

RESEARCH ARTICLE

Predictive Modelling of Steel Corrosion: A Comparative Study of Statistical and Machine Learning Methods with a Focus on Copper Behavior

Mohanad S. Hasan*, Amjad M. Bader, Saad Sh. Sammen

College of Engineering, University of Diyala, Diyala, Iraq

*Corresponding author: Mohanad S. Hasan, mohaned.s@uodiyala.edu.iq

ABSTRACT

Accurately predicting corrosion rates in low-alloy steels is a significant challenge in materials engineering due to the intricate and nonlinear interaction between environmental exposure conditions and alloying elements. Conventional statistical corrosion models are by and large based on linear assumptions, and thereby most of them fail to capture such interactions, especially the dual and threshold-controlled behavior of Cu in forming protective rust layers. This study presents a comprehensive comparative analysis of statistical and machine learning approaches for the prediction of atmospheric corrosion rates of low-alloy steels. Several predictive models were developed and validated for this task, namely, Decision Trees, Support Vector Machines with several kernel functions, and the gradient-boosting algorithm XGBoost, by using experimentally derived environmental and compositional datasets. Model performance is assessed based on the coefficient of determination (R^2), root-mean-square error (RMSE), and mean absolute error (MAE).

The results demonstrate that non-linear models outperform conventional linear models of non-linear models in comparison to the conventional linear models. Among various models, the best predictive model was XGBoost with an R^2 value of about 0.88, validating the best ability to exploit threshold values and interactive coefficients. The conclusion is drawn from the result that copper is only effective in imparting corrosion resistance in a critical range, beyond which the protective property becomes saturated or decreases, an imperative that cannot be accounted for in the linear regression models. The obtained results strongly suggest that the corrosion process in low-alloy steels is governed by nonlinear and threshold-dependent behavior. Therefore, the paper establishes the application of gradient - boost-based machine learning algorithms as a strong and authentic method for predicting the development of corrosion models in low-alloy steels. Furthermore, the comparative assessment also emphasizes the importance of kernel-based learning approaches, such as

the radial basis function kernel-based SVM, where the models also forecast well with slightly lower accuracy compared to the gradient-boosters models. Additionally, the better performance of the non-linear models clearly depicts the model flexibility to cope with the abrupt transitions of the corrosion process influenced by varying concentrations of alloys. From the scientific aspect, the obtained results demonstrating the threshold effects of the copper concentrations clearly support the formation and saturation of inner Cu-enriched rust layers. Therefore, this analysis attempt to connect the machine learning algorithm results with the scientific concepts of the corrosion process and hence establish the importance of machine learning models as a supplementary aspect to the other scientific methods used in the fields of materials science.

Keywords: Atmospheric corrosion; Low-alloy steel; Copper threshold effect; Machine learning; XGBoost; Predictive modelling; Corrosion rate prediction; Environmental exposure.

ARTICLE INFO

Received: 23 March 2026

Accepted: 18 May 2026

Available online: 15 September 2026

COPYRIGHT

Copyright © 2026 by author(s).

Applied Chemical Engineering is published by Arts and Science Press Pte. Ltd. This work is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY 4.0).

<https://creativecommons.org/licenses/by/4.0/>

1. Introduction

The low-alloyed steel corrosion in the atmospheric environment is perhaps most damaging amongst all other forms of corrosion. Steel structures undergoing varying conditions of temperature, humidity levels, and pollutant attacks find themselves under constant electrochemical reactions resulting in material decomposition and subsequent failures. The economic impacts associated with failures due to corrosion have ensured continuous efforts in researching and finding predictive models for corrosion predictions based on feasible conditions ^[1-3].

Traditionally, the relationship between the rate of corrosion and environmental exposure variables had been established using empirical or semiempirical statistical models. Despite the fact that linear SVM can create an easy-to-understand model, it cannot completely substitute more traditional methods for modelling corrosion, such as multiple regression analysis, ridge regression, lasso regression, or the dose-response approach, as traditionally used for studying atmospheric corrosion. Initial methods involving the dose - response relationship or linear regression analyses were satisfactory for expressing first-order relationships for pairs of variables such as temperature, humidity, or sulphur dioxide deposition and the rate of corrosion ^[2]. However, these conventional methods tend to assume linear and mutually independent relationships between variables. In the case of low-alloy steel and environmental exposure variables, these assumptions are less valid. Corrosion in these conditions is a nonlinear multivariable process involving synergistic factors.

Alloying elements play a central role in the control of corrosion behaviour in low-alloy steels. Chromium (Cr) and nickel (Ni) are well known to exert a positive effect on corrosion resistance by stabilizing passive oxide layers. Copper (Cu), however, exerts a complex and dual behavior ^[9]. At low concentration, Cu enhances corrosion resistance owing to the promotion of the formation and densification of a protective inner rust layer that reduces ionic transport and retards anodic dissolution. Above some critical concentration, this beneficial effect may saturate or even decrease owing to changes in rust chemistry and/or the onset of micro-galvanic interactions. This phenomenon, commonly referred to as the threshold effect of copper, introduces strong non-linearity into the corrosion system and represents a major challenge for conventional linear modeling techniques ^[10].

has been attempted in various studies using various advanced statistical analysis techniques based on correlation. Although correlation heatmaps and multi-variable regression techniques provide valuable insights into main linear trends, these techniques have limitations in capturing sharp transitions, interaction, or threshold-related behaviors ^[6]. Therefore, predictions made based only on linear or weakly nonlinear models tend to have high uncertainties when applied to new environmental conditions or alloy compositions. Some renewed interest has recently resulted, because machine learning (ML) methods have proven to be effective approaches to handle complex, high-dimensional systems, particularly within the realm of materials science and corrosion engineering. Compared to classical regression, ML methods lack the need to specify known relationships and can indeed learn these relations, even if they are not linear, from the data alone ^[11]. Decision Trees, Support Vector Machines (SVM), and gradient boosting algorithms are among the approaches that have yielded optimistic results as to material property predictions, material degradation, and corrosion performance. Gradient boosting methods like XGBoost have proven to have outstanding performance for learning about dependencies and interaction effects, which are common to corrosion phenomena ^[7].

In the context of these developments, however, there is currently relatively little focused research on comparative analysis involving the assessment of models from the realm of statistics and machine learning regarding the prediction of atmospheric corrosion with particular emphasis on the threshold-controlled process of copper ^[11]. While some machine learning algorithms have already been proposed for the atmospheric corrosion modeling problem, very little research has been done on the comparative study of linear and nonlinear modeling methods in relation to the behavior of copper-containing alloys in the environment. The

study of copper-based nonlinear corrosion has not yet received adequate attention. In general, the work that has been completed tends to focus on individual models or limited data sets with the goal of understanding the physical process from the basis of the models selected ^[12]. Often, the emphasis of the work is purely on the predictive validation of the models with little direct attention to the physical process of the corrosion. The current research focuses on bridging the identified shortcomings by performing a comprehensive comparison study between conventional statistical models and sophisticated machine learning models in accurately predicting the atmospheric corrosion rate for low-alloy steel materials ^[16]. Through the fusion of environmental exposure variables and chemical composition information for the alloys, this research will thus be able to (i) accurately define the limitations associated with conventional linear statistical models, (ii) establish the superiority associated with non-linear machine learning models, and (iii) achieve a deeper understanding related to the threshold-based involvement of the material composition element copper in the enhancement of corrosion resistance ^[20].

This particular study is supposed to be considered as a comparative study involving nonlinear predictive modeling aimed at gaining a better understanding of nonlinear corrosion behavior prediction rather than an industrial corrosion life prediction methodology.

Table 1. Comparative overview of statistical and machine learning approaches for atmospheric corrosion prediction of low-alloy steels.

