REVIEW ARTICLE OPEN

Reviewing machine learning of corrosion prediction in a data-oriented perspective

Leonardo Bertolucci Coelho¹✉, Dawei Zhang², Yves Van Ingelgem¹, Denis Steckelmacher³, Ann Nowé³ and Herman Terryn¹

This work provides a data-oriented overview of the rapidly growing research field covering machine learning (ML) applied to predicting electrochemical corrosion. Our main aim was to determine which ML models have been applied and how well they performed depending on the corrosion topic considered. From an extensive review of corrosion articles presenting comparable performance metrics, a 'Machine learning for corrosion database' was created, guiding corrosion experts and model developers in their applications of ML to corrosion. Potential research gaps and recommendations are discussed, and a broad perspective for future research paths is provided.

npj Materials Degradation (2022)6:8; <https://doi.org/10.1038/s41529-022-00218-4>

INTRODUCTION

The data-centric analysis presented in this review article derives from the refs^{1–19}. The Materials Genome Initiative (MGI)²⁰ has recently resulted in concentrated efforts to collect materials properties in the form of shared databases²¹. As a result, data-driven approaches have provided effective tools for driving the prediction of properties and the discovery of new materials^{21–23}. In the era of big materials data²⁴, machine learning (ML) has led to a conceptual leap in the field (especially in energy storage applications)^{21,23–27}. Although ML has been gradually applied to corrosion research²⁸, the corrosion community has benefited far less from the progress in Big Data technologies.

As with batteries research data^{21,29}, corrosion data is typically incomplete, noisy, heterogeneous and large in volume (low-value density). Moreover, the in-service corrosive scenarios are complex and changeable, constituting a highly nonlinear system hardly approachable by traditional statistical methods.

ML is the specific area of AI that allows computers to learn from data solving a given task. It consists of a flexible fitting function approach that, in comparison to traditional computational methods, can provide cheap and accurate simulation processes^{21,24,30}. ML aims to acquire knowledge from (very) large datasets by continuously improving their own performance. Despite the lack of prior knowledge of the system, knowledge mining based on ML methods could help the domain expert link outcomes to underlying physical laws or confirm some of the already established concepts^{24,27,31}.

A variety of electrochemical techniques has been employed for the prediction of the corrosion behaviour of metals. Accelerated corrosion tests and weathering tests have been historically used to simulate the working conditions of various systems. However, the simulated environmental factors in accelerated tests are distributed in ranges that often differ from the actual application. While in weathering tests, the environmental factors are time-varying with periodicity randomness^{32,33}. As there are no clear principles on setting fully representative test profiles, the extrapolation of short-term corrosion test results to field exposures in different environments is a challenging topic^{32,33}.

Despite disposing of huge volumes of testing data, the corrosion community still relies on inaccurate predictive models (log-linear³⁴, time-varying function³⁵, dose-response function³⁶, power function^{36,37}, Eyring and Arrhenius³⁸) of the in-service corrosion rate (CR). To the best of our knowledge, only one review referring to ML and corrosion is available³⁹ (this work focused on image recognition and pattern classification).

The selection of the most suitable ML approach is a multifaceted question, depending on the amount of data available, the quality of results desired and the necessity of physical interpretation of the results²⁷. Therefore, the present investigation highlights assessing the predictive power of different machine learning approaches and elaborate on the current status of regression modelling for various corrosion topics.

The data mining approach proposed here also aims at identifying unseen patterns from the intersectional ML and corrosion literature (knowledge mining). We encourage new experimental and computer-oriented scientists to apply ML to corrosion prediction studies by discussing challenges and providing recommendations.

BIBLIOMETRIC DATA MINING

The main tool used for identifying relevant literature for this Review was the Web of Science search engine, a large abstract/citation database covering mainly peer-reviewed journals⁴⁰. The queries employed in the search engine were the following: 'machine learning' AND 'corrosion' ('Topic' search, all types of documents). The 164 records returned from the queries (sent on 1 March 2021) were carefully reviewed and filtered based on domain knowledge.

A co-citation network of publications and their citation relations is presented in Fig. 1 for the 2017–2020 period (the size of the nodes is scaled by the number of citations of each source)^{41,42}. The two largest nodes ('machine learning' and 'corrosion') are highly interconnected and are predominantly related to the 'prediction' node than with the 'design' or 'classification' nodes. The obtained cluster illustrates that research has been conducted on distinct corrosion topics (atmospheric corrosion, pitting

¹Department of Materials and Chemistry, Research Group Electrochemical and Surface Engineering, Vrije Universiteit Brussel, Brussels, Belgium. ²Beijing Advanced Innovation Center for Materials Genome Engineering, National Materials Corrosion and Protection Data Center, Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing, China. ³VUB Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels, Belgium. ✉email: leonardo.bertolucci.coelho@vub.be

Fig. 1 Visualization of the co-citation network of 'machine learning' and 'corrosion' publications. Citation relations from 2017 to 2020 (VOSviewer⁴¹).

corrosion) and materials (alloys, carbon steel, reinforced concrete), using different models (artificial neural network (ANN), random forest (RF), support vector machine (SVM)).

Initially, 35 papers in which supervised learning were explicitly used for corrosion prediction were considered (classification modelling investigations were excluded). From this preselection, 16 references where regression models were employed while providing comparable metrics were selected^{1–10,13–18}. Finally, 3 other works on artificial neural networks (not designated as ML)^{11,12,19} were also included.

Machine learning for corrosion database

The final dataset used for this Review consisted of 19 papers comprising 34 ML works whose models were evaluated by comparable metrics (absolute percentage error (MAPE) and/or the correlation coefficient (R^2)). The dataset was loaded in a Jupyter Notebook (Python language), and a ML and corrosion database was built into a Pandas DataFrame (further details in the Supplementary information).

The database⁴³ disposes of 47 features and 1 indexing column, the 'Reference', the same reference numbering of this article. The features list describes the investigation approach ('Orientation-strategy', 'Targets'). It includes essential descriptors of the data ('Data collection', 'Data availability'), corrosion ('Corrosion topic', 'Material') and ML ('Models description', 'Evaluation method', 'Train-Test split'). All plots presented in this publication were derived from this database⁴³, which is available to download at Mendeley data (related codes available at GitHub).

Based on the 'Selected features' attribute in our ML for corrosion database⁴³, Table 1 displays the predictors organised

by different categories (environment, material, etc.) for all selected papers^{1–19}.

It is essential to remember that our reviewing exercise was conducted with 34 investigations coming from at least 9 broad corrosion topics. A case-by-base judgement of the validity of the predictions was out of our scope. This approach would require in-depth expertise related to all addressed topics and much broader access to the corresponding datasets (including the 'bad training data', which is traditionally seldomly reported).

Data dimension Vs publication year

Figure 2 displays the 'Dataset size' and the 'Initial N° of attributes' features as a function of the publication year. Corrosion has always been a multi-dimensional problem, which is reflected in the plot by the relatively constant 'Initial N° of attributes' (number of corrosion variables considered before data reduction) from 1999 to 2021. Nonetheless, from 2018 onwards, the growing trend in 'Dataset size' identified could be attributed to the 'Big Data Boom'⁴⁰. In parallel to the Big Data Boom, the rise in popularity of ML is expected to reduce the contradiction of high-dimensional problems based on small sample sizes²¹.

PERFORMANCE OF THE ML MODELS

The effect of data dimension

From Fig. 3, it could be seen that the higher the dataset size, the lower tend to be the mean percentage errors, either during training or testing. The associated mean R^2 values were, whenever possible, included next to the mean MAPE data points, indicating the good correlation between both metrics. Although not likely to confirm it without the corresponding mean R^2 , the low mean

Table 1. Predictors organized by different categories (material and environmental variables) from the referenced papers^{1–19}.

Ref.	Electrolytic	Climate	Composition	Metallurgical/ structural	Atomic (physical)/ chemical	Electrochemical	System	Temporal	Spatial
1 ^a		O ₂ _max, S_max, pH_min, T_mean	Content (LAS): Cr, C, Si, P, S		EN _s , K, R _{mv} , R _c , V, λ, EN _{pr} , ΔH _f ⁰			t	
2		T_MAX, T_AVE, RH_MIN, PRECIPIT, SOLAR, Cl ⁻ , SO ₂	Total content of alloying elements (LAS)					t	
3		RH_mean, T_mean, Rainfall, SO ₂ , pH of rain, Cl ⁻	Content (LAS): Mn, S, P, Si, Cr, Cu, Ni					t	
4		T, RH, rainfall				Electrical quantity (Q _{ACM})			Station (Jiangjin, Wuhan)
5				Steel grade (A3, 3C, 20, 08Al)		CRs (1st, 2nd, 4th, 8th year)		Incubation period, duration, duration of exposure, log (duration of exposure)	
6	HCl, H ₂ SO ₄		Contents: B, C, Cr, Cu, Fe, Mn, Mo, Ni, P, S, Si, Ti, V (carbon and alloy steel); Mo (x), Ni (14 -x), (x + 1) (metallic glasses)			Potential, dE			
7	pH, Cl ⁻ , temperature			Alloy 22 treatment (welded or annealed)		Corrosion type (general or localized)		Aging time, exposure time	
8	Dissolved oxygen, salinity, pH, temperature					Oxidation-reduction potential			
9		RH_mean, T_mean, Rainfall, SO ₂ , pH of rain, Cl ⁻	Content (LAS): Mn, S, P, Si, Cr, Cu, Ni, Fe					t	
10	pH, SO ₄ , Cl, Fe, HCO ₃ , Ca, temperature						PS, PCO ₂ , BSW, BOPD, BWPD, MMCFD	Exposure time	
11		SO ₂ , Cl ⁻ , temperature, TOW (time of wetness)							
12	Cl _{TT} (Cl ⁻ concentration), chloride binding, pH, DO (dissolved oxygen concentration)								
13						CR from deterministic models (de Waard, NORSOK)	Elevation, inclination, concavity, conc./ incl., dist_failure, flow regime, hold-up,		

