

## ARTICLE OPEN



# An intelligent framework for forecasting and investigating corrosion in marine conditions using time sensor data

Mohamed El Amine Ben Seghier<sup>1,2</sup><sup>✉</sup>, Ole Øystein Knudsen<sup>3</sup>, Anders Werner Bredvei Skilbred<sup>4</sup> and Daniel Höche<sup>5</sup>

Corrosion of marine steel structures can be regarded as a time-dependent process that might result in critical strength loss and, eventually, failures. The availability of reliable forecasting models for corrosion would be useful, enabling intelligent maintenance program management, and increasing marine structure safety, while lowering in-service expenses. In this study, an intelligent framework based on a data-driven model is developed that employs a group method of data handling (GMDH) type neural network to forecast free atmospheric corrosion as time-series problem. Therefore, data from sensor data with a 30-min interval over a 110 day period that includes free atmospheric corrosion as well as environmental factors are used. In addition, the Shapley additive explanations (SHAP) technique is used to investigate the impact of the surrounding environmental factors on free atmospheric corrosion. For the performance evaluation of the proposed intelligent framework, selected comparative metrics are used. Findings demonstrate the high accuracy and efficiency of the time series data-driven framework for tackling free atmospheric corrosion progression in marine environments.

*npj Materials Degradation* (2023)7:91 | <https://doi.org/10.1038/s41529-023-00404-y>

## INTRODUCTION

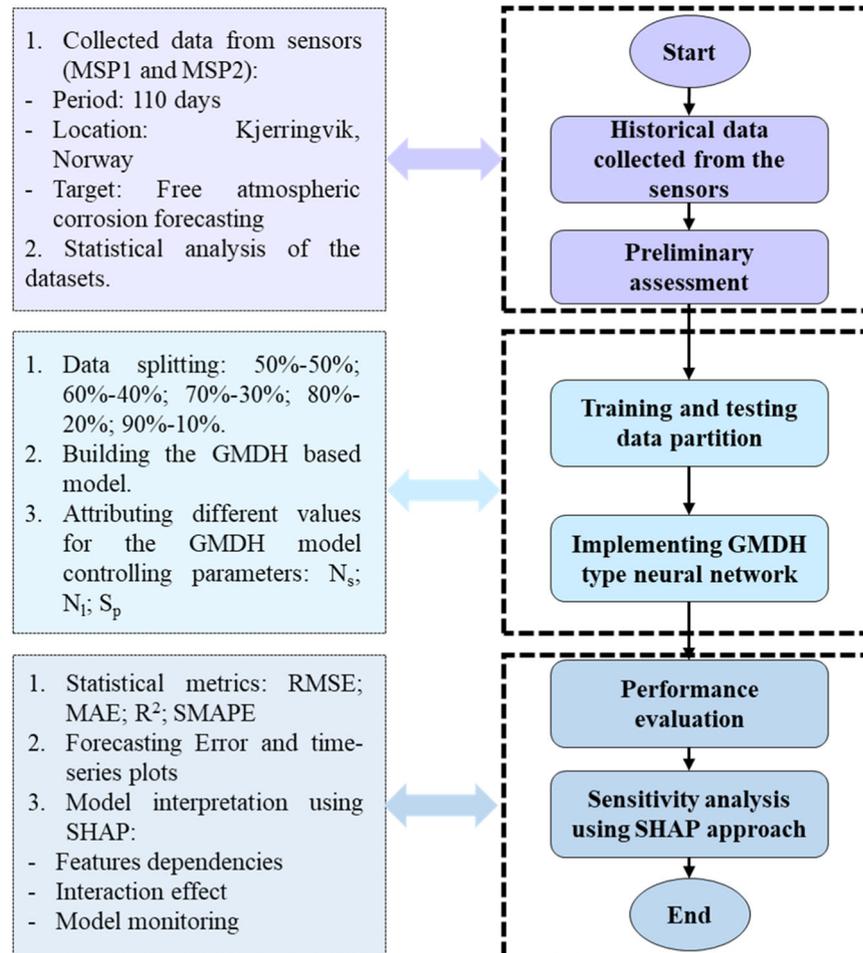
Steel structures located in marine and coastal environments are well-known to suffer from severe corrosion due to the aggressive environment<sup>1,2</sup>. Even though these structures are typically protected with heavy-duty paint coatings, the harsh corrosive environment reduces their durability and lifetime<sup>3–5</sup>. According to statistics, 6.2% of the US GDP is allocated to replace industrial damages caused by corrosion<sup>6</sup>, compared to 3.34% for China<sup>7</sup>. Atmospheric corrosion is a common type of degradation in marine steel structures<sup>8</sup>. A number of factors such as chloride deposition rate, average daily sulfur oxide deposition rate, temperature, time of wetness, and relative humidity will determine the corrosion rate<sup>9–11</sup>. For marine and coastal structures accurate prediction of corrosion behavior will provide a better understanding of the structure's health<sup>12</sup>. The corrosivity concept is yet a complicated challenge to address<sup>13</sup>, but will help to propose maintenance plans to avoid failures and monitor the structure safety<sup>14</sup>. The development of a data-driven forecasting system for steel atmospheric corrosion in marine and coastal environments can result in significant economic benefits, add to that the structure lifetime extension, and the avoidance of catastrophic failures<sup>15</sup>.

Over the last decade, heuristic corrosion modeling, in particular pitting corrosion depth and growth, has been viewed as a prediction-fitting problem. Mathematical formulations in use are based on experimental test results, taking into account a variety of environmental factors, such as empirical, statistical or stochastic-based models<sup>16–21</sup>. Each type among these models demonstrated a number of drawbacks and limitations, such as low adaptation for a wide range of applications and large databases, as well as low dealing with highly complex problems in the case of the statistical, empirical, and stochastic-based analytical models. On the other side, corrosion data are frequently insufficient, noisy, heterogeneous, and large in volume. Furthermore, the marine and coastal environments are complex and changeable resulting in a highly

nonlinear system that is difficult to approach using traditional statistical or analytical methods. Due to their extensiveness and adaptability to various nonlinear and complex data, machine learning models such as artificial neural networks (ANN), support vector regression (SVR), extreme learning machine (ELM), and random forest (RF), among others, have been rapidly developed and employed for the prediction of corrosion process (i.e. rate, morphology and distribution)<sup>1,22–26</sup>. The previous references demonstrated that all machine learning-based models outperform classical models in terms of effectiveness. In addition, it is believed that the corrosion community has a potential benefit from advances in artificial intelligence (AI) technologies, considering the large amount of available corrosion data<sup>1</sup>. Furthermore, the applied machine learning techniques until now have used a surrogate function to predict the target output (i.e. corrosion characteristics such as rate, morphology, distribution or others). However, because corrosion is a time-dependent process, and with the aid of sensors, treating the problem as a time-series can be more effective in forecasting corrosion. To fill the research gap, proposing the concept of machine learning to forecast corrosion as a time series problem will provide new insights in this field.

A two-stage intelligent framework for forecasting free atmospheric corrosion in steel constructions located in marine environments is proposed in this study. The group method data handling (GMDH) type neural network is introduced in the first stage of the framework to solve the corrosion forecasting system as a time series problem, where the employed database is obtained from real-time sensors with 30 min measurement intervals over a 110 day period. Shapley additive explanations (SHAP) are proposed in the second stage to investigate the impact of different surrounding characteristics on free atmospheric corrosion. To demonstrate the forecasting capability, the performance of the proposed models was investigated using various scenarios and analysis metrics.

