Predicting Atmospheric Corrosion Rate in Marine Environment

REPORT SUBMITTED FOR

MSE497 – UNDERGRADUATE PROJECT-II

By

Manini – 220618
Nikhil Yadav – 220711
Rahul Bokan – 220857
Saksham Tripathi - 22108040 (Intern)
Shrasti Sahu – 221025
Srisha Singh – 221087

Under the guidance of

Prof. Amarendra Kumar Singh
Prof. Soumya Sridar

Department of Materials Science and Engineering
Indian Institute of Technology Kanpur
April 2025

ABSTRACT

Marine corrosion continues to be a serious issue for the longevity and safety of metallic structures, necessitating precise and effective predictive models. In this project, different machine learning (ML) algorithms specifically **Linear Regression**, **Decision Tree**, **Random Forest**, **Support Vector Regression**, **XG Boost**, **Gradient Boost** and **Multi-Layer Perceptron (MLP) Regressor** were utilized to predict corrosion rates using environmental and material parameters. Before model training, **Pearson's correlation analysis** and f**eature importance assessment** were conducted to learn about variable relationships and improve model performance via feature selection. Out of all models tested, the **MLP Regressor performed best**, capturing well complex and non-linear relationships. The results prove the potential of data-driven methods, especially neural networks, in facilitating corrosion prediction and decision-making in marine engineering applications. After hyperparameter tuning, the optimized MLP achieved a high R^2 score. Interpretability techniques such as **SHAP** and **Permutation Feature Importance** were employed to identify the most influential features. The results underscore the model's strong predictive capability and its potential application in corrosion forecasting in complex marine environments.

CONTENTS

1. INTRODUCTION

Low-alloy steels are a class of ferrous alloys that contain small amounts (typically less than 5% by weight) of alloying elements such as chromium (Cr), nickel (Ni), molybdenum (Mo), copper (Cu), and vanadium (V). These elements are added in controlled quantities to improve specific properties of the steel.

Unlike carbon steels, which are more prone to rapid corrosion, low-alloy steels offer a balance between mechanical performance and environmental durability, making them an economical and practical choice for structural applications. The alloying elements can influence the resistance to localized corrosion and modify the steel's electrochemical behavior in saline environments.

In marine environments, where structures are constantly exposed to high humidity, salt spray, and fluctuating temperatures, low-alloy steels are widely used in applications (e.g. Ship hulls and superstructures, offshore platforms and oil rigs, marine pipelines and risers, coastal bridges and piers, harbour and port infrastructure).

The combination of mechanical strength and moderate corrosion resistance makes low-alloy steels suitable for these applications, especially when cost constraints make the use of stainless steels or corrosion-resistant alloys impractical.

However, the corrosion performance of low-alloy steels is highly sensitive to both chemical composition and environmental exposure conditions. Therefore, predicting steel behavior under marine atmospheric conditions is crucial.

Nevertheless, existing methods of corrosion prediction have limitations as numerous models are built based on limited environmental and material data, which limit their validity. Most research tends to emphasize mostly linear modeling approaches that do not necessarily reflect the intricate, non-linear interactions involved in corrosion behavior. Also, conventional experimental techniques tend to provide a limited range of results owing to time, cost, and scalability limitations, thus restricting the diversity and quantity of data available for solid analysis.

This project addresses these challenges through a progressive modeling framework that integrates stochastic sampling techniques, statistical feature selection, and regression algorithms. Our approach evolves from semi-empirical models using Latin Hypercube Sampling (LHS) for synthetic data generation to multi-variable regression models leveraging real-world experimental datasets. The methodology encompasses dimensionality reduction through variance thresholding, standardization via z-score normalization (StandardScaler), and both parametric and non-parametric regression techniques including gradient descent

optimization, decision tree ensembles, and feed-forward neural networks with backpropagation.

Through quantitative performance evaluation using coefficient of determination (R^2) metrics across varying train-test splits, we systematically compare the predictive capabilities of ensemble methods (Random Forest, XGBoost, LightGBM), kernel-based approaches (SVR), and neural architectures (MLP). This research demonstrates the potential of advanced machine learning techniques to transform corrosion prediction in marine engineering, providing a mathematical foundation for intelligent systems that can optimize material selection and maintenance strategies for structures in corrosive environments.

2. DATA COLLECTION AND CURATION

The data collection and preparation process for this study involved multiple stages, aimed at identifying a reliable and representative dataset for corrosion rate prediction. Initially, we explored a large corrosion database (Corr-Database), which contained a wide variety of environmental conditions and material compositions. However, due to the high variability and heterogeneity in this dataset, the machine learning models were unable to converge effectively, resulting in poor predictive performance.

To address the problem, we generated a synthetic dataset through Latin Hypercube Sampling (LHS), a statistical method that assists in creating varied sets of input values. The input ranges were derived from parameter distributions in earlier research [1]. This process enabled us to generate a wide and balanced training dataset. But because the data wasn't from actual experiments, we didn't have an accurate method to test how well the model would perform in the real world. Without real-world testing, the value of the results was restricted.

Eventually, we obtained a curated and pre-processed dataset [2] which is a subset from the National Institute for Materials Science (NIMS) CoDS database[3], which offered a well-balanced set of features and corresponding corrosion rate values. This final dataset was used to train and evaluate various machine learning models. It provided the necessary structure and quality to develop reliable prediction models for predicting corrosion rates in marine environments. A correlation analysis was performed to evaluate interdependencies among features and identify those most influential in predicting corrosion rates. The results, shown in a hierarchical ranking plot (Figure 1), guided feature selection for the modelling phase.