Table 1 continued

Ref.	Electrolytic	Climate	Composition	Metallurgical/ structural	Atomic (physical)/ chemical	Electrochemical	System	Temporal	Spatial
14				Concrete mixture design proportions: cement (CE), limestone powder (LSP), coarse aggregate (CA), fine aggregate (FA), water (WA)			pressure, gas flow, total flow, liquid velocity, gas velocity	Exposure period (EXP)	
15	pH, conductivity, temperature			Degree of sensitization - DoS, stress intensity factor - KI (304 SS)		Electrochemical potential			
16	pH			AA alloy grade (2024, 7075)	Molecular descriptors ^c				
17	pH, Total Acidity, TA/pH								
18	Total Cl ⁻ (Cl TT), Cl ⁻ binding, DO, salinity, pH, temperature					Corrosion potential			
19	pH, temperature		SiO ₂ , Al ₂ O ₃ , Fe ₂ O ₃ , CaO, MgO, Na ₂ O, K ₂ O, SO ₃ (cement)	Water/cement ratio, porosity		Electrochemical corrosion potential - ECP (0.5 ppm, 1 ppm, 5 ppm, 8 ppm [O ₂]), passivity breakdown potential - Eb			

These descriptors were extracted from the 'Selected features' attribute in the ML for corrosion database (Supplementary information).

^aReference¹ also deals with 9 principal data components (variable with no physical meaning).

^bEN_e = Sanderson electronegativity; K = Bulk modulus; R_m = molecular single bond covalent radius; R_c = covalent radius; V = molar volume; λ = thermal conductivity; EN_p = Pauling electronegativity; ΔH_f = enthalpy of fusion.

^c1515 molecular descriptors were generated by DRAGON programme. Sparse feature selection (BioModeller) was applied to obtain the most relevant descriptors (number varied between 11 and 31).

Fig. 2 The “Data Boom” effect on the selected ML and corrosion references. The ‘Dataset size’ and the ‘Initial number of attributes’ (derived from the Machine learning for corrosion database) are presented Vs the publication year.

Fig. 3 The mean MAPE as a function of the ‘Dataset size’ (for both training and testing datasets). Corresponding mean R^2 were included whenever possible next to data points. This plot considered all corrosion topics from the ML for corrosion database⁴³. The error bars represent the standard deviation of the mean.

MAPE values obtained for dataset sizes smaller than 10^1 might result from overfitting.

The effect of the Time variable

Figure 4 reveals that the models’ performance (expressed by R^2 ^{1–4,6–17,19}) considerably increased when the Time was considered as an input (refer to the ‘Selected features’ and ‘Time selected?’ features in the ML for corrosion database (Supplementary information)). This trend generally held for both the training and testing datasets and was particularly remarkable for the ANN model type. It is interesting noting that Linear models are not suitable for predicting corrosion, which is essentially characterised by time-series data.

Performance per corrosion topic

The performance of the models (Machine learning for corrosion database⁴³) was evaluated either by the R^2 ^{1–4,6,17,19} and/or by

MAPE^{3,5,9–11,13,19}). With the models’ names and corresponding references, Figs. 5 and 6, respectively, display the achieved R^2 and MAPE values organised by corrosion topic. Moreover, the data points are subdivided by orientation strategy (Fig. 5) or by type of ML model (Fig. 6).

These evaluation metrics derive from models with differences in structure/hyperparameters and represent optimum modelling conditions for each case. Therefore, the reader should be aware that the proposed performance comparisons refer to the modelling approaches in the broad sense (not the models solely).

As displayed in Figs. 5 and 6, atmospheric corrosion has been the most studied corrosion topic^{1–4,6,9,11}. Marine corrosion seems to be a straightforward research topic: not only relatively high predictive accuracies were achieved^{8,18}, but the models demonstrated a high generalisation ability (low difference between training and testing MAPE values (Fig. 6)). On the contrary, the prediction of rebar corrosion has been more scattered, with the highest errors in cases where localised corrosion (pit) occurred (Fig. 6¹⁸).

The ‘Material & environment’ orientation strategy has been the most popular approach, encompassing corrosion-resistant alloys (CRAs)⁷, inhibitors¹⁶ and atmospheric corrosion^{1–3,6,9} investigations. Regardless of the corrosion subject, Fig. 5 shows that diversifying types of variables leads to increased performances. For example, in the case of rebar corrosion, investigations simultaneously considering material, system-related and environmental descriptors¹⁹ clearly outperformed both material¹⁴ and environment-oriented^{12,18} approaches. A system-orientated strategy has been required for pipeline corrosion (including geometrical pipeline characteristics¹³), with increased goodness of fit when instrumental variables (e.g. operating pressures¹⁰) were also considered. Atmospheric corrosion benefited the most when material features were also included in environment-oriented strategies^{1–3,6,9}.

Concerning the types of ML models (Fig. 6), ANN (including BPNN (back propagation neural network), DNN (deep neural network)^{3,8–10}) and tree-based (RF, cForest, etc.^{3,9,10}) have generally modelled atmospheric corrosion^{3,9}, marine corrosion⁸ and pipeline corrosion¹⁰ with reasonable training errors (mostly below 30%). Comparatively, Kernel-based methods (SVR (support vector regression), SVMs) have generated higher prediction scattering across different corrosion topics^{3,5,9,18}. Although only a few works employed Linear (logistic regression³, linear regression¹⁸) or genetic algorithms⁵, these methods tend to produce high and low MAPE values, respectively. Overall, ensemble methods (combination of individual learners) have yielded the best predictive performances^{8–10,19}, except when modelling the pitting of rebar corrosion¹⁸.

More details from each investigation, including a further description of the models employed, are disclosed in the corrosion topic sections hereafter.

REVIEWING ML PER CORROSION TOPIC

Atmospheric corrosion

Atmospheric corrosion data, traditionally acquired by coupon exposure tests, is often associated with small sample sizes and high dimensionality defined by multiple corrosion-influencing factors. The complexity of the corrosive environment makes it highly challenging for traditional predictive models to make optimal decisions in short timescales²⁶.

In 1999, in the pioneering work performed by Cai et al.¹¹, a worldwide dataset on the atmospheric corrosion of steel and zinc was used to train and test neural networks. In contrast to previous works using a linear regression technique, the ANN seemed to be a promising modelling tool for addressing nonlinear/complex systems based on interpolation from past experience.

Fig. 4 The mean R^2 as a function of the 'Type of ML model', depending whether the Time was selected as a feature or not. Both training and testing performances were considered for all references using R^2 as an evaluation metric^{1–4,6–17,19} (ML for corrosion database⁴³). The error bars represent the standard deviation of the mean.

Through sensitivity analysis, ANN could also demonstrate that the CRs of zinc were most likely affected by changes in corrosion products. As the corrosion rates slowed at high TOW (time of wetness), it was suggested that moisture inhibited corrosion due to the formation of protective carbonates. Meanwhile, corrosion significantly increased at high SO_2 concentrations due to sparingly soluble basic sulphates shifting towards soluble sulphates¹¹. Concerning steel, knowledge mining showed that the relation between TOW, SO_2 concentration and atmospheric corrosion was almost linear¹¹.

Regarding the prediction of corrosion depths, ANN achieved the following (testing) performances for steel and Zn, respectively: $R^2 = 0.548$ (MAPE = 39%) and $R^2 = 0.608$ (MAPE = 53%)¹¹. These testing performance values are displayed in Figs. 6 and 7.

The high variances obtained were mainly due to ignorance of other important affecting variables and inherent scatter in the data from several sources. As the datasets were collected from multiple publications (42), more or less noise or bias is typically unavoidable²⁴.

Although the ANN was not suitable to predict very long-term corrosion (lack of datasets), it was expected that its modelling capability could increase by considering more affecting factors and with the availability of more accurate data¹¹.

Nearly a decade later, Kamrunnahr et al.⁶ employed a neural network backpropagation method as a data mining tool to predict the CRs of carbon and alloy steels. According to the authors⁶, the trends that govern long-term behaviour are among the most critical challenges in corrosion studies. Notably, the effect of environmental variables on the outputs is often not well understood.

The ANN employed data organised according to the differences in the environmental conditions, including temperature, pH, etc. (each condition was unique). The supervised neural network could achieve a high prediction accuracy for long periods, despite being trained with short time period data only⁶. The achieved testing performance ($R^2 = 0.987$) is indicated by 'steels' and 'ANN (steels)' in Figs. 7 and 5, respectively.

More recently, Zhi et al.³ developed ML tools to forecast the outdoor atmospheric CR of low alloy steels (LAS). They also applied corrosion-knowledge mining by using a Random Forests (RF) algorithm. By applying this approach on corrosion data of 17 low alloy steels under 6 atmospheric corrosion test stations in China over 16 years⁴⁴, the authors aimed at tackling the three

main challenges with atmospheric corrosion datasets from multiple materials/environments: (1) their non-linearity, (2) their steep-manifold structure and (3) their (often) small sample size.