<sup>1</sup>Institute of Surface Science, Helmholtz-Zentrum Hereon, Max-Planck-Straße 1, 21502 Geesthacht, Germany. <sup>2</sup>Department of Built Environment, Oslo Metropolitan University (OsloMet), Oslo, Norway. <sup>3</sup>SINTEF Industry, Richard Birkelandsvei 2B, 7465 Trondheim, Norway. <sup>4</sup>Jotun AS, Hystadveien 167, 3202 Sandefjord, Norway. <sup>5</sup>Institute of Surface Science, Helmholtz-Zentrum Hereon, Max-Planck-Straße 1, 21502 Geesthacht, Germany. ✉email: moseg7662@oslomet.no



**Fig. 1 Study methodology.** Structure of the proposed free atmospheric corrosion forecasting framework.

## RESULTS AND DISCUSSION

Figure 1 depicts the overall structure of the proposed free corrosion forecasting system. Based on the time-series data of free atmospheric corrosion obtained from the sensors described in the method section (i.e. section “Experimental setup and data analysis”) and detailed in section “Data collection”, the GMDH-type neural network (i.e. see section “Experimental setup and data analysis”) divides the original free corrosion datasets into training (i.e. used to build the model) and testing (i.e. used to validate its performance) datasets, creating several scenarios for the training-testing splitting. Similarly, to achieve optimum performance, multiple values were assigned to the GMDH-type neural network control parameters, and its performance was then evaluated. The forecasting system then aggregates each scenario’s results to produce a result with high accuracy and stability. Furthermore, the developed forecasting system is validated using historically recorded free atmospheric corrosion datasets selected from Norway.

The SHAP approach (i.e. detailed in Methods Section) is used in the second part of the forecasting framework to identify the important variables that contribute to the free atmospheric corrosion in steel structures located in marine environments. In the same way, the model is built from the ground up by randomly dividing the data into a training set and a test set. SHAP is applied to the prediction model (Random Forest (RF) classifier model in this case) to establish additive attributes (Eq. (11)), which are then used to determine the important factors for free atmospheric

corrosion. It should be noted that the RF classifier is a non-parametric method consisting of an ensemble of tree-structures.

### Data collection

To assess the ability of the proposed GMDH-type neural network to forecast free atmospheric corrosion in the marine environment and to use the SHAP technique to explain its relationship to the surrounding environmental factors, historical sensor data from the described experimental setup in Section “Experimental setup and data analysis” are used in this study<sup>27</sup>. The total number of observation data collected over the 110 days period, beginning in May and ending in August, is 5291. Table 1 contains statistical information about the sensor data, which is plotted in Fig. 2. The average temperature was 18 °C (with a maximum of 42 °C and a minimum of 5 °C), which is reasonable given that most of the data was collected during the summer. Similarly, a mean relative humidity of 72% was measured. The average reported corrosion in MSP2 was 0.1361  $\mu\text{A}$   $\in$  [0.6671  $\mu\text{A}$ , 0.0979 $\mu\text{A}$ ], while reported free atmospheric corrosion data by MSP1 were higher, with an average of 0.3015  $\mu\text{A}$   $\in$  [71  $\mu\text{A}$ , 0.1291 $\mu\text{A}$ ]. This can be observed in the irregularity in the period indicated by the red and purple boxes in Fig. 2, where certain high levels of corrosion were reported, which can be attributable to an error in the sensor, which was considered outliers and hence deleted during data processing. The data is organized in the form of a time series for training purposes, with 30 min as input embedding length and the next half-hour of corrosion as the target used to calculate the error.

### Implementation results and analysis

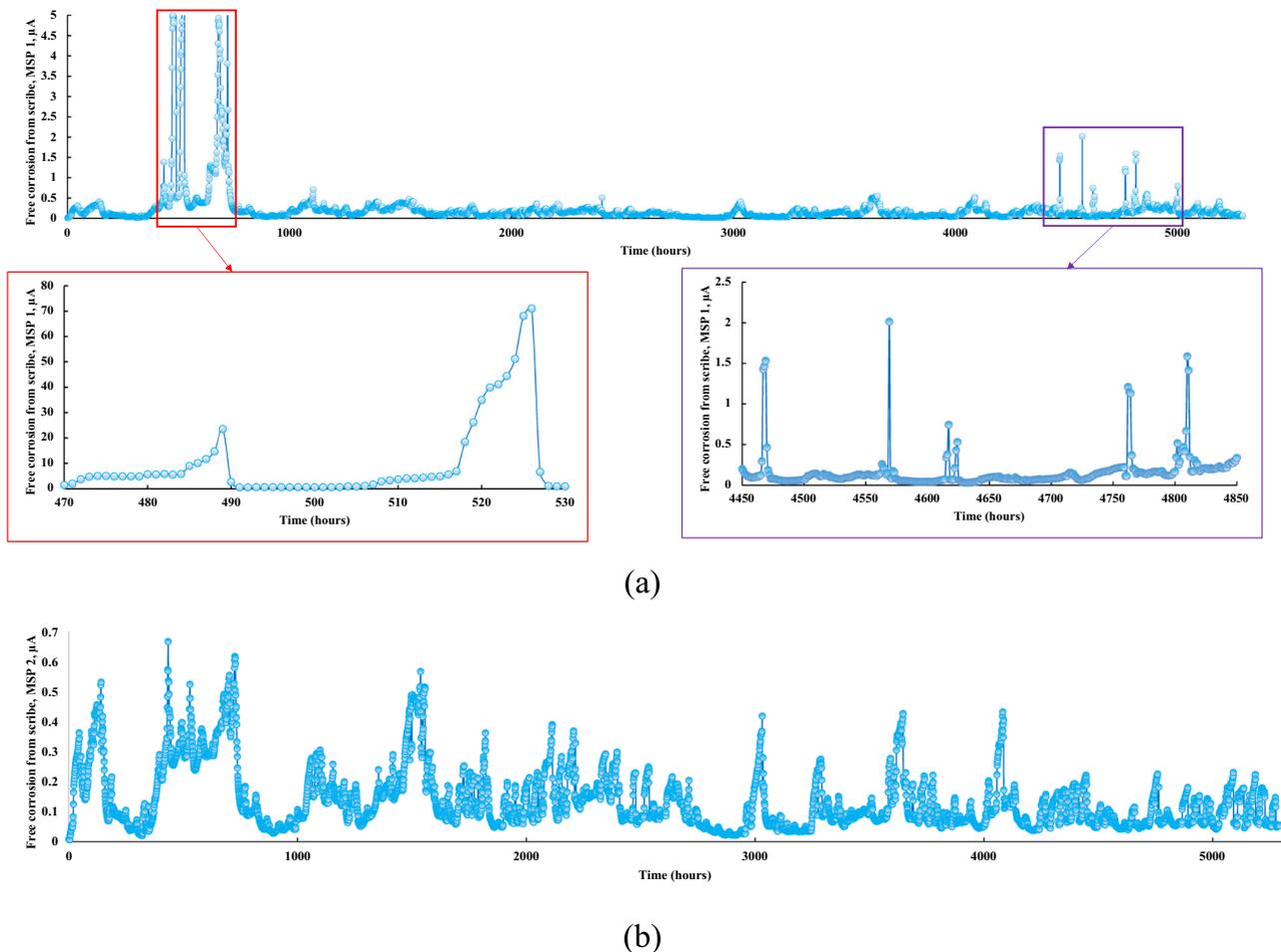
As there are no general rules in machine learning regarding the data-splitting for training and testing the models, it is more logical to investigate several splitting schemes. As a result, the proposed GMDH type neural network model was implemented with different splits as: 50%-50%, 60%-40%, 70%-30%, 80%-20%, and 90%-10% of the data attributed to the training and testing phases, respectively. This investigation will provide us with the optimal free atmospheric corrosion data partitioning. Figure 3 depicts the modeling results in terms of determination coefficient ( $R^2$ ), and it can be seen that the performance during the training phase is usually higher than the testing phase due to the fact that the first

is used to build the GDMH type neural network model (known data), while the second is used to validate it (unknown data). Obviously, the 70%-30% split produced the highest  $R^2$  values in terms of overall performance (Training + Testing). As a result, the 70%-30% partition is designated as the primary splitting for the remainder of this study's modeling phases.

