate and dependable corrosion forecasts. This innovative method shows great potential for the oil and gas sector, providing an advanced tool to actively control and reduce corrosion risks. As a consequence, it enhances safety, operational effectiveness and the lifespan of vital infrastructure in offshore settings. Pseudo code 1 shows the process of PeSO-RF.

Pseudo code 1: Penguins Search Optimized random forest (PeSO-RF)

```
class PeSORF:
def __init__(self, n_estimators, max_depth, peso_parameters):
    self.n_estimators = n_estimators
    self.max_depth = max_depth
    self.peso_parameters = peso_parameters
    self.estimators = []
deftrain(self, X_train, y_train):
for i inrange(self.n_estimators):
        penguins_population = self.generate_penguins_population()
        tree = DT()
        tree.train(X_train[penguins_population], y_train[penguins_population])
        self.estimators.append(tree)
defgenerate_penguins_population(self):
defforecast(self, X_test):
Forecast = []
for tree in self.estimators:
forecasts.append(tree.forest(X_test))
    aggregated_forecasts = aggregate_forecasts (forecasts)
return aggregated_forecasts
class DT:
def __init__(self, max_depth):
            self.max_depth = max_depth
                self.root = None
def train(self, X, y):
def forecast(self, X):
def aggregate_forecasts(forecasts)
```

3. Result

The tasks were executed utilizing the Python programming language on the Windows 10 operating system. The computational capabilities for these activities were facilitated by an Intel i7 10th Generation processor, while the system was equipped with 32 GB of RAM. A laptop was used as the testing device for these processes. In the results section, we assessed several outcome indicators, including the R-squared (R2), mean absolute errors (MAE), root mean square errors (RMSE) and mean square error (MSE). Existing methods, including gradient boosting regression tree (GBRT) [22], light gradient boosting machine (LightGBM) [22] and AdaBoost [22], have been employed. Accuracy and Loss outcomes are illustrated in (Fig. 2 and 3), respectively.

The "Mean Absolute Error (MAE)" quantifies the average absolute difference between predicted and observed values. The results of MAE are presented (Fig 4). The values obtained for AdaBoost, LightGBM and GBRT were 0.430, 0.576 and 0.552, respectively. The suggested method, PeSO-RF, outperformed other methods with an average MAE of 0.326. It demonstrates the superior performance of our proposed PeSO-RF method. In order to measure the average number of errors between the predicted and actual outcomes that provide a numerical evaluation of the accuracy of this approach, The "Root Mean Squared Error (RMSE)" is a commonly used measure in statistics. The RMSE results can be found (Fig. 5). The AdaBoost, LightGBM & GBRT models achieved scores of 0.624, 0.814 & 0.799, respectively.