In their study, the methods of Logistic Regression (LR) and support vector regression (SVR) were not accurate (MAPE (testing) values of 45.63% and 41.35%) to deal with the steep-manifold structure of the data (at least for the number of samples considered). On the other hand, unstable classifiers (tree-based, ANN) produced acceptable performances using datasets with same sizes (MAPE (testing) = 22.12%, for ANN). In particular, RF was the least affected by the relatively small number of training samples (MAPE (testing) = 16.08%). The MAPE values³ are indicated in Fig. 6, while the corresponding R^2 (training) values³ are indicated by purple arrows in Fig. 7.

Furthermore, RF was applied separately to 5 sub-datasets with different exposures times: 1, 2, 4, 8, 16 years (the corresponding performances are present in the 'Sub-datasets with different times' feature in the ML for corrosion database (Supplementary Notes).

From knowledge mining applied to these sub-datasets, it was revealed that the environmental factor was more important than the material factor, with pH being the most important feature at the two corrosion stages (time split at Year 5). At the initial stage, the 'environment' importance was always greater than 80%. Notably, the SO_2 and Cl^- concentrations had higher importance indices than the temperature (T) and the Relative Humidity (RH), while the rainfall was the second most important feature only in Year 1. At the stationary stage (more than 5 years), the corrosion products shielded the LAS from environments, decreasing the importance of the environment factor (the SO_2 importance was much less than Cl^- , T and RH)³.

In another work, Zhi and co-authors⁹ mentioned the current difficulty of making predictions of various materials in diverse and dynamic outdoor environments from modelling approaches based on indoor corrosion tests with just single materials (at least for the commonly used ML models, such as ANN, SVM, power-linear function, hidden Markov). Therefore, the authors proposed a new deep structure, called Densely Connected Cascade Forest-Weighted K-Nearest Neighbours (DCCF-WKNNs), to implement the modelling of LAS atmospheric corrosion from the same dataset previously used in ref. 3. The different models employed in these two works^{3,9} are not directly comparable because the latter⁹ disposed of one more input (Fe element). The DCCF-WKNNs inherited the high generalisation ability (MAPE (testing) = 12.95%)

Fig. 5 Performance of the ML models expressed by the R^2 (refs. 1–4,6–17,19), segmented by ‘Corrosion topic’ and organised by ‘Orientation strategy’. R^2 values are distinguished for training (‘O’) and testing (‘X’) performances. The models’ names of data points referred to in the text were included next to the respective ‘X’ points (other useful descriptors were also included whenever possible).

from RF for modelling the small-size LAS outdoor atmospheric corrosion data, also demonstrating a better feature extraction ability as a representation learning algorithm with a deep structure. The testing performances (MAPE and R^2) obtained with ANN, SVR, RF, RF-WKNNs, cForest (cascade forests), and DCCF-WKNNs⁹ are highlighted in Fig. 6 and indicated by green arrows in Fig. 7. For more information on the developed ensemble model, the reader is referred to the ‘Models description’ column of the ML for corrosion database (Supplementary Notes).

In addition, knowledge mining could provide quantitative thresholds for the environmental variables; for instance, the CR decreased when pH was higher than ~6.3 or the rainfall was lower than ~320 mm per month; the CR increased when the SO_2 concentration was higher than ~0.067 mg cm^{-3} or the Cl^- concentrations were higher than ~0.8 and 1.9 mg cm^{-3} ; and whether the electrolyte was continuous or not and whether the film was saturated or not (RH values equal to 67% and 86%, respectively)⁹.

In the same direction, Yan et al.² proposed a ML modelling method to simulate the atmospheric corrosion behaviour of LAS. While evaluating the correlations between material, environmental factors and CRs, the impact of rust layer formation as a function of the exposure periods was addressed.

The Pearson correlation coefficient (PCC) was used to evaluate the degree of correlation between multiple variables. Features with significant correlation were grouped into one same cluster and were considered multicollinear features. The dominating factors from each cluster were closely correlated with the output (CR) and independent. The achieved Pearson correlation map demonstrated a positive correlation for all the variables analysed ($\text{PCC} > 0$), highlighting their interdependence (refer to the ‘Attributes interdependence’ feature in the ML for corrosion database⁴³).

The RF, GBDT (gradient boosted decision trees), and XGBoost models showed high predictive accuracy from the training set, with R^2 values of 0.94, 0.96 and 0.93, respectively² (points a, b and c in Fig. 7). Next, feature importance analysis revealed that the ELEMENTS (total content of alloying elements) feature was among the most significant during the whole exposure period (the SO_2 feature was selected based on domain knowledge). Concerning the environmental features, the CHLORIDE (chloride deposition rate) and PRECIPIT (precipitation) variables were the most significant ones in the first 3 years, while the RH_MIN (minimum relative humidity) became the most significant factor after 5 years².

From this work, the main shortcoming of the RF, GBDT and XGBoost models was their low generalisation ability: R^2 values (testing) of 0.73, 0.69 and 0.77, respectively² (points a’, b’ and c’ in Fig. 7). The authors believed that the data amount was insufficient to reflect the impact of all factors fully. In particular, the low amount of data with high CR values generated a detrimental effect: the models were only accurate for samples with low CR values (limited extrapolation ability)².

Traditional methods to predict atmospheric corrosion requires the accumulation of environmental data yearly. These annual averages could generate inaccurate predictions, especially in the short term⁴.

Based on this premise, Pei et al.⁴ monitored the atmospheric corrosion of carbon steel by a Fe/Cu type galvanic corrosion sensor for 34 days. Upon application of RF for the ranking of feature importance, the T (temperature), RH and rainfall status were considered the most important features (pollutants were not significant due to their low concentrations). The authors recommend that the deposition of chlorides (monthly recording) be included as a feature in future works, especially for modelling longer-term corrosion in marine atmospheres⁴.

Fig. 6 Performance of the ML models expressed by the MAPE (refs. 3,5,9–11,13,18), segmented by ‘Corrosion topic’ and by ‘Type of ML model’. MAPE values are distinguished for training (‘O’) and testing (‘X’) performances. The models’ names were included right next to the respective ‘X’ points (other useful descriptors were also included whenever possible). Part of the results for the Ensemble models is magnified in the inset image (green borders).

The sensor ended almost fully covered with corrosion products, which affected its response and thus the models’ accuracy over time. This issue was dealt with by considering the rust formation on the sensor with the integral value of instantaneous corrosion current (Q_{ACM}) as an input parameter, resulting in improved prediction accuracy for all models (Q_{ACM} was considered the most crucial factor in the second half of the test). Remarkably, the RF model with charge correction produced the highest accuracy, with a R^2 of 0.940, for both training/testing sets⁴ (indicated as ‘QACM’ and ‘RF-QACM’ in Figs. 7 and 5, respectively).

The application of CR prediction models is often limited to alloys with similar compositions to those used for training (inputs with a fixed number of elements content). Aiming at overcoming this limitation, Diao et al.¹ proposed a feature creation method to convert the chemical composition features from LAS into a set of atomic and physical property features. Thereby, these material-oriented features were used together with environmental factors as inputs, and RF was selected as the modelling algorithm. But firstly, feature reduction methods were applied to select the dominating factors on the data target (GBDT + Kendall provided the best results).

By applying the GBDT + Kendall method, the key factors influencing atmospheric corrosion were identified: Cr, C, Si, P, and S contents, O_2_{max} (maximum dissolved oxygen content),

S_{max} (maximum salinity value of the seawater), pH_{min} (minimum pH), T_{mean} (mean temperature). Next, the feature creation methods proposed an effective way to simulate the influence of elemental contents and their interactions with the steel corrosion resistance¹. Although the feature creation methods (based on physical properties simulating elements) could extend the model’s applicability to other alloys, it should be mentioned that modelling with the top 4 most important physical features produced lower performance than with the top 5 elements¹. This difference in testing performance (R^2 equal to 0.90 or 0.92)¹ is depicted in Fig. 7 as ‘elements’ and ‘phys. features’.

As summarised in Fig. 7, the same predictive tools have successfully modelled both I or CR; namely SVR^{2–4,9}, RF^{1–4,9} and ANN^{3,9,11}. Although more extensive statistics would be needed to confirm this hypothesis, higher goodness of fit was generally observed having the I as data target (CR is an indirect measure often deriving from I).

Finally, Zhi et al.⁵ combined a nonlinear grey Bernoulli model (NGBM(1,1)) with genetic algorithm (GA) for long-term prediction on a specific monotonic data series of atmospheric CR. The collected corrosion data of carbon steels exposed in various outdoor atmospheric environments was provided by the China Gateway to Corrosion and Protection⁴⁴.

Fig. 7 Performance of the ML models expressed by R^2 applied to atmospheric corrosion (refs. 1–4,6,9,11). Data points were segmented by 'Targets' (current I, CR, depth) and by 'ML models'. R^2 values are distinguished for the training ('O') and testing ('X') performances. Meaningful entries from the following features were included in the plot: the 'Material' (Zn, steel¹¹, steels⁶); the 'Selected features' (elements, phys. features¹). Q_{ACM} indicates the RF with charge correction⁴. The purple and green arrows, respectively, indicate the data points from refs. 3 and 9 (modelling on the same dataset). The difference between values a, b, c and a', b', c' highlights a low generalisation ability².

NGBM(1,1) combined with GA had been previously used to model small-sample datasets with high performance^{45,46}. Nevertheless, the authors reported that when trying to apply GANGBM (1,1) to CR series data, the prediction accuracy was only high for historical data, but not for future data (model overfitting)⁵.

