Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ensemble Machine Learning for CO 2 Corrosion Rate Prediction with Heterogeneous Datasets

Joan Ejeta
North Carolina Agricultural and Technical State University
Tolu Emiola-Sadiq
University of Saskatchewan
Robert Eshun
North Carolina Agricultural and Technical State University
Kristen Rhinehardt
klrhineh@ncat.edu

North Carolina Agricultural and Technical State University

Article

Keywords: Machine learning, CO2 corrosion, model evaluation, ensemble machine learning

Posted Date: March 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8896642/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Ensemble Machine Learning for CO_2 Corrosion Rate Prediction with Heterogeneous Datasets

Joan Ejeta[1*], Tolu Emiola-Sadiq[2], Robert Eshun[1], Kristen Rhinehardt[1*]

[1] North Carolina A&T State University, Greensboro, NC, United States. [2] University of Saskatchewan, Saskatoon, Canada

[*]*Corresponding Authors: Joan Ejeta (joejeta@aggies.ncat.edu) and Kristen Rhinehardt (klrhineh@ncat.edu)*

Abstract

Corrosion accounts for billions of dollars in financial losses across the energy industry. However, limited access to quality, publicly available pipeline corrosion data significantly hinders accurate prediction, prevention, and the development of effective, data-driven maintenance strategies. This study develops an ensemble machine-learning framework to predict CO_2 corrosion rates in carbon-steel pipelines. It uses a heterogeneous dataset that integrates simulated data, experimental results, and field measurements to train ensemble machine learning models. Data preprocessing involved removal of outliers and imputation of missing values using simple imputer with Gaussian Mixture Model Expectation–Maximization, which preserves multivariate dependencies. To improve sensitivity to rare, high-consequence corrosion events, Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) was used to address target imbalances. Feature selection identified CO_2 partial pressure, pH, flow velocity, and temperature as dominant predictors. Hyperparameters of four ensemble regressors (Extra Trees, Gradient Boosting Regressor, Random Forest, XGBoost) were tuned using grid search and 3-fold cross-validation. The Gradient Boosting Regressor outperformed other models with accuracy and generalization on the test set (R^2 test=0.70; MSE=6.43 mm/yr). Model validation yields R^2=0.82 across 0–22 mm/yr and median absolute percentage errors below 50% in operationally critical regimes (≥1 mm/yr). The proposed machine learning framework offers a cost-effective, data-driven approach for improving pipeline integrity management on heterogeneous datasets.

Keywords: Machine learning, CO_2 corrosion, model evaluation, ensemble machine learning.

1. Introduction

The global transition toward net-zero CO_2 emissions has accelerated the development of green technologies, including carbon capture and storage (CCS) systems [1] and advanced strategies for monitoring and mitigating CO_2 emissions [2]. These initiatives require extensive infrastructure for CO_2 transport, highlighting the critical importance of pipeline system integrity. Recent studies demonstrate that low-carbon infrastructure solutions must address corrosion, integrity management, and operational constraints to achieve large-scale viability. Therefore, integrated approaches are necessary to ensure safety and reliability throughout the lifecycle of these systems. In this context, carbon-steel pipelines have been the primary focus of corrosion research, with particular emphasis on the investigation of their lifespan [3]. CO_2 corrosion is a significant challenge in petroleum production and transportation systems. This complex corrosion process is influenced by various parameters such as temperature, CO_2 partial pressure, flow velocity and pH [4] [5]. Modeling of CO_2 corrosion process has served as a useful tool to obtain a better understanding of this process and develop approaches to prevent it.

Several theoretical and experimental models; mechanistic, empirical and semi-empirical, have been developed to study CO_2 corrosion [6]. Mechanistic models rely on theoretical background and physical formulas to

represent the mechanisms of corrosion reactions [7]. For example, Gray et al. [8], [9] and Nesle et al. [10] introduced mathematical formulations that use mass, energy, and charge balances to describe chemical and electrochemical phenomena of corrosion. In contrast, empirical models [7], [11] are based solely on experimental regression and best-fit parameters, lacking a theoretical foundation. Models such as Norsok [12], [13] or the one proposed by Dugstad et al. [14] use equations derived from experimental and field data with statistical analysis. Semi-empirical models combine features of both mechanistic and empirical approaches [7]. They utilize parameters and formulas from the literature, guided by researcher experience, and include some basic mechanistic considerations, as seen in models by de Waard et al. [5], [15].

Traditional corrosion prediction models (empirical, semi-empirical, and mechanistic) have fundamental limitations in handling complex, multivariate corrosion phenomena. Laboratory-developed models often fail to generalize to field conditions. This is due to simplified experimental parameters and controlled environments that do not capture operational variability. Field data-based models are calibrated to high accuracy for existing conditions. However, they lack flexibility for adding new variables or phenomena. Substantial re-parameterization or model reconstruction is often needed. Conventional approaches struggle with nonlinear relationships, heterogeneous datasets, and the integration of many simultaneous influencing factors [16] [17] [18]. The scarcity of publicly accessible corrosion data contributes to this challenge. Inconsistent experimental protocols and industrial confidentiality restrictions prevent the creation of comprehensive, unified datasets covering diverse operational regimes[19]. Machine learning methods offer a complementary approach. They learn complex patterns directly from data without explicit mechanistic formulation. This allows integration of information from multiple heterogeneous sources while maintaining predictive accuracy across varied conditions[20], [21].

The increasing availability of computational power has spurred the development of many machine learning models for predicting pipeline corrosion rates. Machine-learning (ML) based corrosion rate prediction starts with identifying the key factors that influence corrosion. Different corrosion mechanisms are governed by distinct parameters. These parameters arise from design characteristics (e.g., diameter, length, material of construction) and operational conditions (e.g., fluid composition, pressure, temperature, flow rate) of the pipeline[22]. Once identified, these factors serve as input features for ML models, where corrosion rate estimation is formulated as a regression task [23]. Traditional statistical methods like linear regression have been employed, in the past. However, more advanced ML techniques such as Random Forest, Support Vector Machines, XGBoost, CatBoost [24], Bagging, and Gradient Boosting[25] offer improved predictive capabilities. The accuracy of these models depends heavily on the quality and quantity of training data. Early applications of machine learning to corrosion prediction primarily used artificial neural networks (ANNs) and support vector machines (SVMs) with limited datasets. Chamkalani et al. [26] used the Dugstad experimental database to train an ANN model using temperature, flow velocity, pH, CO_2 partial pressure, and ferrous ion concentration as input parameters. They found ANNs to be more robust than other algorithms under data-limited conditions. They achieved R^2 values of 0.996 and 0.993 for training and testing datasets, respectively. Aghaaminiha et al. [27] applied supervised machine learning methods to model corrosion rate in steel pipes, considering various dosages of corrosion inhibitors and changing conditions such as temperature, and CO_2 partial pressure. They observed that the corrosion rates depend on the dose schedule of the inhibitors while the eventual rate is affected by other environmental conditions. Fang et al.[28] utilized a hybrid approach. They simulated numerous over 1,000 corrosion scenarios of crude oil and gas pipelines using OLI Studio corrosion Analyser and two machine learning algorithms to predict the internal corrosion rates of the pipelines. The results showed the machine learning algorithms utilized successfully predicted corrosion rates based on factors like temperature and flow rate. Soomro et al. [29] systematically reviewed the integrity assessment of corroded oil and gas

pipelines using machine learning algorithms. They found that most research use experimental and field data, with field data being the most widely used. They also highlighted the importance of key input data as temperature, pH, pressure and flow rate, which were consistent in most studies reviewed. Abbas et al. [30] developed a neural network model for high-pressure CO_2 corrosion prediction. They implemented leave-one-out cross-validation for robust performance assessment and achieving an R^2 of 0.91. Mohammad Zubir et al. [31]augmented limited experimental data through synthetic generation using Gaussian functions combined with the de Waard-Milliams empirical model. Cui et al. [32] compared Deep Belief Networks (DBN), Generative Adversarial Networks, and Transformer models for CO_2 induced corrosion prediction. An optimized DBN achieved maximum relative errors ≤3% (MAE = 0.04, MAPE = 1.89%, RMSE = 0.04). Dong et al. [33] examined temporal corrosion patterns using several algorithms on a 248-sample dataset. Random Forest models best captured corrosion rate variations influenced by temperature, CO_2 partial pressure, and inhibitor effects. Feature selection and dimensionality reduction have proven critical for improving model performance. Yan et al. [34] applied correlation analysis to identify dominant environmental and material factors affecting corrosion, developing an optimized random forest model with R^2 values of 0.94 and 0.73 for training and testing, respectively. Ensemble learning methods have shown promise for small-sample scenarios. Ben Seghier et al. [35]applied ensemble techniques including Random Forest, Gradient Boost, AdaBoost, and XGBoost to a 60-sample subsea pipeline dataset. XGBoost achieved the highest accuracy (R^2 = 0.99, RMSE = 0.03 mm/year) and outperforming traditional empirical correlations [36]. Yang et al. investigated bagging methods integrated with k-nearest neighbors. They achieved a MAE of 0.03 compared to 0.13 for the Norsok model, though they noted that stacking approaches proved inefficient for small datasets. Despite these advances, fundamental challenges remain. Dataset sizes are still limited, experimental data collection is resource-intensive, and models trained on narrow parameter ranges may not generalize to diverse operational conditions.

This study addresses generalization limitations in CO_2 corrosion prediction by integrating experimental, simulated, and field datasets through a systematic data preprocessing and ensemble modeling framework. It presents three methodological contributions. First, a two-step imputation employs Simple Imputer and Gaussian Mixture Model-based Expectation-Maximization (GMM-EM) to handle missing data while preserving multivariate feature dependencies. Second, the Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) is applied to mitigate severe class imbalance in the target variable distribution. Third, ensemble learning algorithms are used to enhance prediction robustness and reduce model-specific bias. This methodology enables reliable corrosion rate prediction across diverse operational conditions in CO_2 transport and processing systems while maintaining predictive sensitivity to rare but operationally critical extreme corrosion events

2. Machine Learning Models

In machine learning, models may perform poorly individually but become more powerful when combined. This is a concept known as ensemble learning [37]. By aggregating the outputs of multiple models, ensemble methods mitigate bias and variance, leading to more robust and accurate predictions [38]. Ensemble learning integrates multiple base models to generate a single predictive outcome. Two primary aggregation strategies are bagging and boosting. In bagging, each individual model contributes equally to the final prediction, typically through democratic voting. This approach commonly decreases prediction variability by stabilizing model outputs [39]. Examples of Bagging are the Random Forest algorithm and Extra Trees method. The Extra Trees Regressor (ETR) is an ensemble regression algorithm built from numerous decision trees. It enhances predictive performance and robustness by aggregating the outputs of these trees. Unlike conventional decision tree methods, ETR incorporates a higher degree of randomness during the training process, thus increasing model

diversity. This approach is commonly referred to as Extremely Randomized Trees (ERT) [40]. In building each decision tree, the ETR randomly chooses a subset of features and cut-points at random, instead of evaluating all possible features and selecting the optimal cut points. This additional randomness helps limit overfitting and enhances the model's ability to generalize to unseen data. The algorithm generates many such trees and averages their predictions to produce the final regression. While the approach resembles that of Random Forest Regression (RFR), ETR incorporates a greater level of randomness during tree construction [41]. Because of this, ETR generalizes well to unseen data, thereby lowering the risk of overfitting. Compared to Random Forests, ETR also trains faster since it does not require finding the best cut-off points [40]..

Boosting is an algorithm used to reduce bias in supervised learning. It works by learning a series of weak classifiers and combines them into a strong classifier. A representative example of Boosting is the AdaBoost (Adaptive Boosting) algorithm [42]. Each training case has equal weight assigned, after which the algorithm trains the dataset for a number of rounds. Higher weights are assigned to misclassified training cases at the completion of each round. Thus, the learning algorithm focuses more on the difficult samples in later rounds, leading to multiple prediction functions [42] . Other methods in the Boosting family include GBR (Gradient Boosting Regressor), XGBoost (Extreme Gradient Boosting), and CatBoost (Categorical Boosting)[43].