Even though the results using the 70%-30% in Fig. 3 are optimistic, the  $R^2$  remains relatively low due to the selection of un-optimal design parameters of the GMDH type neural network model prior to computation. These design parameters are a maximum number of layers ( $N_l$ ), neurons in a layer ( $N_n$ ), and selection pressure ( $S_p$ ). Thus, different scenarios are investigated

**Table 1.** Description of the historical measurements collected from the sensors in Kjerringvik, Norway.

Captured measurements	Symbol	Unit	Statistical attributions					
			Max	Mean	median	Standard deviation	Min	Range
Temperature	$T_s$	°C	42.89	18.2664	17.07	6.0188	5.131	37.759
Temperature	$T_a$	°C	36.21	17.1092	17.06	3.6936	7.373	28.837
Relative Humidity	$RH$	%	100	72.7303	76.19	17.5171	22.76	77.24
Effective humidity	$EH$	%	108.3865	70.3002	72.9977	22.9853	11.2485	97.1379
Conductivity at high frequency	$\sigma$	$\mu S$	10000	2545.0326	6.34	4071.675	5	9995
Free corrosion from scribe, MSP 1,	$C_l$	$\mu A$	71	0.3015	0.1291	2.0292	0.005	70.995
Free corrosion from scribe, MSP 2,	$C_l$	$\mu A$	0.6671	0.1361	0.0979	0.1025	0.0053	0.6618

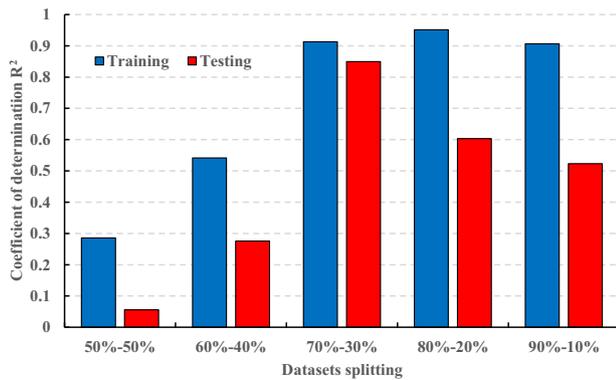


**Fig. 2 Time-series corrosion data.** Historical corrosion data obtained from sensors on the period 110 days (a) MSP1 (b) MSP2.

using numerous design parameter values in order to improve the forecasting accuracy, and after deploying the GMDH type neural network, performance metrics are used to evaluate the models' efficiency. Table 2 summarizes the performance evaluation of the obtained forecasting results based on the statistical metrics (i.e. RMSE, MAE, SMAPE, and  $R^2$ ), whereas nine GMDH type neural network models are explored depending on the design parameters (i.e. no  $S_p$  means the attributed value is 0). Forecasting results revealed that changing the design parameters of a GMDH type neural network has a significant impact on its performance. According to this table, the best performance for MSP1 datasets was obtained using model N<sup>o</sup>. 7 with  $N_b$ ,  $N_n$ , and  $S_p$  equal to 10, 15, and 0.4, respectively, whereas for MSP2 was using model N<sup>o</sup>. 5 with  $N_b$ ,  $N_n$ , and  $S_p$  equal to 10, 15, and 0, in the same respect. Additionally, for MSP2, the results show that the error between the

forecasted and the measured values is lower than the obtained using MSP1 dataset (this conclusion is seen in lower values of RMSE ( $RMSE^{MSP1} = 0.03598 \mu A$ ;  $RMSE^{MSP2} = 0.03008 \mu A$ ) and MSE ( $MAE^{MSP1} = 0.01775 \mu A$ ;  $MAE^{MSP2} = 0.01039 \mu A$ ) for the overall data (70% training+ 30% test)). Overall, the forecasting error based on SMAPE is about 5% for the MSP1 dataset and 3% for the MSP2 dataset, demonstrating that the right design parameters can significantly improve computational performance.

Figure 4 illustrates the forecasted versus measured values using the best GMDH type neural network models (N<sup>o</sup>. 7 and N<sup>o</sup>. 5, respectively, for MSP1 and MSP2), with the results plotted using the testing dataset. Furthermore, the scatterplots with the dashed black-line represent the linear regression equation. The higher the  $R^2$  value, the better the performance of the developed GMDH-model. Forecasting results using MSP1 and MSP2 datasets based on the proposed framework showed high agreement with measured sensor data, with  $R^2 = 0.8942$  for MSP1 datasets and  $R^2 = 0.9456$  for MSP2 datasets. Moreover, in terms of  $R^2$ , the forecasting results obtained by MSP2 outperform MSP1 by 5.47%, which can be attributed to the complexity of collected data from the sensors. In addition, Fig. 5 depicts the histograms of the forecasting error ( $e_k$ ) values of the developed GMDH type neural network models (N<sup>o</sup>. 7 and N<sup>o</sup>. 5 for MSP1 and MSP2, respectively) for the testing dataset. According to the error histograms, the mean error of the GMDH models is approximately  $-0.0068 \mu A$  for MSP1, while it is  $-3.15 \cdot 10^{-4} \mu A$  for MSP2, which these values are relatively close to zero, indicating high forecasting abilities. Furthermore, the standard deviation (StD) of both models is 0.0506 and 0.0132, respectively, indicating that nearly 68% of the forecasting error is around the mean with  $\pm 5\%$  and  $\pm 1\%$  error for MSP1 and MSP2, respectively.

