Machine learning approach to investigate high temperature corrosion of critical infrastructure materials.

MUTHUKRISHNAN, R., BALOGUN, Y., RAJENDRAN, V., PRATHURU, A., HOSSAIN, M. and FAISAL, N.H.

2024

High Temperature Corrosion of Materials
https://doi.org/10.1007/s11085-024-10312-4

ORIGINAL PAPER

Machine Learning Approach to Investigate High Temperature Corrosion of Critical Infrastructure Materials

Ramkumar Muthukrishnan[1] · Yakubu Balogun[1] · Vinooth Rajendran[1] · Anil Prathuru[1] · Mamdud Hossain[1] · Nadimul Haque Faisal[1]

Received: 24 March 2024 / Revised: 13 June 2024 / Accepted: 12 September 2024

Abstract

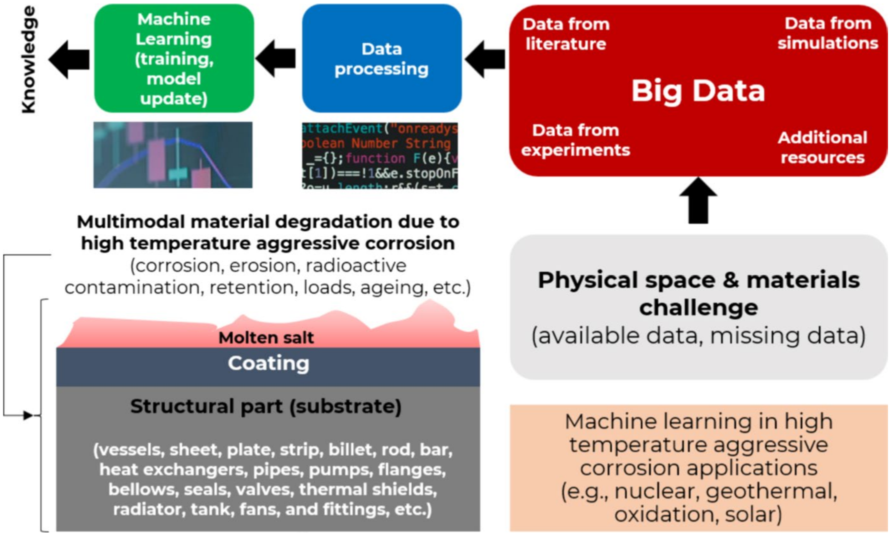
Degradation of coatings and structural materials due to high temperature corrosion in the presence of molten salt environment is a major concern for critical infrastructure applications to meet its commercial viability. The choice of high value coatings and structural (construction parts) materials comes with challenges, and therefore data centric approach may accelerate change in discovery and data practices. This research aims to use machine learning (ML) approach to estimate corrosion rates of materials when operated at high temperatures conditions (e.g., nuclear, geothermal, oxidation (dry/wet), solar applications) but geared towards nuclear thermochemical cycles. Published data related to materials (structural and coatings materials), their composition and manufacturing, including corrosion environment were gathered and analysed. Analysis demonstrated that random forest regression model is highly precise compared to other models. Assessment indicates that very limited sets of materials are likely to survive high temperature corrosive environment for extended period of exposure. While a higher quality and larger dataset are required to accurately predict the corrosion rate, the findings demonstrated the value of ML's regression and data mining capabilities for corrosion data analysis. With the research gap in material selection strategies, proposed research will be critical to advancing data analytics approach exploiting their properties for high temperature corrosion applications.

✉ Nadimul Haque Faisal
N.H.Faisal@rgu.ac.uk

[1] School of Engineering, Robert Gordon University, Garthdee Road, Aberdeen AB10 7GJ, UK

Published online: 22 September 2024

Graphical Abstract


Knowledge
Machine Learning (training, model update)
Data processing
Data from literature
Data from simulations
Big Data
Data from experiments
Additional resources
Physical space & materials challenge
(available data, missing data)
Multimodal material degradation due to high temperature aggressive corrosion
(corrosion, erosion, radioactive contamination, retention, loads, ageing, etc.)
Molten salt
Coating
Structural part (substrate)
(vessels, sheet, plate, strip, billet, rod, bar, heat exchangers, pipes, pumps, flanges, bellows, seals, valves, thermal shields, radiator, tank, fans, and fittings, etc.)
Machine learning in high temperature aggressive corrosion applications
(e.g., nuclear, geothermal, oxidation, solar)