A new RGANGBM(1,1) (regularization genetic algorithm non-linear grey Bernoulli model) was proposed for the long-term prediction of CR vs time series, from the 1st, 2nd, 4th, 8th and 16th year. Despite training the model with data from only one station while testing it on five other datasets, the mean prediction accuracy of this method was relatively high (for instance, MAPE (testing) = 9.15% (16th year)). Such an outcome was attributed to the overfitting optimisation via the regularisation parameter η (knowledge mining showed that η could reflect the corrosivity of the environment)⁵.

BPNN and SVR were shown to be unreliable when applied to comparable datasets (same sizes), yielding MAPE (testing) values of 28.02% and 83.55% (16th years), respectively—as indicated for ref. 5 in Fig. 6.

Corrosion-resistant alloys

Corrosion prediction is challenging for CRAs, as corrosion tests require considerable time periods for reaching steady states and appreciable weight losses⁴⁷. Kamrunnahr et al.⁷ proposed a data mining approach based on weight loss measurements of Alloy 22 in Basic Saturated Water at 60–105 °C (this environment intended to mimic the Yucca Mountain area of the United States, which was of great interest for high-level nuclear waste disposal and storage). The authors were able to predict the general corrosion behaviour of Alloy 22 (canister candidates for nuclear waste disposal) by employing ANN on the training set⁷. The achieved performance ($R^2 = 0.8$)⁷ is indicated in Fig. 5 as 'ANN (Alloy 22)'. One of the main outcomes was the observed validity of the Arrhenius rule: the

model would be applicable outside the ranges considered for the Arrhenius parameters, temperature and CR. This work⁷ was elaborated before the Big Data Boom: more data points would be valuable to test the robustness of this ANN approach.

In their early Data mining approach⁶, Kamrunnahr et al. also employed ANN to investigate the general corrosion of metallic glasses with variable compositions. Their objective was to mimic polarisation curves to calculate corrosion potentials and polarisation resistances then. The developed ANN could learn the underlying functions in mapping corrosion variables, alloy characteristics, and environments (HCl, H₂SO₄) to the different corrosion metrics. Although the train-test split procedure used was not specified, the model showed good agreement with experimental data as a function of the changing environment (average R^2 between 0.87 and 1, depending on the data target)⁶. The achieved performances, with corresponding targets (i , E , R_p) and solutions (HCl, H₂SO₄), are indicated in Fig. 5 as a green 'ANN'⁶.

Marine corrosion

As an early example of ML work, Wen et al.⁸ combined SVR with particle swarm optimisation (PSO) for predictive modelling of marine corrosion. Before 2008, Artificial Neural Networks were the main ML modelling tool employed for corrosion studies and had proven to be effective for predicting atmospheric corrosion of steel and zinc¹¹.

Nonetheless, the weak generalisation ability and tendency to overfitting was already known for ANN modelling with small-sample datasets. For instance, this work⁸ demonstrated that the higher the number of training samples (41 or 42 out of 46), the better the regression performance of BPNN (MAPE (testing) of 5.001% or 4.215%, respectively). In comparison to BPNN, a superior generalisation ability was verified for the SVR-PSO,

which resulted in MAPE (testing) values of 3.840%, 3.180% and 1.360% (for 41, 42 and 45 (LOOCV) training samples, respectively). Nonetheless, from a goodness-of-fit point of view, it is worth noting that the SVR-PSO (LOOCV) produced the lowest R^2 (testing) value reported: 0.836. All reported metrics (MAPE and R^2 (testing)) related to the corresponding models⁸ are highlighted in Figs. 5 and 6.

Using sample datasets considered small by today's standards, SVR-PSO presented a high accuracy for predicting the CR of 3C steel under different seawater environments. Nonetheless, the model also showed a limited extrapolation ability. The collection of more experimental data (with larger data ranges) could enhance the model regression accuracy⁸.

More recently, using the same dataset as in ref. ⁸, Chou et al.¹⁸ also attempted the ML modelling of marine corrosion on carbon steel. The prediction included single and ensemble models constructed from four well-known machine learners, including SVR/SVMs, classification and regression tree (CART), and linear regression (LR). Various ML models are typically used to construct single classifiers, but ensemble systems compensate for errors made by the individual classifiers. The tiering ensemble models were shown to improve further the predictive accuracy of SVR (best individual model)¹⁸.

Nature-inspired metaheuristic optimisation algorithms such as the firefly algorithm (FA) have effectively solved difficult optimisation problems⁴⁸. A hybrid model was recently implemented by integrating a smart firefly algorithm (SFA) with a least-squares SVR⁴⁹. Notably, this hybrid SFA-LSSVR model was shown to be effective for modelling CRs of 3C steel, outperforming all other models¹⁸. The MAPE (testing) values achieved for SFA-LSSVR (1.26%) and for all other tested models (SVR, CART + LR, LR, CART + SVR + LR, SVMs + SVR, SVMs + CART)¹⁸ are highlighted in Fig. 6.

Pipeline corrosion

Understanding and predicting the risk of pipeline corrosion is essential for maintaining oil operations healthy and safe. The first step in the Corrosion Risk Assessment of oil and gas companies is selecting the input parameters for estimating the corrosion defect depth amongst many parameters routinely measured¹⁰.

In a quest for estimating the defect depth growth of aged pipelines, Ossai et al.¹⁰ implemented a data-driven ML approach relying on PSO, Feed-Forward Artificial Neural Network (FFANN), Gradient Boosting Machine (GBM), RF and Deep Neural Network (DNN). The influence of PCA-transformed (Principal Component Analysis) variables on the accuracy of the models was also conducted.

Before the application of the models mentioned above, a multivariate polynomial regression was used to fit the entire dataset coming from 60 wells (R^2 (training) = 0.9819, as included in Fig. 5 as 'MP'¹⁰). The objective was to avoid data bias by having new datasets comprising variable scenarios of corrosion severity. Thus, the multivariate polynomial was used for data generation and provided a larger training dataset for the ML algorithms.

A PCA transformation thus highlighted the improvement in the prediction of future states of corrosion defect depth growth. For PCA non-transformed and transformed inputs (respectively), the PSO-FFANN, GBM, RF and DNN models produced training MAPE (%) values of 34.1329, 31.9266, 32.4267, 23.647 and 7.8588, 6.0082, 7.7421, 6.6813¹⁰. All these models/metrics relationships¹⁰ are displayed in Fig. 6.

Supported by the operating parameters of pipelines, this work contributed to understanding the time-dependent changes in the defect depth growth trajectories (over a 10 year period)¹⁰.

Since internal corrosion of pipelines results from an interaction of different mechanisms, a large degree of uncertainty is associated with the prediction of the damage evolution.

De Masi et al.¹³ developed a predictive model to identify pipeline sections exposed to higher corrosion risks (existing approaches have not reproduced behaviours observed during gauging activity). The authors' premise was that combining data with different natures (geometrical profile of pipelines, fluid dynamic multiphase variables and deterministic models) could enormously improve the performance of predictive tasks¹³.

Thereby, by considering several network inputs, an ANN ensemble model could correctly identify the regions of pipelines where corrosion was most likely. This strategy improved not only the prediction performance obtained with deterministic models but also with single ANN models. For instance, the R^2 (testing) values achieved for the ANN and the ANN ensemble were respectively: 0.689 and 0.884 (for CR) or 0.336 and 0.672 (for volume loss). All these performance metrics are mapped in Fig. 5¹³.

Rebar corrosion

The highly nonlinear nature of embedded steel corrosion and the lack of theoretical basis for some corrosion phenomena render the predictive modelling in this topic difficult. In particular, the corrosion initiation of reinforced concrete is a crucial service life parameter, which depends upon the material, the environment, and the duration of exposure. The early detection of corrosion initiation of embedded steel could save cost and time.

To optimise the predictive modelling of corrosion initiation time, Salami et al.¹⁴ compared the performances of different ML techniques with the corrosion potential (measured in 5% NaCl) as the data target. Before the regression modelling, various dataset transformations were applied, revealing no significant effect on the resulting performance (the raw dataset was considered)¹⁴.

This low sensitivity towards feature engineering techniques is aligned with the more recent models, such as deep learning (provided that the amount of training data is high enough)¹⁴.

This finding is in contraction with the results reported in ref. ¹⁰, where the performance of the DNN model clearly increased with PCA transformation.

Next, RF was shown to combine high flexibility and could generalise with high accuracy the corrosion initiation times (R^2 (testing) = 0.897). The effect of split percentage on the performances of RF revealed that the model variance further reduced with an increase in the training dataset, with the highest performance achieved with an 85/15 dataset percentage split (R^2 (testing) = 0.974). These models and corresponding metrics are displayed in Fig. 5¹⁴.

The modelling of rebar corrosion is further complexified when localised corrosion (pitting) occurs, which is often the case in chloride-laden media. The chloride threshold (Cl_{th}) is defined as the chloride content necessary to sustain the breakdown of the passive film and is an essential parameter for assessing the service life of reinforced concrete¹². The literature reports that the current lack of understanding of the Cl_{th} value of steel in concrete is due to its numerous factors affecting pitting corrosion (pH, E of the steel, physical condition of the steel/concrete interface)¹².

Searching to establish a phenomenological model correlating the pitting risk of steel (characterised by E_{corr} - E_{pit}) and environmental factors, Shi et al.¹² applied a modified back propagation (BP) algorithm for ANN training with laboratory data from simulated concrete pore solutions. Only the training performance of the back propagation ANN was reported (R^2 = 0.8741), as indicated by 'BP ANN' in Fig. 5¹².