The GBR model is built by training a series of weak learners, typically decision trees, and fixing errors from the previous round [44]. Each round concentrates on samples that were initially predicted inaccurately. The model uses gradient descent to optimize the loss function, thereby steadily reducing any prediction error. Each tree models the residuals and updates predictions by weighting them [45]. Although GBR handles a variety of loss functions, including those for regression and classification, it can be quite time consuming with large datasets [46], [47].

XGBR makes many improvements on GBR, through regularization, column sampling, and pruning thus reducing overfitting, memory usage and training duration [48].. The model achieves this by choosing a subset of features through randomization [49]The ability of this model to increase training speed stems on its capacity to support parallel computing [48]. XGBR is flexible when dealing with a wide range of loss functions and evaluation metrics, however, because it may require adjustment during training, it sometimes need more time to detect the optimal configuration [50].

CatBoost deals with categorization features via a unique encoding method (target-based statistics, thereby avoiding the issue of information loss from simplistic encoding in some traditional methods [51]. CatBoost employs symmetric (oblivious) trees in which a single split rule is applied to all nodes at each level. This structure yields balanced trees, enhances model stability, and supports efficient inference [51][52]. Similar to XGBoost, the model is also based on the GBR model but with specific optimizations and improvements. It has the advantage of not requiring manual preprocessing of classification features, thereby simplifying the feature engineering process [52]. It performs well on various datasets, especially when dealing with high-dimensional data. However, the model may require higher computational resources and time due to the need for more complex operations. Since availability of computational resources is not a bottleneck for this research work, CatBoost is used to perform regression analysis in this study.

The justification for the selection of these ensemble algorithms is due to their good performance and use in existing literature [49,52,53]. Hence it establishes as a benchmark in our model comparison with literature.

3. Methodology

In ML, algorithms are processes that take input data and produce outputs. They let computers learn patterns from data enabling predictions. The steps to perform prediction using ML algorithms include: (1) data preprocessing, (2) tuning of hyperparameters for ML models, (3) selecting the best ML algorithm, and (4) training the optimum ML algorithm and testing its performance.

3.1 Data preprocessing: The first step is to ensure that the dataset does not have any missing or incorrect values [45,46]. Data points with missing values are replaced with the simple imputer method from sklearn python library using expectation maximization to establish the threshold distribution of the variables. The EM algorithm is an iterative process. First, it uses other variables to estimate a missing value (expectation step). Then it checks if the imputed value is the most likely estimate (maximization step). The process repeats and refines the estimate until it converges to the most likely value[54]. The EM algorithm assumes an explicit joint distribution between missing and observed data. It aims to find the parameter estimates that maximize the log-likelihood of the observed data, allowing inference of missing values. The EM was used for missing data imputation because it produces unbiased parameter estimates and preserves the original distribution of the dataset[55]. Different variables have a different range of values and so it is important to rescale the variables. Our approach is to rescale the variables so that all the values are using Normalization method. This assumes that the values are normally distributed and so one can convert them into standard normal variables by subtracting out the mean and dividing by the standard deviation. Table 1 lists the variables and their ranges in the combined data of corrosion rate measurements that we have modeled in this work. The dataset includes eight numerical features. In the implementation of our ML strategy, we found that for accurate modeling of the data, some additional data preprocessing steps were needed as discussed in the results section. Feature handling which includes outlier removal, class balancing, missing values imputation, scaling and removal of high correlated features was done on the dataset before model predictions.

3.2 Tuning hyperparameters: Generally, K-fold cross-validation is performed on the training set to determine the best set of hyperparameters. Once the best set of hyperparameters is found, then the performance of the models is tested on the testing set. Prior to the hyperparameter tuning, the top performing ML models from AdaBoost, ETR, GBR, RFR, XGBR, CatBoost are chosen. Each ML algorithm (ETR, GBR, RFR and XGBR) requires optimal hyperparameters. These are hyperparameters that will result in the smallest mean squared error (MSE) and largest R^2 value [44]. The best hyperparameters were found by performing a grid search over a range of hyperparameter values. For each set of hyperparameter values, a 3-fold cross-validation method was used on the data. In 3-fold cross-validation, the data is divided into 3 subsets. The ML algorithm is then trained using the data in two subsets, and the performance of the trained algorithm is tested on the third subset. This procedure is repeated 3 times, each time choosing a different subset for testing the performance [33,34]. Then, the average performance of the ML algorithm (in our case, MSE, RMSE and R^2 score) is determined. Using this methodology, the best hyperparameter values of the different ML models (ETR, GBR, RFR, XGBR) were determined. For XGBR and GBR models, 10 iterations of 3-fold cross-validation were performed to pick the best as they had close results.

3.3 Selecting the optimum ML model: After tuning the hyperparameters, we selected the best ML model and divided the entire dataset into a training set, comprising 70% of the data, and the remaining as the testing set 30% using GridSearch CV. The ML models were trained on the training set. Then we evaluate their performance on the testing set. The model with the lowest MSE and highest R^2 value, was selected as the optimum ML model.

3.4 Training the optimum ML model and testing its performance: A practical use of our ML modeling is to predict corrosion rates for different operating conditions. To evaluate the performance of the proposed approach, we used 3 standard evaluation metrics for the regression task, mean absolute error (MAE), root mean square error (RMSE) and goodness of fit (R^2), to assess the output of the predicted results. The mathematical expressions of these performance metrics are provided as equations 1, 2 and 3 [56]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y^i - \hat{y}^i\right| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y^i - \hat{y}^I\right)^2} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y^i - \hat{y}^i\right)^2}{\sum_{i=1}^{n}\left(\bar{y}^i - \hat{y}^i\right)^2} \quad (3)$$

Assuming a set of samples $y = \{y_1, y_2, \ldots, y_n\}$ and the predicted dataset $y = \{\widehat{y_1}, \widehat{y_2}, .., \widehat{y_n}\}$, where n is the number of a sample dataset. For training and testing datasets, the smaller the MAE and RMSE values, the closer the R^2 value is to 1, indicating that the model's prediction results are closer to the real experimental results, the better the prediction ability.

4. Results and Discussion

4.1 Data Preprocessing

Three distinct types of pipeline corrosion rate data from the literature were combined for algorithmic prediction. The dataset consists of 243 simulated corrosion data points from reference [57] and [58] generated using the M-506 corrosion rate model, which is an empirical model developed by NORSOK. It also includes 50 experimental data from reference[30] where corrosion rates for carbon steel were determined by weight loss using autoclaves. Additionally, 60 field data points from reference [36] and [35], come from a submarine oil and gas pipeline in eastern Hainan, China. Another 74 field data points from reference [22] were collected under a water oil-gas multiphase flow conditions in the Field Egua-2 (NDANE-2) flow line in the Niger Delta of Nigeria. The merged dataset contains 427 observations with eight features and corrosion rate as the target variable. These datasets have been combined and the link to the combined dataset is available in the data availability section of this paper. Table 1 shows the statistical characteristics of the input and output variables for the presented case study at mathematical accuracy.

Table 1: Descriptive Statistical Parameters of The Internal CO_2 Corrosion of Combined Data

	Temperature (°C)	Flow Velocity (m/s)	CO2 Pressure (bar)	Internal Pressure (bar)	Corrosion Inhibitor Efficiency (%)	Shear Stress (Pa)	pH	Corrosion Rate (mm/yr)
Count	427	427	427	376	243	303	427	427
Mean	47.4	2.63	12.08	56.82	50	3.58	4.34	4.60
Std Dev	17.35	7.74	63.94	15.57	8.18	1.78	0.69	5.24
Min	13	0	0.9	18.12	40	0.385	3.1	0.02
25%	40	1.2	1.5	56	40	3	3.9	2.72
50% (Median)	45.5	1.5	2	63	50	4	4	3.6
75%	52.7	1.8	2.50	67	60	5	4.91	5.15
Max	150	82.3	827	73	60	11.3	6.5	48.25

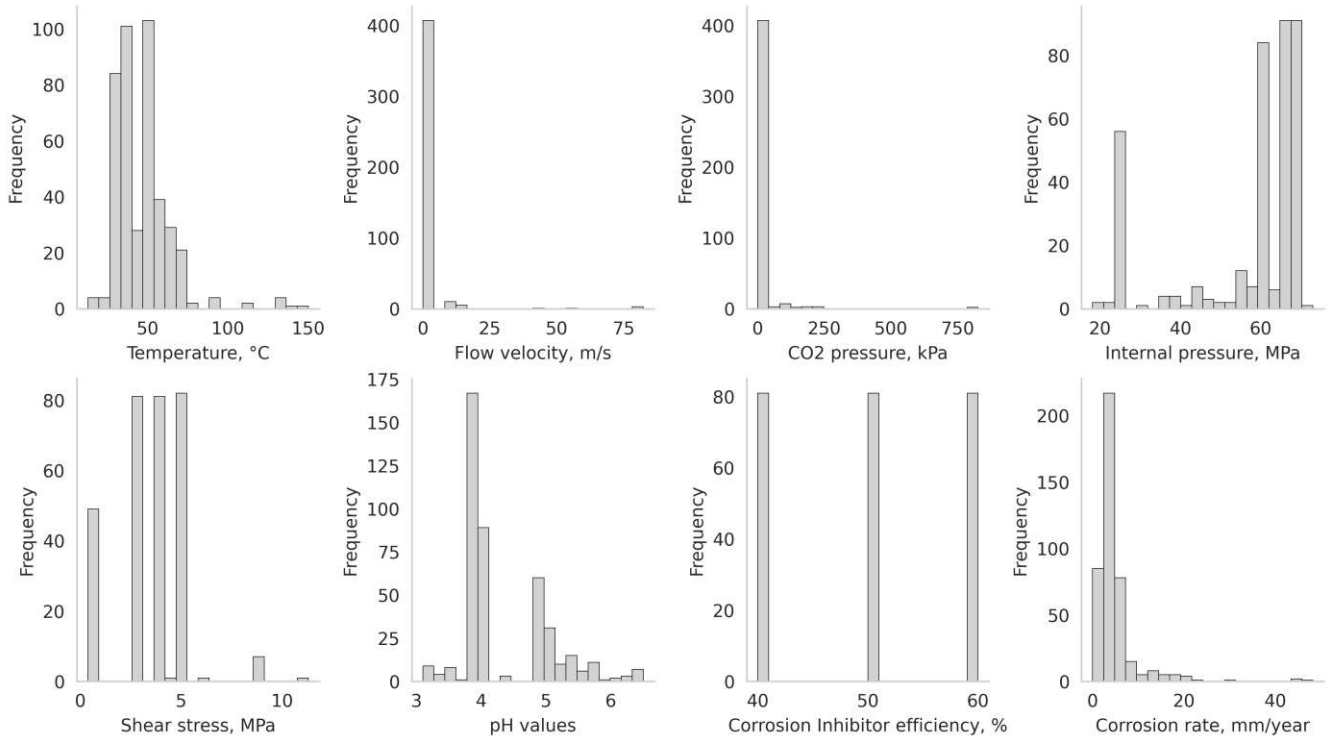


Figure 1: Distribution Histograms of the Inputs and Output Variables of the Internal CO_2 Data.

In addition, the histograms of the data distributions related to each variable of the merged datasets are illustrated in Fig. 1. It is clear from Fig. 1 that the histograms indicate the presence of non-normal distributions at various scales. As an example, the temperature varies in a range of 13-150 ^0C. Similarly, the flow velocity goes from 0 to 82.3 m/s, while the output target, corrosion rate, ranges from 0.02 to 48.3 mm/yr. The high standard deviation relative to the mean indicates heterogeneous corrosion conditions across the merged datasets. Internal pressure had 51 missing values (11.9%) with available data ranging from 18.1 to 73.0 MPa. Corrosion Inhibitor efficiency had 184 missing values and exhibited discrete levels at 40%, 50%, and 60%. Shear stress contained 124 missing values (29.0%), with observed values ranging from 0.39 to 11.3 MPa. Given the limited dataset size typical of corrosion research, complete case deletion would have resulted in substantial information loss, reducing the effective sample size to approximately 150-200 observations. Therefore, an imputation approach was deemed necessary to preserve statistical power and maintain the integrity of the dataset.