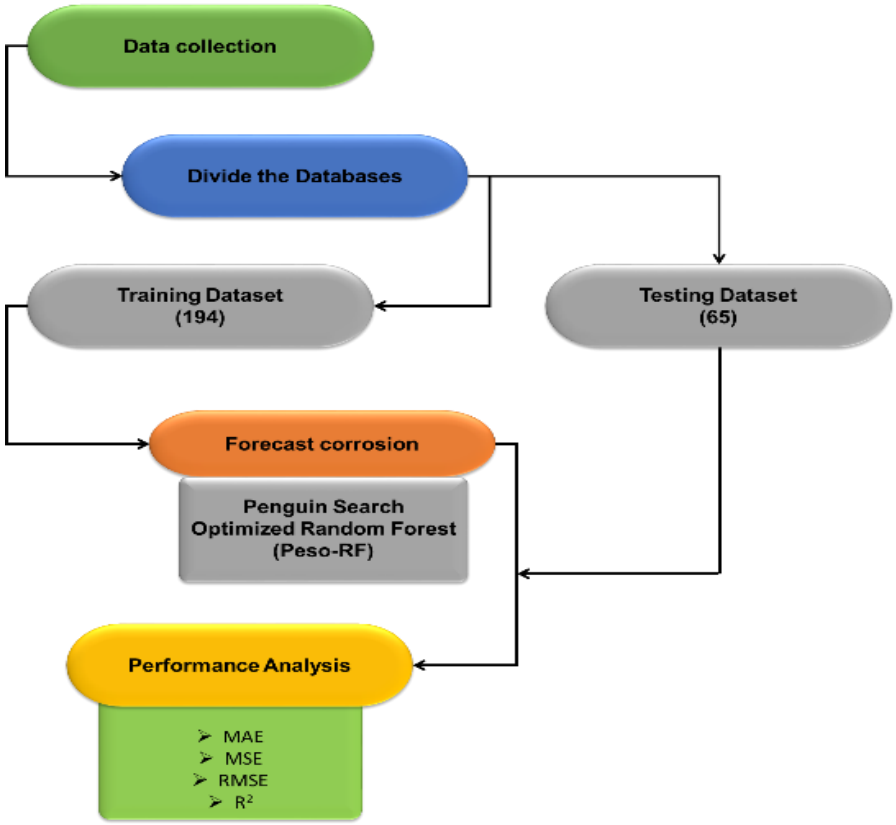

Figure 1. Flow diagram of the proposed PeSO-RF method (Source: Author)

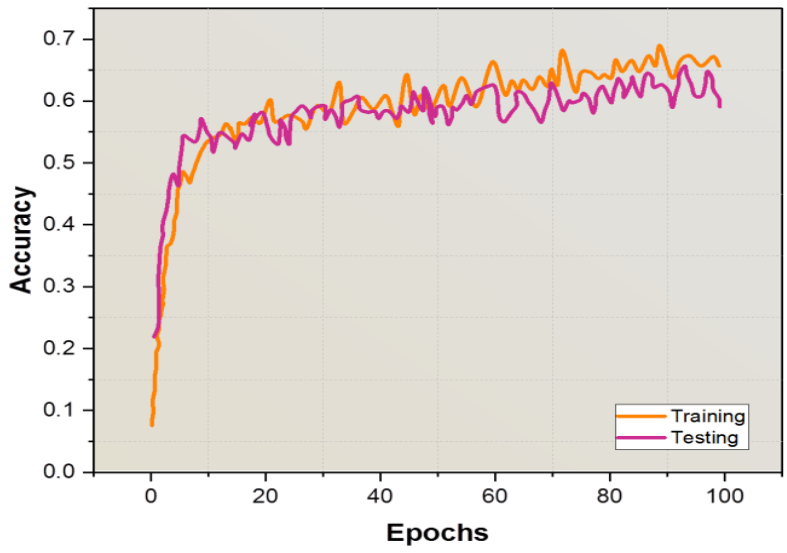

Figure 2. Outcome of Accuracy (Source: Author)

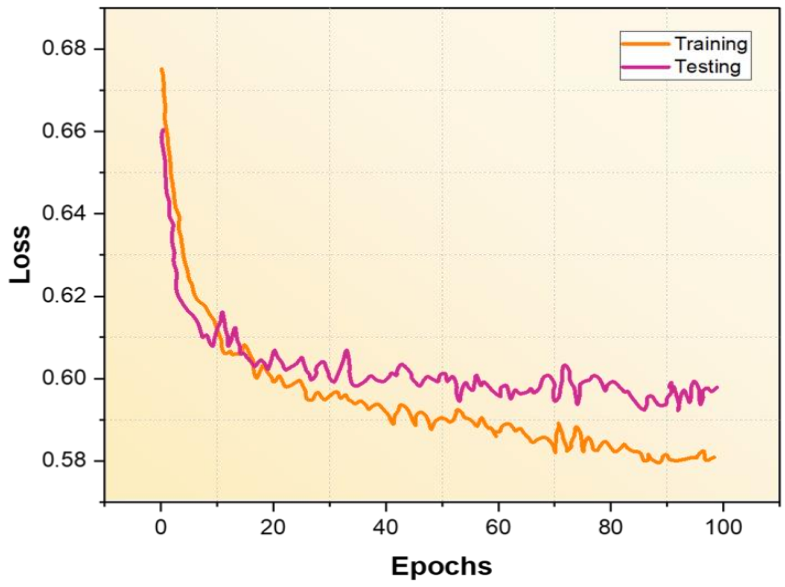

Figure 3. Outcome of Loss (Source: Author)