From knowledge mining, the authors elaborated the following measures to enhance the Cl_{th} of rebar: the initial [Cl], as well as the DO, should be limited, while the percentage of bound Cl⁻ and the pH of concrete should be maximised. However, they cautioned that only the overall trends were observed with independent variables maintained at specific levels¹².

Six years after the publication of the 'pitting corrosion risk of steel rebar' dataset¹², Chou et al.¹⁸ derived this data for testing their advanced SFA-LSSVR model. Unfortunately, no performance comparison could be elaborated between both works^{12,18}, as different evaluation metrics were used.

The SFA-LSSVR model presented a lower predictive performance for rebar pitting risk than the obtained for marine CR reported in the same work (shown in the Marine corrosion topic)¹⁸. Here, as illustrated in Fig. 6 as 'SFA-LSSVR (pit)', the model produced a MAPE (testing) value of 5.60%¹⁸. Such as for the marine corrosion case, although the tiering ensemble models outperformed all individual models, none of them reached reasonable percentage errors such as the one achieved by the hybrid model¹⁸. Further investigations would be envisaged for testing the prediction accuracy of SFA-LSSVR applied to different scenarios.

In 2021, Zhu et al.¹⁹ combined ANN with Kohonen Self-Organized Mapping (KSOM) for modelling the chloride threshold (CT) in reinforced concrete from a sparse literature database. Their KSOM-ANN approach was first applied to find the most probable values for missing data, including primary (pH, corrosion potential, breakdown potential, temperature) and secondary independent variables (cement composition, concrete porosity, water/cement ratio) ('Attributes interdependence' feature in the ML for corrosion database⁴³). The trained KSOM-ANN was validated by comparing the predicted values with the experimental ones in the evaluation database for CR (%TotalCl/cem). Modelling a dataset of 1242 vectors yielded a high training performance ($R^2 = 0.99$, as indicated by 'K-ANN' in Fig. 5)¹⁹.

Next, the network revealed quantitative correlations between the major dependent (CT in three possible definitions) and the corresponding independent variables. For instance, it was found that CT was only poorly correlated with porosity, water/cement ratio, and the Al_2O_3 , Fe_2O_3 , MgO , K_2O and SO_3 oxide contents of the cement. These findings are generally not supported by experimental investigations, thus claiming that CT is a highly scattered property and ambiguous in capturing all variables concerned. Particularly in the case of porosity, the weak correlation seemed to result from an interplay between the volume available to accommodate the same amount of Cl^- and the migration/diffusion effect, which increases with porosity. Nonetheless, for a comprehensive understanding of the correlations obtained, one should consider that they depend on the definition of CT used. Indeed, a given definition of CT does not provide a unique measure of the aggressiveness of a particular concrete system because of simultaneous changes in pH and pK_w (this issue was further addressed in Part II of the series⁵⁰). Interestingly, the CT was independent of the w/cem ratio in all three definitions, for which the lack of correlation was attributed to the weak effect of the w/cem ratio on the bound Cl^- .

Crevice corrosion

Besides being effective when dealing with atmospheric and general corrosion, the ANN model developed by Kamrunnahar et al.⁶ also proved to be efficient for modelling localised corrosion. Based on crevice corrosion data on Ti-2, the ANN predicted the alloy's penetration depth over much longer times than the timescale of the experimental data available. The ANN could perfectly fit the datasets for the 4 outputs considered⁶. The corresponding performance values (R^2 (training) = 1) are illustrated in Fig. 5 as 'ANN (Ti-2)'⁶.

The ANN model developed by Kamrunnahar et al. was not only able to predict the weight loss of the corrosion-resistant Alloy 22 (Nuclear corrosion topic) but also the alloy's localised behaviour⁷. Remarkably, the model mapped the environment (pH, $[Cl^-]$), the alloy condition, and the sample treatment to the crevice repassivation potential. Through ANN weighting, the input

variable with the highest importance was the chloride concentration, which was a very realistic outcome based on domain knowledge. As only the average R squared value was provided (2-fold cross-validation (CV))⁷, this was considered as a training (R^2) performance point in Fig. 5 (indicated by 'ANN'⁷).

Stress corrosion cracking

In 2014, Shi et al.¹⁵ derived a database from the literature to train an ANN to predict the crack growth rate (CGR) in 304 SS. The primary motivation of the authors was to reduce the dispersion data (often higher than two orders of magnitude) associated with the modelling estimation of CGR from sensitised stainless steels. First, ANN sensitivity analysis was done to evaluate the contribution of each independent variable to the dependent variable. Next, the ANN-predicted CGR values were shown to be virtually identical to the ones obtained by the (deterministic) extended Coupled Environment Fracture Model (CEFM) (R^2 (training) = 0.95, as shown in Fig. 5 by 'ANN'¹⁵). Moreover, both the AI model and the deterministic one predicted that CGR increased strongly with increased concentration in O_2 .

Corrosion inhibitors

Benign and efficient corrosion inhibitors are needed for overpassing the restrictions of toxic chromate-based technologies in the aerospace industry.

Winkler et al.¹⁶ combined high throughput experiments with ML to determine the inhibition score (0–10) of corrosion inhibitors for aluminium alloys. The prediction of inhibition performance was conducted over an extensive library (100) of organic molecules tested on AA2024 and AA7075, at pH 4 and 10.

First, ML showed a negligible correlation between quantum chemical properties (calculated by DFT) and inhibition, suggesting that published structure-inhibition models available in the literature were incorrect (these were presumably based on minimum numbers of molecules).

Then, different robust models (including the BRANNGP (feed forward Bayesian regularised neural networks using a Gaussian prior)) were developed, linking molecular properties calculated by non-quantum chemical methods and inhibition. The BRANNGP model overperformed the MLREM model (multiple linear regressions), increasing the R^2 (training, 80/20 split) values from 0.79, 0.56, 0.75 to 0.82, 0.60, 0.82, for AA7075 (pH 4), AA7075 (pH 10), AA2024 (pH 4), respectively. The only exception was the AA2024 (pH 10) system, where the MLREM produced a higher R^2 (training, 80/20 split) than the BRANNGP (0.76 and 0.74, respectively)¹⁶. All these metric relationships are exhibited in Fig. 5 by the models' names (with respective alloy and solution)¹⁶.

In terms of the number and diversity of inhibitors, range of alloys and pHs, this work¹⁶ stands out as one of the most comprehensive corrosion modelling studies. Besides, many descriptors (~2000) were obtained, and the ones made available were employed in a subsequent ML work on the classification of inhibitors⁵¹. Nevertheless, the authors referred to the relatively opaque or arcane nature of the ensemble methods, which renders difficult the mechanistic interpretation of the phenomena¹⁶.

Conversion coatings

In a recent design-oriented work¹⁷, a new designing strategy, i.e. the ratio of the total acidity (TA) and pH value (TA/pH) was proposed to validate the chemical formulation of a phosphate-based conversion coating. The results indicated that ANN could not fit the corrosion performance solely with the pH or the TA predictor, presenting a relatively accurate prediction based on TA/pH¹⁷.

In Fig. 5, the average R squared obtained are presented as R^2 (training) values, with 0.19, 0.49 and 0.86 for pH, TA and

TA/pH, respectively¹⁷. Furthermore, the corrosion performance of the phosphate conversion coatings (PCC) increased with the decrease in TA/pH. However, it was impossible to produce a coating with good corrosion performance in pure water (TA/pH equal to 0). The authors explained that a critical value of TA/pH should be the boundary of the proposed theory of TA/pH. They also pointed out the complexity of decreasing TA/pH under the premise of maintaining the stability of the conversion bath and concluded that the validity of using TA/pH in large pH ranges should be further discussed¹⁷.

High-temperature oxidation

Although the data analysis conducted in this Review is entirely based on research papers focusing on electrochemical corrosion, two recent articles covering a correlated topic (ML applied to high-temperature oxidation)^{52,53} are worth the assessment.

In the first of these two publications⁵², ML was used to predict the parabolic rate constant (k_p) for high-temperature oxidation of Ti alloys. Three different regression ML models (Gradient Boosting, Random Forest, k-Nearest Neighbours) were applied to a dataset of k_p values (115 data points) built from experimental studies reported in the literature. Several descriptors (alloys' composition, constituent phase, temperature of oxidation, time for oxidation, O_2 content, moisture content, remaining atmosphere, mode of oxidation) were selected as independent variables. The gradient boosting model produced the highest achieved accuracy ($R^2 = 0.92$). The learning from this work could be extended to design novel Ti alloys with enhanced resistance against high-temperature oxidation⁵².

In the second publication⁵³, 237 unique k_p values were extracted from publications, calculated from corrosion product data and structured in a comprehensive database. The high-temperature oxidation dataset comprised entries from various environments (exposed temperatures between ~500 and 1700 °C) from 75 alloys, including low- and high-Cr ferritic/austenitic steels, nickel superalloys, aluminide materials. Several modelling strategies (including supervised ML) were implemented to explore the relationships between composition and oxidation kinetics. The modelling of materials based on their k_p values presented the lowest errors achieved when supervised ML (logistic regression, linear regression, RF, decision tree, neural network) was employed. Furthermore, the dominant elements encountered to control the oxidation kinetics were Ni, Cr, Al and Fe (Mo and Co were also considered relevant descriptors)⁵³.