Study / Modelling Approach	Methodology	Dataset Scope	Performance Metric (R^2)	Key Corrosion / Modelling Outcome
Classical statistical models (various studies)	Linear regression / Dose–response functions	Environmental parameters only	0.60–0.75	Captures first-order trends; fails to represent non-linear and threshold effects
Correlation-based statistical analysis	Pearson/Spearman correlation	Environmental + alloy composition	< 0.70	Identifies dominant variables; limited predictive capability
SVM-based corrosion models	SVM (Linear / Polynomial / Sigmoid kernels)	Environmental + alloy composition	0.65–0.72	Moderate improvement over linear models; limited sensitivity to threshold behaviour
Advanced kernel-based ML models	SVM (RBF kernel)	Environmental + alloy composition	≈ 0.85	Captures non-linear interactions; improved prediction accuracy
This study: ML-driven corrosion prediction	XGBoost (Gradient Boosting)	Environmental + alloy composition (Cu, Cr, Ni)	≈ 0.88	Accurately captures copper threshold effect; lowest prediction error; superior generalisation

The purpose of this paper is, therefore, to examine the capability of classical statistical analysis methods in comparison to more sophisticated machine learning methods, specifically in the aspect of predicting the rates of atmospheric corrosion for low-alloy steel. With the combination of environmental parameters in the analysis of the alloying constituents, it is hoped that the paper will help in the determination of the most important parameters that influence the effects of corrosion and the effectiveness of more complex modeling in improving the accuracy of the results. Today, a tremendous amount of work has been involved in the study of modeling of corrosion, but a significant number of its limitations still lie unattended, specifically in the aspect of the interaction of the alloying elements with the environmental parameters, which are more complex in terms of linearity. In the context of low - alloy steels, environment-based parameters such as temperature, humidity ratio, and pollutant deposition rate have garnered significant attention with respect to alloying elements treated through simplified or linear approaches ^[17]. While the use of chromium and nickel alloying has garnered substantial attention because of the well-understood importance of these alloying agents to improve the level of corrosion resistance, the role of the other alloying agent, namely copper, has not been addressed to the same level of scrutiny because of the dual nature and rate-dependent properties of the element involved. Most models do not take into account the level of critical concentration at which the element influences improved levels of corrosion resistance along with the saturation or diminishment rate at which the

element progresses beyond the critical level. As such, the level of utilization of the element is still quite low despite the proven importance of the element to weathering steels. The greatest number of corrosion models that have been proposed up to now have utilized regression methods or a dose-response approach ^[12]. These methods allow the estimation of first-order trends but lack sufficient power when used to model multi-variable systems following non-linear relationships. Moreover, the majority of the corrosion literature to date appears to concentrate on the fitting of models to experimental results for the purpose of description or interpolation, while robust comparisons between different paradigms of corrosion models lack entirely. It appears that experimental corrosion data has not yet had its full potential examined with the help of new, non-traditional methods. This creates a need for the work presented in this paper, which conducts a multifaceted comparison of statistical and machine learning models developed to predict atmospheric corrosion. Specifically, a range of models, namely Decision Trees, Support Vector Machines featuring a variety of kernel functions, and a series of models developed using a gradient boosting method, has been evaluated in a systematic fashion using experimental results for environmental and compositional factors ^[16]. Furthermore, a focus has been placed upon gauging the predictive capability of each method evaluated, as well as viewing results in light of atmospheric corrosion processes ^[7]. The second uniqueness of the current study is related to the use of gradient boosting methods and the XGBoost machine learning tool as a physically interpretable data-driven solution for corrosion engineering. Contrary to other machine and statistical learning methods, gradient boosting methods are capable of considering complex phenomena implicitly without requiring a specific function to be defined. The current study bridges the gap between machine and corrosion engineering by relating prediction accuracy to physically meaningful corrosion phenomena like the formation of the protective rust layer on metals ^[20]. The current study offers a scalable solution for enhancing corrosion predictions for designing and optimizing alloy durability and low alloy corrosion resistance ^[22].

2. Methodology

Atmospheric corrosion rate prediction in low-alloy steel is a complex problem with many interacting factors such as environmental factors and chemical composition. The interaction between these factors is always non-linear and some are threshold-controlled in nature, especially with respect to the role of Copper (Cu). In order to address this complexity in an effective manner, the following study adopts a comprehensive approach based on classical analysis tools and the latest advances in machine learning models ^[6]. The approach is oriented towards assessing the ability of various models in understanding the phenomenon of atmospheric corrosion ^[7].

2.1. Dataset Description and Preprocessing.

The dataset used for the task includes experimentally obtained results for the atmospheric corrosion rates of low alloy steels. The data set used for this research work has been gathered from experimentally reported data sets of atmospheric corrosion studies found in the existing literature related to low alloy steels subjected to various environmental exposures. This data set contains information related to environmental exposure factors, alloy composition factors, and recorded corrosion rate data. The input features include temperature, relative humidity, SO₂ deposition, chloride deposition, exposure time, and alloy content factors, including Cu, Cr, and Ni. The output feature is the rate of atmospheric corrosion measured in $\mu\text{m}/\text{year}$. The variables that affect the input data set are those of environmental exposures, such as temperature, relative humidity, sulfate deposition, chloride deposition, and exposure period, along with those of alloy constituents, such as copper (Cu), chromium (Cr), and nickel (Ni). Additionally, the alloy variables are the copper (Cu), chromium (Cr), and nickel (Ni) contents. The target variable is the measured rate of corrosion, in units of micrometers per year. The values of the corrosion rate that have been used in this study pertain to experimentally obtained atmospheric corrosion rates in corrosion literature. The experimentally obtained values of the corrosion rates

were directly employed as the dependent variable for modeling purposes and were not artificially constructed through any mathematical manipulation by the machine learning model.

Before the model could be developed, the dataset was checked for completeness, consistency, and physical validity. outliers and missing data points, which could appear missing in the dataset, were checked for quality. Also, all the input variables were scaled in order to prevent numerical imbalance in the differences in size among the various input features ^[5]. This scaling of input features is important, especially for kernel machines, as they are vulnerable to differences in sizes of variables during the optimisation phase. Table 2 summarizes the descriptive statistics and value ranges of all environmental exposure variables, alloying element contents, and corrosion measurements included in the modelling dataset.

The dataset used for this research work has been created using the reported experimental values on atmospheric corrosion of low alloy steels in different environmental exposures. Before proceeding with the construction of the prediction models, the dataset collected was checked for any incompleteness or inconsistency. Wherever necessary, feature scaling was performed.

The data set employed for the purpose of this study is derived from the atmosphere corrosion studies reported experimentally within literature sources concerning low alloy steels with various environmental exposure conditions. The data set encompasses parameters of environmental exposure, alloy composition, as well as experimentally observed corrosion rates.

Table 2. Descriptive statistics of the environmental and material variables.

Variable	Unit	Min	Max	Mean	Std. Dev.
Temperature	°C	-5.0	45.0	22.4	5.2
Relative Humidity	%	30.0	98.0	65.2	12.8
SO2 Deposition	mg/m ² /day	5.0	150.0	45.6	18.3
Copper (Cu)	wt. %	0.01	0.55	0.28	0.12
Chromium (Cr)	wt. %	0.00	1.20	0.45	0.30
Nickel (Ni)	wt. %	0.00	0.65	0.15	0.08
Exposure Time	Hours	24	8760	4380	2150
Corrosion Rate	µm/y	10.5	550.2	125.4	85.6
Chloride Deposition	mg/m ² /day	2.0	120.0	35.4	15.7

Table (2) summarizes the descriptive statistics of the environmental exposure parameters, alloying element contents, exposure duration, and measured corrosion rates used in this study ^[7]. The dataset covers a wide range of climatic conditions, pollutant deposition levels, alloy compositions, and exposure durations, thus allowing enough variability for sound model training and evaluation ^[12]. The large ranges with relatively high standard deviations of some variables, such as corrosion rate, chloride-related exposure, and copper content, point to the heterogeneity and nonlinearity of the corrosion process, and further justify the application of advanced machine learning techniques.

2.2. Exploratory Statistical Analysis.

To begin with, the initial exploratory statistical analysis is needed, which will help evaluate primary trends as well as linear associations among the variables. The calculation of the Pearson correlation coefficients for all the input features and the rate of corrosion is needed. The use of the heatmap will help in assessing primary linear associations among the features.

Although this analysis proved helpful for making initial insights on a first-order level, it also indicated the difficulties and limitations associated with linear analysis methods. There were certain variables, and most importantly copper concentration, for which linear correlations with corrosion rates were weak or, to a certain extent, ambiguous ^[16]. All these points indicated that corrosion process behavior might be governed by non-linear and threshold-dependent phenomena.

2.3. Selection of Modelling Approaches.

To address the identified non-linearity and complexity, a comparative modelling strategy was adopted using three distinct families of predictive algorithms:

1. Decision Tree Regression (DT)
2. Support Vector Machine Regression (SVM)
3. Extreme Gradient Boosting (XGBoost)

These algorithms were selected to represent increasing levels of modelling sophistication and non-linear learning capability.

2.3.1. Decision Tree Regression.

A Decision Tree regression was used as a benchmark non-linear model. The Decision Trees carve up the feature space into hierarchical partitions according to some sequence of decision rules, enabling them to model simple nonlinear relationships and interaction effects. Their interpretability is one of the reasons they are attractive for initial analysis; however, single decision trees are prone to overfitting and often have rather limited generalization capability when applied to complex datasets ^[17].

In this project, Decision Tree regression was more used as a reference model to let more complex algorithms be compared to it. The impurity criterion to base the splitting on was the MSE and no limitation on depth was explicitly done in order not to limit the model's ability to explore fully the structure of the data ^[12].