4.2 Missing Data and Outlier Removal: Expectation-Maximization (EM) algorithm combined with Gaussian Mixture Models (GMM) was used for missing value imputation. This approach was selected for three reasons. First, it can capture complex multivariate relationships among correlated features. Second, it models the underlying data distribution through mixture components. Third it provides probabilistic imputation that accounts for uncertainty in missing values. Model selection for the Gaussian Mixture Model was performed by systematically evaluating the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) across a range of 1 to 17 mixture components in Fig. 5 (see supporting information). Both criteria exhibited a monotonic decrease from 1 to 14 components. The optimal number of components was 14. The agreement between these independent information criteria strengthens confidence in the model selection. Identifying 14 distinct mixture components reflects substantial heterogeneity within the dataset. This is likely due to diverse corrosion regimes arising from varying experimental conditions such as differential inhibitor concentrations,

pressure ranges, and mechanical stress states. The two-stage imputation framework included an initial simple imputer mean imputation followed by iterative refinement via the Expectation-Maximization algorithm. This allowed the model to capture complex multivariate dependencies among internal Pressure, shear stress, and Corrosion Inhibitor Efficiency. It also preserved correlations critical to corrosion mechanism representation. This approach ensures that the training dataset remains large enough for robust machine learning model development and preserves the integrity of inter-feature relationships, essential for accurate corrosion rate prediction. Figure 2 presents the distributional characteristics of the three imputed variables before and after the EM imputation procedure.

Corrosion Inhibitor Efficiency exhibited a trimodal distribution with peaks at approximately 40%, 50%, and 60%. This distribution was preserved post-imputation with enhanced continuity as the EM populated missing values according to the underlying mixture distribution. This process yielded a dominant peak at 51% efficiency reflecting the most prevalent operational condition. Internal Pressure demonstrated bimodal clustering at 25 bar and 60-70 bar representing distinct operational regimes. Post-imputation kernel density estimates showed smoother transitions between modes while maintaining the bimodal structure, confirming appropriate assignment of missing values based on multivariate correlations. Shear Stress displayed characteristic right-skewed distribution with primary concentration between 0-5 Pa and secondary peak at 3-4 Pa. This pattern was preserved post-imputation, reflecting physically realistic pipeline stress conditions where moderate stress predominates. The preservation of multimodal and skewed structures across all variables validates the 14-component selection. Univariate imputation would have artificially inflated central tendencies and reduced distributional variance, whereas the probabilistic EM framework maintained complex statistical properties essential for accurate corrosion rate modelling. Outlier detection was performed using the Interquartile Range (IQR) method, defining outliers as observations falling outside Q1 - 1.5×IQR or Q3 + 1.5×IQR bounds. Removal of these outliers reduced the dataset from 427 to 394 samples (8.9% reduction), eliminating measurement errors or experimental anomalies while preserving adequate sample size for robust model development and improving generalization performance on unseen data (see supporting document).

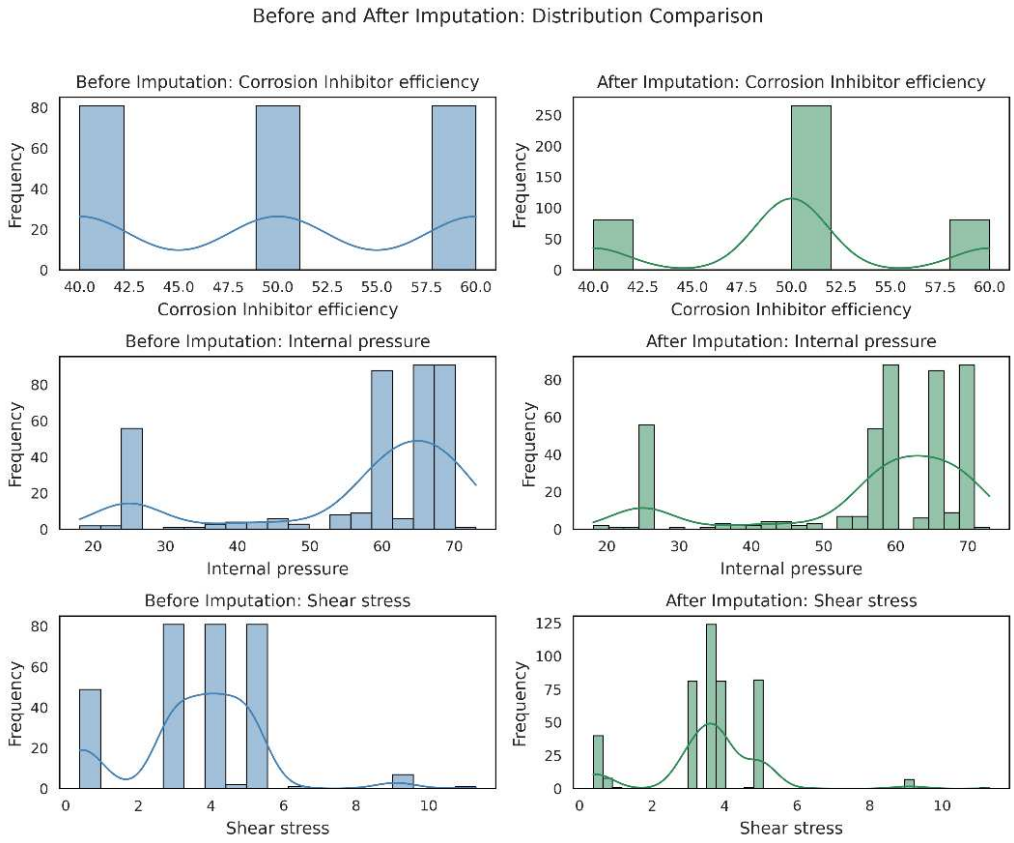


Figure 2: Distribution Comparison of Pipeline Integrity Parameters Before And After Multiple Imputations. *Histograms with overlaid kernel density estimate curves show the frequency distributions of corrosion inhibitor efficiency (%), internal pressure (bar), and shear stress (Pa). Left panels display the original data with missing values (before imputation, shown in blue), while right panels show the complete dataset after multiple imputations (shown in green).*

Although the EM-GMM imputation framework effectively preserved the dataset's multivariate distributional characteristics, several methodological limitations should be acknowledged. First, the dataset comprises simulated, experimental, and field data, which may possess heterogeneous statistical properties. Selecting a 14-component GMM aimed to address this heterogeneity; however, the assumption of Gaussian mixture distributions may not fully capture the underlying data-generating processes across all data sources. Secondly, the probabilistic imputation procedure can generate parameter combinations that, while statistically consistent with the fitted mixture model, are not physically realizable under actual pipeline operating conditions. For example, certain combinations of internal pressure and shear stress may fall outside the physically plausible range for corrosion scenarios. Thirdly, GMM-based imputations are sensitive to initialization and component selection. This sensitivity was mitigated by employing a fixed random state (random_state=42) to ensure reproducibility and by systematically evaluating BIC and AIC criteria across multiple component configurations. Finally, although the imputed variables (internal pressure, shear stress, and corrosion inhibitor efficiency) are continuous, which avoids issues related to treating discrete variables as continuous, the imputation procedure does not explicitly enforce physical constraints or domain-specific bounds. Despite these limitations, the validation of distributional preservation (Figure 2) and the agreement between BIC and AIC in model selection support the robustness of the imputation approach for the objectives of this study.

4.3 Imbalanced Data Handling: Analysis of the target variable distribution revealed pronounced class imbalance, with 58.88% of observations concentrated in the moderate corrosion range (1-5 mm/year), 19.54%

in the high range (5-10 mm/year), 17.51% in the low range (<1 mm/year), and only 4.06% representing severe corrosion cases (>10 mm/year) as seen in Figure 3a-c.

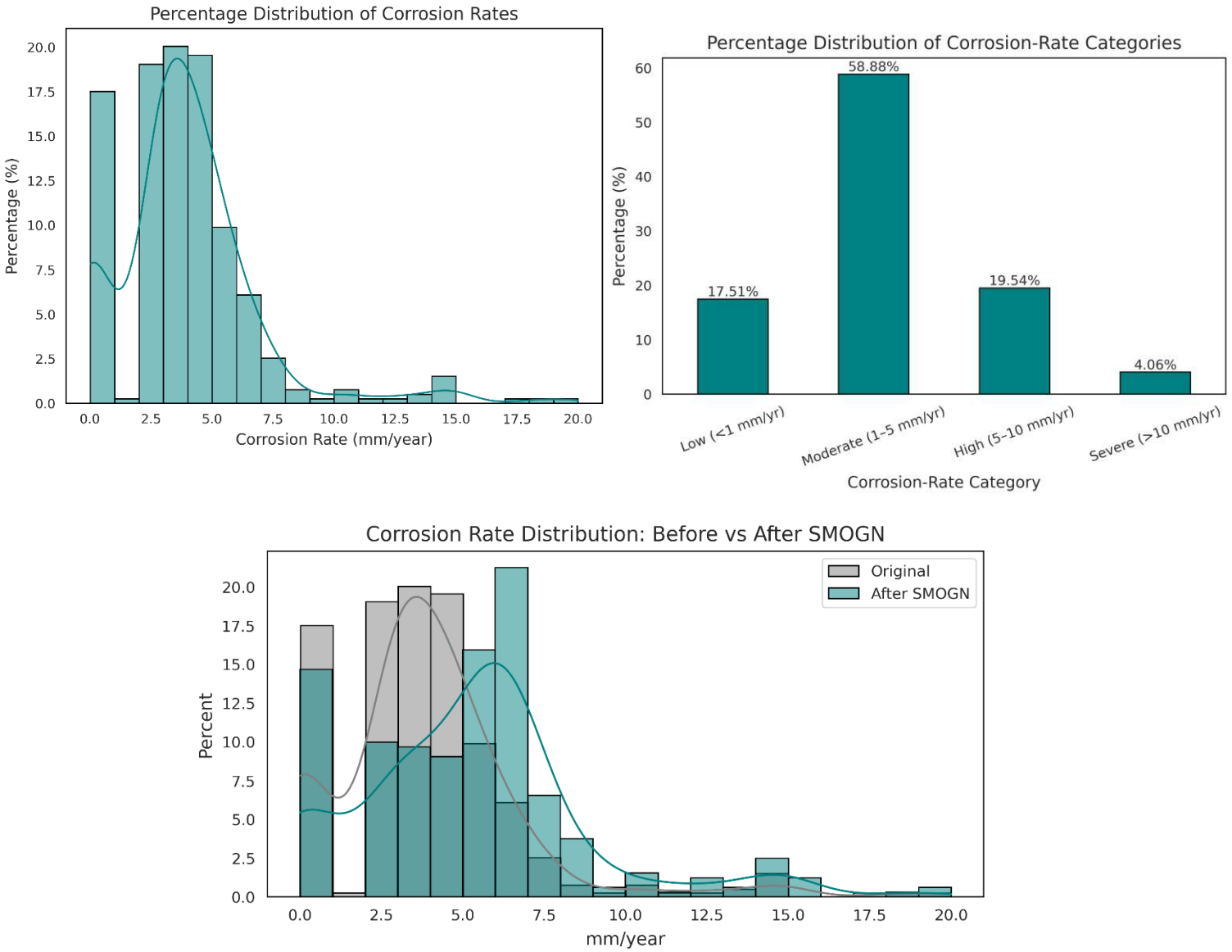


Figure 3: Target Variable Distribution: Corrosion Rate Characteristics *(a) Histogram with kernel density overlay showing right-skewed distribution of continuous corrosion rates (mm/year), with peak frequency between 2.5-5.0 mm/year. (b) Bar chart displaying categorical breakdown of corrosion severity, with moderate rates (1-5 mm/year) comprising the largest proportion at 58.88%, followed by high (19.54%), low (17.51%), and severe (4.06%) categories.* ***(c)*** *Distribution comparison before (gray) and after (teal) SMOGN application, showing enhanced representation of rare high corrosion values while preserving the original distribution pattern*

Application of SMOGN reduced the dataset from 394 to 320 observations. It also fundamentally restructured the class distribution: severe corrosion representation increased from 4.06% to 10.94%, high corrosion from 19.54% to 44.69% while the over-represented moderate class decreased from 58.88% to 28.44%, and the low corrosion class declined from 17.51% to 15.94%. Figure 3c shows the distributional transformation, demonstrating successful rebalancing toward extreme values without introducing artificial multimodality or discontinuities. This preprocessing ensures that subsequent machine learning models are exposed to all corrosion regimes during training. As a result, predictive accuracy for high-consequence scenarios improves while maintaining model generalization. This approach prevents overlooking rare but impactful events.