Keywords Thermochemical cycles · Materials · Structural parts · Coatings · Degradation · Electrolysis

Introduction

Molten salts have considerable potential in energy applications, including generation and storage [1]. An example of energy generation systems could be a thermochemical cycle reactor where high temperature molten salt and steam could be used for water splitting leading to hydrogen production. However, one of the critical challenges facing the development of high temperature thermochemical cycle-based hydrogen production pilot plant is the identification of suitable materials for fabricating structural and coating parts of such plant [2]. Current nuclear reactors operate on a 12–18 months cycle of operation (limited by refueling, reactor design and operation duration) [3], where structural parts, coatings and auxiliary components exposed to molten salts in a thermochemical cycle will have to maintain integrity for at least this period before they could be replaced without impacting plant availability. It is an important challenge which needs to be addressed as such plants (i.e., structural parts) need to sustain a high temperature molten salt corrosive environment for a prolonged duration (Fig. 1).

Various thermochemical cycles such as cerium–iodine (Ce–I) [4], iron–chloride (Fe–Cl) [5], iodine–sulphur (I–S) [6], and Cu–Cl [7, 8] can be used for splitting water by using heat sources (nuclear or in some cases solar). Depending on the specific cycle, some common components in nuclear thermochemical cycles include nuclear reactor, heat exchanger, chemical reactors, separation units, recuperator, heat

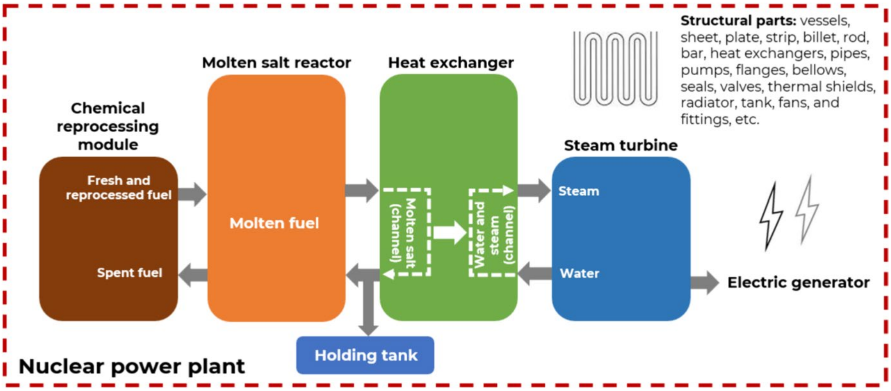

Structural parts: vessels, sheet, plate, strip, billet, rod, bar, heat exchangers, pipes, pumps, flanges, bellows, seals, valves, thermal shields, radiator, tank, fans, and fittings, etc.
Molten salt reactor
Heat exchanger
Chemical reprocessing module
Fresh and reprocessed fuel
Spent fuel
Molten fuel
Molten salt (channel)
Water and steam (channel)
Steam turbine
Steam
Water
Electric generator
Holding tank
Nuclear power plant

Fig. 1 An illustration of a molten salt reactor (MSR) nuclear power plant and location of molten salt-based thermochemical system

rejection system, and containment structure. Since thermochemical cycles operate at high temperatures with corrosive environment, the structural parts require special considerations. The accelerated degradation results in economic problems for the thermochemical cycle operation since the maintenance costs are greatly high. As the temperature of the thermochemical cycle is increased for higher efficiency, the severity of the materials degradation is even further accelerated. Overall, the design and composition of multi-material layers can be tailored to specific thermochemical cycle requirements, aiming to optimise performance, stability, and cost-effectiveness. The choice of materials plays a crucial role in achieving desired properties and overall performance. Addressing material issues requires a combination of material selection and manufacturing strategies, and optimisation of the composition and operating conditions.

There are numerous examples that have been published where the application of advanced alloys (as structural parts for heat exchangers, reactors, or containment structures) and coatings has been deployed in thermochemical molten salt at high temperature environment. Coatings of specialised feedstock materials have been proposed for such harsh environments [9]. These include ceramics, refractory metals, superalloys, and graphite-based materials, as the most suitable overlaid materials (on base metals or substrates) for high temperature corrosion. Research has shown how to improve materials performance through development of new candidate feedstock materials and using a range of advanced manufacturing routes. However, more research is needed to achieve durability of structural and coating parts [2].