MODELLING APPROACHES: CRITICAL ANALYSIS

One of the most commonly encountered findings against good practices in ML is the use of small training datasets, which potentially leads to overfitting. For instance, BPNN and SVR were unreliable when applied to datasets with low sizes⁵. In ref. ⁷ ('Crevice corrosion' section), the small sample dataset was the main drawback of the study. The small size of the training sample for the network modelling was also critical in ref. ¹³. Conversely, the MP regression was most likely only efficient because dealing with such a large input dataset (8300 samples)¹⁰. Regrettably, the final data dimension upon application of PCA was not presented¹⁰.

Another issue with small datasets is that appropriate data splitting methods become less feasible. For example, due to the small sample size (only 8 samples), the data was not divided into training and testing sets in ref. ⁶ ('Crevice corrosion' section). The training-testing split method was unfortunately not clear⁶ ('Atmospheric corrosion' section): only one type of steel seems to have been included in the testing set (the training performance was not presented). In ref. ⁷ ('Corrosion-resistant alloys' section), the same

data was used for training-testing (the employed cross-validation method was not specified).

Not only the model's generalisation ability is essential, but also its extrapolation ability. Hence, to improve the prediction accuracy, more relevant data within appropriate ranges would be needed. In the case of ref. ², this would imply more samples with corrosion rates higher than 0.050 mm per year. As reported in ref. ¹³, the prediction accuracy could be enhanced by considering larger datasets comprising several pipeline cases and flow conditions. In ref. ¹⁴, tests were performed under laboratory conditions only: the validation of the RF method for the prediction of concrete corrosion would be recommended in in-field samples. Similarly, further study would be required to explore the broader applicability of the SFA-LSSVR¹⁸. Unfortunately, the electrochemical methods employed for data acquisition in ref. ⁸ are unknown.

In ref. ¹, features such as the flow velocity or microorganisms, also seriously influencing the corrosion rate of LAS in marine environments, were not explored. The investigation³ lacks features describing the material microstructure. Unfortunately, the paper⁹ did not discuss the effects of the composition or the microstructure. An illustrative example is the pitting risk of steel in simulated concrete pore solutions¹²: modelling from the referred predictors would be more complex than the initially conceived. Although the flow velocity was not included in this analysis¹⁵, theory indicates that this operational variable should be considered in SCC modelling. When assumptions are oversimplified, the model might be susceptible to underfitting. Increasing the data dimension with properly selected variables (based on domain knowledge) might improve further the model performance.

Finally, only one referenced paper¹⁹ formally addressed the critical property of attributes interdependence. In there, as the corrosion potential (ECP) is also a function of T , pH and $[Cl^-]$, the trends in CT with these independent variables must consider their interdependence with ECP¹⁹. Input independence leads to a more accurate prediction when there is enough information to describe data distribution in feature space. However, adding variables that depend on others and contain newer and richer information would improve the model accuracy and robustness.

PERFORMANCE OF THE ML MODELS: CONCLUSIONS

Not all articles used the same training, validation, and testing datasets, bringing extra uncertainty to comparing performances. As clearly visible in the plots dividing the training and testing performances, the testing performances are sometimes not reported, contrary to the good practices in the field. For the sake of easing comparisons, it would also be recommended to clearly report the number of learnable parameters/hyperparameters between models⁵⁴.

Nonetheless, one could apprehend from Figs. 6 and 7 that the linear regression algorithms have difficulties in dealing with nonlinear datasets and/or with high dimensionality data. Kernel-based algorithms are effective when dealing with nonlinear and/or sparse data but have a low training efficiency with large datasets^{26,27}. The SVR models' performances are associated with points with relatively high scatter, reflecting their sensitivity to the Kernel selection (which could lead to training failure).

On the other hand, neural networks are more training expensive (vast number of tuning parameters) but provided there is enough data available, they might yield the best possible quality fitting²⁷. This observation corresponds well with the performances observed for BPNN and/or ANN (Figs. 6 and 7). Neural networks were developed for modelling sequence data, and the recurrent method (most of the ANN models in the ML for corrosion database⁴³) is especially suitable for time series data (Fig. 4). For example, in Fig. 7, ANN presented a lower predictive performance on corrosion depth (static data)¹¹ than on CR (temporal data)^{3,9}.

The Tree-based methods can handle missing data and datasets with multiple features but can be prone to instabilities^{24,26,40}. As explicit in Figs. 6 and 7, the considerable divergence between training and testing performances might reflect their overfitting characteristic.

Deep learning models (for example, ANN with multiple hidden layers) have high robustness, accuracy, and reliability when dealing with highly nonlinear and complex relationships between the input and output data^{26,40}. The main drawback of deep learning, which requires a large amount of training data, is most likely to be surmounted as we enter the Big Data and the internet of things (IoT) era⁴⁰.

The ensemble (hybrid) methods cited accomplished predictive tasks with superior performance by combining various weak learners. Worthy mentioning the regression trees (weak learner) was efficiently used to build models (GBDT, XGBoost) for predicting the weight loss of LAS².

Modelling with ensemble methods (SVR-PSO, RF-WKNNs, PSO-FFANN, SFA-LSSVR, KSOM-ANN) or deep learning (cForest, DCCF-WKNNs, DNN, and the most ANNs here addressed) generally produced the highest observed performances. Both approaches could certainly solve possible contradictions between the high dimensionality and the small sample often observed in electrochemical data^{21,40}.

CHALLENGES

A central challenge in machine learning is the conflict between the complexity and accuracy of the models²¹. This point is particularly true in the case of deep learning because these models usually have more complicated structures (structural designing becomes vital)²⁶. Such as illustrated in Fig. 7, hybrid models with high complexity (XGBoost, RF-WKNNs, cForest, DCCF-WKNNs) could yield better performance than all the others. However, concerns related to their scalability exist, especially when non-parametric methods (tree-based, CARTs, SVMs, KNNs) are included⁴⁰.

From a Review on ML approaches for energy systems⁴⁰, the need to develop models that are generalisable to wider settings is claimed. The authors also point out the lack of modelling details across a vast part of the reviewed literature. Notably, the number of times the models were trained are rarely reported⁵⁴.

As it is complicated to translate the ML black box mechanics to formulas, many efforts have been devoted to developing more interpretable algorithms^{24,55}, including the rule extraction (express the implied knowledge with patterns that are easy to understand)²¹ and argument based machine learning (ABML), which improves the comprehensibility of the learned concepts³¹.

The learning achieved by ML frequently conflicts with domain expert knowledge. As much as for the batteries field, it is also imperative to establish AI approaches encompassing domain knowledge in the corrosion field²¹. For example, from this work on atmospheric corrosion², each feature's significance to the CR was ranked based on the Pearson correlation coefficient and MIC (maximal information coefficient) methods. Although the SO₂ feature was not indicated as dominating factor, the authors decided to include it based on domain knowledge.

Although ML is a powerful tool for finding statistical patterns hidden behind multi-dimensional data, condensing the extracted knowledge into scientific laws for general cases is challenging²⁴. When the expertise from the corrosion field is incorporated into the modelling already during the definition of problems, it often leads to the best results³¹.

Typical algorithms that could integrate domain knowledge and ML methods include Bayesian network, fuzzy learning²¹ and argument based machine learning (ABML enables the knowledge to be articulated easily)³¹.

The availability of high-quality datasets is a primary challenge in ML for corrosion and is separately addressed in the Data sharing section.

GUIDELINES

The success of ML modelling most likely relies on the use of large datasets (Fig. 3), but even more on the proper selection of features^{23,24}. Incorporating basic descriptors based on domain knowledge often simplifies the representation and improves the interpretability of machine learning modelling²¹. As highlighted in this Review (Fig. 4), the selection of temporal variables positively affects the overall models performance. Alternatively, ML could be applied to capture the remaining divergence between the target properties and the outputs from deterministic modelling²⁷.

Such as in the case of this addressed study¹³, where data from different natures were considered (in-field measurements, software simulation and deterministic models), ML should be coupled more tightly with other approaches (experiments, numerical simulations, deterministic models)²⁶. For instance, as previously described⁶, ANN could validate the high fidelity prediction of a deterministic model (CEFM) in capturing the electrochemical/mechanical/microstructural phenomena of crack growth rates.

In a Review paper on energy materials published in 2021²⁶, the authors emphasised that the application of ML should not be limited to the predictive tasks but should encompass all data-related stages, including data-preprocessing and data mining. Text mining technologies, including full-text publisher application programming interfaces (APIs) and natural language processing (NLP)⁵⁶, have been extensively used in materials science and chemistry²⁴.

Moreover, ML could be an intermediate step for processing the collected datasets, thus reducing redundant experiments (or simulations)²⁶. Ensemble learning strategies could be applied in the training phase to increase the model efficiency and robustness.

DATA SHARING: INITIATIVES AND RECOMMENDATIONS

Unlike traditional hard-coded modelling based on human expertise, ML approaches rely on available datasets for building reliable models. For accelerated progress of corrosion predictive modelling, collaborative culture of sharing of knowledge and resources (data, models, software) between the computer science and corrosion fields should be fostered⁵⁷.

In the era of big materials data, considerable efforts have been made to collect and build materials properties databases^{24,58}. In their review on machine learning for molecular and materials science, Butler et al.⁵⁹ listed the publicly accessible machine learning resources and tools. The available ML tools related to molecules and materials might be of great interest for corrosion research. The Materials Project and the Open Quantum Materials Database (OQMD) (referred to in ref. ²⁷) also provides an ideal platform for machine learning.