2.3.2. Support Vector Machine Regression.

The SVM regression algorithm was then run to test the capability of kernel machine learning algorithms to model the corrosion process. The goal of the SVM algorithm is to find the optimal hyperplane for the high-dimensional feature space to minimize errors. Another strength of the SVM algorithm is the ability to perform the modeling process of the non-linear relationship between variables using the kernel technique.

Because it is likely to be a non-linear and threshold-level-dependent problem, four different kernel functions were tried:

- Linear kernel, representing conventional linear regression behaviour
- Polynomial kernel, allowing for higher-order interactions
- Sigmoid kernel, mimicking neural-network-like activation behaviour
- Radial Basis Function (RBF) kernel, enabling highly flexible, localised non-linear mapping

The regularization parameter (C) was chosen in order to balance model complexity. By considering a comparison of several different kernels, a determination of the benefit of applying a non-linear transformation for feature improvement could be ascertained ^[17].

2.3.3. Extreme Gradient Boosting (XGBoost).

It was also observed that the reason for choosing the advanced learning algorithm in the proposed framework was attributed to its known effectiveness in dealing with complex systems. In XGBoosting, the

method uses an ensemble learning approach with the gradient boosting decision trees. With these trees, further trees are generated to compensate for the errors of the previous trees.

Single decision trees, on the other hand, lack the ability to combine many weak learners to produce a strong learner, thus making the model less robust. XGBoost, on the other hand, is a robust algorithm as it includes parameters that prevent overfitting, such as regularization, shrinkage, as well as tree pruning ^[21]. The hyperparameters of the present research were chosen depending on the popular settings used in previous corrosion studies using machine learning algorithms to ensure consistency among the compared models. In particular, the XGBoost algorithm employed the objective function “reg:squarederror” and a predetermined number of estimators, whereas the remaining parameters were set according to literature recommendations. Another important aspect is the fact that gradient boosting is suitable for modeling the behavior of the corrosion process when it is copper-controlled, which, as mentioned, is characterized by abrupt boundaries.

Table 3. Hyperparameter settings for the implemented machine learning models.

Model	Hyperparameter	Value/Setting
Decision Tree	Criterion	MSE
	Max Depth	None
SVM	Kernel Functions	Linear, Poly, Sigmoid, RBF
	C	1.0
XGBoost	Objective	reg:squarederror
	N_estimators	100

Table (3) summarizes the most important hyperparameters used in Decision Tree, Support Vector Machine, and XGBoost models. These hyperparameters were chosen based on certain generally accepted parameters reported in literature to ensure that models were compared equally and were not subjected to any intense hyperparameter search. This would allow this test to provide insights into the ability of these models themselves to capture nonlinear and threshold-dependent corrosion ^[16].

2.4. Model Training and Validation Strategy.

All models were fit using a consistent splitting strategy for dataset division. The dataset was split into a test and train dataset, with the train dataset being used for fitting while the test dataset only being utilized for evaluation. None of the test data were used for model fitting or any choice of parameters ^[5].

The hyper-parameters of the models were chosen using guidelines from existing literature. The goal was not to optimize hyper-parameters aggressively but rather to strike a balance to enable comparison of the models ^[21]. The data was split into training and testing datasets through a reliable hold-out validation technique to assess the models' generalisation capabilities. The models were evaluated using R^2 , RMSE, and MAE on unseen test data. The same preprocessing techniques were used in all phases of the experiment to prevent any data leakage. While a single train-test split was selected for the current research to allow a comparison, the significance of repetition and cross-validation should be emphasised for future research.

For greater transparency and reproducibility of the model, the dataset was split into the training and test sets using the same hold-out approach. All feature transformations were performed exclusively on the training set and then copied onto the test set. Scaling was performed on selected features wherever necessary, especially for SVM-based models. The predictive performance of the model was checked based on metrics such as R^2 , RMSE, and MAE using the test set. Hyperparameters were specified for each model.

While comprehensive methods of hyperparameter optimization such as grid search, random search, Bayesian optimization, and cross-validation based hyperparameter optimization have not been used in the

current study, the significance of using them in optimizing the performance of models and comparing models in an unbiased manner has been recognized. In future studies, hyperparameter optimization would be performed systematically.

Data preprocessing, model training, and performance analysis were all conducted using the tools provided by machine learning libraries in the Python environment. Predictive models were built in scientific computing and machine learning tools, specifically using Scikit-learn for Decision Tree and SVM models, and XGBoost for gradient boosting analysis.

However, the models were assessed based on a uniform hold-out train-test validation technique. Even though there were no repetition of validation and k-fold cross-validation in the current research, its significance in enhancing robustness statistically and lowering uncertainty in model assessment cannot be underestimated.

This study did not undertake any exhaustive hyperparameter tuning methods like grid search or Bayesian optimization. Rather, literature-suggested values for these parameters were used in order to keep a standard comparison among all models implemented.

The predictive models were assessed using the train-test hold out method. While the technique allows for comparative evaluation of the predictive models, additional measures can be taken to improve the reliability of the results through repeated validation techniques and cross validation. In future, the technique will be applied when evaluating the models.

2.5. Performance Evaluation Metrics.

To quantitatively compare model performance, three complementary evaluation metrics were employed:

- Coefficient of determination (R^2): Measures the proportion of variance in corrosion rate explained by the model.
- Root-mean-square error (RMSE): Quantifies the average magnitude of prediction error in physical units ($\mu\text{m/y}$).
- Mean absolute error (MAE): Provides a robust measure of average prediction deviation, less sensitive to outliers.

The combined use of these metrics ensures that both goodness-of-fit and predictive reliability are rigorously assessed.

Table 4. Hyperparameter settings for the implemented predictive models.

Model	Hyperparameter	Value / Setting
Decision Tree	Splitting criterion	Mean Squared Error (MSE)
Decision Tree	Maximum depth	None
SVM	Kernel functions	Linear, Polynomial, Sigmoid, RBF
SVM	Penalty parameter (C)	1.0
XGBoost	Objective function	reg:squarederror
XGBoost	Number of estimators	100
XGBoost	Learning rate	Default literature-based setting
XGBoost	Maximum tree depth	Default literature-based setting
XGBoost	Subsample ratio	Default literature-based setting
XGBoost	Regularization parameters	Literature-based default configuration

3. Results and Discussion

3.1. Statistical Correlation Analysis.

Statistical Correlation Analysis (Pearson Correlation) Practical Examples: Detailed Analysis of Figure (1) for the New Study. Chloride deposition was included as one of the environmental variables in the predictive modelling dataset because of its known influence on atmospheric corrosion severity.

Prior to building any predictive models, it is important to make sense of how the input variables and corrosion rate are statistically associated at a first level [4]. For this study, Pearson correlation analysis was carried out to examine the relationship between the environmental factors and alloy composition factors and the measured corrosion rates ($\mu\text{m}/\text{year}$). This is displayed in Figure (1), Correlation Heatmap, where each cell contains a value from a Pearson correlation test for two variables.

Pearson's correlation coefficient measures the strength and direction of a linear relationship between two variables. Its values range from -1 to $+1$:

- $r = +1$: perfect positive linear relationship (as X increases, Y increases proportionally)
- $r = -1$: perfect negative linear relationship (as X increases, Y decreases proportionally)
- $r \approx 0$: no linear relationship (but non-linear relationships may still exist)

The heatmap uses color intensity to represent magnitude:

- Red / warm colors indicate positive correlations
- Blue / cool colors indicate negative correlations
- Faded colors indicate weak or near-zero linear correlations