Table 1: Pre-processed NIMS CoDS dataset

Si	Mn	P	S	Al	Cu	Cr	Ni	ELEMENTS	T_MAX	T_MIN	T_AVE	RH_MIN	RH_AVE	SUNSHINE	TOW	PRECIPIT	WIND_MA	WIND_AVI	SOLAR	UV	CHLORIDE	SO2	TIME	Corrosion rate
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	35	-7.2	14.5	37	77.5	1457.6	4670	1344.5	12	2.5	5740	246.9	3.3	5.3	1	0.0161
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	32.2	-5.5	15.3	50.5	79	1828.9	4908.6	1511.5	16.1	3.5	4193.3	301.7	32.3	5.1	1	0.0484
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	33.4	9.2	24	57	79	1701.4	5248.6	2314	20.9	4.7	5229	338.1	45.8	2.1	1	0.0583
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	35.9	-6.3	15.1	37	77.5	1457.6	4670	1138.5	14	2.3	5740	246.9	3.3	5.3	2	0.0132
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	32.4	-5.5	15.1	45.7	78.7	1880.3	4835	1484.8	15.8	3.4	4352	307.2	31.2	5.5	2	0.0445
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	33.2	9.8	23.9	58	79.7	1677.6	5369.8	2321	27.8	4.6	5207.2	351.7	42.3	2.3	2	0.0544
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	36.3	-6.4	15.3	37	77.5	1457.6	4670	1103.1	11.7	2.1	5740.6	246.9	3.3	5.3	3	0.0127
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	32.3	-5.9	14.9	44.3	78.5	1913	4788.5	1515.4	17	3.3	4539.7	304.1	33.4	6.1	3	0.0451
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	33.4	9.7	23.9	56.8	79.3	1706	5275.2	2302.3	29	4.5	5295.5	355.8	43.9	2.2	3	0.0916
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	36.7	-5.3	15.5	33.5	75.8	1457.6	4339.2	1160.5	9.1	1.7	6274.2	272.9	2.8	4.3	7	0.0088
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	32.4	-6.5	14.6	42.9	77.9	1952.9	4670.2	1687.1	20.8	3.1	4875.1	303.1	34.7	5.3	7	0.0454
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	33.6	8.5	23.8	55.9	79.1	1775.4	5235.5	1914.5	33.5	4.3	5212	357.3	47.9	2.3	7	0.1471
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	36.6	-4.4	15.3	26.9	74.6	1457.6	4088.1	1270.1	10.6	1.5	4182.8	181.9	2.8	3.7	10	0.0067
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	32.6	-6.1	14.7	35.2	77.6	1894	4629.3	1791.1	23.6	3	4901.4	294	32	4.9	10	0.0396
0.003	0.003	0.0006	0.0007	0.001	0.009	0.005	0.003	0.0253	33.9	8.6	23.9	50	78.5	1759.4	5113.3	1898.7	32.3	4.2	5260.1	349.3	49.2	2.4	10	0.1984
0.003	0.01	0.0003	0.0001	0.001	0.009	0.005	0.98	1.0084	35	-7.2	14.5	37	77.5	1457.6	4670	1344.5	12	2.5	5740	246.9	3.3	5.3	1	0.0114
0.003	0.01	0.0003	0.0001	0.001	0.009	0.005	0.98	1.0084	32.2	-5.5	15.3	50.5	79	1828.9	4908.6	1511.5	16.1	3.5	4193.3	301.7	32.3	5.1	1	0.0371
0.003	0.01	0.0003	0.0001	0.001	0.009	0.005	0.98	1.0084	33.4	9.2	24	57	79	1701.4	5248.6	2314	20.9	4.7	5229	338.1	45.8	2.1	1	0.0385
0.003	0.01	0.0003	0.0001	0.001	0.009	0.005	0.98	1.0084	35.9	-6.3	15.1	37	77.5	1457.6	4670	1138.5	14	2.3	5740	246.9	3.3	5.3	2	0.0087