For a range of applications, analysis of corrosion data is important. Particularly, corrosion time-series modelling, and analysis may be used to forecast failure times and the remaining lifetimes of equipment and materials [10]. To date, no model has been established to model the corrosion behaviour of materials used in thermochemical cycle structural and coating parts and correctly predicts the corrosion rates, largely due to corrosion rate modelling complexity. Literature shows that machine learning (ML) methods could be able to accurately reproduce first principles data for a range of applications, which refers to data from fundamental principles, laws, or equations. It has been extensively employed in material investigations because of its enhanced modelling libraries and increased functionality [11]. ML approach can help in identification of patterns, correlation, and trends in the datasets that may not be apparent otherwise. ML approach can make informed predictions about performance of materials. It could assist in identifying the key factors influencing the materials' performance. It can complement physical models by providing data-driven insights and can enable more accurate predictions of materials' behaviour. It can help in learning from multi-scale datasets and make predictions with different levels of complexity.

Machine learning can obtain complicated interactions between objectives and multi-dimension parameters [12, 13]. Researchers have been using neural networks (NN) with artificial intelligence (AI) and other ML methods since the 1980's to forecast materials corrosion and create novel corrosion-resistant metals, making significant advancements in corrosion issue [14, 15]. This has aided in forecasting the local corrosion performance of any material with fixed corrosion circumstances, temperatures, preparation procedures, various compositions, and corrosion periods. It has also assisted with anticipating the corrosion rate in substrates and feedstock-applied substrates [11, 16]. ML-based forecasting of electrochemical corrosion is an expanding area of research, and it also shares a data-driven outline of this realm [17]. They emphasize evaluating the forecast efficacy of several methods and elaborate on the state of regression modelling for numerous corrosion subjects. Recently random forest (RF) approach was utilised to model corrosion rate observations of carbon steel [18]. And still, conventional statistical analysis techniques like gradient boosting regression (GBR), support vector machines (SVM), and random forests (RF) operate under the underlying presumption that the datasets have self-sufficient and equal distributions [19].

Moreover, systems for time-series analysis of data, such as the popular long short-term memory (LSTM) neural network, may perform dependency mining on a series of data and train functions that convert a succession of previous inspections from a feed input to an outcome observation [20]. In this way, the time-series data analysis technique can offer a chance for the development forecast of pitting generation [21, 22]. To forecast corrosion rates and analyse data, several research have so far employed support vector regression (SVR) and back-propagation neural networks (BPNN), which are dependent on statistical principles and need a lot of data to assure model accuracy [23, 24]. By utilising the SVR and BPNN approaches, it is challenging to create a consistent corrosion time-series model [10]. Furthermore, measurements of corrosion data from multiple publications are often made under numerous testing circumstances [25, 26]. It is important to note that the corrosion data is challenging to employ for modelling research due to the varied testing conditions [27, 28]. These drawbacks of experiential modelling severely restrict how corrosion data may be used [16].

Previous research demonstrates that their approach can successfully handle a range of situations. Unfortunately, these models' predictive ability to anticipate corrosion rates is sporadic at best. For evaluating the data series on corrosion rate, a unique model is required. The latest model, which has greater flexibility than the existing model, may be used to implement the modelling and fitting of nonlinear series [10].

The overall goal of this research was to investigate the prospective uses of ML in the domain of corrosion and more needs to be investigated. This becomes important as the global cost of corrosion is estimated to be US$2.5 trillion, and on a global basis between US$375 billion and US$875 billion can be saved by using corrosion control practices [29]. Among many practices, the application of coatings on structural parts is common practice to control corrosion, a market which is estimated to be US$25.80 billion in 2024 and is expected to reach US$32.01 billion by 2029 [30]. In this work, corrosion data was gathered from various published literature. Although the overarching subject is of current research interest, published data in relation to nuclear thermochemical materials corrosion is very sparse. Where some work in this area has been published, very limited information is given on the materials used, or corrosion data which is needed for this research. Therefore, due to limited datasets available in relation to nuclear thermochemical corrosion of materials, the datasets from other high temperature aggressive corrosion of materials (e.g., geothermal, oxidation (wet/dry), solar) were included in the analysis. Also, as typical of any data analytics work, overfitting and inaccurate results are bound to occur when models are trained on very limited data [31]. We therefore aim to avoid such problems by choosing and training the most suitable regression model. Exploratory analysis of the collated dataset was carried out before the predictive analytics was carried out by implementing different regression models to predict the corrosion rate.