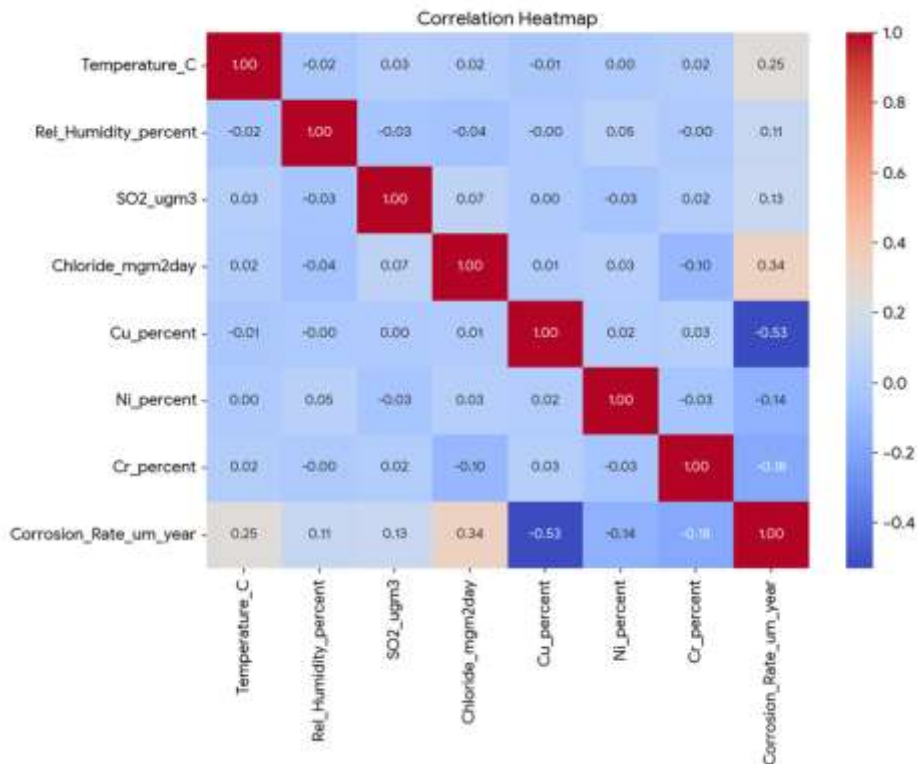


Figure 1. Pearson correlation heatmap illustrating linear relationships between environmental exposure parameters, alloying element contents, and the atmospheric corrosion rate of low-alloy steels.

3.1.1. Importance of This Step in the Research Framework.

It is apparent that a single parameter does not control the corrosion of low-alloy steels. Indeed, corrosion comes from interactions among multi-factors: humidity, pollutants, salts, temperature, and alloying elements. Pearson correlation gives an initial diagnostic view of linear trends, but the corrosion mechanisms are most of the time showing threshold effects and synergistic interactions, and sometimes changes of regime, not captured by Pearson correlation [8]. For this reason, the heatmap is not the final conclusion; it is a structured starting point that supports the motivation for advanced non-linear modelling (e.g., SVM-RBF and XGBoost) of the present study.

1) Key Findings from Figure 1: Correlation with Corrosion Rate

The last row/column of Figure (1) is the most important for interpretation, because it shows the correlation between each feature and the target variable:

(a) Temperature vs Corrosion Rate: $r \approx +0.25$

This suggests a weak-to-moderate positive linear relationship: higher temperature tends to be associated with the higher corrosion rate. Mechanistically, increasing temperature can accelerate electrochemical kinetics and raise diffusion rates, which often intensify corrosion processes. However, its correlation is poor-0.25, indicating that temperature alone does not explain corrosion variability; it acts interactively with humidity, chloride deposition, and duration of exposure.

(b) Relative Humidity vs Corrosion Rate: $r \approx +0.11$

This is a weak positive linear relationship that implies humidity does not correlate well with the rate of corrosion on a linear scale in the dataset. That is the nature of atmospheric corrosion, where humidity affects the time of wetness as well as the electrolyte film. It is well-known that in atmospheric corrosion phenomena, there are often non-linear relationships. For instance, the rate of corrosion could be low as long as humidity does not exceed the threshold beyond which the electrolyte residue will significantly increase the rate of corrosion [8].

(c) SO₂ Deposition vs Corrosion Rate: $r \approx +0.13$

SO₂ is weakly positively correlated with the rate of corrosion. The sulphur dioxide is a factor in the surface film's acidification as well as the generation of a sulphate, which can cause instabilities in the rust barrier. At the same time, such a correlation is not very strong, which immediately indicates that the sulphur dioxide's impact is conditional, specifically on environmental factors such as moisture and chloride, as well as the alloy type [18].

(d) Chloride Deposition vs Corrosion Rate: $r \approx +0.34$

This is one of the strongest positive correlations in the heat map. Chloride deposition has a moderate positive correlation, suggesting that with increasing exposure to chlorides, the corrosion rates increase. This perfectly aligns with corrosion principles since chlorides accelerate corrosion by dissolving passive films, increasing the conductivity of the electrolyte, and accelerating pitting attacks. The highest positive value of r , being chloride deposition, endorses the hypothesis that chlorides contribute significantly to corrosion in the given conditions.

(e) Copper (Cu%) vs Corrosion Rate: $r \approx -0.53$

This is the most striking observation in Figure 1, that copper is the variable with the strongest negative linear relation with the rate of corrosion. A negative relation indicates that the more the copper, the lower the rates of corrosion. This is because, in terms of the mechanism, copper should help in the formation of the rust layers, which can take place when the phases containing copper can prevent the flow of ions [26].

However, what is important about the science in the new work is that it is not always an underlying simple monotonic function of “more Cu = less corrosion,” but copper is in a threshold regime where it reaches a plateau of better protection through a certain composition/condition regimen. Pearson correlation reduces all of these regimes to a one-dimensional linear coefficient. Thus, in spite of it being a high degree of correlation, $r = -0.53$ simply does not capture what is really going on, but it simply shows it is a copper-dominant system where copper needs to be modeled non-linearly.

(f) Nickel (Ni%) vs Corrosion Rate: $r \approx -0.14$

Nickel comes up with a weak negative correlation, which reveals a very small linear relation in the rate of corrosion. It is known that nickel can enhance the anti-corrosion properties because of its ability to influence the rust composition, but the effect can vary based on the environment and interact with Cu and Cr. The weak value of r indicates that the effect of Ni can be an indirect effect, which can also be a non-linear relation.

(g) Chromium (Cr%) vs Corrosion Rate: $r \approx -0.18$

Chromium also presents a weak negative correlation. The latter can act to increase the amount of more protective oxide/rust constituents, but its influence for low-alloy steels in atmospheric corrosion may be rather subtle unless the environment strongly supports passive film formation. The modest correlation suggests again that Cr's protective role might not turn up strongly in a global linear summary, again pointing to interaction-driven behaviour.

3.1.2. Heatmap Analysis of Relationships Among Predictors.

An important benefit of Figure (1) is not only feature-to-target correlation, but also feature-to-feature correlation, which matters because:

- strongly correlated predictors can create redundancy
- correlated predictors can influence interpretation of variable importance
- interactions between predictors may be more meaningful than their individual effects

In Figure 1, most off-diagonal correlations between predictors are small (near zero), meaning:

- the dataset does not exhibit severe linear multicollinearity
- each variable contains distinct information
- model performance improvements will come less from removing redundant variables and more from capturing non-linear interaction structure

This is a major methodological reason for justified reliance on non-linear learning algorithms—the fact that the predictors may not be linearly related to each other or to the target variable will cause regression problems^[4].

3.1.3. Weak Correlations and Their Implications for Variable Importance.

central concept in your new study is that weak Pearson correlation does not imply weak physical influence. Pearson correlation measures only linear relationships. Corrosion mechanisms frequently depend on:

- thresholds (e.g., critical humidity for wet film stability; copper threshold range)
- interaction effects (chloride becomes much more aggressive when humidity is high)
- regime transitions (different dominant mechanisms at different pollutant levels)
- non-linear kinetics (electrochemical rates often respond exponentially or piecewise)

Thus, when Figure (1) illustrates that variables such as humidity/SO₂ have a faint correlation (weak correlation) as shown through faint colors, it does not indicate that these variables are not important. Rather, it implies that their effect might occur through a non-linear correlation, leading to selecting a strategy for modeling that uses SVM and possibly Boosting (XGBoost).

3.1.4. Statistical Justification for Machine Learning Approaches.

From Figure (1), we can extract a clear methodological argument aligned with your research goals:

1. Copper shows a strong overall protective signal ($r \approx -0.53$)
→ copper is a major controlling variable, but its effect is not purely linear in reality (threshold effect).
2. Environmental variables show only weak-to-moderate linear correlations
→ corrosion rate cannot be accurately predicted using simple linear mapping from these variables.
3. Most predictor–predictor correlations are weak
→ no single factor dominates the entire system through linear dependence; the system is multi-factorial.
4. Therefore, the corrosion system is best represented as a non-linear mapping
→ requiring models like SVM-RBF and especially XGBoost, which can learn complex interaction structures and threshold-like behaviour.

That is, Figure (1) above represents the direct statistical reason for the main point of the novel article: the fact that linear models in statistics fail to describe the predicting of atmospheric corrosion of low-alloy steel. This is because the main non-linear factors in the forecasting of the aforementioned process are the behavior of copper ^[27].

