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 materials21–23.

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 data21,29, 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 processes21,24,30. 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 experimental and computer-oriented scientists to apply ML to 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...
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 (R²)). 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 database [13] 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 database [43], which is available to download at Mendeley data (related codes available at GitHub).

Based on the ‘Selected features’ attribute in our ML for corrosion database [43], 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-case 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 sizes [21].

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 trend to be the mean percentage errors, either during training or testing. The associated mean R² 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 R², the low mean

---

**Fig. 1** Visualization of the co-citation network of ‘machine learning’ and ‘corrosion’ publications. Citation relations from 2017 to 2020 (VOSviewer [41]).
| Ref. | Electrolytic | Climate | Composition | Metallurgical/structural | Atomic (physical)/chemical | Electrochemical System | Temporal | Spatial |
|------|--------------|---------|-------------|-------------------------|---------------------------|------------------------|----------|---------|
| 1a   | O₂_max, S_max, pH_min, T_mean | Content (LAS): Cr, C, Si, P, S | EN₀, K, R_m, R_p | V, λ, EN₀, ΔH₀ | t | |
| 2    | T_MAX, T_AVE, RH_MIN, PRECIPIT, SOLAR, Cl⁻, SO₂ | Total content of alloying elements (LAS) | | | t | |
| 3    | RH_mean, T_mean, Rainfall, SO₂, pH of rain, Cl⁻ | Content (LAS): Mn, S, P, Si, Cr, Cu, Ni | | | t | |
| 4    | T, RH, rainfall | | Electrical quantity (Q_{ACM}) | | | |
| 5    | | Steel grade (A3, 3C, 20, 08Al) | CRs (1st, 2nd, 4th, 8th year) | | Station (Jiangjin, Wuhan) | |
| 6    | HCl, H₂SO₄ | 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, SO₂, pH of rain, Cl⁻ | Content (LAS): Mn, S, P, Si, Cr, Cu, Ni, Fe | | t | |
| 10   | pH, SO₄, Cl, Fe, HCO₃, Ca, temperature | | | PS, PCO₂, BSW, BOPD, BWPD, MMCFD | Exposure time | |
| 11   | SO₂, Cl⁻, temperature, TOW (time of wetness) | | | | |
| 12   | Cl⁻ (Cl⁻ concentration), chloride binding, pH, DO (dissolved oxygen concentration) | CR from deterministic models (de Waard, NORSOK) | Elevation, inclination, concavity, conc./ind., dist_failure, flow regime, hold-up, | | |
| 13   | | | | | |
| Ref | Electrolytic | Climate | Composition | Metallurgical/structural | Atomic (physical)/chemical | Electrochemical System | System | Temporal | Spatial |
|-----|--------------|---------|-------------|-------------------------|--------------------------|------------------------|--------|----------|---------|
| 14  | Concrete mixture design proportions: cement (CE), limestone powder (LSP), coarse aggregate (CA), fine aggregate (FA), water (WA) | | | | | | | 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<sup>a</sup> | | | |
| 17  | pH, Total Acidity, TA/ pH | | | | | | | |
| 18  | Total Cl<sup>-</sup> (Cl<sub>TT</sub>) Cl<sup>-</sup> binding, DO, salinity, pH, temperature | | | | | | | Corrosion potential |
| 19  | pH, temperature | SiO<sub>2</sub>, Al<sub>2</sub>O<sub>3</sub>, Fe<sub>2</sub>O<sub>3</sub>, CaO, MgO, Na<sub>2</sub>O, K<sub>2</sub>O, SO<sub>3</sub> (cement) | Water/cement ratio, porosity | | | Electrochemical corrosion potential – ECP (0.5 ppm, 1 ppm, 5 ppm, 8 ppm [O2]), passivity breakdown potential - Eb | | | |

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

<sup>a</sup>Reference<sup>1</sup> also deals with 9 principal data components (variable with no physical meaning).
<sup>b</sup>EN<sub>s</sub> = Sanderson electronegativity; K = Bulk modulus; R<sub>m</sub> = molecular single bond covalent radius; R<sub>c</sub> = covalent radius; V = molar volume; λ = thermal conductivity; EN<sub>p</sub> = Pauling electronegativity; ΔHf = enthalpy of fusion.
<sup>c</sup>1515 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).
MAPE values obtained for dataset sizes smaller than $10^3$ might result from overfitting.

**The effect of the Time variable**

Figure 4 reveals that the models’ performance (expressed by $R^2$) 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 database) was evaluated either by the $R^2$ 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. 6).\(^{18}\)

The ‘Material & environment’ orientation strategy has been the most popular approach, encompassing corrosion-resistant alloys (CRAs)\(^7\), inhibitors\(^16\) 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 descriptors\(^19\) clearly outperformed both material\(^14\) and environment-oriented\(^12,18\) approaches. A system-orientated strategy has been required for pipeline corrosion (including geometrical pipeline characteristics\(^3\)), with increased goodness of fit when instrumental variables (e.g. operating pressures\(^10\)) 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), SVMs) have generated higher prediction performances\(^3,6,9\). 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).