4.4 Feature selection: Correlation analysis was conducted to examine linear relationships among predictor variables and identify potential multicollinearity issues that could compromise model performance. Figure 4 presents the Pearson correlation matrices after SMOGN resampling. Before resampling, the dataset showed several notable correlations: Internal Pressure demonstrated strong positive correlation with Shear Stress (r = 0.74) and moderate negative correlation with pH (r = -0.51). pH had a moderate negative correlation with both Shear Stress (r = -0.37) and Corrosion Rate (r = -0.49). After SMOGN resampling, the correlation structure changed moderately but the main patterns remained. The Internal Pressure-Shear Stress correlation decreased from 0.74 to 0.66, staying below the typical of |r| > 0.7 multicollinearity threshold [35]. The pH-Corrosion Rate correlation grew stronger from -0.49 to -0.55 and the CO_2 Pressure-Corrosion Rate correlation remained the same (0.47 vs. 0.47). The resampling procedure slightly altered feature relationships, particularly affecting correlations among extreme values. The generation of SMOGN synthetic samples in high-corrosion regions caused subtle distribution shifts.

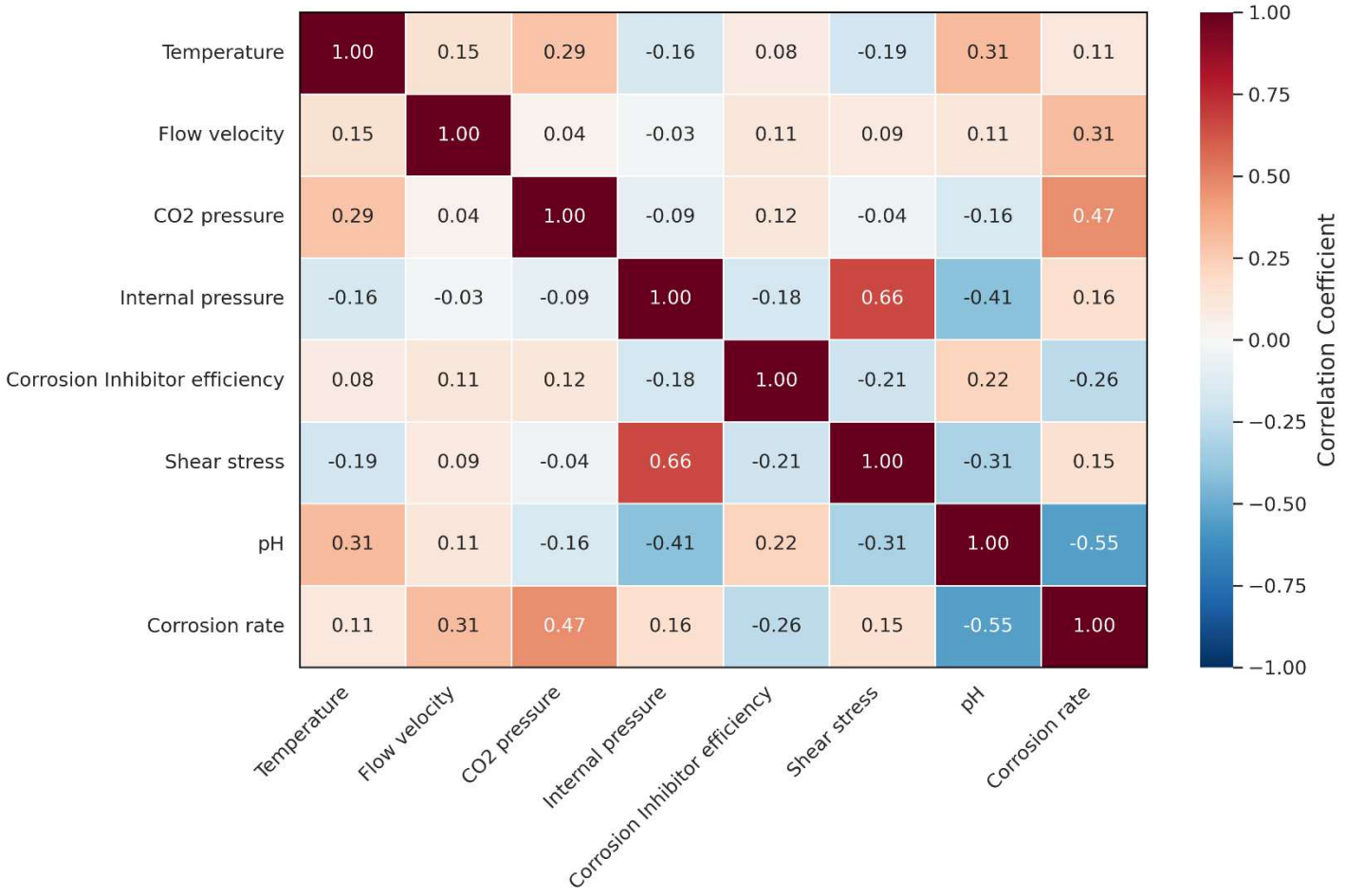


Figure 4: *Pearson Correlation Matrix of the CO_2 Data for All Variables After Data Processing.*

Feature importance was assessed using the Extra Trees Regressor, an ensemble method that builds multiple decision trees and averages their feature importance scores based on impurity reduction. This approach provides a model-agnostic assessment of each predictor's contribution to target variable prediction. It accounts for non-linear relationships that may not be captured by linear correlation analysis alone. Figure 5 presents the feature importance of rankings derived from the Extra Trees model. CO_2 pressure emerged as the most influential predictor (27.61%), followed by pH (26.12%), Flow Velocity (20.15%), and Temperature (11.30%). These four

features collectively accounted for 85.18% of the model's predictive capacity. This indicates their dominant role in corrosion rate determination. In contrast, Corrosion Inhibitor Efficiency (8.08%), Internal Pressure (5.08%), and Shear Stress (1.66%) exhibited substantially lower importance scores. These features contributed minimally to model predictions despite their physical relevance to corrosion processes. Based on these scores, three features were removed: Shear Stress (1.66%), Internal Pressure (5.08%), and Corrosion Inhibitor Efficiency (8.08%). This reduced the feature space from seven to four predictors. This decision was informed by both the low importance scores and the previous correlation analysis. The retention of CO_2 Pressure, pH, Flow Velocity, and Temperature preserves the core physicochemical drivers of CO_2 corrosion.

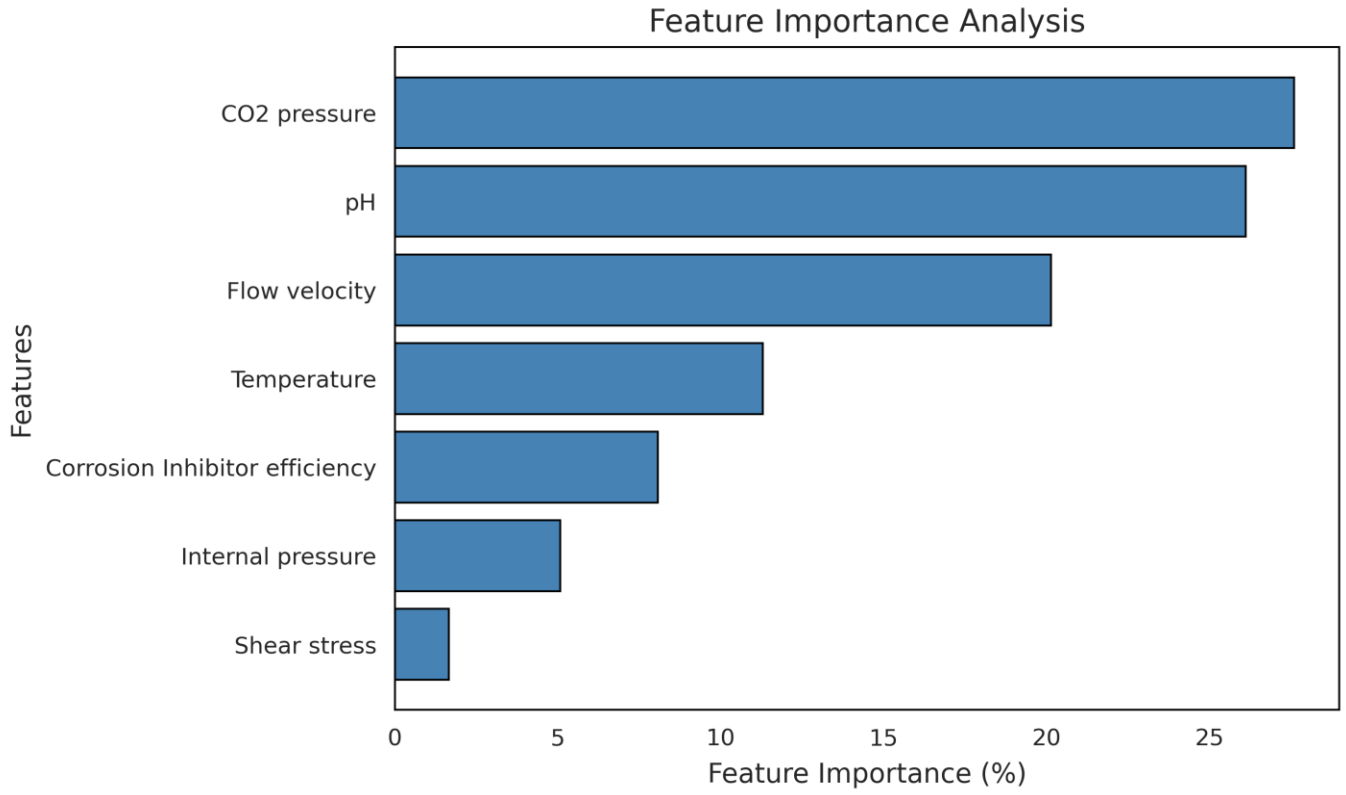


Figure 5: *The Relative Importance Weights Of 8 Features According to Extra Tree Regressor*

While dimensionality reduction can improve model efficiency and reduce overfitting risk, the removal of Corrosion Inhibitor Efficiency is particularly concerning from a practical standpoint. Despite its modest importance score (8.08%), inhibitor efficiency is a directly controllable operational parameter with established protective effects in corrosion mitigation. Its low importance ranking may reflect limited variability in the dataset (concentrated at discrete levels: 40%, 50%, 60%) rather than true lack of predictive value, or potential interactions with other features that tree-based importance metrics may not fully capture. Similarly, the removal of Internal Pressure eliminates a fundamental operating parameter that governs CO_2 solubility and partial pressure effects. To assess the validity of this feature selection approach, comparative modeling was performed using both the full seven-feature set and the reduced four-feature set. This allows quantitative assessment of prediction accuracy trade-offs against model simplicity, providing empirical evidence for whether the dimensionality reduction enhances or compromises model performance.

4.5 Model Comparison: To validate the effectiveness of different ensemble learning approaches the baseline training parameters of each model was used. Figures 6 and 7 illustrate prediction performance comparisons of six ensemble models (AdaBoost, ETR, GBR, RFR, XGBR, and CatBoost) on both training and testing datasets. In these scatter plots, blue circles represent training data points while yellow diamonds indicate testing data points, with the diagonal line (X=Y) representing perfect prediction. The proximity of data points to this diagonal line indicates superior model performance.