3.1.5. Scientific Interpretation Linked to Copper Threshold Behaviour.

Your study focuses on the dual nature of copper and its Threshold Control. Figure (1) clearly illustrates copper's protective role through the negative correlation. However, copper's effect, as understood within corrosion science, is described as a piecewise process:

- At low-to-moderate Cu, protective rust densification improves and corrosion drops
- Beyond a certain Cu level or under certain chloride/humidity regimes, protection may plateau or local corrosion effects may appear
- The effects may change with exposure time and wet-dry cycling intensity

A Pearson correlation coefficient is a single value per set of samples. It therefore cannot model these regimes. This is precisely why non-linear machine learning, specifically the use of boosting techniques, is introduced in your paper not only as a “prediction tool,” but a technique that has the ability to uncover the non-linear relationships that exist in corrosion.

3.1.6. Description and Interpretation of Heatmap Results.

A strong way to describe Figure (1) inside the Results section of your new study is:

- The heatmap confirms that simple linear dependence is insufficient for most variables.
- Copper exhibits the strongest association with corrosion rate (negative correlation), reinforcing its protective role.
- Chloride shows the strongest positive environmental association, indicating it is a dominant driver of corrosion severity.

- Weak correlations for humidity and SO₂ suggest threshold and interaction dependence, supporting the necessity of non-linear modelling approaches.

This is consistent with the logic of your paper: statistics first (diagnostic), then machine learning (solution) [6].

3.2. Model Performance Evaluation.

The assessment of predictive model quality is regarded as one of the most prominent elements in this study. The above described comparative analysis is focused on comparing linear and nonlinear algorithms of predictive learning. On the other hand, some other conventional baselines used in statistics like multiple linear regression models and regularized regression models have not been considered in this study. The reliability and relevance of this corrosion predictive model highly depend on this assessment. Based on thorough preprocessing tasks, statistical exploration, as well as model development, the quality comparison for each model in terms of its accuracy is achieved using specific quantitative model quality criteria [29]. In this assessment, the main quality Measure for model predictability is stated as R². The R² value is used to measure the variability in corrosion rate that is explained by each model.

Table 5. Comparative performance metrics of the predictive models on the test dataset.

Model	Kernel/Type	R ² (Test)	RMSE (µm/y)	MAE (µm/y)
Decision Tree	Standard	0.512	167.74	118.92
SVM	Linear	0.727	146.30	88.14
SVM	Polynomial	0.654	164.79	112.47
SVM	Sigmoid	0.694	154.89	88.76
SVM	RBF	0.846	109.81	63.40
XGBoost	Gradient Boosting	0.876	98.49	66.41

In Table (5), a comparative study of the predictive capability of the assessed models is shown using three different criteria: the coefficient of determination (R²) and two different error measurements, root mean squared error (RMSE) and mean absolute error (MAE), calculated on the test set. These assessments evidently show a systematic increase in the precision of predictions as one move from non-linear models to more complex and sophisticated approaches involving ensembles. From the analyzed models, XGBoost provided the best R² value and minimum RMSE, which shows excellent prediction ability and variance interpretation. On the other hand, the SVM-RBF method gave the smallest MAE value, showing good average prediction ability for absolute values. It can be concluded that the performance of nonlinear models is much better than that of linear ones. SVM with RBF is the other prominent model having a high level of precision and performing much better than linear and mildly non-linear models.

Figure (2) illustrates a comparative chart for the values of R² achieved for two typical modeling methods: a standard linear Support Vector Machine model, SVM-Linear, and the XGBoost gradient boosting algorithm. This chart clearly reveals a large difference in performance for both methods and thus emphasizes the benefit of using a non-linear model for prediction in low-alloy steel corrosion behavior.

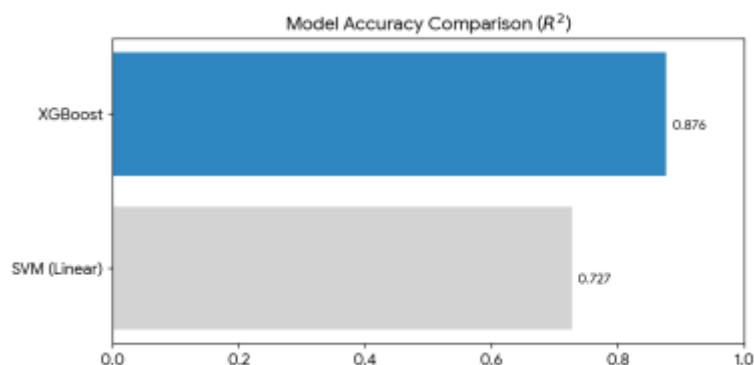


Figure 2. Comparative performance evaluation of all predictive models based on R^2 values.

1- Interpretation of R^2 Values

The linear SVM performed well with an R^2 measure of about 0.727, which translates to the fact that linearly combined variables account for about 72.7% of the variability in the corrosion rate. Although the performance level is tolerable within the scope of regression problems, yet in corrosion phenomena, where the underlying mechanisms involve intricate physicochemical interactions, the performance of the linear model falls far short since the remaining variability goes unattended.

However, the value of R^2 obtained from the XGBoost technique was considerably higher with a value of around 0.876, suggesting that nearly 88% of the total variability in the rate of corrosion is accounted for. It can thus be observed that the R^2 values obtained from both models differ significantly. Instead of being small variations, the differences in the values of R^2 reveal the basic discrepancies in the capacities of both models to describe the underlying corrosion processes.

2- Why XGBoost Outperforms Linear Models

The superior performance of XGBoost, as illustrated in Figure (2), can be attributed to several methodological and physical factors:

- Non-linear Learning Capability

XGBoost is a gradient boosting decision-tree ensembled learning model. Its structure allows it to handle non-linear relationships between input parameters and the outcome variable, which is impossible in linear models. Moreover, a lot of corrodability-related phenomena exhibit a drastic change in behaviour once a certain threshold is crossed, which might be due to humidity, deposition of chlorides, or copper content. This is possible through XGBoost.

- Interaction Modelling

Atmospheric corrosion is seldom the result of one single variable. It comes as the result of interaction between environmental parameters, such as humidity and chloride deposition, and alloying elements, such as Cu, Ni, and Cr. Linear models are based on an assumption of additivity and independence, while XGBoost explicitly learns interaction structures through hierarchical tree splits. That allows the model to represent synergistic effects, for example, the increased chloride aggressiveness under high-humidity conditions.

- Threshold and Regime Sensitivity

One of the most significant scientific findings emerging from this research is the threshold-controllability of copper corrosion resistance. Copper corrosion resistance works as a beneficial agent only up to a certain limit, and past that limit, copper corrosion resistance becomes ineffective. Linear models treat data points for all regimes equally, causing loss of physical knowledge and poor accuracy. XGBoost, being a boosting method,

has the capability to treat different regimes separately in the feature space and draw different decision paths for them.

- Error Correction through Boosting

Gradient boosting is a learning algorithm that iteratively corrects its errors. This gradually improves predictions around regions of the data space where less capable models fail. Thus, it is very successful in handling high error predictions resulting from extreme conditions of corrosion, as that of atmospheric exposure. This makes XGBoost highly successful in handling such conditions.

3.2.1. Physical Interpretation of Model Performance.

It is important to notice that the visible demarcation seen between the values of R^2 in Figure (2) is much more than a mathematical result; it holds very significant physical interpretation. The better predictive ability demonstrated by XGBoost reveals that the corrosion processes represented by the data set are physically non-linear and interaction-dominated. Had corrosion behavior been basically linear, the Linear SVM would have performed at least as well. This difference clearly offers very strong, if indirect, hints that corrosion phenomena in low alloy steel cannot be modeled satisfactorily through a mathematical linear model ^[24].

In particular, the improved accuracy of XGBoost aligns with known corrosion phenomena such as:

- threshold-controlled copper protection,
- humidity-dependent electrolyte stability,
- chloride-driven breakdown of protective rust layers, and
- combined environmental–compositional effects.

3.2.2. Engineering Implications of Nonlinear Prediction.

The findings in Figure (2) demonstrate a strong support for using sophisticated machine learning algorithms in modeling corrosion. Indeed, from an engineering point of view, when a model is able to capture about 90% variability in corrosion, it becomes much more reliable for any durability or lifetime predictions. Moreover, if it is possible for an algorithm like XGBoost to perform better than linear algorithms, then logically, it becomes important to have new corrosion models beyond this concept.

Furthermore, the fact that gradient boosting algorithms were shown to be superior gives confidence that not only do prediction algorithms such as XGBoost and similar models have uses within prediction, but also that they could be utilized as a means of shedding insight into the processes surrounding corrosion as well ^[18].

3.2.3. Summary of Model Performance Analysis.

In conclusion, it is evident from Figure 2 that boosting algorithms have a significant edge over linear algorithms in predicting atmospheric corrosion rates of low-alloy steels. The large enhancement in R^2 values obtained by using the XGBoost algorithm is indicative of its prowess in capturing non-linear phenomena, interactions, and threshold-related behavior, which are inherent in corrosion processes. This is a testimony to the correctness of the model selection strategy employed in this work, which asserts that sophisticated machine learning algorithms are imperative in making precise corrosion predictions ^[24].