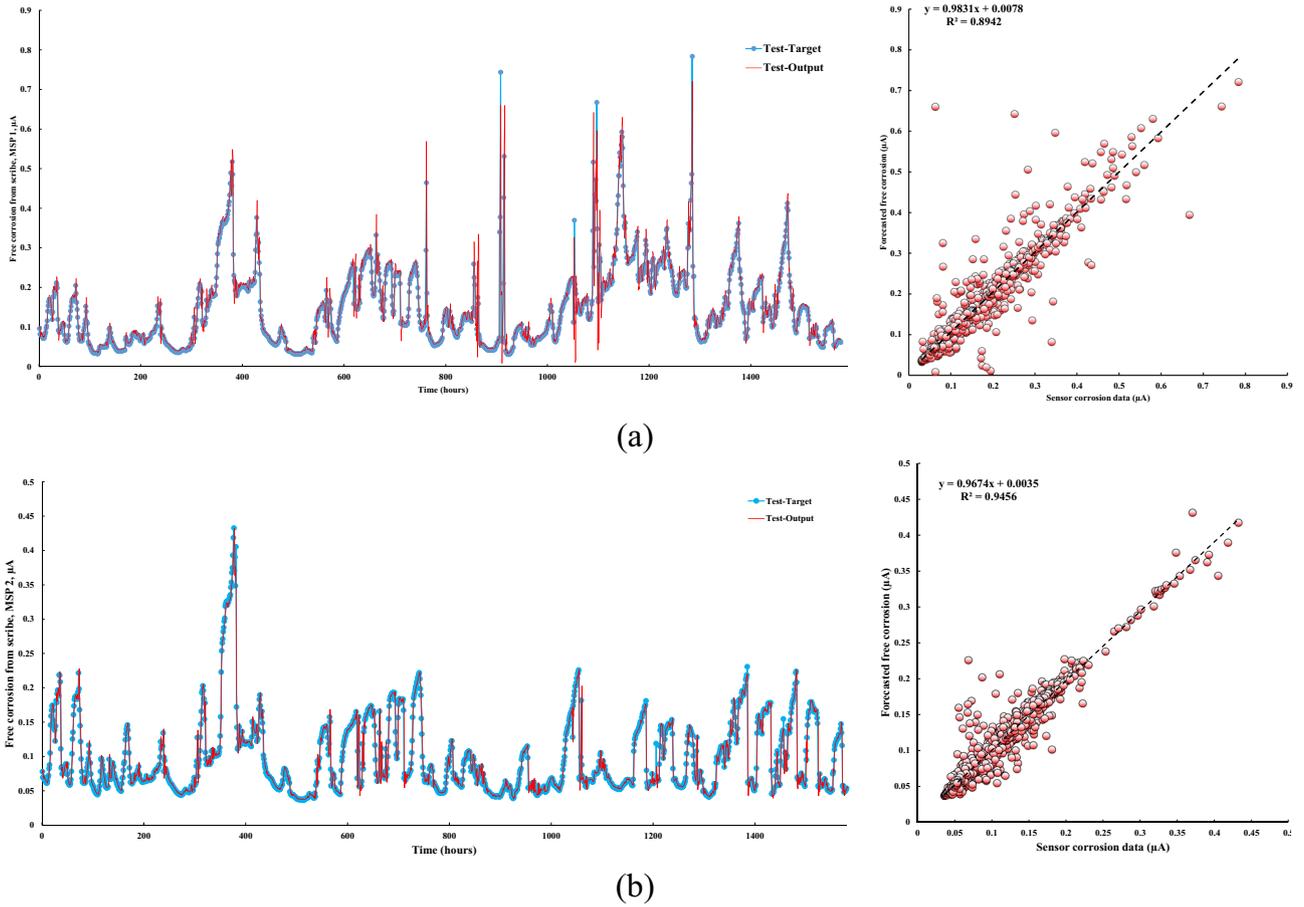


**Fig. 3 GMDH-type neural network performance.** Impact of the data splitting on the modeling performance.

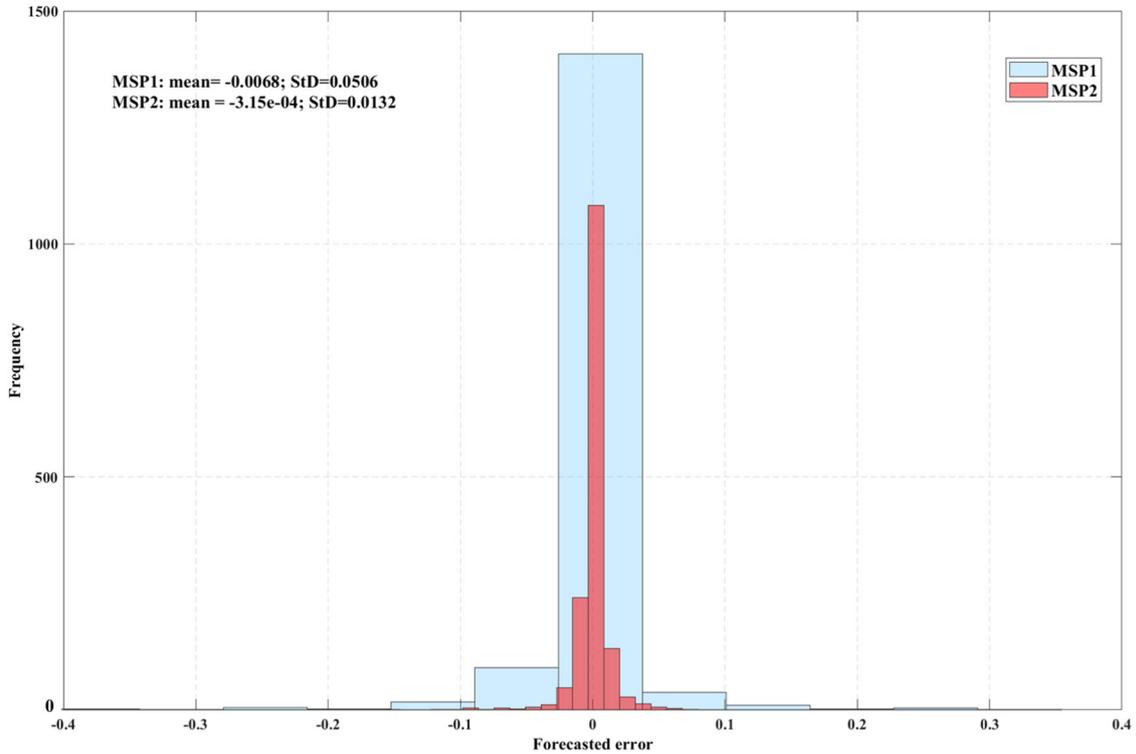
**Table 2.** Performance evaluation of GMDH type neural network with different design parameters for forecasting corrosion.