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 timescales\(^{26}\).

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.
testing performance values are displayed in Figs. 6 and 7. Concerning steel, knowledge mining showed that the relation between TOW, SO₂ concentration and atmospheric corrosion was almost linear. By applying this approach on corrosion data of 17 low alloy steels under atmospheric corrosion test stations in China over 16 years, 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. Notably, the SO₂ and Cl⁻ concentrations had higher importance 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 SO₂ importance was always greater than 80%). Notably, the SO₂ and Cl⁻ concentrations had higher importance than the temperature (T) and the Relative Humidity (RH). 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 values are indicated in Fig. 6, while the corresponding R² (training) values 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 SO₂ 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 SO₂ importance was much less than Cl⁻, T and RH). In another work, Zhi and co-authors 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 are not directly comparable because the latter disposed of one more input (Fe element). The DCCF-WKNNs inherited the high generalisation ability (MAPE (testing) = 12.95%).
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-WKNNs 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 (PCC $> 0$), highlighting their interdependence (refer to the ‘Attributes interdependence’ feature in the ML for corrosion database$^{23}$).

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, respectively$^2$ (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 variables, 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 years$^2$.

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, respectively$^2$ (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 require the accumulation of environmental data yearly. These annual averages could generate inaccurate predictions, especially in the short term$^4$.

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 atmospheres$^4$. 

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).
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 sets (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. 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 provided an effective way to simulate the influence of elemental contents and their interactions with the steel corrosion resistance. 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 elements. This difference in testing performance ($R^2$ equal to 0.90 or 0.92) 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, RF, and ANN. 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. 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 Protection.
NGBM(1,1) combined with GA had been previously used to model small-sample datasets with high performance. 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). 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 \( \eta \) (knowledge mining showed that \( \eta \) could reflect the corrosivity of the environment).

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 losses. 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). 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 \( \eta \) (knowledge mining showed that \( \eta \) could reflect the corrosivity of the environment).

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.

**Marine corrosion**

As an early example of ML work, Wen et al. 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 zinc. Nonetheless, the weak generalisation ability and tendency to overfitting was already known for ANN modelling with small-sample datasets. For instance, this work 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 models\(^8\) 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 accuracy\(^8\).

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)\(^8\).

Nature-inspired metaheuristic optimisation algorithms such as the firefly algorithm (FA) have effectively solved difficult optimisation problems\(^48\). A hybrid model was recently implemented by integrating a smart firefly algorithm (SFA) with a least-squares SVR\(^49\). Notably, this hybrid SFA-LSSVR model was shown to be effective for modelling CRs of 3C steel, outperforming all other models\(^18\). 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 measured\(^10\).

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 proxy the actual 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, 23.6467 and 7.8588, 6.0082, 7.7421, 6.6813\(^10\). All these models/metrics relationships\(^10\) 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 tasks\(^13\).

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. 5\(^13\).

### 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 contradiction 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. 5\(^14\).

The modelling of rebar corrosion is further complicated when localised corrosion (pitting) occurs, which is often the case in chloride-laden media. The chloride threshold (Cl\text{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 concrete\(^12\). The literature reports that the current lack of understanding of the Cl\text{th} value of steel in concrete is due to the numerous factors affecting pitting corrosion (e.g. 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 Ecorr - Epit) 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. 5\(^12\).

From knowledge mining, the authors elaborated the following measures to enhance the Cl\text{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 levels\(^12\).
Six years after the publication of the ‘pitting corrosion risk of steel rebar’ dataset\textsuperscript{12}, Chou et al.\textsuperscript{18} derived this data for testing their advanced SFA-LSSVR model. Unfortunately, no performance comparison could be elaborated between both works\textsuperscript{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)\textsuperscript{18}. Here, as illustrated in Fig. 6 as ‘SFA-LSSVR (pit)’, the model produced a MAPE (testing) value of 5.60%\textsuperscript{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 model\textsuperscript{18}. Further investigations would be envisaged for testing the prediction accuracy of SFA-LSSVR applied to different scenarios.

In 2021, Zhu et al.\textsuperscript{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 database\textsuperscript{43}). 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)\textsuperscript{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 $\text{Al}_2\text{O}_3$, $\text{Fe}_2\text{O}_3$, $\text{MgO}$, $\text{K}_2\text{O}$ and $\text{SO}_3$ 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 $\text{Cl}^-$ 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 pkw (this issue was further addressed in Part II of the series\textsuperscript{11}). 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 $\text{Cl}^-$.