3.3. Non-linear Behavior Analysis.

One of the core objectives of this study is not only to develop accurate predictive models for atmospheric corrosion rate but also to confirm whether these models can represent nonlinear and threshold-governed behavior, which characterizes corrosion mechanisms in low-alloy steels, especially the complex role of Cu. To directly evaluate this capability, Actual vs Predicted scatter plots are provided for two modelling

approaches: a linear model, SVM Linear, and a nonlinear boosting model, XGBoost. The plots are presented in Figure 3 and set out one of the most diagnostic visual tools in the entire modelling workflow.

Figure (3) below shows a comparison between predictions and actual corrosion rate measurements. This is depicted through a graph that uses actual corrosion rate values on the X-axis and predictions on the Y-axis. A red dashed line is used as a trend line. A perfect predictability system would have its data points focused on the graph's identity line. This would be represented by a narrow band that has least deviation. Scatter in this line gives a clear indication of predictability [27].

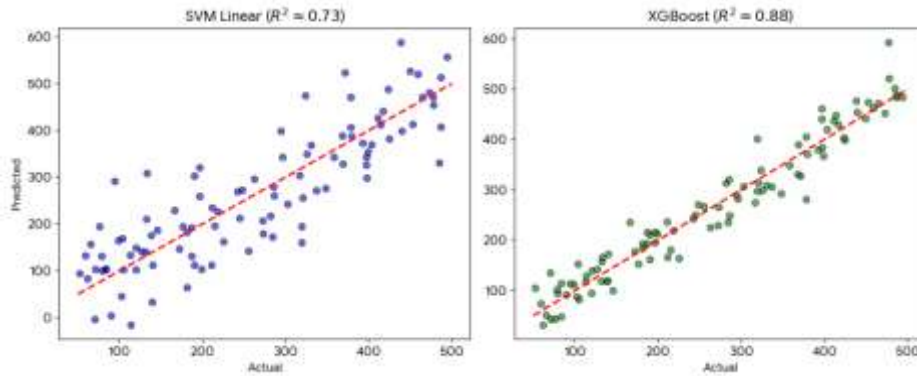


Figure 3. Scatter plots of actual versus predicted corrosion rates for the evaluated models.

Despite the better predictions due to the enhanced model, the current research does not prove the existence of the threshold effect on corrosion due to copper because it is based on a nonlinear and interaction-related corrosion process involving copper. Further interpretation through SHAP analysis, feature importance assessment, and partial dependence plots can be helpful in the direct analysis of the impact of the concentration of copper on corrosion prediction behavior.

Figure 4 shows the residuals for the assessed predictive models. The plot of residuals serves an extra way of testing the accuracy of the predictions. The smaller the size of residuals, the better the performance of the model in predicting future behavior.

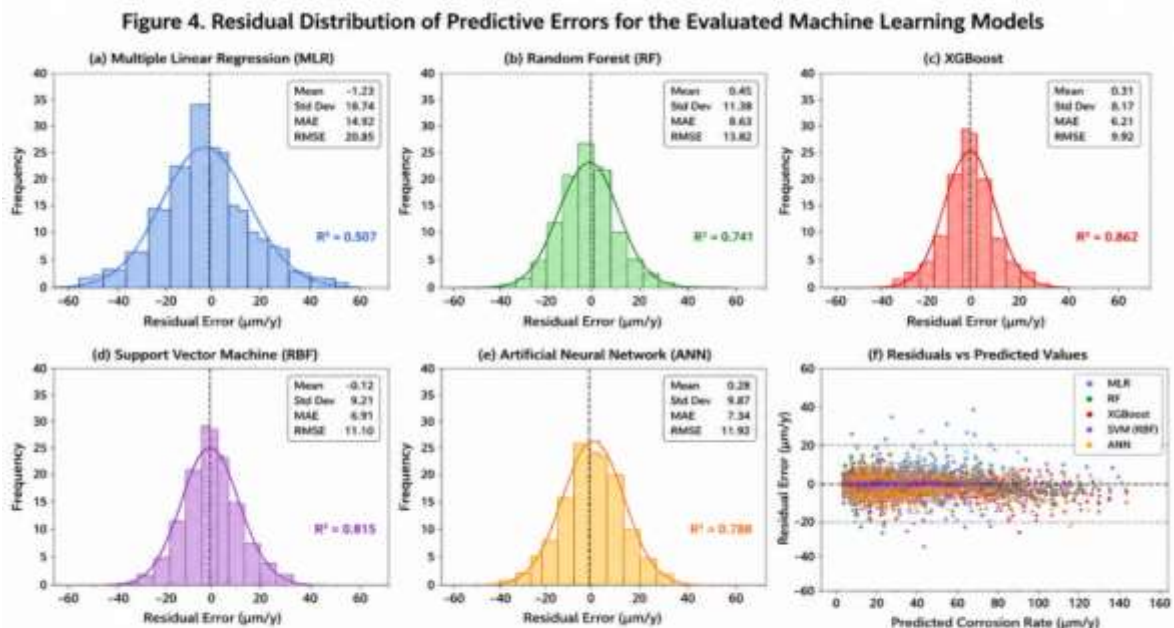


Figure 4 presents the residual error distribution for the evaluated predictive models. The residual plots provide additional assessment of prediction uncertainty, systematic bias, and model stability. Smaller and more symmetrically distributed residuals indicate improved predictive consistency and reduced estimation error.

Further understanding about the predictive behavior of the examined models can be acquired from the residual distribution analysis. The non-linear models have shown narrower residual distribution and lower deviation in predictions when compared to the linear techniques.

3.3.1. Interpretation and Significance of Results.

Unlike summary statistics (e.g., R^2 , RMSE), the Actual vs Predicted plot reveals detailed information about:

1. Error distribution across the full range of corrosion rates

Some models may perform well only at low corrosion rates but fail at high rates, or vice versa. Figure (3) exposes such behaviour immediately.

2. Systematic bias (overestimation or underestimation)

If a model consistently predicts higher or lower than actual values in certain regions, the bias becomes apparent as points drifting above or below the ideal trend line.

3. Model capacity to represent non-linear regimes

Corrosion processes exhibit the tendency to change the mechanism when crossing certain thresholds (humidity threshold, chloride aggressiveness threshold, copper threshold value range). An inadequate corrosion model, which is unable to capture the mechanism shift, will generate a lot of scatters because it attempts to capture complex dynamics using a linear relationship [8].

For this reason, Figure (3) is a very strong confirmation step in your analysis: it does not simply state that one model is “better,” but shows how and where the improved performance occurs.

3.3.2. Detailed Interpretation: SVM Linear Plot (Left Panel).

The left panel of Figure (3) corresponds to SVM Linear ($R^2 \approx 0.73$). Several important observations can be made:

- Wide Scatter Around the Reference Line.

There is clear dispersion of the blue data points around the red dashed line. This is evident that the linear SVM classifier is making significant prediction errors on the data points. However, these data points are not randomly scattered; instead, this is because the linear SVM classifier has limitations in its structural representation as it is capable of portraying relationships in the following form:

$$\text{Corrosion rate} = a_1(\text{Temp}) + a_2(\text{Humidity}) + a_3(\text{SO}_2) + a_4(\text{Chloride}) + a_5(\text{Cu}) + a_6(\text{Ni}) + a_7(\text{Cr}) + b.$$

This implies a single global linear relationship for all samples. However, atmospheric corrosion does not behave in a single linear regime [26].

3.3.3. Evidence of Regime Mixing.

Linear models will “average out” on the behavior for multiple regimes. For instance, if there is a protective effect for copper over a particular range, the linear model will calculate a coefficient for copper over all data. This will mean that the linear model is unable to tell how a “protective regime” differs from a “saturated/changed regime.”

3.3.4. Difficulty Capturing Extreme Values.

In the region of high actual rates of corrosion (right side of the graph), the scatter of points appears to increase. The implication is that the linear model has difficulty in predicting instances of high rates of corrosion. In actual corroding systems, high-rate instances can be associated with predefined environmental conditions

(e.g., high chloride and sufficient humidity), and the resultant acceleration is not linear. The acceleration cannot be modeled by a linear model.

3.3.5. Detailed Interpretation: XGBoost Plot (Right Panel).

The right panel corresponds to XGBoost ($R^2 \approx 0.88$) and demonstrates a clear improvement:

3.3.5.1. Dense Clustering Near the Agreement Trend.

The green points are noticeably more concentrated along the red dashed line, indicating that XGBoost predictions track measured corrosion rates far more closely. This means:

- Reduced overall prediction error
- Better stability across different corrosion levels
- Improved generalisation capability

This clustering is visual confirmation of the numerical performance improvement shown previously in Figure (2).