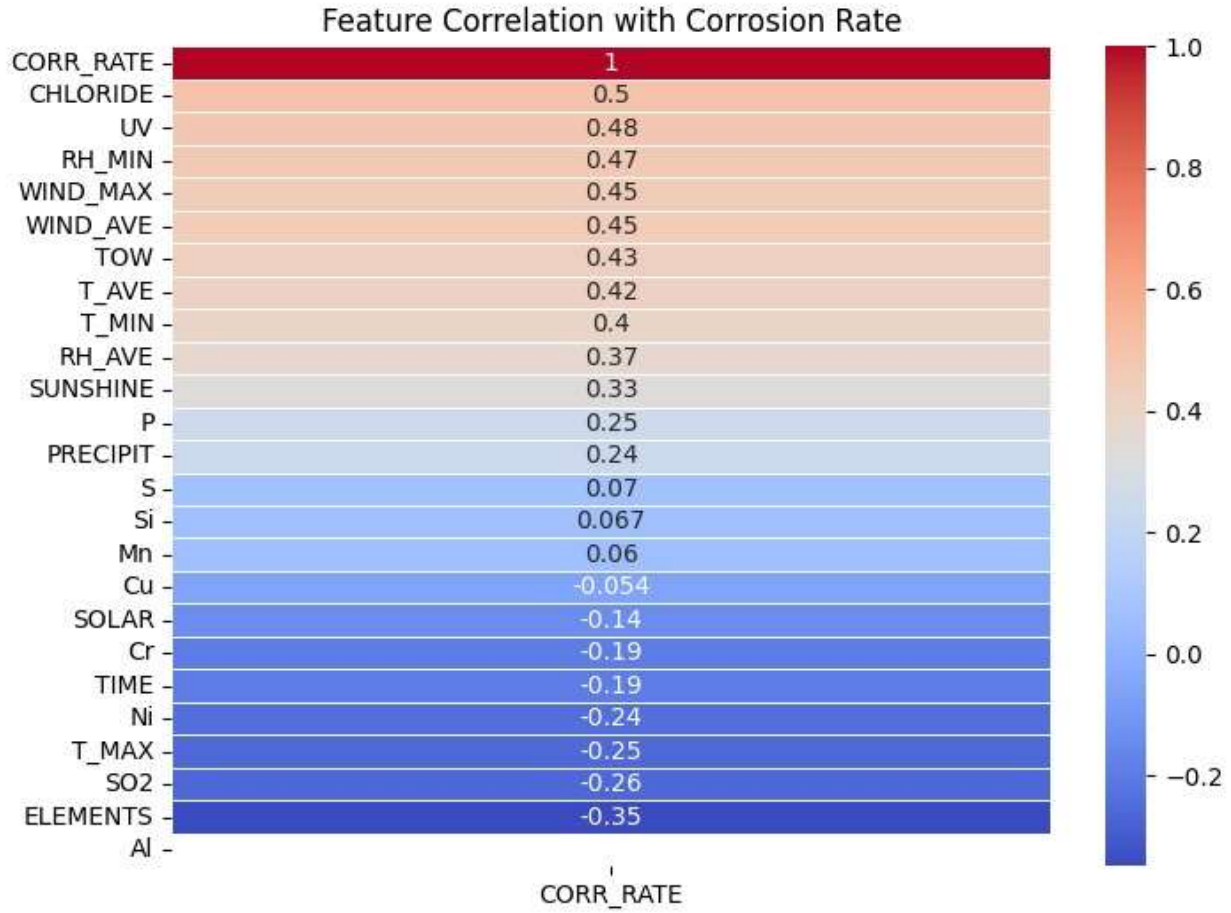


Figure1.: Feature correlation with Corrosion rate of NIMS pre-processed data

3. MODELING FRAMEWORK

i) EMPIRICAL CORROSION MODEL

Our initial approach to corrosion rate prediction was grounded in a semi-empirical model, designed to estimate corrosion depth based on a limited set of environmental factors—specifically, temperature, dissolved oxygen content, seawater velocity, and

exposure time. This model was inspired by existing formulations in the literature and used correction factors to adjust for deviations from standard reference conditions.

In order to train and test machine learning models, a synthetic dataset was generated using **Latin Hypercube Sampling (LHS)** over specified ranges of parameters. This allowed wide coverage of the input space and facilitated creation of an artificial neural network (ANN) model to approximate the correlation between chosen input variables and corrosion depth.

While the ANN model trained on this artificial data had a low Mean Squared Error

ii) RFR, MLR, SVR AND HYBRID MODEL

Building on insights from prior attempts, our second modelling phase was inspired by the study conducted by Thanush et al. (2022)[4], which demonstrated the application of machine learning models to predict the atmospheric corrosion rate of low-alloy steels using real-world data from the NIMS MatNavi database.

In this approach, a structured and experimentally validated dataset was used,

consisting of 23 independent variables and one dependent variable—the corrosion rate (in μm/year). The data included both material composition features (such as percentages of phosphorus, silicon, manganese, etc.) and environmental parameters (like temperature, humidity, chloride and SO_2 deposition, precipitation, and time of wetness), collected from multiple test sites across Japan under open and sheltered exposure conditions.

(MSE), its use was still limited. Most importantly, the model did not take into account many significant factors known to affect corrosion behavior, including salinity, pH, and chemical composition of the material. Consequently, it did not have the complexity to accurately represent real-world corrosion mechanisms.

Additionally, since the dataset was not grounded in real experimental data, we lacked any mechanism to test the predictions made by the model within real-world ocean environments. This emphasized the necessity for a more extensive dataset encompassing a greater number of influencing factors and derived from genuine corrosion research.

Data Preparation and Feature Selection

Initial preprocessing involved handling missing values through mode and mean imputation for categorical and numerical features, respectively. Feature importance was assessed using a Random Forest Regression model to identify the most influential parameters. Less significant variables were removed based on a defined importance threshold, helping to reduce noise and enhance model generalizability.

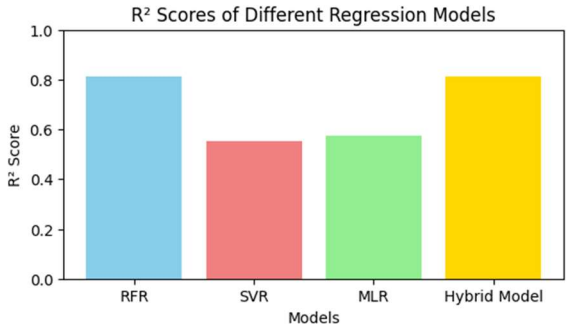
The final selected features included key alloying elements (e.g., Manganese, Phosphorus, Sulphur, Silicon and Chromium), environmental factors (e.g., chloride deposition, SO_2 deposition, and precipitation), and exposure time.