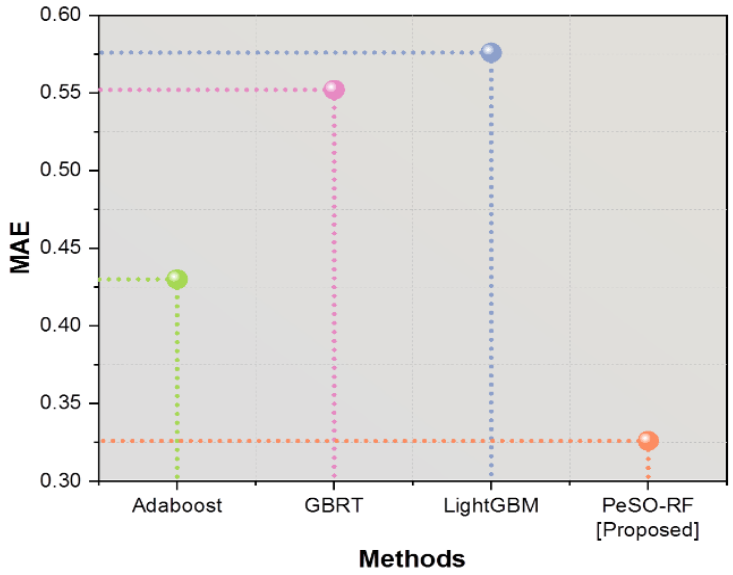

Figure 4. The result of MAE (Source: Author)

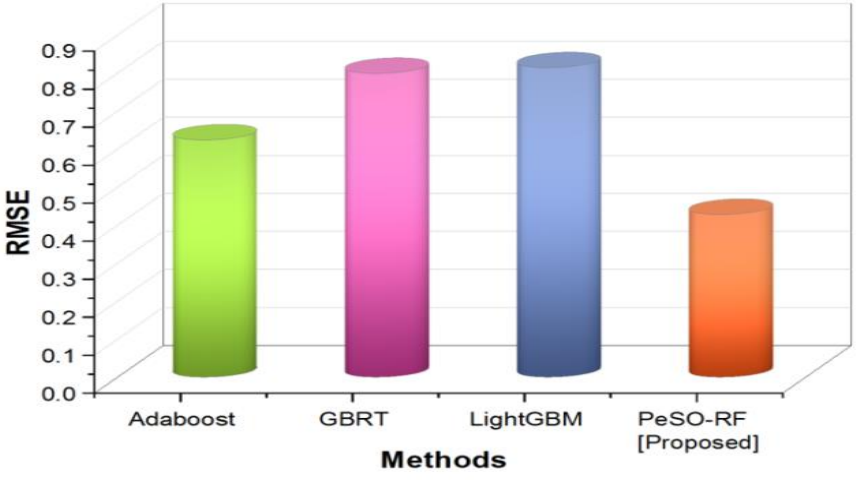

Figure 5. The result of RMSE (Source: Author)