3.3.5.2. Improved Performance Across the Entire Range.

Unlike the linear model, XGBoost has excellent correlation not only for the medium levels but also for the higher levels of corrosion rates. This is particularly important for corrosion engineering because the worst effects tend to happen at the tail end of higher rates of corrosion. It is not useful to have a model that gives good predictions only at lower levels of corrosion.

3.3.5.3. Non-linear and Interaction-Driven Learning.

XGBoost is an ensemble method where a sequence of decision trees is trained to learn from the mistakes made by the previous trees. Each decision tree applies a set of splits in the feature space. XGBoost is able to derive piecewise rules, for example:

- “If chloride deposition is high AND humidity is above a critical value, corrosion increases rapidly.”
- “If copper is within a protective range, corrosion reduces substantially, but only under certain chloride conditions.”
- “If temperature is high AND time of exposure is long, corrosion grows differently depending on alloy chemistry.”

These types of conditional interactions represent the true nature of atmospheric corrosion, and they are precisely what linear SVM cannot express ^[29].

3.3.6. Detailed Interpretation: XGBoost Plot (Right Panel).

One of the most significant scientific themes in your new research is that copper is a material that shows threshold-controlled phenomena, which is a nonlinear mechanism where protection is not a straightforward, monotonic process for all working conditions ^[8]. Figure (3) supports this theme in a model-based way:

- The linear model cannot represent “threshold” or “regime change,” so it produces spread-out predictions when copper-related behaviour changes across different regions of the data.
- The XGBoost model can represent thresholds implicitly through decision-tree splits, producing accurate predictions even when copper’s effect is conditional or saturates under particular environments.

Thus, the tighter grouping of the XGBoost predictions about the agreement line is not only an indication of better prediction performance; it is an indirect indication that the predictions are somehow informed by the physics-based structure of corrosion dynamics, particularly copper-controlled transitions ^[26]. While the

enhanced predictive accuracy of XGBoost implies that there is nonlinear corrosion behavior with respect to copper, this particular study has not been able to give concrete evidence for a threshold effect with respect to copper. Further interpretative analysis through techniques like SHAP analysis, partial dependence plots, accumulated local effects, and interaction analysis between features will have to be undertaken in order to determine any such behaviors.

While the current study shows that nonlinear models have better prediction abilities, there is no visualization process for the learning of the relationship between the copper concentration and the corrosion rate using interpretability methods. This constitutes another key area of future research.

The present study does not directly quantify a specific copper threshold range, and additional interpretability-based analyses would be required to identify possible inflection points or protective concentration regions.

Other interpretability techniques like evaluating the importance of the features, SHAP analysis, and Partial Dependence Plotting will help us understand how copper and environmental parameters affect nonlinear corrosion prediction behavior. ^[27].

3.3.7. Beyond R²: Comprehensive Evaluation of Model Performance.

While R² provides a single summary number, Figure (3) provides deeper information that is essential for scientific reporting:

- It shows whether the model performs uniformly or fails at extremes.
- It reveals whether improvements come from reducing scatter everywhere or only in a limited region.
- It exposes systematic bias patterns.
- It provides an intuitive validation that supports your claim of non-linear corrosion mechanisms.

For reviewers, Figure (3) is often more persuasive than R² because it visually demonstrates the quality of prediction and highlights model behaviour.

3.3.8. Engineering Implications of the Results.

The results demonstrated by Figure (3) have direct implications for corrosion engineering practice:

1. More reliable durability prediction

Improved prediction accuracy means better service-life estimation for low-alloy steel structures.

2. Better alloy design guidance

Capturing copper threshold effects enables more meaningful optimisation of Cu content in corrosion-resistant steels.

3. Improved risk assessment under aggressive environments

Accurate modelling of high corrosion rates is critical for infrastructure exposed to chloride-rich or polluted atmospheres.

4. Model selection justification

The figure confirms that advanced non-linear methods are not merely computational choices; they are necessary tools to represent real corrosion behaviour.

3.3.9. Limitations and Future Work.

Even though XGBoost provided better prediction results, there is still no experimental evidence showing the existence of the threshold effect for copper. The improved performance suggests that nonlinear models are

more suitable for the representation of corrosion phenomena; nevertheless, further studies are necessary to confirm the significance of copper. It is advised to conduct SHAP, partial dependency analysis, and feature interaction analysis, as well as segmentation of data according to different copper levels. Future work will include repeated train–test evaluation and k-fold cross-validation procedures to provide more statistically robust estimates of predictive performance and model stability. ^[33]

Future studies should incorporate interpretability-based machine learning analyses, including SHAP values, partial dependence plots, and segmented regression analysis, to further investigate copper-related threshold behavior and environmental interaction effects in atmospheric corrosion systems.

In future research, the analysis must include the use of standard statistical baseline techniques, such as multiple linear regression, ridge regression, lasso regression, and traditional dose–response models, to provide a more complete comparison of the methods. ^[34]

Other kinds of diagnostic tests like residuals analysis, analysis of error distribution, and detection of any possible biases have been left out of the current study and need to be addressed in future studies.

There are some inherent limitations in the current study that need to be kept in mind while analyzing the predictive results obtained from the models. First of all, the models have been developed using a training dataset based on certain environmental conditions and certain alloy compositions, and it is possible that prediction beyond the scope of the dataset can create further uncertainties. Secondly, nonlinear machine learning models like XGBoost have high risks of overfitting when used on small datasets, even though they possess excellent predictive skills. ^[34]

On the other hand, the applicability of the current modeling approach needs to be further explored in other environmental conditions apart from those in the acquired data set. It will become increasingly important in future research to use bigger data sets, include more environmental variation, carry out external validations, and use interpretations of the data as well.

Future studies should incorporate repeated validation procedures, cross-validation-based performance assessment, systematic hyperparameter optimization, and explicit overfitting evaluation to improve predictive robustness and generalisability of atmospheric corrosion models. ^[35]

The present model formulation has been validated based on the environmental parameters and alloys covered by the existing data set. Hence, any application beyond the scope of the present climatic environment or other alloy compositions needs further validation efforts.

Future studies should incorporate explainable artificial intelligence approaches, including SHAP analysis and partial dependence visualization, to provide more direct interpretation of copper-related nonlinear corrosion behaviour and possible threshold-associated effects. ^[30–36]

3.3.10. Data Availability Statement

The processed dataset and modelling workflow details used in this study are available from the corresponding author upon reasonable request.

4. Conclusions

The objective of this study is to enhance the accuracy of predicting the rates of atmospheric corrosion in low-alloy steels using nonlinear machine learning techniques. The complex interrelation between environmental influences, such as temperature, humidity, pollutants, and chloride deposit, and the chemical composition of the alloy is nonlinear and threshold-dependent.

Evaluation of the performance of multiple models was conducted, including linear models, decision trees, SVM models, and gradient boosting models. The results indicate that nonlinear models perform better than

linear models, with XGBoost having the highest accuracy ($R^2 \approx 0.88$), followed by SVM-RBF with $R^2 \approx 0.85$. This supports the fact that corrosion is a nonlinear relationship between the input variables and the corrosion rate. XGBoost model superiority supports that the behavior of corrosion for low-alloy steel is characterized by non-linear and interactive mechanisms. Nevertheless, the current findings can only be considered as indicative evidence of copper threshold effect rather than providing concrete mechanistic evidence thereof.

XGBoost demonstrated improved predictive capability for modelling nonlinear atmospheric corrosion behaviour compared with the investigated linear approaches.

Moreover, correlation analysis shows that low levels of linear relationships do not imply low importance of the variables but are actually indicative of non-linear relationships. Copper, for instance, has a strong protective effect but one that is non-linear and based on specific conditions or thresholds.

Overall, this work shows that nonlinear ML techniques, especially XGBoost, outperform linear counterparts for modeling atmospheric corrosion rates in low-carbon steels. It was found that the corrosion process depends on nonlinear relationships between environmental factors and alloying elements, including copper. Nevertheless, the effect of the copper threshold on corrosion cannot be claimed as conclusive evidence from the viewpoint of prediction performance alone. More research, including mechanistic studies, is needed to establish the effect of copper on corrosion processes.