Dataset	GMDH type neural network-models N <sup>o</sup>	Training results				Testing results			
		RMSE $\mu A$	MAE $\mu A$	SMAPE %	R <sup>2</sup>	RMSE $\mu A$	MAE $\mu A$	SMAPE %	R <sup>2</sup>
Free corrosion datasets 1 (MSP1)	1 $N_n = 5; N_l = 5$	0.10894	0.03351	14.61749	0.49532	0.11569	0.04385	14.65221	0.43206
	2 $N_n = 5; N_l = 10$	0.04106	0.02032	12.17580	<b>0.91819</b>	<b>0.09995</b>	<b>0.03143</b>	10.40972	<b>0.68271</b>
	3 $N_n = 5; N_l = 15$	<b>0.03620</b>	<b>0.01753</b>	<b>6.92449</b>	0.91190	0.11879	0.03205	<b>8.07397</b>	0.44577
	4 $N_n = 5; N_l = 10$	0.04106	0.02031	12.15976	0.91819	0.09552	0.03079	10.38546	0.73779
	5 $N_n = 10; N_l = 10$	0.06402	0.04335	25.27315	0.87918	0.10501	0.05168	23.84227	0.58145
	6 $N_n = 15; N_l = 10$	<b>0.04011</b>	<b>0.01315</b>	<b>8.44835</b>	<b>0.92603</b>	<b>0.06115</b>	<b>0.02433</b>	<b>8.70761</b>	<b>0.82193</b>
	7* $N_n = 15; N_l = 10; S_p = 0.4$	0.06618	0.02216	9.26640	0.74295	0.10530	0.03245	9.82060	0.71024
	8 $N_n = 15; N_l = 10; S_p = 0.6$	0.03621	0.01557	7.13647	0.91532	0.07108	0.02120	7.07970	0.75154
	9 $N_n = 15; N_l = 10; S_p = 0.8$	0.04866	0.02811	13.36761	0.86016	0.08803	0.03385	12.64693	0.66598
Free corrosion datasets 2 (MSP2)	1 $N_n = 5; N_l = 5$	0.02034	0.00944	3.19179	0.96749	0.02108	0.00798	3.67979	0.87300
	2 $N_n = 5; N_l = 10$	<b>0.02016</b>	<b>0.00944</b>	<b>3.09216</b>	<b>0.96808</b>	<b>0.01709</b>	<b>0.00708</b>	<b>3.47375</b>	<b>0.91036</b>
	3 $N_n = 5; N_l = 15$	0.02034	0.00949	3.19662	0.96749	0.01820	0.00744	3.64240	0.89974
	4 $N_n = 5; N_l = 10$	0.02016	0.00944	3.09216	0.96808	0.01709	0.00708	3.47375	0.91036
	5* $N_n = 10; N_l = 10$	0.02037	0.00940	3.12103	0.96739	0.01594	0.00697	3.48140	0.92129
	6 $N_n = 15; N_l = 10$	<b>0.02010</b>	<b>0.00920</b>	<b>3.04229</b>	<b>0.96823</b>	<b>0.01316</b>	<b>0.00640</b>	<b>3.34344</b>	<b>0.94556</b>
	7 $N_n = 20; N_l = 10$	0.02020	0.00928	3.10425	0.96791	0.01599	0.00682	3.38230	0.92111
	8 $N_n = 15; N_l = 10; S_p = 0.4$	<b>0.02037</b>	<b>0.00943</b>	<b>3.13245</b>	<b>0.96739</b>	<b>0.01601</b>	<b>0.00695</b>	<b>3.49901</b>	<b>0.92111</b>
	9 $N_n = 15; N_l = 10; S_p = 0.6$	0.02057	0.00958	3.16392	0.96674	0.01606	0.00699	3.50363	0.92023
9 $N_n = 15; N_l = 10; S_p = 0.8$	0.02059	0.00962	3.18399	0.96666	0.01609	0.00704	3.53589	0.91991	

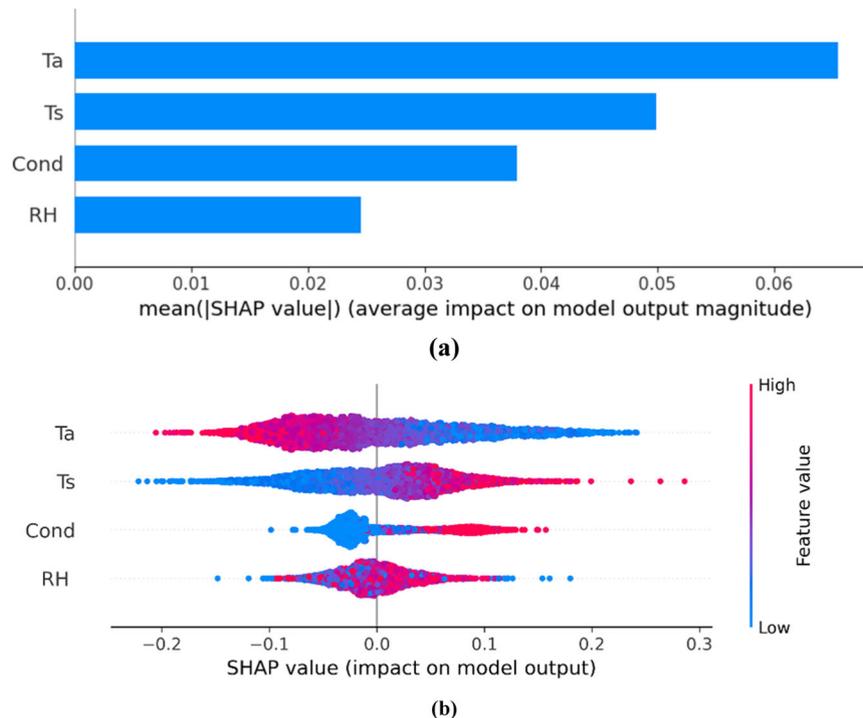
The bold numbers represent the best results in each sub-category, while \*represents the best GMDH-type neural network model among all others.



**Fig. 4 Scatterplots of the forecasted data.** Forecasted results versus the measured values during the testing phase, (a) MSP1 dataset and (b) MSP2 dataset.



**Fig. 5 Modeling uncertainties.** Forecasting error based on the testing phase.



**Fig. 6 SHAP results.** Summary plots for explaining the influence of the corrosion sensors' collected variables. **a** Bar plot of mean SHAP values; **(b)** Beeswarm plot.

### SHAP based factor impact discussions

The effect of the collected data related to the environmental factors, in addition to the corrosion measurements from the CorRES sensors, on the final output (i.e. free atmospheric corrosion), as stated in the framework section, has been quantified using the SHAP technique. The primary goal of this approach is to assign a SHAP value to each environmental variable (i.e.  $T_s$ ,  $T_a$ ,  $RH$ , and Conductivity), which can summarize the factor's impact on the final prediction results. A positive SHAP value indicates that the variable has a positive effect on the output, while a negative value indicates that the variable has a negative effect. In other words, positive values indicate that the variable will hasten or significantly contribute to the corrosion process, whereas negative values indicate that the corresponding variable will decrease or insignificantly contribute to the corrosion process. The impact of a variable on the collected free atmospheric corrosion measurement from the sensor, on the other hand, is proportional to the magnitude of its SHAP value.

Figure 6a, b show the results of the SHAP model. Figure 6a depicts the magnitude of the SHAP value, demonstrating that all environmental variables have an impact on the resulting free atmospheric corrosion output, with a higher influence of temperatures (air and surface temperatures) followed by conductivity and relative humidity, respectively. Figure 6b explains the impact of each variable on corrosion-model in terms of the SHAP Beeswarm plot, making the results of the SHAP values for these variables more interpretable. According to Fig. 6b, higher  $T_s$  values lead to higher SHAP values and the opposite for  $T_a$ , which can be explained by indicating that a high-temperature gap accelerates the corrosion process, with results not clearly showing the temperature impact and this can vary depending on the time. It is worth mentioning that the temperature-corrosion process relationship is complicated and difficult to address without significant experimentation. On the other hand, point distribution can be useful. We can notice a dense cluster of low conductivity values with small yet negative conductivity SHAP values. Higher

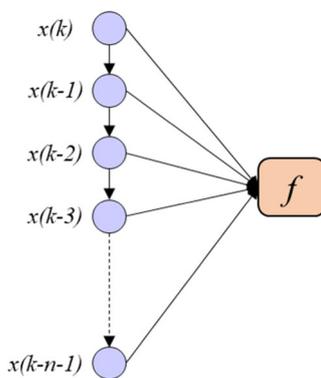
conductivity extends further to the right, implying that high conductivity has a significant impact on increasing the atmospheric corrosion process. The surface conductivity parameter represents chloride concentration, where Chlorides accelerate corrosion by changing the nature of the surface oxide, making it less protective. It is worth emphasizing that  $SO_2$  has little influence in our study because the sensors were placed very far away from the emission locations, and the  $SO_2$  content in the air in Norway is less than  $0.1 \mu\text{g}/\text{m}^3$ . Similarly, a dense cluster of relative humidity was observed with lower SHAP values, where higher humidity values can further contribute to the atmospheric corrosion process. On the other hand, the results are limited to only four variables and more parameters should be included to provide a general model and explanation of the variables impacting the free atmospheric corrosion process in marine environments, such as wind intensity and direction, dew point, and precipitation, all of which may contribute to this process.