The comparative evaluation across all six machine learning models is presented in Figure 8(a-c). The results demonstrate that ETR consistently outperforms all other ensemble models across all evaluation metrics, achieving the lowest error values with RMSE = 2.47 mm/year, MAE = 1.24 mm/year, and MSE = 6.09 mm/year. Following ETR in performance ranking, RFR (Random Forest Regressor) shows competitive results with RMSE = 2.60 mm/year and MAE = 1.32 mm/year. The XGBR (XGBoost) and GBR (Gradient Boosting Regressor) models exhibit comparable performance with RMSE values of 2.69 and 2.76 mm/year, respectively. The CatBoost and AdaBoost models show relatively higher error metrics, with AdaBoost recording the highest RMSE of 2.90 mm/year and MAE of 2.08 mm/year among all tested models.

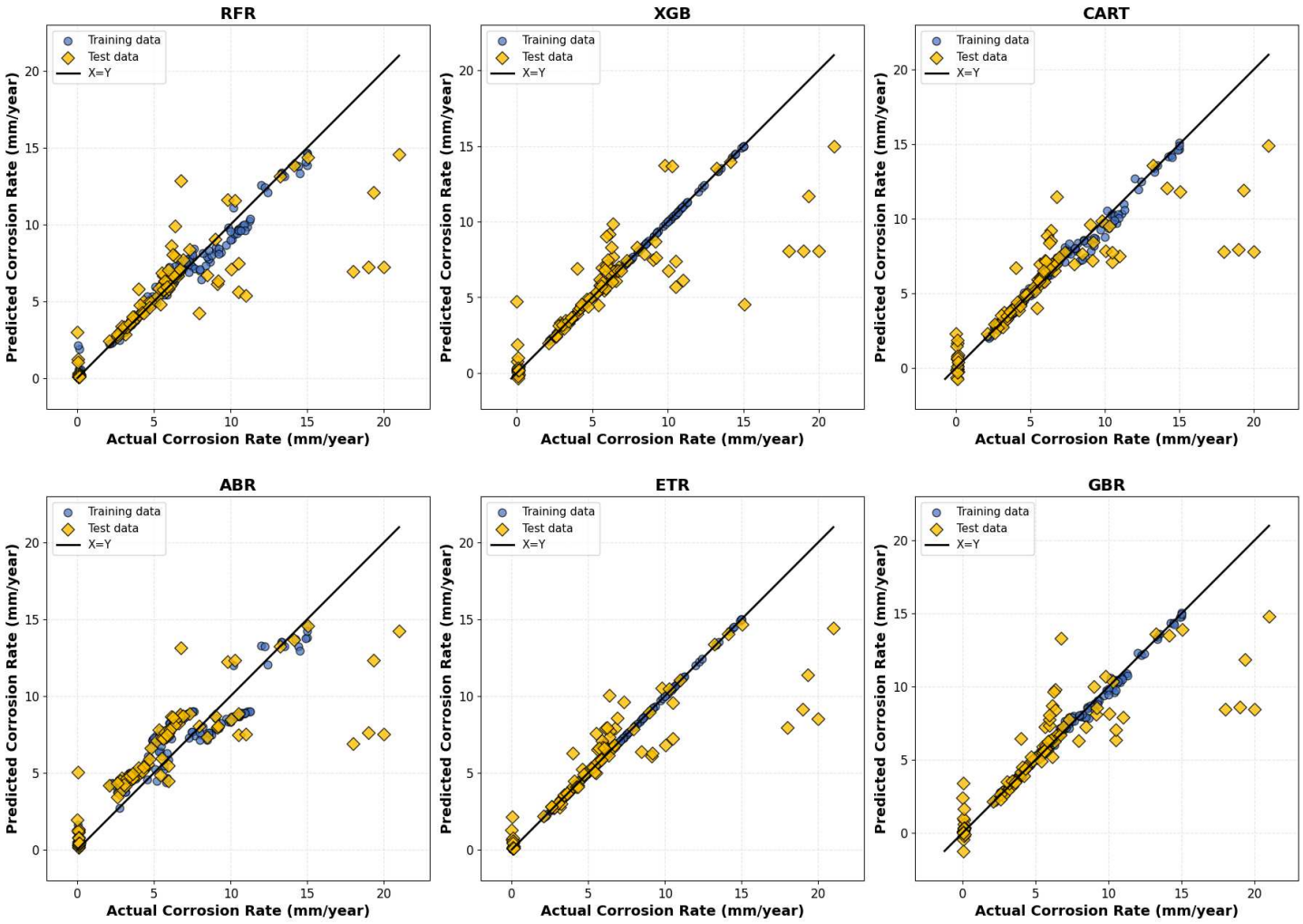


Figure 6: *Model Prediction Results of Different Ensemble Learning Models Without Feature Selection. Blue Dots Indicate Training Data, and Yellow Dots Indicate Testing Data.*

To investigate the influence of feature selection on ensemble learning model performance, a comparative analysis was conducted between models trained with and without feature selection. Figure 6 presents the prediction performance of all ensemble models without feature selection. In these models, the complete original feature set was used as input variables. The scatter plots reveal that prediction performance is much worse for all ensemble models without feature selection compared to those with feature selection (Figure 7). The degradation in model performance without feature selection appears as increased scatter of data points away from the diagonal line and less clustering. This effect is particularly noticeable in the testing dataset (yellow diamonds) and consistent across all six models. It shows that irrelevant or redundant features introduce noise. This noise compromises model accuracy and generalization ability.

Figure 7 shows substantial performance improvements for all models when feature selection is implemented. For example, the ETR model's performance metrics improved significantly after feature selection. Thus, intelligent feature selection is essential for optimizing machine learning model performance in corrosion rate prediction applications. The consistent performance enhancement observed across the models highlights the universal benefit of feature selection. This effect is seen in tree-based methods (RFR, ETR), boosting algorithms (XGBR, GBR, AdaBoost), and individual decision trees (CatBoost). Feature selection reduces model complexity, minimizes overfitting, and improves predictive accuracy for pipeline corrosion rate estimation.

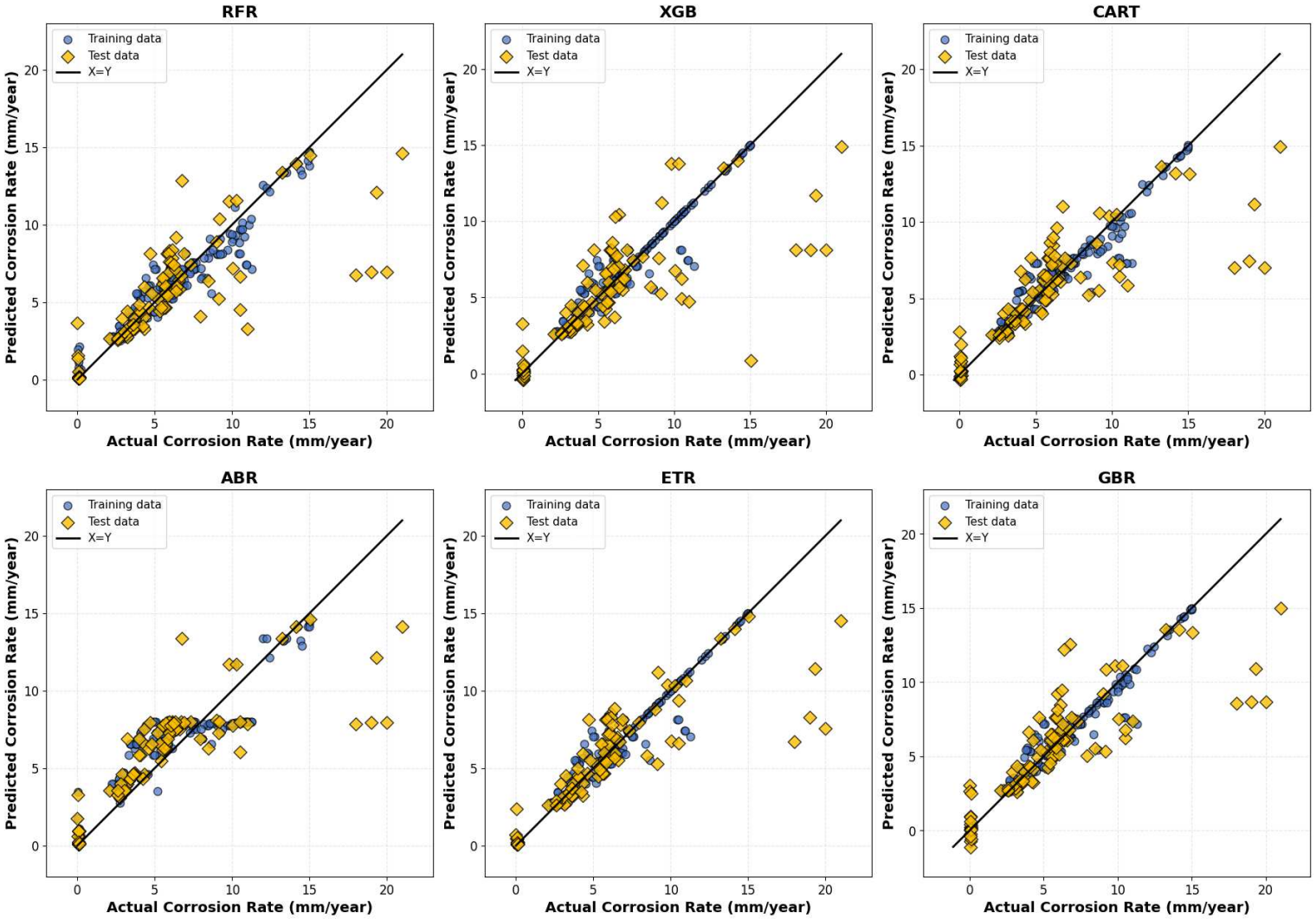


Figure 7: Model Prediction Results of Different Ensemble Learning Models with Feature Selection. Blue Dots Indicate Training Data, and Yellow Dots Indicate Testing Data.

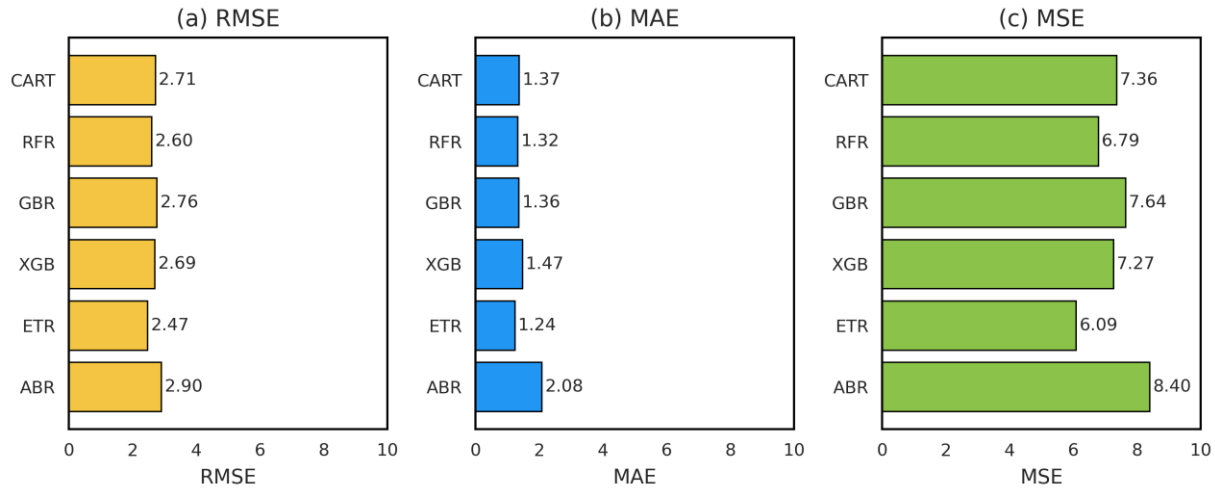


Figure 8: *Comparison of the RMSE% (A), MAE (B) And MSE (C) Values of the Six Ensemble ML Models for Modelling the CO_2 Corrosion Rate*

A comparative analysis was performed to assess the impact of SMOGN resampling on model performance, using all features both before and after SMOGN (Table 2). This approach directly addresses concerns regarding the necessity of preprocessing steps and the risk of introducing unintended bias.