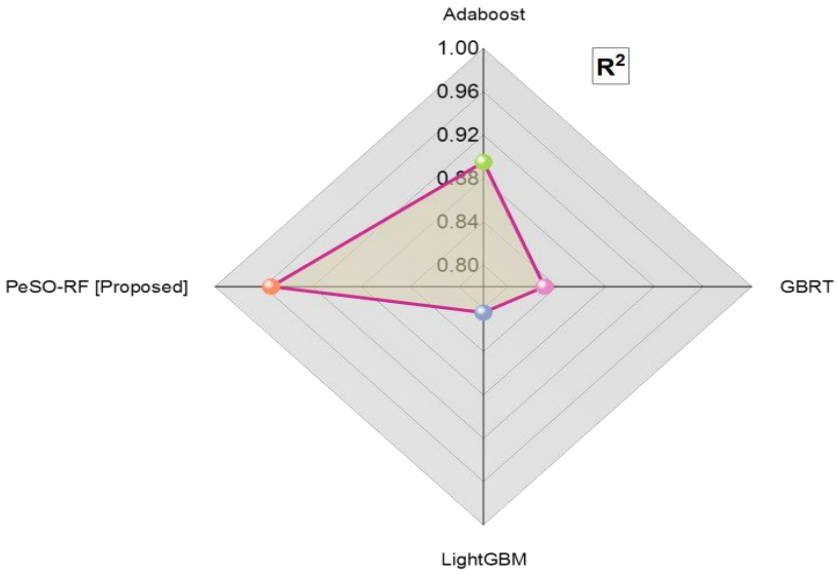

Figure 6. The result of R-squared (R2

)

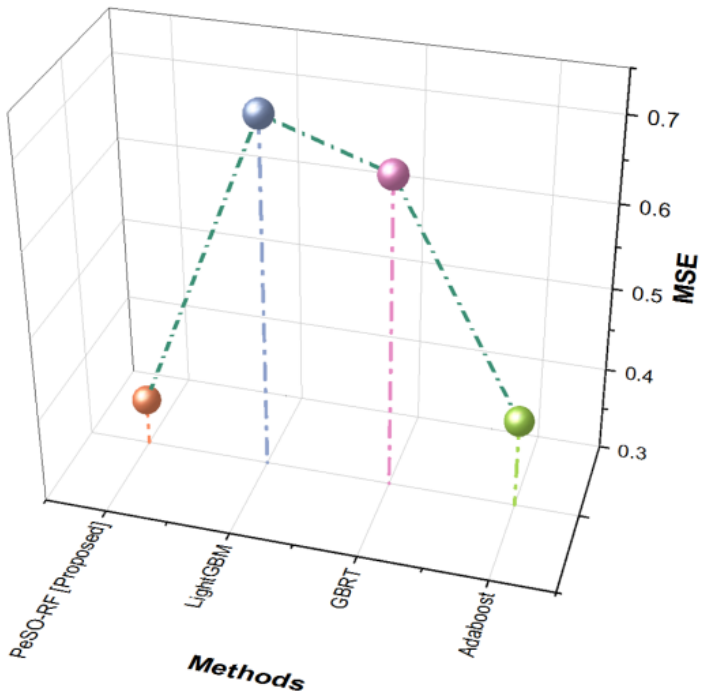

Figure 7. The result of MSE (Source: Author)

Table 1. Overview of the dataset

Variables	Time Unit (years)	pH	Pipe/soil potential (V)	Resistivity (W m)	Water content (%)	Bulk density (g/mL)	Chloride (ppm)	Bicarbonate (ppm)	Sulphate (ppm)	Redox potential (mV)(b)	Coating type	Maximum depth (mm)
All datasets	259	259	259	259	259	259	259	259	259	259	259	259
Minimum	5	4.14	-1.97	1.9	8.8	1.1	0.99	0.99	0.99	2.1	0.3	0.41
Maximum	50	9.88	-0.42	399.5	66	1.56	672.7	195.2	1370.2	348	1	13.44
Mean	22.988	0.6.130	-0.877	50.148	23.869	1.303	47.728	19.668	152.965	167.048	0.768	2.024
SD	9.118	928	0.239	55.919	6.659	0.088	75.159	25.33	168.182	85.484	0.128	2.046

Table 2. The metrics values for existing techniques and our proposed methodology

Methods	MAE	RMSE	R^2	MSE
Adaboost	0.430	0.624	0.895	0.406
GBRT	0.552	0.799	0.830	0.669
LightGBM	0.576	0.814	0.804	0.716
PeSO-RF [Proposed]	0.326	0.428	0.954	0.356