In summary, this study addresses the forecasting of free atmospheric corrosion in marine structures using an intelligent time-series ML framework. The decision to handle the corrosion problem as a time-series problem stems from the fact that corrosion data was measured over a specified time step by the sensors, allowing us to tackle the process as time-dependent. As a result, this framework was tested on time-series corrosion datasets obtained from real-time sensors with 30 min intervals in Norway over a 110-day period. The following are the study's main findings:

- The training-testing percentage splitting was discovered to have an effect on the modeling process's outcome, with different scenarios (i.e. 50%-50%, 60%-40%, 70%-30%, 80%-20%, and 90%-10%) yielding different results. This latter showed that the best performance was obtained with 70-30% splitting in terms of  $R^2$  when compared to the other cases.
- Implementing the GMDH type neural network with different design variables demonstrated the importance of selecting these variables optimally to improve the forecasting system's performance. However, the forecasted data pattern using



**Fig. 7 The experimental setup.** View of a sensor installed toward the south east sea in Kjerringvik, Norway<sup>27</sup>.



**Fig. 8 Time-series forecasting model.** Explanation of the forecast process in time-series models.

GMDH type neural network was found to closely match the actual sensor data pattern, with little disagreement.

- The GMDH technique performed the best with  $N_n = 15$ ;  $N_l = 10$ ;  $S_p = 0.4$  for the first set of data from MSP1 with  $RMSE = 0.051 \mu A$  and  $SMAPE = 5.06\%$ ; and with  $N_n = 15$ ;  $N_l = 10$ ;  $S_p = 0.4$  for the second set of data obtained from MSP2 with  $RMSE = 0.0069 \mu A$  and  $SMAPE = 3.49\%$ .
- The SHAP approach provides an effective way to explain the impact of some environmental factors on the corrosion model, demonstrating that higher values of conductivity can increase corrosion currents, indicating the presence of corrosive ions, and that higher values of humidity can do the same.

Future research should focus on optimizing the selection of the machine algorithm (in this case, GMDH) rather than the manual approach, or on the implementation of advanced data-driven models as deep learning techniques for time-series data. Furthermore, further climatic and marine factors on the corrosion process and data using the SHAP approach can provide new insights in this field.

## METHODS

### Experimental setup

As shown in Fig. 7, the experimental setup in this study consists of installing two sensors (referred to as MSP1 and MSP2, respectively) provided by Luna Innovations Company (i.e. CorRES sensors) at a coastal site in Kjerringvik, southern Norway. Each sensor has two sensor panels on a docking platform, with one panel reporting weather data such as surface temperature ( $T_s$ ) and air temperature ( $T_a$ ), where measurements are done with thermistors; Surface conductivity, which is measured using a gold interdigitated sensor;

and relative humidity ( $RH$ ) measured using capacitive. The other panel is used to detect carbon steel-free atmospheric corrosion using LPR on an interdigitated steel sensor (steel type UNS G10080)<sup>27,28</sup>. All of the reported data was collected continuously during 2022 (i.e. From May to August over a period of 110 days).

### Machine learning models

In this part, a review of the theories relating to the developed forecasting and investigation framework based on the GMDH and SHAP methods was conducted. The proposed GMDH framework is based on previous measurements of free atmospheric corrosion. This type of prediction can be described by a possible linear or non-linear autoregressive process, which can be expressed as<sup>29</sup>:

$$x(k+h) = f[x(k), x(k-1), x(k-2), \dots, x(k-n-1)] \quad (1)$$

where  $f$  is a function of the  $x$ 's past and present values.

The main directions that can be applied to such problems (forecasting the  $h$  next values in time-series) are using an iterative method, which consists of repeating one-step-ahead predictions to the desired horizon as described in Fig. 8, and independent value prediction training the direct model to forecast  $x(k+h)$ . A machine learning algorithm is generally defined as a function  $f(\mathbf{x})$  that takes an input vector  $\mathbf{X}$  and produces an output vector  $\mathbf{Y}$ , making them excellent tools for dealing with time series problems. In this study, GMDH-type neural network is developed to predict the free atmospheric corrosion in a marine environment, and the SHAP approach is used to investigate the relationship between environmental factors and this phenomenon, as described in the subsection that follows.

**Group method of data handling (GMDH) type neural network.** The GMDH type neural network is a powerful data-driven modeling technique that is built based on mathematical functions to obtain complex nonlinear relationships between given input-output datasets<sup>30</sup>. The search for a function  $f$  that uses an input vector  $\mathbf{X} = (x_1, x_2, \dots, x_n)$  to calculate the output  $\mathbf{Y}$  based on a  $M$  observation using multi-input-single-output data, as described in Eq. (2), is a difficult task<sup>31</sup>. The goal is to train the GMDH to generate an approximated function  $\hat{f}$  rather than the actual  $f$  in order to achieve the closest output values  $\hat{y}$  to the actual output  $y$  as described in Eq. (3)<sup>29</sup>.

$$y_i = f(x_{i1}, x_{i2}, \dots, x_{in}) \quad (i = 1, 2, \dots, M) \quad (2)$$

$$\hat{y}_i = \hat{f}(x_{i1}, x_{i2}, \dots, x_{in}) \quad (i = 1, 2, \dots, M) \quad (3)$$

Thus, GMDH seeks to minimize the square of the difference between predicted and actual data, as expressed in Eq. (4).

$$\sum_{i=1}^M [\hat{f}(x_{i1}, x_{i2}, \dots, x_{in}) - y_i]^2 \rightarrow \min \quad (4)$$

The Kolmogorov-Gabor polynomial approach (also known as the Volterra functional series) is used in GMDH type neural networks to create a connection between the inputs and outputs, as shown in Eq. (5):

$$y = a_0 + \sum_{i=1}^n a_i x_i + \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j + \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n a_{ijk} x_i x_j x_k + \dots \quad (5)$$

To simplify, a system of partial quadratic polynomials composed of two neurons can be used to represent the previous equation as follows:

$$\hat{y} = G(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2 \quad (6)$$

In Eq. (6), the quadratic form is used to construct three polynomials, with coefficients  $a_i$  determined using a regression

**Table 3.** Comparative error metrics.

Metric	Abbreviation	Expression	Role/Utilization
Mean Absolute Error	MAE	$\frac{1}{N} \sum_{k=1}^N  e_k $	Used to determine the degree of similarity between predicted and observed values.
Root Mean Square Error	RMSE	$\sqrt{\frac{1}{N} \sum_{k=1}^N  e_k }$	Used to assess the overall disparity between observed and predicted values.
Mean Absolute Percentage Error	SMAPE	$\frac{1}{N} \sum_{k=1}^N \frac{ e_k }{ C_i  +  \hat{C}_i } \%$	Based on the percentage absolute error, it is used to calculate the accuracy of forecasted results.
Coefficient of determination	R <sup>2</sup>	$1 - \frac{\sum_{k=1}^N (-e_k)}{\sum_{k=1}^N (C_{i,k} - \bar{C}_i)}$	Used to compare the agreement between the forecasted data and real measurements

$C_i$  : The actual measured sensor data of the free atmospheric corrosion.  
 $\hat{C}_i$  : The forecasted free atmospheric corrosion.  
 $\bar{C}_i$  : The average of the free atmospheric corrosion measurements.