Model Development and Evaluation

Several regression algorithms were applied to predict the corrosion rate:

- Multiple Linear Regression (MLR)
- Support Vector Regression (SVR)
- Random Forest Regression (RFR)

To overcome individual model limitations, a Hybrid Model was developed using a weighted average of predictions from the three base models. The weights were fine-tuned to optimize overall prediction accuracy.



iii) COMPARATIVE MODELING AND OPTIMIZATION

Following earlier modeling trials, our third and final modeling phase focused on building a **highly optimized and generalizable model** to predict corrosion rates with high accuracy under varying environmental and material conditions

Data Preparation and Preprocessing

The initial dataset contained a wide array of environmental and material parameters. Columns with constant values were identified and removed to avoid redundancy and improve model generalizability. Any missing values in the dataset were imputed using the **mean strategy**.

Feature variables and the target variable (corrosion rate, in µm/year) were scaled using the **StandardScaler**, standardizing the data to ensure better training stability and faster convergence across machine learning models. The input features included key environmental parameters such as **chloride deposition**, **relative humidity**, **precipitation**, **UV exposure**, and **Sulphur dioxide levels**, along with compositional data like **Si**, **Mn**, **P**, **S**, **Cu**, **Cr**, and **Ni** content.

Table 2. List of considered material and environmental features.

Features		Data range	Descriptions
Material	ELEMENTS	0.5–9.2 wt.%	Total content of alloying elements
Environmental	T_MAX	31.0–37.0 °C	Maximum air temperature
	T_MIN	−8.0–9.5 °C	Minimum air temperature
	T_AVE	14.2–24.0 °C	Mean air temperature
	RH_MIN	15.0–55.0%	Minimum relative humidity
	RH_AVE	72.5–79.5%	Mean relative humidity
	SUNSHINE	1450 – 1990 h	Duration of sunshine
	TOW	3700 – 5300 h	Time of wetness
	PRECIPIT	1100 – 2300 mm	Precipitation
	WIND_MAX	5.5–39.5 m/s	Maximum velocity of wind
	WIND_AVE	1.1–4.7 m/s	Mean velocity of wind
	SOLAR	4100 – 6600 MJ/m^2	Solar radiation
	UV	180 – 370 MJ/m^2	Ultraviolet radiation
	CHLORIDE	2 – 55 mg $NaCl/m^2 \cdot d$	Chloride deposition rate
	SO_2	1.8–6.1 mg $SO_2/m^2 \cdot d$	SO_2 deposition rate
	TIME	1, 2, 3, 5, 7, 10 years	Exposure period
Target property	Corrosion rate	0.0003–0.1995 mm/a	Annual corrosion depth, millimeter per year

Model Development and Evaluation

A comprehensive comparison was performed using the following 15 machine learning regression algorithms:

- **Random Forest Regressor**
- **Gradient Boosting Regressor**
- **Extra Trees Regressor**
- **Support Vector Regression (SVR)**
- **Ridge Regression**
- **Lasso Regression**
- **Multi-Layer Perceptron (MLP) Regressor**
- **XGBoost Regressor**
- **LightGBM Regressor**
- **CatBoost Regressor**
- **Histogram-based Gradient Boosting Regressor**
- **AdaBoost Regressor**
- **Bagging Regressor**
- **Decision Tree Regressor**
- **K-Neighbors Regressor**

Each model was trained and evaluated using an appropriate train-test split. And evaluated using the **R^2 score** as the primary performance metric. Evaluation results were visualized through comparative bar plots, facilitating a clear understanding of each model's predictive power.

Among all models, the **Multi-Layer Perceptron (MLP) Regressor** stood out as the most effective. It captured the complex, non-linear dependencies between features and the corrosion rate, outperforming traditional ensemble and linear models.

Hyperparameter Optimization

To maximize the performance of the MLP Regressor, a detailed **hyperparameter tuning** process was carried out using RandomizedSearchCV. The search space included:

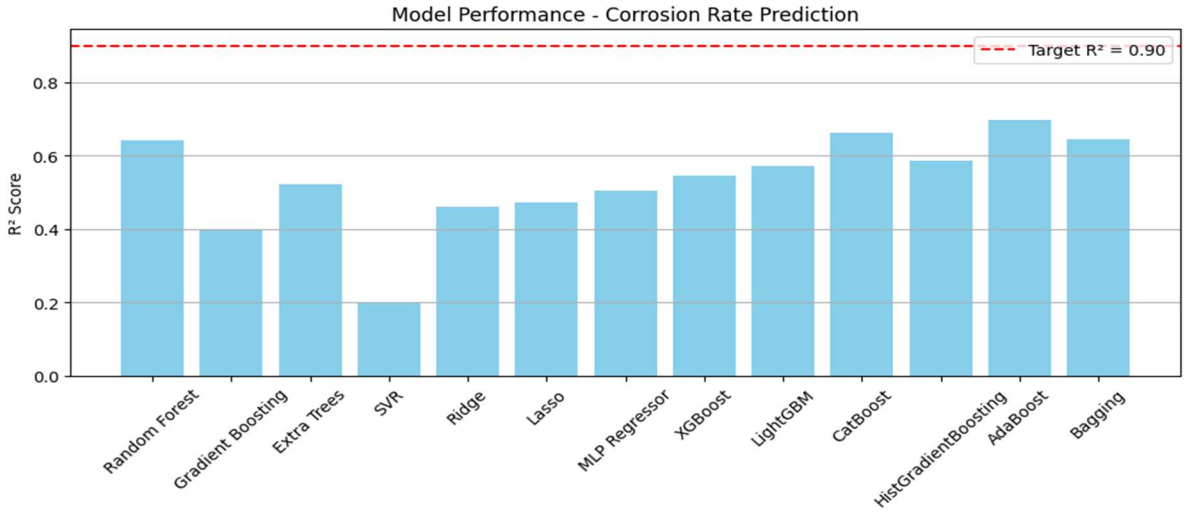
- Hidden layer architectures (e.g., (100,), (100, 100), (150, 100))
- Activation functions (relu, tanh)
- Regularization strengths (alpha)
- Learning rates (constant, adaptive)
- Maximum iterations for convergence