The proposed technique (PeSO-RF) yield superior outcomes, with an average RMSE of 0.428. The performance of our suggested PeSO-RF method is improved. "R-squared (R2)" is a measure of the extent to which an independent variable can explain the variability in the dependent variable. The range spans from 0 to 1. R2 analysis findings are displayed (Fig. 6). The values obtained for AdaBoost, LightGBM and GBRT were 0.895, 0.804 and 0.830, respectively. The proposed strategy, PeSO-RF, showed superior outcomes with an R2 value of 0.954. It demonstrates a significant improvement in the performance of our proposed PeSO-RF approach.

$$e_{j,i,r}^{f} = e\left(h_{j,i,r}^{f}\right) \quad (4)$$

The "Mean Square Error (MSE)" is a statistical measure that quantifies the average of the squared differences between predicted & actual values. A metric offers an arbitrary measure of the precision for a model, where smaller numbers signify superior performance. The results of the MSE analysis are shown in (Fig. 7). The results showed that GBRT, AdaBoost and LightGBM had respective values of 0.669, 0.406 and 0.716, respectively. The PeSO-RF method demonstrated superior results, as evidenced by a mean squared error (MSE) value of 0.356. Our proposed PeSO-RF technique exhibits a substantial enhancement in performance. Table 2 displays the metric values (MAE, RMSE, R2 and MSE) for existing approaches and our suggested method.

4. Conclusion

Accurate corrosion prediction in gas and oil platforms is crucial for improving structural integrity and ensuring safety. In this work, Penguins Search Optimized Random Forest (PeSO-RF) was introduced as a unique approach for improving the accuracy of corrosion rate prediction for oil and gas platforms. The suggested approach was executed utilizing the Python programming tool. To evaluate metrics such as MAE, RMSE, R2, & MSE, our proposed methodology outperforms existing methods. The computed values for RMSE, MAE, R2 & MSE were 0.428, 0.326, 0.954 & 0.356, respectively. A significant need for comprehensive corrosion data is one of the limitations, making it less applicable in situations where datasets are limited. The future involves improving PeSO-RF by incorporating Internet of Things (IoT) technology to obtain real-time data. Such information will be combined with advanced corrosion models and utilized for continuous learning.

References

[1] A. Groysman. (2017). Corrosion problems and solutions in oil, gas, refining and petrochemical industry. KOM–Corrosion and Material Protection Journal. 61(3): 100-117.

[2] G. Mubarak, C. Verma, I. Barsoum, A. Alfantazi K.Y. Rhee. (2023). internal corrosion in oil and gas wells during casings and tubing: Challenges and opportunities of corrosion inhibitors. Journal of the Taiwan Institute of Chemical Engineers. 150: 105027.

[3] M. Taleb-Berrouane, F. Khan, K. Hawboldt, R. Eckert, T.L. Skovhus. (2018). Model for microbiologically influenced corrosion potential assessment for the oil and gas industry. Corrosion Engineering, Science and Technology. 53(5): 378-392.

[4] A.S.H. Makhlouf, V. Herrera, E. Muñoz. (2018). Corrosion and protection of the metallic structures in the petroleum industry due to corrosion and the techniques for protection. In Handbook of Materials Failure Analysis. 107-122.

[5] M. Iannuzzi, A. Barnoush, R. Johnsen. (2017). Materials and corrosion trends in offshore and subsea oil and gas production. Materials Degradation. 1(1): 2.

[6] S.J. Hashemi, N. Bak, F. Khan, K. Hawboldt, L. Lefsrud, J. Wolodko. (2018). Bibliometric analysis of microbiologically influenced corrosion (MIC) of oil and gas engineering systems. Corrosion. 74(4): 468-486.