The 'Data availability' feature of the Machine learning for corrosion database (Supplementary Notes) provides links to high quality publicly available datasets that can be downloaded for ML modelling. Atmospheric corrosion is the most studied topic, thanks to the good availability of atmospheric/climatic data obtained following standardised approaches.

Besides sharing datasets, researchers should ideally open their modelling codes and libraries as much as possible, thus providing a viable path for the transferability of their models^{25,30}. Among the papers considered in the Machine learning for corrosion database⁴³, only one investigation¹⁶ made available related codes (see the 'Libraries' feature (Supplementary information)).

However, electrochemical data is often not shared in a systematic and machine-readable way^{25,27}. The research community is also

concerned with how to build/optimize such databases in a more automated fashion.

Data standards and tools for automated organisation are needed to structure the data in a reusable fashion. Such initiatives have paved the way for innovation in battery technologies (through wider availability of high-quality databases)^{25,27,30}. It would be recommended not to let corrosion data buried in figures only but also to present it within structured Tables that are readily accessible. In particular, reporting more failed data used during training and data from negative experiments would accelerate this process^{24,27}.

RESEARCH EVOLUTION AND PERSPECTIVES

Training ML models with large datasets presenting comparable features across wide ranges of values is paramount for accurate prediction^{2,24,26,54}.

In a Review paper on rechargeable batteries²¹, virtual sample generation (VSG) methods are cited as a solution for filling in data gaps. As a similar example from the selected papers on pipeline corrosion, the multivariate polynomial employed by Ossai et al.¹⁰ generated defect depth values to complement their training dataset.

Another method for constructing training sets that might be particularly useful for corrosion studies based on small sample data is active learning (selecting effective samples by iterative sampling that ensures the highest possible accuracy²¹). Active learning has been considered a promising approach for learning the high-dimensional data of battery materials²¹. Similarly, in a recent review paper on ML and energy storage, Gao et al.²⁶ refer to an unsupervised learning (generative adversarial network, GAN) capable of reconstructing/repairing datasets, thus enhancing their quality. The GAN approach might be beneficial for in-field corrosion measurements, where the quality and integrity of the data are often compromised.

Finally, transfer learning technology has been demonstrated to improve the applicable range of predictive models^{23,26}. From the review on deep learning on materials degradation³⁹, Nash et al. mentioned the capability of transfer learning in fine-tuning models for tasks other than the ones they were trained on.

Although not the focus of this paper, machine learning has accelerated the discovery of new materials^{21–23}. In the development of batteries, Attia et al.⁵² employed a ML method that could reduce the time of performance tests by 98%. A similar approach would be instrumental in corrosion research because testing in simulated environments usually lasts weeks or even months (corrosion acceleration tests). This time-consuming aspect has become a limiting factor for the long-term study of corrosion, particularly atmospheric corrosion, where testing protocols often take years to complete.

Seeking at discovering the links between material composition, structure, properties and performance, scientists have input various property features to machine learning^{21,28}. For example, ML methods have been applied to property prediction for energy conversion materials, potentially overcoming the shortcomings of DFT (high consumption of computational resources)²⁴. For instance, ML of quantum-mechanical reference data has been used for fast and accurate interatomic potentials of battery materials³⁰. ML has indeed the potential to bridge new synergies between experimental data and DFT modelling.

In their comprehensive study on the machine learning of corrosion inhibitors, Winkler et al.¹⁶ assumed that molecular properties permitted more reliable models than quantum chemical properties of the individual compounds. In an ensuing investigation based on classification algorithms⁵¹, the structure-property relationships in aluminium alloy inhibitors were further elucidated. Galvão et al.⁵¹ employed new descriptors related to the self-association between inhibitors (dimerisation enthalpies

and Gibbs energies), gaining further insights into the properties that distinguish these molecules.

This bibliometric data mining revealed not only the low number of ML works focusing on corrosion inhibitors (or other vital topics such as protective coatings or localised corrosion) but also identified several research gaps in which ML could be envisaged: EIS, local electrochemical techniques, microbial corrosion, corrosion of electronics, tribocorrosion.

To decrease the usage complexity of ML, automated (Auto) ML has attempted to limit human intervention in the model implementation²¹. In this context, meta-learning or 'learning to learn' methods have emerged and shown to be particularly useful in image recognition²⁴. The task of hyperparameters' optimisation of nonlinear models (ANN, SVR) is complex and often relies on trial and error. Hyperparameter optimisation methods have been proposed to simplify these tuning processes²¹, including Auto ML⁶⁰. From the selected papers, two investigations implemented distinct approaches to address this issue: the SFA was proven to be effective in dynamically optimise the LSSVR hyperparameters¹⁸, and the genetic algorithm NGBM (1,1) included a regularisation term η to avoid the overfitting problem⁵.

A variation of ensemble method called Temporal Ensemble Learning (TEL), in which temporal features portion the dataset for the forecast at specific time ranges⁴⁰, might be a promising alternative for time series data in corrosion.

Although probabilistic (Gaussian processing regression) models have shown promising forecasting results in the energy and battery domains^{26,40}, as far as the authors' knowledge on supervised learning goes, probabilistic regression tools have not yet been considered for corrosion research. All in all, hybrid models and holistic modelling approaches transcending time, length and mechanism scales point towards fruitful directions.

Unlike supervised and unsupervised learning, Reinforcement Learning (RL) can consider the trade-off between exploration and exploitation. RL is one of the most promising computational ML approaches by interacting with an unknown environment while explicitly focusing on goal-oriented learning⁴⁰.

CONCLUSIONS

This data-driven review work depicts the current status of the emerging research on predictive machine learning applied to corrosion.

A bibliometric data mining approach led to creating a Machine learning for corrosion database, covering 34 works where relative metrics did the performance evaluation. Exploratory data analysis demonstrated the positive effects of increasing data dimension and including time as an input on the models performance. A comparative performance analysis was conducted, segmenting the references by corrosion topic, orientation strategy, type of ML model and data targets. In particular, expanding the types of variables (orientation strategies comprising data from multiple relevant sources) likely increases the models performance. The plots bridging the multi-dimensional data to the references serve as a roadmap for future regression modelling of corrosion. More studies focusing on the prediction of localised corrosion and inhibition efficiency should be considered in the future.

Accurate and reliable modelling requires large sets of training (and validation) data of high quality. Moreover, the process of labelling corrosion features should be based on the experience and intuition of domain researchers. Despite the challenge of establishing direct comparative analysis on the performances, this reviewing exercise allowed identifying research gaps and evaluating common practices in the field. For more systematic incorporation of ML within the corrosion community, concerted international efforts are needed to promote data sharing practices. Domain experts would benefit far more from the literature if the modelling datasets were made available alongside the related publications.

DATA AVAILABILITY

The processed data required to reproduce these findings are available to download from Bertolucci Coelho, Leonardo (2021), 'Machine learning for corrosion database', Mendeley Data, V1, <https://doi.org/10.17632/jfn8yhrphd.143>, permanently stored at the repository via the link <https://data.mendeley.com/datasets/jfn8yhrphd/1>.

CODE AVAILABILITY

The code required to reproduce these findings is available to download from GitHub: https://github.com/bcoelho-leonardo/Machine-learning-for-corrosion/blob/77c05253447d4bca737a126f678391e0f98108df/Machine%20learning%20for%20corrosion%20database_codes.ipynb.

Received: 7 September 2021; Accepted: 22 December 2021;
Published online: 26 January 2022