**Crevice corrosion**

Besides being effective when dealing with atmospheric and general corrosion, the ANN model developed by Kamrunnahar et al.\textsuperscript{6} also proved to be efficient for modelling localised corrosion. Based on a crevice corrosion dataset 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 considered\textsuperscript{6}. The corresponding performance values ($R^2$ (training) $= 1$) are illustrated in Fig. 5 as ‘ANN (Ti-2)’\textsuperscript{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 behaviour\textsuperscript{7}. Remarkably, the model mapped the environment ($\text{pH}$, $[\text{Cl}^-]$), 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)\textsuperscript{7}), this was considered as a training ($R^2$) performance point in Fig. 5 (indicated by ‘ANN’\textsuperscript{7}).

**Stress corrosion cracking**

In 2014, Shi et al.\textsuperscript{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’\textsuperscript{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 overcoming the restrictions of toxic chromate-based technologies in the aerospace industry.

Winkler et al.\textsuperscript{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 outperformed 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 work$^{16}$ 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 inhibitors$^{11}$. Nevertheless, the authors referred to the relatively opaque or arcane nature of the ensemble methods, which renders difficult the mechanistic interpretation of the phenomena$^{16}$.

**Conversion coatings**

In a recent design-oriented work\textsuperscript{17}, 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/pH\textsuperscript{17}.

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, respectively. 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 discussed.

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) are worth the assessment.

In the first of these two publications, 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 oxidation.

In the second publication, 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).

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 sizes. In ref. 7 (Defence 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) was not presented.

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 (Defence corrosion’ section). The training-testing split method was unfortunately not clear (‘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 (Defence corrosion’ 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. 14, 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-LSSVR. Unfortunately, the electrochemical methods employed for data acquisition in ref. 6 are unknown.

In ref. 7, features such as the flow velocity or microorganisms, also seriously influencing the corrosion rate of LAS in marine environments, were not explored. The investigation lacks features describing the material microstructure. Unfortunately, the paper did not discuss the effects of the composition or the microstructure. An illustrative example is the pitting risk of steel in simulated concrete pore solutions. Modelling from the referred predictors would be more complex than the initially conceived. Although the flow velocity was not included in this analysis, 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 paper 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 ECP. 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 models.

Nonetheless, one could apprehend from Figs. 6 and 7 that the linear regression algorithms had 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. 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 fitting. 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 methodology could be extended to training failure.
The Tree-based methods can handle missing data and datasets with multiple features but can be prone to instabilities. 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. 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) era.

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 LAS.

Modelling with ensemble methods (SVR-PSO, RF-WKNNs, PSO-FFANN, SFA-LSSVR, KSON-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.

CHALLENGES
A central challenge in machine learning is the conflict between the complexity and accuracy of the models. This point is particularly true in the case of deep learning because these models usually have more complicated structures (structural design becomes vital). 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 included.

From a Review on ML approaches for energy systems, 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 reported.

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

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 field. For example, from this work, we pointed out that the atmosphere corrosion, 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 challenging. When the expertise from the corrosion field is incorporated into the modelling already during the definition of problems, it often leads to the best results.

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

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. 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 modelling.

Such as in the case of this addressed study, 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). For instance, as previously described, 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 2021, 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 applications and natural language processing (NLP), have been extensively used in materials science and chemistry.

Moreover, ML could be an intermediate step for processing the collected datasets, thus reducing redundant experiments or simulations. 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 fostered.

In the era of big materials data, considerable efforts have been made to collect and build materials properties databases. In their review on machine learning for molecular and materials science, Butler et al. 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) 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. Among the papers considered in the Machine learning for corrosion database, only one investigation made available related codes (see the ‘Libraries’ feature).

However, electrochemical data is often not shared in a systematic and machine-readable way. The research community is also
RESEARCH EVOLUTION AND PERSPECTIVES

Training ML models with large datasets presenting comparable features across wide ranges of values is paramount for accurate prediction. In a Review paper on rechargeable batteries, 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. generated defec 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 accuracy). Active learning has been considered a promising approach for learning the high-dimensional data of battery materials. Similarly, in a recent review paper on ML and energy storage, Gao et al. 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. From the review on deep learning on materials degradation, 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. In the development of batteries, Attia et al. 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. For example, KL models have been applied to property prediction for energy conversion materials, potentially overcoming the shortcomings of DFT (high consumption of computational resources). For instance, ML of quantum-mechanical reference data has been used for fast and accurate interatomic potentials of battery materials. 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. assumed that molecular properties permitted more reliable models than quantum chemical properties of the individual compounds. In an ensuing investigation based on classification algorithms, the structure-property relationships in aluminium alloy inhibitors were further elucidated. Galvão et al. 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 implementation. In this context, meta-learning or ‘learning to learn’ methods have emerged and shown to be particularly useful in image recognition. 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 processes, including Auto ML. 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 hyperparameters, and the genetic algorithm NGBM included a regularisation term to avoid the overfitting problem.

A variation of ensemble method called Temporal Ensemble Learning (TEL), in which temporal features portion the dataset for the forecast at specific time ranges, 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, 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 learning.

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.

L.B. Coelho et al.
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., Malyz, 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.

© The Author(s) 2022, corrected publication 2022