[7] N. Sridhar, R. Thodla, F. Gui, L. Cao, A. Anderko. (2018). Corrosion-resistant alloy testing and selection for oil and gas production. Corrosion Engineering, Science and Technology. 53(1): 75-89.

[8] M. Mahmoodian, C.Q. Li. (2017). Failure assessment and safe life prediction of corroded oil and gas pipelines. Journal of Petroleum Science and Engineering. 151: 434-438.

[9] S. Peng, Z. Zhang, E. Liu, W. Liu, W. Qiao. (2021). A new hybrid algorithm model for prediction of internal corrosion rate of multiphase pipeline. Journal of Natural Gas Science and Engineering. 85: 103716.

[10] K. Zakikhani, F. Nasiri, T. Zayed. (2021). A failure prediction model for corrosion in gas transmission pipelines. Proceedings of the Institution of Mechanical Engineers. Journal of Risk and Reliability. 235(3): 374-390.

[11] A.S. Dyer, D. Zaengle, J.R. Nelson, R. Duran, M. Wenzlick, P.C. Wingo, J.R. Bauer, K. Rose, L. Romeo. (2022). Applied machine learning model comparison: Predicting offshore platform integrity with gradient boosting algorithms and neural networks. Marine Structures. 83: 103152.

[12] O. Shabarchin, S. Tesfamariam. (2017). Risk assessment of oil and gas pipelines with consideration of induced seismicity and internal corrosion. Journal of Loss Prevention in the Process Industries. 47: 85-94.

[13] S. Adumene, S. Adedigba, F. Khan, S. Zendehboudi. (2020). An integrated dynamic failure assessment model for offshore components under microbiologically influenced corrosion. Ocean Engineering. 218: 108082.

[14] W. Chalgham, K.Y. Wu, A. Mosleh. (2020). System-level prognosis and health monitoring modeling framework and software implementation for gas pipeline system integrity management. Journal of Natural Gas Science and Engineering. 84: 103671.

[15] X. Miao, H. Zhao, B. Gao, F. Song. (2023). Corrosion leakage risk diagnosis of oil and gas pipelines based on semi-supervised domain generalization model. Reliability Engineering & System Safety. 238: 109486.

[16] N.L. Tran, T.H. Nguyen, V.T. Phan, D.D. Nguyen. (2021). A machine learning-based model for predicting atmospheric corrosion rate of carbon steel. Advances in Materials Science and Engineering. 1-25.

[17] Y. Yang, P. Zheng, F. Zeng, P. Xin, G. He, K. Liao. Metal Corrosion Rate Prediction of Small Samples Using an Ensemble. CMES-Computer Modeling in Engineering and Sciences. 134(1): 267-291

[18] H. Yin, C. Liu, W. Wu, K. Song, Y. Dan, G. Cheng. (2021). An integrated framework for criticality evaluation of oil & gas pipelines based on fuzzy logic inference and machine learning. Journal of Natural Gas Science and Engineering. 96: 104264.

[19] P. Vuttipittayamongkol, A. Tung, E. Elyan. (2021). A data-driven decision support tool for offshore oil and gas decommissioning. IEEE Access. 9: 137063-137082.

[20] A.Z. Djeddi, A. Hafaifa, N. Hadroug, A. Iratni. (2022). Gas turbine availability improvement based on long short-term memory networks using deep learning of their failures data analysis. Process Safety and Environmental Protection. 159: 1-25.

[21] M.E.A.B. Seghier, B. Keshtegar, M. Taleb-Berrouane, R. Abbassi, N.T. Trung. (2021). Advanced intelligence frameworks for predicting maximum pitting corrosion depth in oil and gas pipelines. Process Safety and Environmental Protection. 147: 818-833.

[22] Y. Song, Q. Wang, X. Zhang, L. Dong, S. Bai, D. Zeng, Z. Zhang, H. Zhang, Y. Xi. (2023). Interpretable machine learning for maximum corrosion depth and influence factor analysis. Materials Degradation. 7(1): 9.