REFERENCES

- Diao, Y., Yan, L. & Gao, K. Improvement of the machine learning-based corrosion rate prediction model through the optimization of input features. *Mater. Des.* **198**, 109326 (2021).
- 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).
- Zhi, Y., Fu, D., Zhang, D., Yang, T. & Li, X. Prediction and knowledge mining of outdoor atmospheric corrosion rates of low alloy steels based on the random forests approach. *Metals* **9**, 383 (2019).
- Pei, Z. et al. Towards understanding and prediction of atmospheric corrosion of an Fe/Cu corrosion sensor via machine learning. *Corros. Sci.* **170**, 108697 (2020).
- Zhi, Y. et al. Long-term prediction on atmospheric corrosion data series of carbon steel in China based on NGBM(1,1) model and genetic algorithm. *Anti-Corros. Method M* **66**, 403–411 (2017).
- Kamrunnagar, M. & Urquidi-Macdonald, M. Prediction of corrosion behavior using neural network as a data mining tool. *Corros. Sci.* **52**, 669–677 (2010).
- Kamrunnagar, M. & Urquidi-Macdonald, M. Prediction of corrosion behaviour of Alloy 22 using neural network as a data mining tool. *Corros. Sci.* **53**, 961–967 (2011).
- Wen, Y. F. et al. Corrosion rate prediction of 3C steel under different seawater environment by using support vector regression. *Corros. Sci.* **51**, 349–355 (2009).
- Zhi, Y., Yang, T. & Fu, D. An improved deep forest model for forecast the outdoor atmospheric corrosion rate of low-alloy steels. *J. Mater. Sci. Technol.* **49**, 202–210 (2020).
- Ossai, C. I. A data-driven machine learning approach for corrosion risk assessment —a comparative study. *Big Data Cogn. Comput.* **3**, 28 (2019).
- Cai, J., Cottis, R. A. & Lyon, S. B. Phenomenological modelling of atmospheric corrosion using an artificial neural network. *Corros. Sci.* **41**, 2001–2030 (1999).
- Shi, X., Anh Nguyen, T., Kumar, P. & Liu, Y. A phenomenological model for the chloride threshold of pitting corrosion of steel in simulated concrete pore solutions. *Anti-Corros. Method M* **58**, 179–189 (2011).
- De Masi, G., Gentile, M., Vichi, R., Bruschi, R. & Gabetta, G. Machine learning approach to corrosion assessment in subsea pipelines. in *OCEANS 2015 - Genova 1–6* (IEEE, 2015). <https://doi.org/10.1109/OCEANS-Genova.2015.7271592>
- Salami, B. A., Rahman, S. M., Oyehan, T. A., Maslehuddin, M. & Al Dulaijan, S. U. Ensemble machine learning model for corrosion initiation time estimation of embedded steel reinforced self-compacting concrete. *Measurement* **165**, 108141 (2020).
- Shi, J., Wang, J. & Macdonald, D. D. Prediction of crack growth rate in Type 304 stainless steel using artificial neural networks and the coupled environment fracture model. *Corros. Sci.* **89**, 69–80 (2014).
- Winkler, D. A. et al. Using high throughput experimental data and in silico models to discover alternatives to toxic chromate corrosion inhibitors. *Corros. Sci.* **106**, 229–235 (2016).
- Chunyan, Z. et al. Ratio of total acidity to pH value of coating bath: a new strategy towards phosphate conversion coatings with optimized corrosion resistance for magnesium alloys. *Corros. Sci.* **150**, 279–295 (2019).
- Chou, J., Ngo, N. & Chong, W. K. The use of artificial intelligence combiners for modelling steel pitting risk and corrosion rate. *Eng. Appl. Artif. Intell.* **65**, 471–483 (2017).
- Zhu, Y., Macdonald, D. D., Qiu, J. & Urquidi-Macdonald, M. Corrosion of rebar in concrete. Part III: Artificial Neural Network analysis of chloride threshold data. *Corros. Sci.* **185**, 109438 (2021).
- Materials Genome Initiative. <https://www.mgi.gov/> (2021).
- Liu, Y., Guo, B., Zou, X., Li, Y. & Shi, S. Machine learning assisted materials design and discovery for rechargeable batteries. *Energy Storage Mater.* **31**, 434–450 (2020).
- Scully, J. R. & Balachandran, P. V. Future frontiers in corrosion science and engineering, part III: the next “Leap Ahead” in corrosion control may be enabled by data analytics and artificial intelligence. *Corrosion* **75**, 1395–1397 (2019).
- Luo, Z. et al. A survey of artificial intelligence techniques applied in energy storage materials R&D. *Front. Energy Res.* **8**, 1–12 (2020).
- Chen, A., Zhang, X. & Zhou, Z. Machine learning: accelerating materials development for energy storage and conversion. *InfoMat* **2**, 553–576 (2020).
- Aykol, M., Herring, P. & Anapolsky, A. Machine learning for continuous innovation in battery technologies. *Nat. Rev. Mater.* **5**, 725–727 (2020).
- Gao, T. & Lu, W. Machine learning toward advanced energy storage devices and systems. *iScience* **24**, 1–33 (2021).
- Ng, M.-F., Zhao, J., Yan, Q., Conduit, G. J. & Seh, Z. W. Predicting the state of charge and health of batteries using data-driven machine learning. *Nat. Mach. Intell.* **2**, 161–170 (2020).
- Wei, X. et al. Data mining to effect of key alloying elements on corrosion resistance of low alloy steels in Sanya seawater environment Alloying Elements. *J. Mater. Sci. Technol.* <https://doi.org/10.1016/j.jmst.2020.01.040> (2020).
- Li, S., Li, J., He, H. & Wang, H. Lithium-ion battery modeling based on Big Data. *Energy Procedia* **159**, 168–173 (2019).
- Deringer, V. L. Modelling and understanding battery materials with machine-learning-driven atomistic simulations. *J. Phys. Energy* **2**, 041003 (2020).
- Mozina, M., Guid, M., Krivec, J., Sadikov, A. & Bratko, I. *Fighting Knowledge Acquisition Bottleneck With Argument Based Machine Learning*. 234–238, <https://doi.org/10.3233/978-1-58603-891-5-234> (2008).
- Cai, Y., Xu, Y., Zhao, Y. & Ma, X. Atmospheric corrosion prediction: a review. *Corros. Rev.* **38**, 299–321 (2020).
- Cai, Y., Xu, Y., Zhao, Y. & Ma, X. Extrapolating short-term corrosion test results to field exposures in different environments. *Corros. Sci.* **186**, 109455 (2021).
- Feliu, S., Morcillo, M. & Feliu, S. The prediction of atmospheric corrosion from meteorological and pollution parameters—I. Annual corrosion. *Corros. Sci.* **34**, 403–414 (1993).
- Chan, V. Degradation-based reliability in outdoor environments. <https://doi.org/10.31274/rtd-180813-12114> (Iowa State University, Digital Repository, 2001).
- Mikhailov, A. A., Tidblad, J. & Kucera, V. The classification system of ISO 9223 standard and the dose–response functions assessing the corrosivity of outdoor atmospheres. *Prot. Met.* **40**, 541–550 (2004).
- Klinesmith, D. E., McCuen, R. H. & Albrecht, P. Effect of environmental conditions on corrosion rates. *J. Mater. Civ. Eng.* **19**, 121–129 (2007).
- David, P. K. & Montanari, G. C. Compensation effect in thermal aging investigated according to Eyring and Arrhenius models. *Eur. Trans. Electr. Power* **2**, 187–194 (2007).
- Nash, W., Drummond, T. & Birbilis, N. A review of deep learning in the study of materials degradation. *npj Mater. Degrad.* **2**, 37 (2018).
- Antonopoulos, I. et al. Artificial intelligence and machine learning approaches to energy demand-side response: a systematic review. *Renew. Sustain. Energy Rev.* **130**, 109899 (2020).
- van Eck, N. J. & Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **84**, 523–538 (2010).
- Hashemi, S. J. et al. Bibliometric analysis of microbiologically influenced corrosion (MIC) of oil and gas engineering systems. *Corrosion* **74**, 468–486 (2018).
- Bertolucci Coelho, L. Machine learning for corrosion database, Mendeley Data. <https://doi.org/10.17632/jfn8yhrphd.1> (2021).
- China Gateway to Corrosion and Protection. <http://data.ecorr.org/> (2021).
- Hsu, L.-C. A genetic algorithm based nonlinear grey Bernoulli model for output forecasting in integrated circuit industry. *Expert Syst. Appl.* **37**, 4318–4323 (2010).
- Ma, D. & Bai, H. Groundwater inflow prediction model of karst collapse pillar: a case study for mining-induced groundwater inrush risk. *Nat. Hazards* **76**, 1319–1334 (2015).
- Donaldson, L. Metallic glass-based materials in wearable energy storage devices. *Mater. Today* **36**, 3 (2020).
- Yang, X.-S. Firefly Algorithms for Multimodal Optimization. 169–178, https://doi.org/10.1007/978-3-642-04944-6_14 (2009).
- Chou, J.-S. & Pham, A.-D. Smart artificial firefly colony algorithm-based support vector regression for enhanced forecasting in civil engineering. *Comput. Civ. Infrastruct. Eng.* **30**, 715–732 (2015).
- Zhu, Y., Macdonald, D. D., Yang, J., Qiu, J. & Engelhardt, G. R. Corrosion of rebar in concrete. Part II: Literature survey and statistical analysis of existing data on chloride threshold. *Corros. Sci.* **185**, 109439 (2021).
- Galvão, T. L. P., Novell-Leruth, G., Kuznetsova, A., Tedim, J. & Gomes, J. R. B. Elucidating structure-property relationships in aluminum alloy corrosion inhibitors by machine learning. *J. Phys. Chem. C* **124**, 5624–5635 (2020).
- Attia, P. M. et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature* **578**, 397–402 (2020).

53. Bhattacharya, S. K., Sahara, R. & Narushima, T. Predicting the parabolic rate constants of high-temperature oxidation of Ti alloys using machine learning. *Oxid. Met.* **94**, 205–218 (2020).
54. Vidal, C., Malysz, P., Kollmeyer, P. & Emadi, A. Machine learning applied to electrified vehicle battery state of charge and state of health estimation: state-of-the-art. *IEEE Access* **8**, 52796–52814 (2020).
55. Yang, L., Wang, P., Jiang, Y. & Chen, J. Studying the explanatory capacity of artificial neural networks for understanding environmental chemical quantitative structure–activity relationship models. *J. Chem. Inf. Model.* **45**, 1804–1811 (2005).
56. Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
57. Li, X. et al. Materials science: share corrosion data. *Nature* **527**, 441–442 (2015).
58. Liu, S. et al. An infrastructure with user-centered presentation data model for integrated management of materials data and services. *npj Comput. Mater.* **7**, 1–8 (2021).
59. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
60. Huang, J., Sun, W. & Huang, L. Deep neural networks compression learning based on multiobjective evolutionary algorithms. *Neurocomputing* **378**, 260–269 (2020).

ACKNOWLEDGEMENTS

The author L.B. Coelho is a Postdoctoral Researcher of the Fonds de la Recherche Scientifique – FNRS which is gratefully acknowledged.

AUTHOR CONTRIBUTIONS

L.B.C.: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing - original draft. D.Z.: Validation, investigation, writing - review & editing. Y.V.I.: Validation, writing - review & editing. D.S.: Software, validation, data curation, writing - review & editing. A.N.: Validation. H.T.: Conceptualization, writing - review & editing, supervision.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41529-022-00218-4>.

Correspondence and requests for materials should be addressed to Leonardo Bertolucci Coelho.

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) 2022, corrected publication 2022