After optimization, the final tuned model achieved an impressive test-set **R^2 score of 0.9775**, highlighting its ability to generalize well across unseen data and validating its suitability for practical applications in marine corrosion forecasting

4. RESULTS AND DISCUSSION

A comparative analysis of multiple machine learning models was conducted using key performance metrics including R^2, MSE, RMSE, and MAE. The results are summarized in Table below, highlighting the strengths and weaknesses of each approach. This evaluation facilitated the selection of the most accurate and generalizable model for corrosion rate prediction.

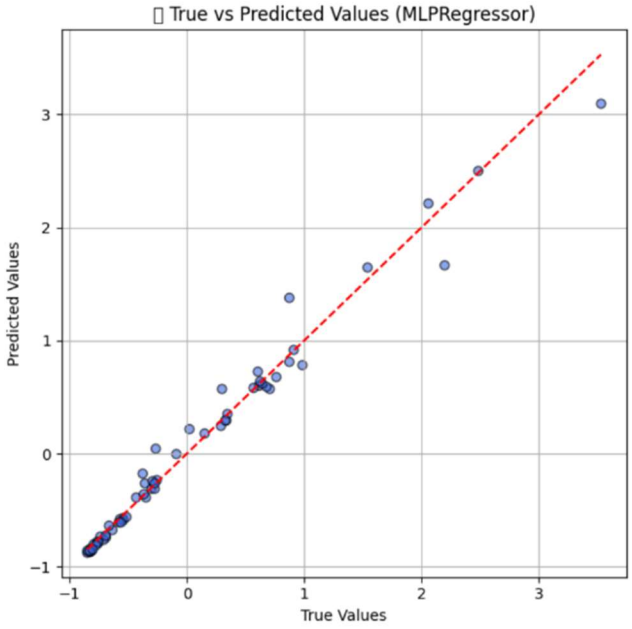
The predictions of the optimized MLP Regressor were compared with those from the best-performing ensemble model (Random Forest). The MLP model demonstrated **superior predictive alignment**, particularly in regions with complex, non-linear interactions between input feature.



After hyperparameter optimization, the MLP Regressor emerged as the most effective model for corrosion rate prediction. Post-tuning, it achieved an impressive R^2 score of 0.9775, indicating excellent predictive capability. The best-performing configuration included the Adam solver, ReLU activation, and a hidden layer architecture of (150, 100), with a constant learning rate and an alpha value of 0.001. This combination allowed the model to learn complex patterns in the data while maintaining good generalization.

Metric	Formula	What It Means	MLP Result
MAE (Mean Absolute Error)	$MAE = (1/n) \times \Sigma \lvert y_i - \hat{y}_i \rvert$	Measures the average absolute error in predictions; lower values indicate better accuracy. A value of **0.0732** shows the model makes very small errors in µm/year units.	**0.0732**
MSE (Mean Squared Error)	$MSE = (1/n) \times \Sigma (y_i - \hat{y}_i)^2$	Penalizes larger errors more than MAE. A low value like **0.0116** reflects strong consistency and few large deviations.	**0.0116**
RMSE (Root Mean Squared Error)	$RMSE = \sqrt{[(1/n) \times \Sigma (y_i - \hat{y}_i)^2]}$	Expresses average error magnitude in original units. A result of **0.1077 µm/year** is very low and interpretable.	**0.1077**

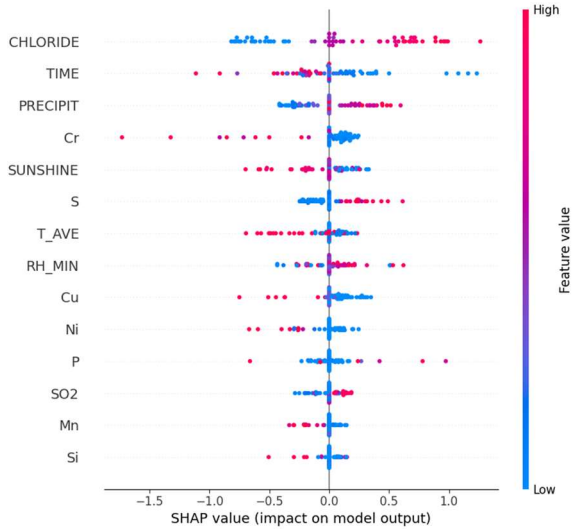
MAPE (Mean Absolute Percentage Error)	MAPE = (100/n) × Σ | $(y_i - \hat{y}_i) / y_i$ |	Indicates average prediction error as a percentage. A **26.20%** MAPE is acceptable given the variability in corrosion behavior.	**26.20%**
R^2 Score (Coefficient of Determination)	$R^2 = 1 - [\Sigma (y_i - \hat{y}_i)^2 / \Sigma (y_i - \bar{y})^2]$	Measures how well the model explains target variance. A score of **0.9775** signifies excellent model fit and generalization.	**0.9775**



The scatter plot illustrates the predicted vs. true values for the MLPRegressor model after hyperparameter tuning. Most data points lie close to the red dashed line, which represents the ideal case where predicted values equal true values. This indicates strong model performance and minimal prediction error. A few outliers exist, but overall, the model generalizes well across the dataset.