method, in which the difference between the predicted and actual output is minimized for each pair of  $(x_i, x_j)$  input variables. As a result, maintaining the coefficient of each quadratic function  $G_i$  is required to optimally fit the output in the entire set of input-output data pairs using the expression:

$$E = \frac{\sum_{i=1}^M (y_i - G_i)}{M} \rightarrow \min \quad (7)$$

To generate Eq. (7), the basic GMDH model considers all possible combinations of the two independent variables out of a total of  $n$  input variables. So, from the observations  $\{(y_i, x_{ip}, x_{iq}); (i = 1, 2, \dots, M)\}$  for different  $q, p \in \{1, 2, \dots, n\}$ ,  $\binom{n}{2} = \frac{n(n-1)}{2}$  neurons will be generated in the first hidden layer of the feed-forward network. Respectively, Eq. (8) represents the feasibility of generation  $M$  data triples  $\{(y_i, x_{ip}, x_{iq}); (i = 1, 2, \dots, M)\}$  from observations using such  $q, p \in \{1, 2, \dots, n\}$ .

$$\begin{bmatrix} x_{1p} & x_{1q} & \cdots & y_1 \\ x_{2p} & x_{2q} & \cdots & y_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{Mp} & x_{Mq} & \cdots & y_M \end{bmatrix} \quad (8)$$

The matrix equation  $\mathbf{Aa} = \mathbf{Y}$  can be calculated using the quadratic sub-expression Eq. (4) for each row of  $M$  data triples, where  $\mathbf{a}$  ( $\mathbf{a} = \{a_0, a_1, a_2, a_3, a_4, a_5\}$ ) is the vector of unknown coefficients of the quadratic polynomial in Eq. (6), while  $\mathbf{Y} = \{y_1, y_2, \dots, y_M\}^T$  is vector of output's value from measurements. Therefore  $\mathbf{A}$  can be expressed as follows:

$$\mathbf{A} = \begin{bmatrix} 1 & x_{1p} & x_{1q} & x_{1p}x_{1q} & x_{1p}^2 & x_{1q}^2 \\ 1 & x_{2p} & x_{2q} & x_{2p}x_{2q} & x_{2p}^2 & x_{2q}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{Mp} & x_{Mq} & x_{Mp}x_{Mq} & x_{Mp}^2 & x_{Mq}^2 \end{bmatrix} \quad (9)$$

The solution to the normal equations can be obtained by performing a multiple-regression analysis based on the least-squares technique in the form of Eq. (10), which will determined the vector of the best coefficient of Eq. (6) for the entire set of  $M$  data triples:

$$\mathbf{a} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Y} \quad (10)$$

*Shapley additive explanations (SHAP).* SHAP is an interpretable game theory-based approach for describing the performance of data-driven models based on an additive feature attribution method, in which an output model is defined as a linear addition of input variables<sup>32,33</sup>. Given a vector with  $p$  input variables  $\mathbf{X} = (x_1, x_2, \dots, x_p)$ , and based on an original model  $f(\mathbf{X})$  with simplified input  $\mathbf{X}'$ , the explanation model  $g(\mathbf{X}')$  is as follows:

$$f(\mathbf{X}) = g(\mathbf{X}') + \Phi_0 + \sum_{i=1}^M \Phi_i x'_i \quad (11)$$

where  $M$  is the number of input features, and  $\Phi_0$  is a constant value if all inputs are missing. The inputs  $\mathbf{X}'$  and  $\mathbf{X}$  are linked by a mapping function,  $\mathbf{X} = h_{\mathbf{X}}(\mathbf{X}')$ .

Lundberg and Lee<sup>34</sup> stated that the single solution for Eq. (11) should have three desirable properties: local accuracy, missingness, and consistency. When  $\mathbf{X} = h_{\mathbf{X}}(\mathbf{X}')$ , the local accuracy ensures that the function output is the sum of the feature attributions, which requires the model to match the output of  $f$  for the simplified input  $\mathbf{X}'$ . Missingness guarantee that no weights are given to missing features. In another word, missingness is satisfied, as  $\mathbf{X}'_i = 0$ . Changing a larger impact feature will not reduce the attribution assigned to that feature because of the consistency. When  $\mathbf{z}'_i = 0$ , for a setting  $\mathbf{z}' \setminus i$ ,  $f'_x(\mathbf{z}') - f'_x(\mathbf{z}' \setminus i) \geq f_x(\mathbf{z}') - f_x(\mathbf{z}' \setminus i)$  implies  $\Phi_i(f', \mathbf{X})$ , in which the only possible model that meets these requirements is as follows<sup>35,36</sup>:

$$\Phi_i(f', \mathbf{X}) = \sum_{\mathbf{z}' \subseteq \mathbf{X}'} \frac{|\mathbf{z}'|!(M - |\mathbf{z}'| - 1)!}{M!} [f_x(\mathbf{z}') - f_x(\mathbf{z}' \setminus i)] \quad (12)$$

In Eq. (12),  $|\mathbf{z}'|$  denotes the number of non-zero entries in  $\mathbf{z}'$ , and  $\mathbf{z}' \subseteq \mathbf{X}'$ , while  $\Phi_i$  represent the Shapely values. As a result, Lundberg and Lee<sup>34</sup> proposed a solution to Eq. (12), where  $f_x(\mathbf{z}') = f(h_x(\mathbf{z}')) = E[f(\mathbf{z}) | \mathbf{z}_S]$ , with  $S$  denotes the set of non-zero indices in  $\mathbf{z}'$ , known as SHAP values.

*Evaluation metrics.* The prediction evaluation of forecasting results in terms of accuracy and stability is crucial task<sup>37,38</sup>. Thus, the forecasting accuracy is evaluated using the forecasted error ( $e_k$ ), which is the difference between the actual measured sensor data ( $C_i$ ) and the forecasted atmospheric corrosion ( $\hat{C}_i$ ) for the  $k$ -th forecast as follows:

$$e_k = C_{i,k} - \hat{C}_{i,k} \quad (13)$$

Table 3 summarizes the statistical metrics used in this study (i.e. MAE, RMSE, SMAPE, and R<sup>2</sup>) as well as their intended use, where  $N$  denotes the total number of forecasted samples. For MAE ( $\mu\text{A}$ ), RMSE ( $\mu\text{A}$ ), and SMAPE (%), the lower the value, the higher the

model performance accuracy, whereas for  $R^2$ , the higher the value, the better the model efficiency.

## DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## CODE AVAILABILITY

The code script is available from the corresponding author upon reasonable request.