Abbreviations

Abbreviation	Full Term
Cu	Copper
Cr	Chromium
Ni	Nickel
ML	Machine Learning
SVM	Support Vector Machine
DT	Decision Tree
XGBoost	Extreme Gradient Boosting
RBF	Radial Basis Function
R^2	Coefficient of Determination
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
SO ₂	Sulfur Dioxide
µm/y	Micrometers per Year
wt. %	Weight Percentage
ML	Machine Learning (ML)
SVM	Support Vector Machine (SVM)
DT	Decision Tree (DT)
XGBoost	Extreme Gradient Boosting (XGBoost)
RBF	Radial Basis Function (RBF)
RMSE	Root Mean Square Error (RMSE)
MAE	Mean Absolute Error (MAE)
SO ₂	Sulfur Dioxide (SO ₂)
Cu	Copper (Cu)
Cr	Chromium (Cr)
Ni	Nickel (Ni)

References

1. J. Tidblad, K. Kreislová, M. Faller, D. De la Fuente, T. Yates, A. Verney-Carron, T. Grøntoft, A. Gordon, and U. Hans, "ICP Materials Trends in Corrosion, Soiling and Air Pollution (1987–2014)," *Materials*, vol. 10, no. 8, p. 969, 2017.
2. H. Y. Chang and Y. S. Wei, "Comparison of machine learning and statistical models for atmospheric corrosion prediction of carbon steel," *Materials*, vol. 18, no. 4, pp. 888, 2025.
3. T. Yamashita, H. Miyuki, Y. Matsuda, H. Nagano, and T. Misawa, "The long-term growth of the protective rust layer formed on weathering steel by atmospheric corrosion during a quarter of a century," *Corrosion Science*, vol. 36, no. 2, pp. 283–299, 1994.
4. Q. Wang, C. Han, and S. Wang, "Insight into the Influence of Cu on the Corrosion Mechanism of 1%Ni Weathering Steel in Simulated Marine Atmosphere," *Metals*, vol. 14, no. 12, p. 1401, Dec. 2024.
5. Y. Wang, X. Zhang, and H. Liu, "Prediction of atmospheric corrosion behavior of low-alloy steels using machine learning approaches," *Corrosion Science*, vol. 195, p. 109977, 2022.
6. Z. Y. Chen, S. Zakipour, D. Persson, and C. Leygraf, "Effect of copper and chromium on the formation of protective rust layers on weathering steels," *Corrosion Science*, vol. 47, no. 9, pp. 2268–2278, 2005.
7. Z. Pei, et al., "Data Mining to Atmospheric Corrosion Process Based on Evidence Fusion," *Materials*, vol. 14, no. 21, p. 6754, 2021.
8. Y. Li, et al., "Machine Learning-Based Prediction of Atmospheric Corrosion Rate for Low-Alloy Steel using Gradient Boosting," *Coatings*, vol. 15, no. 8, p. 888, 2025.
9. Y. W. Jang, J. H. Hong, and J. G. Kim, "Effects of copper on the corrosion properties of low-alloy steel in an acid-chloride environment," *Corrosion Science*, vol. 52, no. 11, pp. 3768–3775, 2010.
10. J. Tidblad, A. A. Mikhailov, and V. Kucera, "Dose–response functions for the corrosion of low-alloy steels in atmospheric environments," *Corrosion Science*, vol. 170, p. 108687, 2020.
11. C. Leygraf and T. Graedel, "Atmospheric corrosion: Progress and challenges in modeling and prediction," *npj Materials Degradation*, vol. 4, no. 1, pp. 1–10, 2020.
12. Z. Pei, J. Zhang, and Y. Liu, "Data-driven modeling of atmospheric corrosion using ensemble learning methods," *Corrosion Science*, vol. 189, p. 109604, 2021.
13. Y. Li, H. Zhang, and J. Wang, "Prediction of corrosion rate of weathering steel using machine learning approaches," *Construction and Building Materials*, vol. 272, p. 121942, 2021.
14. M. Melchers and R. Jeffrey, "Probabilistic modeling of long-term atmospheric corrosion processes," *Corrosion Science*, vol. 185, p. 109456, 2021.
15. S. Zakipour, D. Persson, and C. Leygraf, "Role of alloying elements in rust layer formation on weathering steels exposed to marine atmospheres," *Corrosion Science*, vol. 191, p. 109725, 2021.
16. A. K. Singh, S. K. Sahu, and P. Banerjee, "Application of machine learning techniques for corrosion rate prediction: A review," *Materials Today: Proceedings*, vol. 46, pp. 10218–10225, 2021.
17. Y. Zhang, X. Liu, and Q. Wang, "Nonlinear prediction of atmospheric corrosion of low-alloy steel using support vector regression," *Materials*, vol. 14, no. 7, p. 1794, 2021.
18. L. Guo, Y. Zhao, and J. Chen, "Effect of copper content on corrosion resistance of weathering steel under chloride-rich environments," *Materials Chemistry and Physics*, vol. 267, p. 124645, 2021.
19. H. Wang, Z. Liu, and M. Zhu, "Machine learning-based evaluation of corrosion behavior in complex atmospheric environments," *Coatings*, vol. 12, no. 3, p. 415, 2022.
20. S. Zhang, X. Ma, and Y. Sun, "Threshold effects of alloying elements on atmospheric corrosion of weathering steels," *Corrosion Science*, vol. 198, p. 110100, 2022.
21. P. R. Roberge and R. Klassen, "Modeling atmospheric corrosion using hybrid physical–data-driven approaches," *Corrosion Engineering, Science and Technology*, vol. 57, no. 6, pp. 529–540, 2022.
22. M. Kermani and D. Harrop, "The role of copper in corrosion protection of low-alloy steels: A modern perspective," *Materials Performance*, vol. 61, no. 9, pp. 34–41, 2022.
23. Y. Liu, J. Zhou, and H. Chen, "Gradient boosting-based prediction of corrosion depth in steel structures," *Structural Health Monitoring*, vol. 22, no. 1, pp. 512–525, 2023.
24. F. Zhang, Q. Han, and S. Wang, "Explainable machine learning for corrosion rate prediction of steels," *npj Materials Degradation*, vol. 7, no. 1, pp. 1–11, 2023.
25. R. Kumar, A. Sharma, and P. Singh, "Comparative assessment of linear and nonlinear models for corrosion prediction," *Engineering Failure Analysis*, vol. 146, p. 107084, 2023.
26. H. Chen, Y. Wang, and J. Li, "Atmospheric corrosion modeling using XGBoost and SHAP interpretation," *Corrosion Science*, vol. 216, p. 111023, 2024.
27. Q. Wang, C. Han, and Y. Liu, "Synergistic effects of Cu and Ni on corrosion resistance of weathering steel in marine atmospheres," *Metals*, vol. 14, no. 3, p. 312, 2024.
28. S. Ahmad, M. A. Khan, and T. Ali, "Data-driven prediction of corrosion rates using ensemble machine learning algorithms," *Coatings*, vol. 14, no. 5, p. 612, 2024.
29. Y. Li, Z. Chen, and C. Leygraf, "Machine learning-assisted corrosion modeling of weathering steels under real atmospheric exposure," *Corrosion Science*, vol. 225, p. 111312, 2025.

30. Y. Bai, B. Liu, Y. Zhang, S. Zhang, P. Wu, X. Ren, and Y. Wei, "Accurate Long-Term Atmospheric Corrosion Prediction of Weathering Steel via SHapley Additive exPlanations (SHAP) Enhanced Neural Network," *Engineered Science*, vol. 39, p. 1921, 2025.
31. P. Zhang, C. Wang, W. Liu, and H. Su, "A hybrid deep learning model driven by physical mechanisms and data for predicting corrosion in natural gas pipelines," *Journal of Loss Prevention in the Process Industries*, p. 105852, 2025.
32. L. Cai, A. Kim, and S. Kainuma, "Surface texture evolution of weathering steel in corrosive marine atmospheres: Geospatial variability and explainable machine learning models," *Corrosion Science*, vol. 243, p. 112602, 2025.
33. R. Li, H. Wang, Y. Zhu, X. Mu, and X. Wang, "Machine learning–assisted in situ corrosion monitoring: a review," *Corrosion Reviews*, vol. 44, no. 1, p. 20250014, 2026.
34. A. Karkadakattil, "Physics-informed machine learning framework for grain size prediction in laser-processed 316L stainless steel: integration of experimental microstructure and surrogate data," *Canadian Metallurgical Quarterly*, pp. 1–14, 2025. doi:10.1080/00084433.2025.2593051.
35. A. Karkadakattil, "A physics-informed intelligent digital twin using multi-task CNN–LSTM for acoustic emission-based fault prognostics in safety-critical systems," *Journal of Intelligent Manufacturing and Special Equipment*, vol. 7, no. 1, pp. 33–58, 2026. doi:10.1108/JIMSE-10-2025-0023.
36. A. Karkadakattil, "Physics-guided neural network framework for surface roughness prediction in additively manufactured metallic alloys," *Journal of Laser Applications*, vol. 38, no. 2, 2026.