In the absence of SMOGN, models produced high R^2 values (0.77–0.98), with XGB reaching R^2 = 0.98 and GBR achieving R^2 = 0.98. The associated error metrics were low, with RMSE values ranging from 0.38 to 0.63 mm/year and MAE values from 0.14 to 0.33 mm/year. These high-performance metrics suggest overfitting to the imbalanced dataset, in which 58.9% of observations fell within the moderate corrosion range (1–5 mm/year), and only 4.06% represented severe corrosion cases (>10 mm/year). Models trained on such imbalanced data tend to minimize overall error, resulting in poor sensitivity to rare but operationally significant extreme corrosion events.

Table 2: Comparison of Model Performance Metrics Before and After SMOGN Resampling (All features).

Without SMOGN	**R2 score**	**RMSE**	**MAE**	**MSE**
ABR	0.77	1.29	1.03	1.66
ETR	0.95	0.61	0.19	0.37
GBR	0.98	0.42	0.18	0.18
XGB	0.98	0.38	0.14	0.15
RFR	0.97	0.49	0.18	0.24
CART	0.95	0.63	0.33	0.40
With SMOGN	**R2 score**	**RMSE**	**MAE**	**MSE**
ABR	0.63	2.80	1.81	7.85
ETR	0.73	2.42	1.05	5.86
GBR	0.71	2.52	1.22	6.34
XGB	0.64	2.79	1.32	7.78
RFR	0.64	2.79	1.29	7.79
CART	0.70	2.56	1.28	6.53

After SMOGN resampling, R^2 values decreased to more moderate levels (0.63–0.73), with corresponding increases in RMSE (2.42–2.80 mm/year) and MAE (1.05–1.81 mm/year). While these metrics appear less favorable, the reduction reflects improved generalization capability rather than a decline in performance. Before SMOGN, the models were biased toward predicting moderate corrosion rates, which dominated the training distribution and reduced accuracy for extreme values critical to integrity management decisions. SMOGN resampling redistributed class representation: the proportion of severe corrosion cases increased from 4.1% to 10.9%, high corrosion from 19.5% to 44.7%, while the over-represented moderate class decreased from 58.9% to 28.4%, and the low corrosion class declined from 17.5% to 15.9%. This preprocessing step was implemented to address dataset imbalance (approximately 134 field samples, 50 experimental samples, and 243 simulated samples) and to mitigate the risk of overfitting associated with a limited dataset (427 data points). The rebalancing process ensures that subsequent models are exposed to all corrosion regimes during training, thereby enhancing predictive accuracy in high-consequence scenarios and maintaining generalization.

4.6 Hyperparameter tuning: Table 3 presents the comprehensive results of the grid search cross-validation process conducted across the top four (ETR, RFR, GBR, XGBR) ensemble learning models after feature selection to enhance predictive performance for pipeline CO_2 corrosion rate. The selection of optimal hyperparameters was systematically performed by minimizing error metrics while maximizing the coefficient of determination (R^2 score) across both training and testing datasets. **RFR** achieved R^2=0.88/0.65 (train/test) with MSE=1.51/7.59 mm/year, showing moderate overfitting. **GBR** demonstrated the best testing performance with R^2=0.87/0.70 (test/train) and MSE=1.72/6.43 mm/year, indicating superior generalization among boosting algorithms. XGBR exhibited the highest training accuracy (R^2=0.90) but poorest generalization (R^2=0.59, MSE=8.83 mm/year), with significant overfitting (ΔR^2=0.31). ETR showed the most balanced performance (R^2=0.78/0.67, ΔR^2=0.11), demonstrating minimal overfitting and the best generalization capability despite lower absolute accuracy. Overall, GBR provided the optimal balance of accuracy and robustness for corrosion rate prediction.

Table 3:*The Optimal Hyperparameters for Six Models with Grid Search.*

ML models	Grid search	Best parameters	R2 score_ train	R2 score_ test	MSE train	MSE test
RFR	'max_depth': [30, 40, 50, 80, 90, 100], 'max_features': [2, 3, 4, 5, 6, 8, 10], 'min_samples_leaf': [2, 3, 4, 5, 6, 7, 8, 9, 10], 'min_samples_split': [2, 4, 6, 8, 10, 12], 'n_estimators': [100, 200, 300, 400, 500]	max_depth=90, max_features=2, min_samples_leaf=2, min_samples_split=10 n_estimators=400	0.88	0.65	1.51	7.59
GBR	'loss': ['squared_error', 'absolute_error', 'huber', 'quantile'], 'learning_rate': [0.001, 0.1, 0.2, 1, 10], 'n_estimators': [100, 150, 200, 300, 400], 'max_depth': [3, 5, 7, 10, 15, 20]	learning_rate=0.2, loss='absolute_error', max_depth=7, n_estimators=400)	0.87	0.7	1.72	6.43
XGB	'n_estimators': [100, 200, 300, 400, 500, 600], 'learning_rate': [0.01, 0.02, 0.03, 0.05, 0.1, 0.001, 0.1, 0.2, 1, 10], 'booster': ['gbtree', 'gblinear'], 'gamma': [0, 0.5, 1], 'reg_alpha': [0, 0.2, 0.5, 1], 'reg_lambda': [0.5, 1, 2, 3, 4, 5], 'base_score': [0.2, 0.1, 0.4, 0.5, 0.8, 1]	base_score=0.4, booster='gbtree', gamma=1, learning_rate=1, n_estimators=100	0.90	0.59	1.26	8.83
ETR	n_estimators': range(50,500,50), 'max_features': range(50,250,50), 'min_samples_leaf': range(5,50,5), 'min_samples_split': range(5,40,5)}	max_features=200, min_samples_leaf=5, min_samples_split=10 n_estimators=50	0.78	0.67	2.82	7.19

These findings highlight that choosing optimal hyperparameter is critical for improving model generalization. Hyper parameter tuning showed that model complexity parameters particularly tree depth, number of estimators, and regularization terms strongly affect the balance between model accuracy and overfitting. All other hyperparameters were kept at defaults to ensure reproducibility and consistency across comparisons.

A key distinction between this study and previous research lies in the dataset composition strategy. Most prior studies utilized either pure experimental data [26] purely simulated data [28], or field data [29] [35] This study integrates experimental, simulated, and field datasets through systematic preprocessing. This approach addresses the "scarcity of publicly accessible corrosion data" and the challenge of "inconsistent experimental protocols and industrial confidentiality restrictions". These issues prevent creation of "comprehensive, unified datasets covering diverse operational regimes" [19]. While this integration strategy results in more conservative performance metrics compared to studies using homogeneous datasets, it provides the critical advantage of prediction reliability across the full spectrum of operational conditions encountered in CO_2 transport and processing systems. The moderate testing R^2 values of our model should not be seen as indicative of inferior performance but rather as realistic predictive capability. They reflect a model trained on truly diverse data that represents industrial variability. In contrast, extraordinarily high R^2 values (>0.95) reported in some studies [35], may reflect overfitting to narrow experimental conditions rather than genuine generalization. As Mohammad Zubir et al. [31] demonstrated through synthetic data augmentation, achieving high accuracy requires careful data processing. This study applies the principle of using SMOGN-based minority class enrichment with Gaussian augmentation. Consequently, the apparent reduction in performance metrics following SMOGN application should be interpreted as a methodological strength, indicating reduced overfitting and improved sensitivity to operationally significant extreme corrosion events

4.7 Model Validation: The optimized Gradient Boosting Regression model was validated on an independent stratified test set (20% of total samples) to see how well it generalizes to different levels of corrosion rates. The GBR model achieved $R^2 = 0.83$ on the validation set, showing strong predictive ability for CO_2-induced corrosion rates between 0 and 22 mm/year (Figure 9). Predictions exhibit strong agreement with actual values across low-to-moderate ranges (0-8 mm/year), with increased scatter observed for extreme corrosion rates (>10 mm/year). This is due to having a few severe corrosion rates in training (4.1% of all samples) and natural variation in corrosion mechanisms under extreme conditions. Importantly, the model maintains directional accuracy without systematic bias, avoiding under-prediction of high-consequence failure scenarios.

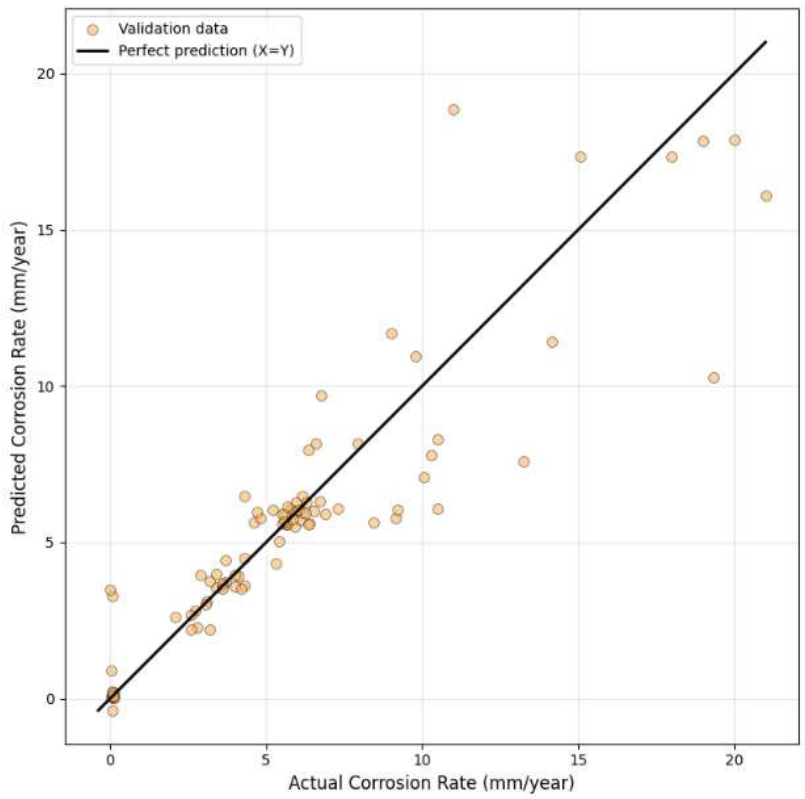


Figure 9: *Cross Plots Illustrating Predicted CO_2 Corrosion Rate vs Actual CO_2 Corrosion on Test Dataset.*

The absolute percentage error (APE) analysis grouped by corrosion severity reveals distinct performance characteristics (Figure 10). For Moderate (1-5 mm/year), High (5-10 mm/year), and Severe (>10 mm/year) categories, median APE approaches 0%. The interquartile ranges below 100%, indicating excellent predictive accuracy with minimal variability. The compact distributions and absence of significant outliers confirm reliable prediction across operationally critical corrosion regimes. The Low category (<1 mm/year) exhibits elevated median APE (~150%) with extreme outliers (2800-4500%), attributable to percentage-based metric amplification near zero actual values. For example, predicting 0.3 mm/year versus actual 0.1 mm/year yields 200% APE despite absolute error of only 0.2 mm/year negligible for integrity assessment over typical inspection intervals (5-10 years). Absolute prediction errors in this category remain operationally insignificant, as corrosion rates below 1 mm/year pose minimal integrity risk regardless of proportional accuracy. The category-specific analysis confirms reliable model performance (median APE <50%) precisely where predictive accuracy governs engineering decisions—in moderate-to-severe corrosion regimes driving inspection scheduling, maintenance planning, and replacement strategies.