Residual plots indicated **tight clustering around zero**, and a significant reduction in prediction bias — further reinforcing the model's robustness.

Additionally, **Permutation Feature Importance** and **SHAP (SHapley Additive exPlanations)** analysis were conducted to interpret the model's behavior. These tools revealed the relative importance of input features and explained how each contributed to the predicted corrosion rate.



Feature Selection and Justification

To ensure model efficiency and scientific relevance, we performed **feature selection** using a combination of **statistical correlation analysis** and **domain-specific material science knowledge**.

Correlation-Based Feature Clustering

Here's how we systematically analyzed correlations to reduce redundancy and group features for selection

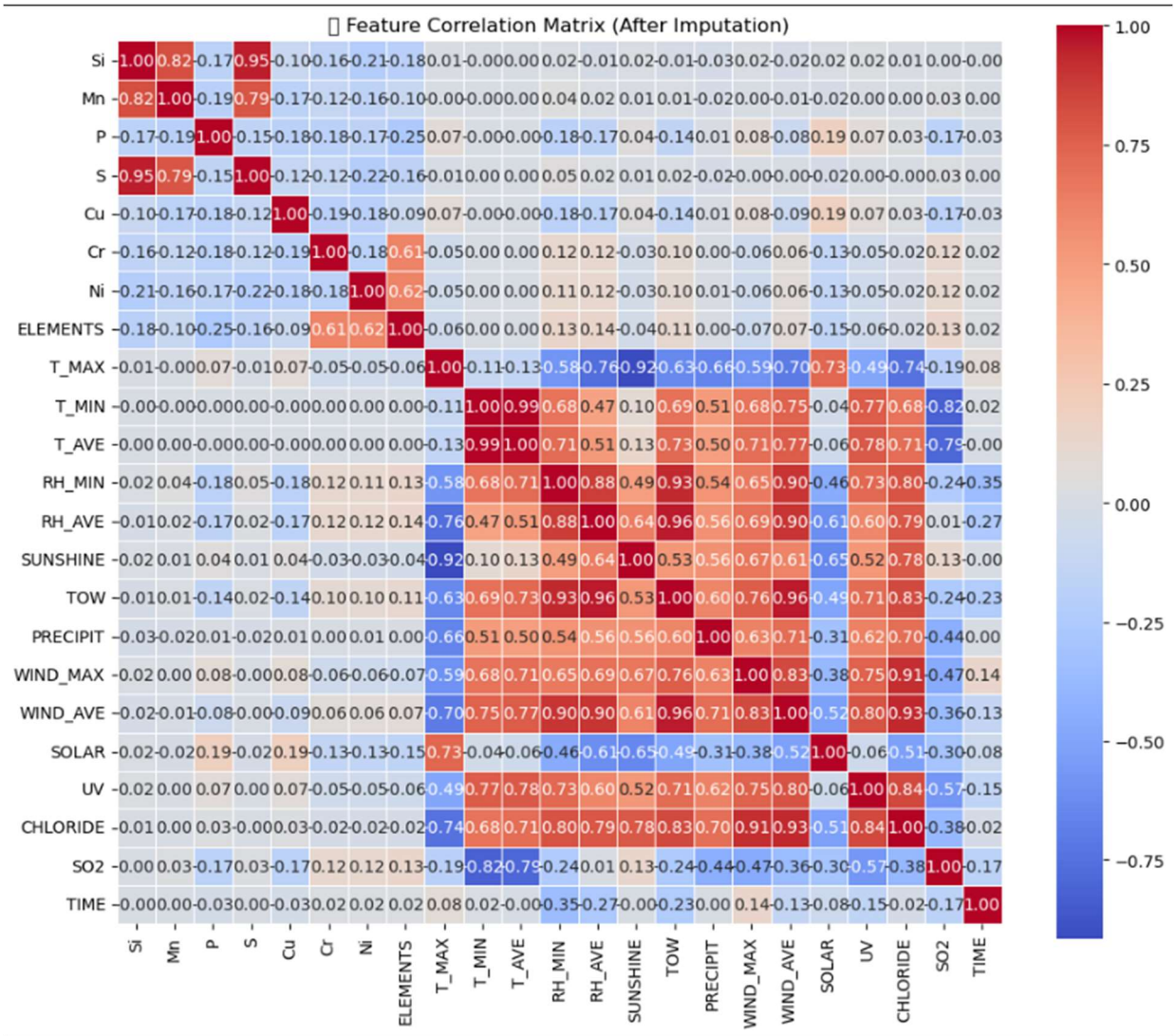


Figure a: Pearson correlation matrix of all input features (material and environmental) after missing value imputation. High correlations (|r| > 0.8) indicate multicollinearity, particularly among humidity, temperature, and wind-related variables, supporting the clustering-based feature reduction approach

The **Pearson correlation matrix** shows the linear relationships among the 22 original features, with values close to ±1 indicating strong correlation. Based on the heatmap, we identified four primary **clusters of correlated environmental variables**, suggesting potential multicollinearity:

- in urban coastal regions with industrial emissions. Thus, increased sunlight and temperature often coincide with higher pollutant levels.