Received: 9 July 2023; Accepted: 11 October 2023;

Published online: 24 November 2023

## REFERENCES

- Roy, A. et al. Machine-learning-guided descriptor selection for predicting corrosion resistance in multi-principal element alloys. *NPJ Mater. Degrad.* **6**, 9 (2022).
- Ben Seghier, M.E.A., Mustafa Z., Zayed, T. Reliability assessment of subsea pipelines under the effect of spanning load and corrosion degradation. *J. Nat. Gas Sci. Eng.* 2022: 104569.
- Fielding, T. ISO 12944: Recent Revisions. *J. Prot. Coat. Linings* **37**, 36–38 (2020).
- Diamantino, T. C., Gonçalves, R., Nunes, A., Páscoa, S. & Carvalho, M. J. Durability of different selective solar absorber coatings in environments with different corrosivity. *Sol. Energy Mater. Sol. Cells* **166**, 27–38 (2017).
- Bea, R. G. Evaluation of alternative marine structural integrity programs. *Mar. Struct.* **7**, 77–90 (1994).
- El-Sherik, A.M. *Trends in oil and gas corrosion research and technologies: production and transmission*. Woodhead Publishing; 2017.
- Hou, B. et al. The cost of corrosion in China. *NPJ Mater. Degrad.* **1**, 4 (2017).
- Xia, D.-H., Song, S., Qin, Z., Hu, W. & Behnamian, Y. Electrochemical probes and sensors designed for time-dependent atmospheric corrosion monitoring: fundamentals, progress, and challenges. *J. Electrochem. Soc.* **167**, 37513 (2019).
- Morcillo, M., Chico, B., Diaz, I., Cano, H. & De la Fuente, D. Atmospheric corrosion data of weathering steels. A review. *Corros. Sci.* **77**, 6–24 (2013).
- De la Fuente, D., Díaz, I., Simancas, J., Chico, B. & Morcillo, M. Long-term atmospheric corrosion of mild steel. *Corros. Sci.* **53**, 604–617 (2011).
- De la Fuente, D., Castano, J. G. & Morcillo, M. Long-term atmospheric corrosion of zinc. *Corros. Sci.* **49**, 1420–1436 (2007).
- Xia, D.-H. et al. Electrochemical measurements used for assessment of corrosion and protection of metallic materials in the field: a critical review. *J. Mater. Sci. Technol.* **112**, 151–183 (2022).
- Farh, H.M.H., Ben Seghier, M.E.A., Zayed, T. A comprehensive review of corrosion protection and control techniques for metallic pipelines. *Eng. Fail Anal.* 2022:106885.
- Xia, D.-H., Ma, C., Song, S. & Xu, L. Detection of atmospheric corrosion of aluminum alloys by electrochemical probes: theoretical analysis and experimental tests. *J. Electrochem. Soc.* **166**, B1000 (2019).
- Momber, A. W., Langenkämper, D., Möller, T. & Nattkemper, T. W. The exploration and annotation of large amounts of visual inspection data for protective coating systems on stationary marine steel structures. *Ocean Eng.* **278**, 114337 (2023).
- ISO D. Paints and varnishes—Corrosion protection of steel structures by protective paint systems—Part 5: Protective paint systems. 2018.
- Morcillo, M., Almeida, E., Chico, B., Fuente, La.D. Analysis of ISO standard 9223 (classification of corrosivity of atmospheres) in the light of information obtained in the Ibero-American Micat project. In: *Outdoor Atmospheric Corrosion*. ASTM International; 2002.
- Vera, R. et al. Tropical/non-tropical marine environments impact on the behaviour of carbon steel and galvanised steel. *Mater. Corros.* **69**, 614–625 (2018).
- Feliu, S., Morcillo, M. & Feliu, S. Jr The prediction of atmospheric corrosion from meteorological and pollution parameters—II. Long-term forecasts. *Corros. Sci.* **34**, 415–422 (1993).
- Kihira, H. et al. A corrosion prediction method for weathering steels. *Corros. Sci.* **47**, 2377–2390 (2005).
- Valor A, Caleyó F, Alfonso L, Velázquez JC, Hallen JM. Markov chain models for the stochastic modeling of pitting corrosion. *Math. Probl. Eng.* 2013;2013.
- Li, Q. et al. Long-term corrosion monitoring of carbon steels and environmental correlation analysis via the random forest method. *NPJ Mater. Degrad.* **6**, 1 (2022).
- Ben Seghier MEA, Keshtegar B, Taleb-Berrouane M, Abbassi R, Trung N-T Advanced intelligence frameworks for predicting maximum pitting corrosion depth in oil and gas pipelines. *Process Saf. Environ. Prot.*
- Ben Seghier MEA, Höche D, Zheludkevich M. Prediction of the internal corrosion rate for oil and gas pipeline: implementation of ensemble learning techniques. *J. Nat. Gas. Sci. Eng.* 2022: 104425.
- Yan, L., Diao, Y., Lang, Z. & Gao, K. Corrosion rate prediction and influencing factors evaluation of low-alloy steels in marine atmosphere using machine learning approach. *Sci. Technol. Adv. Mater.* **21**, 359–370 (2020).
- Coelho, L. B. et al. Reviewing machine learning of corrosion prediction in a data-oriented perspective. *NPJ Mater. Degrad.* **6**, 8 (2022).
- Daneshian, B., Höche, D., Knudsen, O. Ø. & Skilbred, A. W. B. Effect of climatic parameters on marine atmospheric corrosion: correlation analysis of on-site sensors data. *NPJ Mater. Degrad.* **7**, 10 (2023).
- Buxton, P. A. & Mellanby, K. The measurement and control of humidity. *Bull. Entomol. Res* **25**, 171–175 (1934).
- Khosravi, A., Machado, L. & Nunes, R. O. Time-series prediction of wind speed using machine learning algorithms: a case study Osorio wind farm, Brazil. *Appl. Energy* **224**, 550–566 (2018).
- Nikolaev, N. Y. & Iba, H. Polynomial harmonic GMDH learning networks for time series modeling. *Neural Netw.* **16**, 1527–1540 (2003).
- Abdel-Aal, R. E., Elhadidy, M. A. & Shaahid, S. M. Modeling and forecasting the mean hourly wind speed time series using GMDH-based abductive networks. *Renew. Energy* **34**, 1686–1699 (2009).
- García, M. V. & Aznarte, J. L. Shapley additive explanations for NO2 forecasting. *Ecol. Inf.* **56**, 101039 (2020).
- Nohara, Y., Matsumoto, K., Soejima, H., Nakashima, N. Explanation of machine learning models using improved shapley additive explanation. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019:546.
- Lundberg, S.M., Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process Syst.* 2017;30.
- Lundberg, S.M., Erion, G.G., Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *arXiv Prepr arXiv180203888*. 2018.
- Feng, D.-C., Wang, W.-J., Mangalathu, S. & Tacioglu, E. Interpretable XGBoost-SHAP machine-learning model for shear strength prediction of squat RC walls. *J. Struct. Eng.* **147**, 4021173 (2021).
- Ben Seghier, M.E.A., et al. On the modeling of the annual corrosion rate in main cables of suspension bridges using combined soft computing model and a novel nature-inspired algorithm. *Neural Comput. Appl.* 2021: 1–17.
- El, M. et al. Prediction of maximum pitting corrosion depth in oil and gas pipelines. *Eng. Fail Anal.* **112**, 104505 (2020).

## ACKNOWLEDGEMENTS

This work was financed by Bundesministerium für Wirtschaft und Klimaschutz, Germany (reference 03SX521B), Jotun AS, Research Council of Norway (reference 311714), and Vlaio, Belgium (reference HBC.2019.2654.) in a joint research project in the MarTERA program. The first author also acknowledges the support of the Alexander Von Humboldt Foundation and the European Union's Horizon 2021 research and innovation programme under the Marie Skłodowska-Curie project No-101061320.

## AUTHOR CONTRIBUTIONS

Methodology, formal analysis, and writing—original draft preparation: M.E.I., A.B.S.; writing—review and editing: O.Ø.K.; data preparation: A.W.B.S.; writing—review and editing: D.H.

## FUNDING

Open Access funding enabled and organized by Projekt DEAL.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Correspondence** and requests for materials should be addressed to Mohamed El Amine Ben Seghier.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023