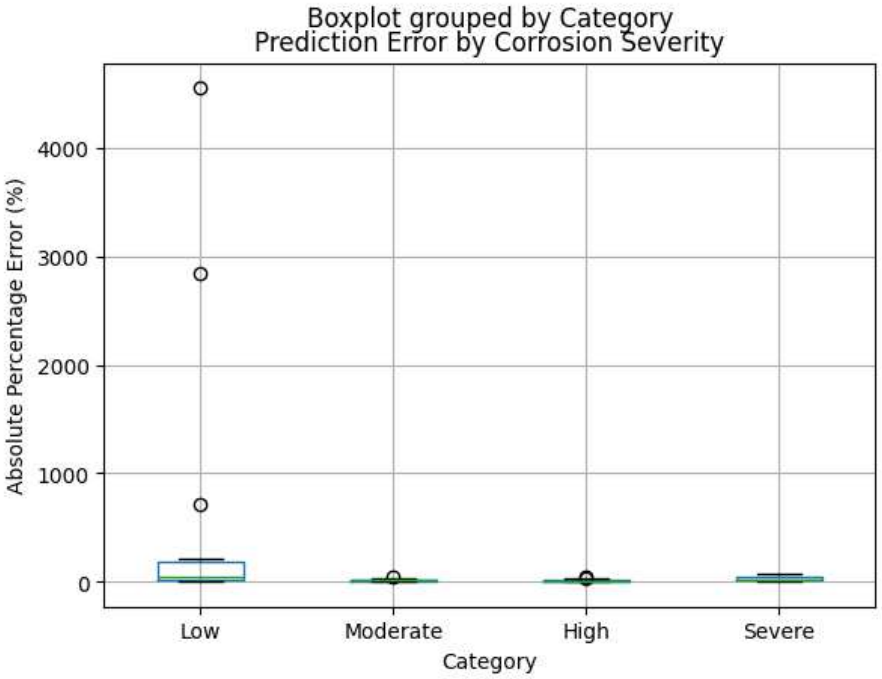


Figure 10: *The Boxplot for The Evaluation Metrics Of 10-Fold Cross Validation with GBR Model on Testing Dataset*

4.8 Discussion on adaptability and limitations: The ensemble learning framework demonstrates broad applicability across diverse oil and gas pipeline systems, particularly in complex operational environments with variable service conditions. The comparative analysis of six algorithms provides flexibility in model selection prioritizing either maximum accuracy (GBR: R^2_test = 0.70) or optimal generalization (ETR: $\Delta R^2 = 0.11$) based on specific requirements. Integrating experimental, simulated, and field datasets through GMM-EM imputation enables reliable prediction across the full operational spectrum, from controlled laboratory to variable field conditions. However, several fundamental limitations warrant consideration for practical implementation. They are:

a. Model performance and generalization: The moderate testing R^2 values (0.59–0.70) result from prioritizing generalizability across heterogeneous industrial conditions over maximizing predictive accuracy. This approach intentionally accepts lower peak accuracy to ensure model robustness in varied real-world scenarios. In contrast, some literature report $R^2 > 0.95$[23], [35], [60], [61], but these often arise from overfitting or from using homogeneous datasets that lack real-world variability. Although the integrated dataset's heterogeneity and limited size constrain absolute predictive performance, this approach yields more realistic estimates of the model's ability to generalize to diverse operational conditions.

b. Data quality and quantity dependency: Model accuracy is fundamentally dependent on the quality, quantity, and representativeness of training data. While GMM-EM imputation preserves multivariate dependencies essential for corrosion mechanism representation, it cannot correct systematic measurement errors or biased sampling inherent in source datasets. The probabilistic imputation procedure may also generate parameter combinations that are statistically consistent with the fitted mixture model but not physically realizable under actual pipeline operating conditions. These data-driven limitations could affect overall model reliability and highlight the need for expanded curated training databases.

c. Algorithmic constraints: Tree-based algorithms excel at piecewise-constant approximations and are well-suited for problems with abrupt changes or discrete boundaries. However, they struggle to model smooth,

continuous transitions in electrochemical corrosion with complex parameter interactions, where change is gradual. XGBR's severe overfitting ($\Delta R^2 = 0.31$) shows that even advanced boosting can fail to generalize complex corrosion behavior with limited, heterogeneous data. While alternative approaches such as deep neural networks or physics-informed machine learning might better capture smooth nonlinearities, these methods typically reduce interpretability, a critical consideration for engineering applications where understanding model behavior supports risk-based decision-making.

d. Limited feature set: The feature selection reduced predictors to CO_2 partial pressure, pH, flow velocity, and temperature based on importance ranking and multicollinearity analysis. While this dimensionality reduction improved prediction accuracy for the current dataset, excluding corrosion inhibitor efficiency, internal pressure, and shear stress may limit its practical applicability in operational contexts where these parameters are directly controlled. The low importance score for corrosion inhibitor efficiency (8.08%) may reflect limited variability in the dataset (concentrated at discrete levels: 40%, 50%, 60%) rather than a true lack of predictive value. For deployment scenarios where inhibitor dosing is a key control parameter, the full feature set may be preferable despite modest accuracy penalties.

e. Hyperparameter sensitivity: Optimal hyperparameters identified through grid search with 3-fold cross-validation reflect characteristics of the current training dataset and may not remain optimal under different operational conditions or with new data sources. While cross-validation mitigates overfitting to specific hyperparameter configurations, practical deployment in dynamic operational environments may benefit from adaptive tuning methods (e.g., Bayesian optimization, automated machine learning pipelines) or periodic re-calibration as new operational data becomes available.

Despite these limitations, the framework is highly valuable for pipeline integrity management. Accurate corrosion prediction (RMSE = 2.47-2.90 mm/year for top models) helps move from reactive to proactive risk-based inspection programs. This optimizes resource allocation to high-risk locations.

5. Conclusion

This study presents a systematic ensemble learning framework for predicting CO_2 corrosion rates in oil and gas pipelines. The approach integrates experimental, simulated, and field datasets. Six ensemble machine learning algorithms (RFR, GBR, XGBR, ETR, AdaBoost, CatBoost) were evaluated following rigorous preprocessing. Two-step imputation used Simple Imputer and Gaussian Mixture Model-based Expectation-Maximization (GMM-EM). Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise (SMOGN) addressed class imbalance. Grid search optimization identified Gradient Boosting Regressor (learning_rate=0.2, loss='absolute_error', max_depth=7, n_estimators=400) as optimal, achieving R^2_test = 0.70, MSE = 6.43mm²/year², and minimal generalization gap ($\Delta R^2 = 0.17$). Validation of model yielded $R^2 = 0.83$ across 0-22 mm/year with median absolute percentage errors below 50% for critical severity categories (>1 mm/year).

GMM-EM imputation preserved multivariate dependencies in incomplete datasets, while SMOGN corrected severe distribution imbalance (58.9% moderate, 4.1% severe cases), enhancing sensitivity to rare extreme events. Feature selection confirmed that dimensionality reduction substantially improves prediction accuracy. The framework addresses limitations of traditional mechanistic and empirical models. It learns nonlinear patterns from multi-source data without domain-specific re-parameterization. The validated model provides adequate precision for integrity management at typical inspection intervals, supporting risk-based inspection strategies and loss-of-containment risk mitigation. Based on the limitations highlighted in Section 4.8, recommendations for future work, which would further benefit this research include the expansion of training databases by collecting more field data from diverse operational regimes, especially underrepresented scenarios

such as multiphase flow and high temperature/high-pressure conditions. Secondly, integration of time-series monitoring data to enable temporal prediction using long short-term memory (LSTM) or recurrent neural network (RNN) architectures. Thus, exploring alternative model types like deep learning could improve capturing complex corrosion dynamics while addressing interpretability tradeoffs Lastly, the development of physics-informed features that use knowledge from corrosion science is recommended for future work.

.

Data availability

The data that supports the findings of this study are openly available in Dataverse at https://doi: 10.6084/m9.figshare.31386829 as well as Figshare at https://doi.org/10.6084/m9.figshare.31386829. Codes used in EM imputation and class imbalance are included in the supporting information. The full code script of the work is available only on request.

Funding Declaration

This work is supported by the Department of Energy EFRC Center of Electrochemical Dynamics and Reactions on Surfaces (CERDAS) DE-SC0023415

References

[1] M. Safarzadeh Khosrowshahi *et al.*, "Recent progress on advanced solid adsorbents for CO2 capture: From mechanism to machine learning," *Materials Today Sustainability*, vol. 27, p. 100900, Sep. 2024, doi: 10.1016/j.mtsust.2024.100900.

[2] H. Zhang *et al.*, "Catalyzing net-zero carbon strategies: Enhancing CO2 flux Prediction from underground coal fires using optimized machine learning models," *J. Clean. Prod.*, vol. 441, p. 141043, Feb. 2024, doi: 10.1016/j.jclepro.2024.141043.

[3] T. He, I. A. Karimi, and Y. Ju, "Review on the design and optimization of natural gas liquefaction processes for onshore and offshore applications," Apr. 01, 2018, *Institution of Chemical Engineers*. doi: 10.1016/j.cherd.2018.01.002.

[4] Z. Chen, Y. Jin, X. Wang, H. Chen, B. Zhu, and W. Li, "Advancements in prediction models for corrosion in oil and gas pipelines," 2025, *Walter de Gruyter GmbH*. doi: 10.1515/corrrev-2024-0119.

[5] A. Kahyarian, M. Achour, and S. Nesic, "Mathematical modeling of uniform CO2 corrosion," in *Trends in Oil and Gas Corrosion Research and Technologies: Production and Transmission*, Elsevier Inc., 2017, pp. 805–849. doi: 10.1016/B978-0-08-101105-8.00034-6.

[6] R. Elgaddafi *et al.*, "Modeling and experimental study of CO2 corrosion on carbon steel at elevated pressure and temperature," *J. Nat. Gas Sci. Eng.*, vol. 27, pp. 1620–1629, Nov. 2015, doi: 10.1016/j.jngse.2015.10.034.

[7] R.C. Woollam and S.E. Hernandez, "Assessment and comparison of CO2 corrosion prediction," May 2006, doi: https://doi.org/10.2118/100673-MS.

[8] J. Nesic Srdjan and Postlethwaite, “Modelling of CO2 Corrosion Mechanisms,” in *Modelling Aqueous Corrosion: From Individual Pits to System Management*, P. R. Trethewey Kenneth R. and Roberge, Ed., Dordrecht: Springer Netherlands, 1994, pp. 317–335. doi: 10.1007/978-94-011-1176-8_15.

[9] J. Huang and D. N. Sormaz, “09572 - AN OPEN SOURCE MECHANISTIC MODEL FOR CO2 / H2S CORROSION OF CARBON STEEL,” 2009. [Online]. Available: https://www.researchgate.net/publication/235645650

[10] M. Nordsveen, S. Nes ˇic´,ˇic´, R. Nyborg, and A. Stangeland, “A Mechanistic Model for Carbon Dioxide Corrosion of Mild Steel in the Presence of Protective Iron Carbonate Films-Part 1: Theory and Verification.”

[11] Y. P. Asmara and M. C. Ismail, “Efficient design of response surface experiment for corrosion prediction in CO 2 environments,” *Corrosion Engineering Science and Technology*, vol. 47, no. 1, pp. 10–18, Feb. 2012, doi: 10.1179/1743278211Y.0000000013.

[12] S. H. Olsen, “CO2 Corrosion Prediction by Use of the NORSOK M-506 Model - Guidelines and Limitations,” *Corrosion*, 2003, [Online]. Available: https://api.semanticscholar.org/CorpusID:136381254

[13] Rolf. Nyborg, *Guidelines for prediction of CO_2 corrosion in oil and gas production systems*. Institute for Energy Technology, 2009.

[14] A. Dugstad, L. Lunde, and K. Videm, *Parametric study of CO{sub 2} corrosion of carbon steel*. United Kingdom: NACE International, Houston, TX (United States), 1994. [Online]. Available: https://www.osti.gov/biblio/70026

[15] C. De Waard, U. Lotz, and D. E. Milliams, “Predictive Model for C02 Corrosion Engineering in Wet Natural Gas Pipelines*.” doi: https://doi.org/10.5006/1.3585212.