- **Cluster B**: RH_MIN, RH_AVE, TOW, PRECIPIT
  These variables describe the **moisture regime** of the exposure environment. High average or minimum humidity (RH_MIN, RH_AVE) promotes longer time of wetness (TOW), which is the duration when relative humidity remains above the dew point, enabling a conductive electrolyte film to persist. PRECIPIT` (precipitation) further amplifies surface wetting, especially under high-humidity conditions, reinforcing the presence of a corrosive layer.

- **Cluster C**: WIND_MAX, WIND_AVE, UV, CHLORIDE
  This group captures **salt transport and deposition mechanisms**. Wind (both average and gusts) facilitates the movement and deposition of chloride-rich aerosols from the sea onto metal surfaces. Concurrently, strong UV radiation enhances **evaporation rates**, concentrating deposited salts and **accelerating corrosion onset**, especially during dry-wet cycling.

- **Cluster A**: T_AVE, T_MAX, SUNSHINE, SO2
  These features are interconnected through **seasonal and diurnal thermal dynamics**. High sunshine exposure increases atmospheric temperature (T_MAX, T_AVE), which in turn influences **photochemical reactions** that elevate SO2 concentrations, especially

- **Cluster D**: Si, Mn, P, S, Cu, Cr, Ni
  Although statistical correlation exists due to alloy design practices, each element has a **chemically distinct role**. These elements collectively influence **passivation behavior**, **inclusion formation**, and the **electrochemical stability** of the steel. For instance, Cr, Cu, and Ni enhance corrosion resistance, while Mn, P, and S may form localized sites that **initiate pitting** or **intergranular corrosion.**

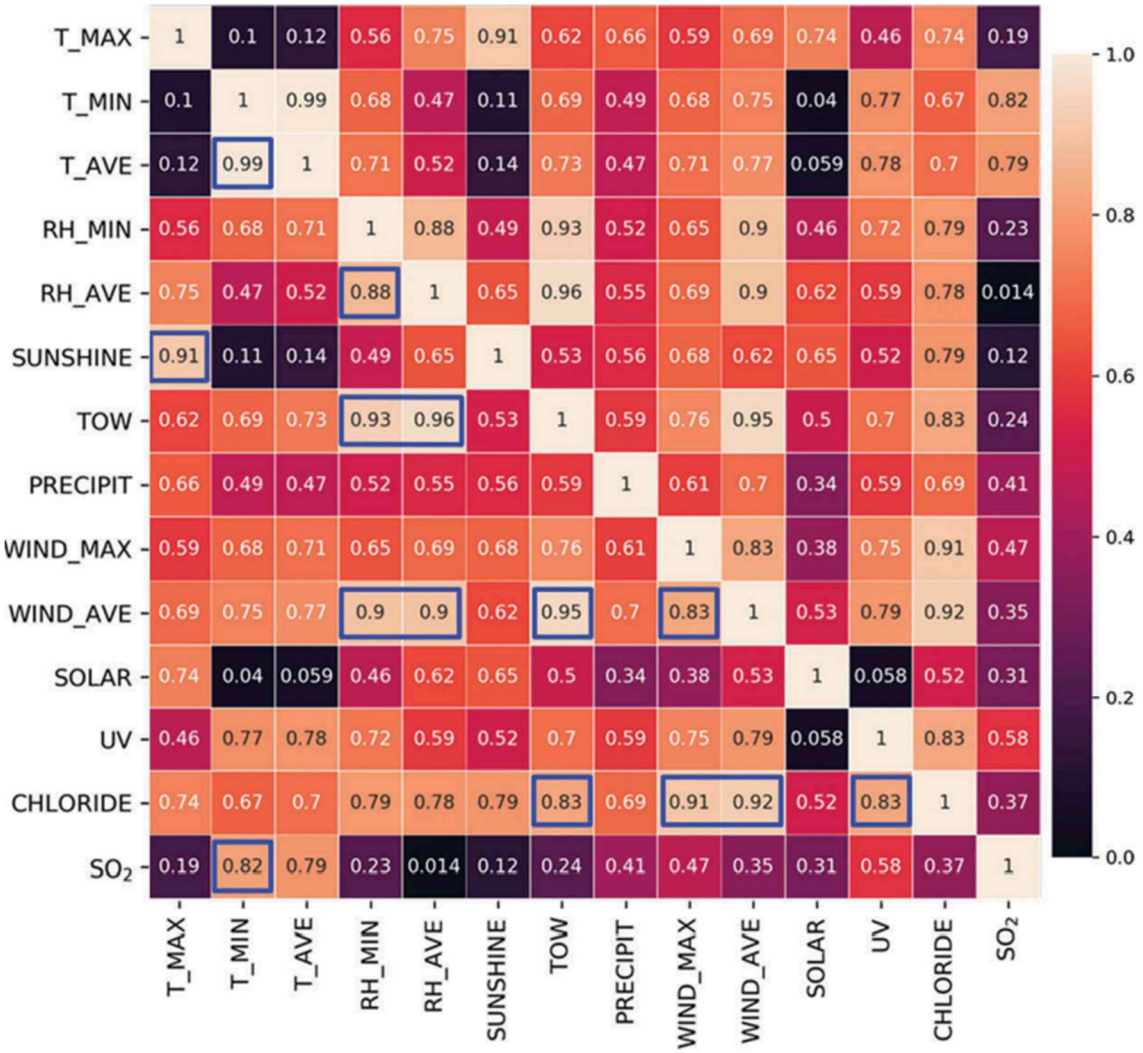


Figure b: Correlation matrix among environmental features. Strong correlations (e.g., `T_AVE` with `T_MAX`, `RH_MIN` with `RH_AVE`, and `CHLORIDE` with `WIND_AVE`) highlight interdependencies in climate variables, motivating the formation of clusters for simplified model input.