[16] Y. Leong, F. Alia, and T. Kurniawan, “High temperature oxidation behavior of T91 steel in dry and humid condition,” *Indonesian Journal of Science and Technology*, vol. 1, no. 2, pp. 232–237, 2016, doi: 10.17509/ijost.v1i2.3806.

[17] F. F. Alia, T. Kurniawan, Y. P. Asmara, M. H. B. Ani, and A. B. D. Nandiyanto, “High temperature oxidation in boiler environment of chromized steel,” in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Nov. 2017. doi: 10.1088/1757-899X/257/1/012086.

[18] Y. P. Asmara, A. Juliawati, A. Sulaiman, and Jamiluddin, “Mechanistic model of stress corrosion cracking (scc) of carbon steel in acidic solution with the presence of H2s,” in *IOP Conference Series: Materials Science and Engineering*, 2013. doi: 10.1088/1757-899X/50/1/012072.

[19] H. Lu, X. Ma, K. Huang, L. Fu, and M. Azimi, “Carbon dioxide transport via pipelines: A systematic review,” Sep. 01, 2020, *Elsevier Ltd*. doi: 10.1016/j.jclepro.2020.121994.

[20] A. Kahyarian, M. Singer, and S. Nesic, “Modeling of uniform CO2 corrosion of mild steel in gas transportation systems: A review,” *J. Nat. Gas Sci. Eng.*, vol. 29, pp. 530–549, Feb. 2016, doi: 10.1016/j.jngse.2015.12.052.

[21] A. Rachman, T. Zhang, and R. M. C. Ratnayake, “Applications of machine learning in pipeline integrity management: A state-of-the-art review,” *International Journal of Pressure Vessels and Piping*, vol. 193, p. 104471, Oct. 2021, doi: 10.1016/j.ijpvp.2021.104471.

[22] M. Obaseki, P. B. Alfred, P. T. Elijah, and S. O. Okuma, “Corrosion Rate Prediction in Oil and Gas Pipelines Based on Multiphase Flow Modelling,” *International Journal of Engineering Research in Africa*, vol. 67, pp. 27–48, 2023, doi: 10.4028/p-brqaI1.

[23] Q. Wang *et al.*, “Evolution of corrosion prediction models for oil and gas pipelines: From empirical-driven to data-driven,” Apr. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.engfailanal.2023.107097.

[24] M. R. Youcefi, W. Wei, F. M. Alqahtani, H. Djema, M. Nait Amar, and M. Ghasemi, “Robust Ensemble Learning Models for Predicting Hydrogen Sulfide Solubility in Brine,” *Energy and Fuels*, vol. 38, no. 21, pp. 21174–21188, Nov. 2024, doi: 10.1021/acs.energyfuels.4c04354.

[25] S. Pourrahimi, S. Hakimian, A. H. Bouzid, and L. A. Hof, “On the Use of Machine Learning Algorithms to Predict the Corrosion Behavior of Stainless Steels in Lactic Acid,” *Metals (Basel).*, vol. 13, no. 8, Aug. 2023, doi: 10.3390/met13081459.

[26] A. Chamkalani, M. A. Nareh’ei, R. Chamkalani, M. H. Zargari, M. R. Dehestani-Ardakani, and M. Farzam, “Soft computing method for prediction of co2 corrosion in flow lines based on neural network approach,” *Chem. Eng. Commun.*, vol. 200, no. 6, pp. 731–747, Jun. 2013, doi: 10.1080/00986445.2012.717311.

[27] M. Aghaaminiha *et al.*, “Machine learning modeling of time-dependent corrosion rates of carbon steel in presence of corrosion inhibitors,” *Corros. Sci.*, vol. 193, Dec. 2021, doi: 10.1016/j.corsci.2021.109904.

[28] J. Fang, X. Cheng, H. Gai, S. Lin, and H. Lou, “Development of machine learning algorithms for predicting internal corrosion of crude oil and natural gas pipelines,” *Comput. Chem. Eng.*, vol. 177, Sep. 2023, doi: 10.1016/j.compchemeng.2023.108358.

[29] A. A. Soomro, A. A. Mokhtar, J. C. Kurnia, N. Lashari, H. Lu, and C. Sambo, “Integrity assessment of corroded oil and gas pipelines using machine learning: A systematic review,” Jan. 01, 2022, *Elsevier Ltd*. doi: 10.1016/j.engfailanal.2021.105810.

[30] M. H. Abbas, R. Norman, and A. Charles, “Neural network modelling of high pressure CO2 corrosion in pipeline steels,” *Process Safety and Environmental Protection*, vol. 119, pp. 36–45, Oct. 2018, doi: 10.1016/j.psep.2018.07.006.

[31] W. M. A. Mohammad Zubir, I. Abdul Aziz, and J. Jaafar, “Evaluation of machine learning algorithms in predicting CO2 internal corrosion in oil and gas pipelines,” in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2019, pp. 236–254. doi: 10.1007/978-3-030-00211-4_22.

[32] J. Cui, K. Fang, X. Sun, Y. Gao, and Y. Shen, “ Deep learning-based modeling of CO 2 corrosion rate prediction in oil and gas pipelines ,” *Applied Mathematics and Nonlinear Sciences*, vol. 10, no. 1, Jan. 2025, doi: 10.2478/amns-2025-0415.

[33] Z. Dong *et al.*, “Development of a Predictive Model for Carbon Dioxide Corrosion Rate and Severity Based on Machine Learning Algorithms,” *Materials*, vol. 17, no. 16, p. 4046, Aug. 2024, doi: 10.3390/ma17164046.

[34] T. Yan, L. C. Xu, Z. X. Zeng, and W. G. Pan, “Mechanism and anti-corrosion measures of carbon dioxide corrosion in CCUS: A review,” Jan. 19, 2024, *Elsevier Inc.* doi: 10.1016/j.isci.2023.108594.

[35] M. E. A. Ben Seghier, D. Höche, and M. Zheludkevich, “Prediction of the internal corrosion rate for oil and gas pipeline: Implementation of ensemble learning techniques,” *J. Nat. Gas Sci. Eng.*, vol. 99, Mar. 2022, doi: 10.1016/j.jngse.2022.104425.

[36] S. Peng, Z. Zhang, E. Liu, W. Liu, and W. Qiao, “A new hybrid algorithm model for prediction of internal corrosion rate of multiphase pipeline,” *J. Nat. Gas Sci. Eng.*, vol. 85, p. 103716, Jan. 2021, doi: 10.1016/j.jngse.2020.103716.

[37] A. Mohammed and R. Kora, “A comprehensive review on ensemble deep learning: Opportunities and challenges,” Feb. 01, 2023, *King Saud bin Abdulaziz University*. doi: 10.1016/j.jksuci.2023.01.014.

[38] F. Alserhani and A. Aljared, “Evaluating Ensemble Learning Mechanisms for Predicting Advanced Cyber Attacks,” *Applied Sciences (Switzerland)*, vol. 13, no. 24, Dec. 2023, doi: 10.3390/app132413310.

[39] J. L. Garrido-Labrador, A. Serrano-Mamolar, J. Maudes-Raedo, J. J. Rodríguez, and C. García-Osorio, “Ensemble methods and semi-supervised learning for information fusion: A review and future research directions,” *Information Fusion*, vol. 107, Jul. 2024, doi: 10.1016/j.inffus.2024.102310.

[40] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[41] A. R. Panhalkar and D. D. Doye, “A novel approach to build accurate and diverse decision tree forest,” *Evol. Intell.*, vol. 15, no. 1, pp. 439–453, Mar. 2022, doi: 10.1007/s12065-020-00519-0.

[42] R. E. Schapire, “The Boosting Approach to Machine Learning: An Overview,” in *Nonlinear Estimation and Classification*, M. H. and H. C. C. and M. B. and Y. Bin Denison David D. and Hansen, Ed., New York, NY: Springer New York, 2003, pp. 149–171. doi: 10.1007/978-0-387-21579-2_9.

[43] S. Mahmood, H. Sun, A. Ali Alhussan, A. Iqbal, and E. S. M. El-kenawy, “Active learning-based machine learning approach for enhancing environmental sustainability in green building energy consumption,” *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-70729-4.

[44] J. H. Friedman, “999 REITZ LECTURE GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE 1,” 2001.

[45] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front. Neurorobot.*, vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.

[46] P. Florek and A. Zagdański, “Benchmarking state-of-the-art gradient boosting algorithms for classification,” May 2023, [Online]. Available: http://arxiv.org/abs/2305.17094

[47] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017. [Online]. Available: https://github.com/Microsoft/LightGBM.

[48] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Jun. 2016, doi: 10.1145/2939672.2939785.

[49] Z. Arif Ali, Z. H. Abduljabbar, H. A. Tahir, A. Bibo Sallow, and S. M. Almufti, “eXtreme Gradient Boosting Algorithm with Machine Learning: a Review,” *Academic Journal of Nawroz University*, vol. 12, no. 2, pp. 320–334, May 2023, doi: 10.25007/ajnu.v12n2a1612.

[50] S. Nadarajah, J. C. Mba, N. M. V. Ravonimanantsoa, P. Rakotomarolahy, and H. T. J. E. Ratolojanahary, "Empirical Calibration of XGBoost Model Hyperparameters Using the Bayesian Optimisation Method: The Case of Bitcoin Volatility," *Journal of Risk and Financial Management*, vol. 18, no. 9, Sep. 2025, doi: 10.3390/jrfm18090487.

[51] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *32nd Conference on Neural Information processing Systems (NeurIPS 2018)*, Montreal, 2018. [Online]. Available: https://github.com/catboost/catboost

[52] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J. Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00369-8.

[53] U. Singh, M. Rizwan, M. Alaraj, and I. Alsaidan, "A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments," *Energies (Basel).*, vol. 14, no. 16, Aug. 2021, doi: 10.3390/en14165196.

[54] S. K. Ng, T. Krishnan, and G. J. McLachlan, "The EM algorithm," in *Handbook of Computational Statistics: Concepts and Methods: Second Edition*, Springer Berlin Heidelberg, 2012, pp. 139–172. doi: 10.1007/978-3-642-21551-3__6.

[55] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm." [Online]. Available: https://academic.oup.com/jrsssb/article/39/1/1/7027539

[56] X. Li, L. Zhang, F. Khan, and Z. Han, "A data-driven corrosion prediction model to support digitization of subsea operations," *Process Safety and Environmental Protection*, vol. 153, pp. 413–421, Sep. 2021, doi: 10.1016/j.psep.2021.07.031.

[57] S. Khakzad, M. Yang, A. Lohi, and N. Khakzad, "Probabilistic failure assessment of oil pipelines due to internal corrosion," *Process Safety Progress*, vol. 41, no. 4, pp. 793–803, Dec. 2022, doi: 10.1002/prs.12364.

[58] M. Dai, "In situ mathematically simulation for CO2 internal corrosion in wet natural gas gathering pipelines system by HYSYS," *Eng. Fail. Anal.*, vol. 122, Apr. 2021, doi: 10.1016/j.engfailanal.2021.105265.

[59] P. Branco, R. P. Ribeiro, L. Torgo, B. Krawczyk, and N. Moniz, "SMOGN: a Pre-processing Approach for Imbalanced Regression," 2017.

[60] J. Fang, X. Cheng, H. Gai, S. Lin, and H. Lou, "Development of machine learning algorithms for predicting internal corrosion of crude oil and natural gas pipelines," *Comput. Chem. Eng.*, vol. 177, Sep. 2023, doi: 10.1016/j.compchemeng.2023.108358.

[61] J. Hu, "Prediction of the internal corrosion rate for oil and gas pipelines and influence factor analysis with interpretable ensemble learning," *International Journal of Pressure Vessels and Piping*, vol. 212, p. 105329, Dec. 2024, doi: 10.1016/j.ijpvp.2024.105329.

Authors contributions statement

E.J. developed the research methodology and carried out the data analysis. E.T. provided the corrosion dataset and literature support. E.R. provided the contextual background and literature support. R.K. conceived and designed the study. All authors reviewed and approved the final manuscript.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformationUpdatedmanuscript1.pdf