From each environmental cluster, we selected the **most physically meaningful feature** based on corrosion science:

- `T_AVE` — average temperature governs electrochemical kinetics.
- `RH_MIN` — minimum humidity determines electrolyte persistence.
- `CHLORIDE` — key aggressive ion in marine corrosion.
- `SO2` — contributes to acidic deposition and atmospheric attack.

These selections maintain diversity across weathering parameters while minimizing multicollinearity.

Why are these columns retained?

T_AVE was retained as the most representative and physically interpretable feature, influencing:

- Electrochemical reaction rates
- Evaporation-condensation cycles

SO2 was also retained **despite correlation**, due to its **distinct role in forming acidic deposits**, making it a chemically significant pollutant in corrosion.

RH_MIN was selected:

- It controls the **threshold condition** for moisture film formation.
- More critical than average humidity (RH_AVE) in determining when corrosion can occur.

PRECIPIT was retained separately due to its role in **sustained electrolyte** presence, despite mild correlation.

CHLORIDE was retained as the most directly influential feature: High chloride deposition significantly accelerates pitting and crevice corrosion.

In contrast to prior studies (e.g., Yan et al., 2020) that reduced feature sets purely through statistical clustering, we intentionally **preserved composition-level granularity**. Each retained element has a distinct impact on corrosion behavior:

Element	Corrosion Influence	Effect
Si	Promotes passive oxide formation	▼ Negative correlation (resists corrosion)
Mn	May form MnS inclusions	▲ Positive (can initiate pits)
P	Enhances intergranular attack	▲ Positive
S	Forms sulfide inclusions, localizes corrosion	▲ Positive
Cu	Improves atmospheric corrosion resistance	▼ Negative
Cr	Key passivator forming Cr_2O_3 layer	▼ Negative
Ni	Stabilizes passive layer in chloride-rich conditions	▼ Negative

Additionally, TIME was retained as it directly captures cumulative corrosion exposure.

Final Feature Set

Combining physical insight with statistical evidence, we retained the following **14 features**:

- **Environmental**: `T_AVE`, `RH_MIN`, `CHLORIDE`, `SO2`, `SUNSHINE`, `PRECIPIT`, `TIME`
- **Material Composition**: `Si`, `Mn`, `P`, `S`, `Cu`, `Cr`, `Ni`

This selection ensures the model receives a rich, physically interpretable feature set capable of capturing the multifactorial nature of atmospheric corrosion.

5. CONCLUSIONS

This project successfully developed a highly accurate predictive model for corrosion rates of low-alloy steels in marine environments through a progressive modeling approach. Key findings include:

1. A progressive modeling approach evolving from **semi-empirical models to advanced machine learning techniques** demonstrated significant improvements in predictive accuracy for corrosion rate estimation.
2. Synthetic data generation using **Latin Hypercube Sampling** (LHS) proved useful for initial model development but highlighted critical limitations in representing real-world corrosion mechanisms without experimental validation.
3. Real-world data from the **NIMS MatNavi database** [2] significantly enhanced model performance by incorporating crucial material composition features and environmental parameters that influence corrosion behavior.
4. Feature selection identified key corrosion determinants: specific alloying elements (particularly Mn, P, S, Si, and Cr) and environmental factors (chloride deposition, SO_2 levels, precipitation, and exposure time) emerged as the most influential predictors.
5. Among 13 tested algorithms, the **Multi-Layer Perceptron (MLP) Regressor outperformed** all others, including **ensemble** and **kernel-based methods**, demonstrating its capability to model the complex, non-linear interactions inherent in corrosion processes.
6. Hyperparameter optimization of the MLP model resulted in metrics (**R^2 score of 0.9775**), confirming its ability to accurately predict corrosion rates across diverse conditions.
7. The optimal MLP configuration utilized the **Adam solver**, **ReLU activation function**, and a **hidden layer architecture** of (150, 100), with constant learning rate and an alpha value of **0.001**.
8. Interpretability methods such as **Permutation Feature Importance** and **SHAP analysis** highlighted key environmental and material features driving corrosion behavior.
9. The optimized model offers a reliable basis for intelligent corrosion prediction, supporting improved decision-making in material selection and maintenance strategies for marine structures

6. REFERENCES

0.1] Guedes Soares, C., Garbatov, Y., & Zayed, A. (2011). Effect of environmental factors on steel plate corrosion under marine immersion conditions. *Corrosion Engineering, Science and Technology*, *46*(4), 524–541. https://doi.org/10.1179/147842209X12559428167841

[2] Yan, L. (2019). *Processed data for modeling - corrosion science* (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/nh5fmdcxkd.1

[3] https://cods.nims.go.jp/

[4] Thanush, A. A., Chitra, P., Kasinath, J., & Prakash, R. S. (2022). Atmospheric corrosion rate prediction of low-alloy steel using machine learning models. *IOP Conference Series: Materials Science and Engineering, 1248*(1), 012050. https://doi.org/10.1088/1757-899X/1248/1/012050

7. APPENDIX

All the related codes and datasets can be found at:

https://github.com/UGP-Group-D/Corrosion-rate-prediction