The aim of this research is to use ML approach to investigate high temperature aggressive corrosion of materials, and meaningfully analyses complex datasets to extract valuable insights, accelerate materials discovery, and enhance our understanding of materials behaviour. The outcome from this research is a first step towards development of material informatics for applications in high temperature aggressive corrosion, such as nuclear, geothermal, oxidation, and solar sectors which can address the above challenges, as well as enable opportunities for thermochemical cycle electrolysis and hydrogen production.

Data Analytics Methodology

Data Background

Corrosion, which is an electrochemical process, has always been a multi-dimensional problem, as materials deteriorate due to reactions with their environment. Further on, corrosion at high temperature exposes the structures to new challenges such as material compatibility and operation durations. Various features and factors influence the rate and extent of corrosion, which can be broadly categorised into material properties (composition, microstructure, surface condition, mechanical properties), environmental conditions (moisture, temperature, chemical composition, oxygen concentration), electrochemical factors (electrode potential, electrolyte conductivity), and corrosion types (uniform, pitting, crevice, stress corrosion cracking, and galvanic corrosion), etc. Considering multiple features and factors, identification and selection of suitable materials is one of the major difficulties facing the development of high temperature corrosion resistant critical infrastructure that can sustain the corrosive and harsh environment for a prolonged use [32].

Overall, the accurate and reliable performance of the ML models either during training or testing requires large sets of data of high quality, and the process of labelling corrosion features should be based on the know-how of domain expert [17]. In the current analysis, the dataset was created which included comprehensive and well-structured data that can be effectively used for training and testing ML models. The dataset included a CSV file with each row (total 155) representing a single observation, and each column (total 16) representing a feature related to the observation (Note: The CSV file can be found at the GitHub link, which is provided in the data availability section. We primarily used the data file for rudimental analysis and corrosion rate prediction using the Jupyter notebook (Corrosion_rate_prediction_using_ML.ipynb), which is also available on the GitHub link. Both files can be found in the 'code notebook' folder).

Key components of the dataset included substrates (structural part), such as steel and Ni-based alloys (steel-stainless steel, 4340 steel, low-carbon steel, ferritic stainless steel, Ti-stabilised high-carbon stainless steel, SAN25 steel, SA516 steel, Ni-based superalloys-12H18N10T, HN80MT, HN80MTY, HN80M-VI, MONICR, HN80MTW, AP164, Hastelloy, Nimonic alloy 263, Inconel 713LC). Coatings of specialised materials have been considered for harsh environments. These include superalloys, ceramics, refractory metals, and others as the most suitable coating materials (on base metal or substrate) for high temperature corrosive environments. Noteworthy, coating materials can fail due to reasons not related to layer adhesion to substrates, but other factors can be important. These can be geometry of the specimen, factors associated with coating integrity (chemical resistance, thermo-mechanical durability, surface preparation), as well as failure of equipment that failed to prevent leaking of oxygen, etc. [32]. Similarly, key components of the feedstock (coating) materials included Ni, Ni–Cr, Inconel 625 (Ni–Cr–Fe–Mo–Nb–Co alloy), NiCoCrAlY + YSZ, NiCoCrAlTaY + ScYSZ, LZ-LZC/YSZ/NiCoCrAlY, ([$La(NO_3)_3.6H_2O$], [$Ce(NO_3)_3.6H_2O$], and [$Zr(C_2H_4O_2)_4$]), SHS9172 (Fe–25Cr–15W–12Nb–6Mo), Diamalloy 4006 (Ni-based superalloys), YSZ, and Al_2O_3. Manufacturing (coating deposition) processes included high velocity oxy-fuel (HVOF), air plasma spray (APS), and cold gas dynamics spray (CGDS).

Corrosive environment (electrolyte or salt) included NaCl, HCl, LiCl-KCl-CsCl, 46.5LiF-11.5NaF-42KF, 92$NaBF_4$-8NaF, 71.7LiF-16BeF_2-12ThF_4-0.3UF_4 + Te, 66LiF-34BeF_2 + UF_4, 15LiF-58NaF-27BeF_2 + PuF_3, 15LiF-58NaF + 27BeF_2 + Cr3Te_4, 73LiF-5BeF_2-20ThF_4-2UF_4 + Cr_3Te_4, 71LiF-27BeF_2-2UF_4 + Cr_3Te_4, LiCl-Li_2O-Li, 71LiF-29BeF_2, 53LiF-46BeF_2-1UF_4, 62LiF-36.5BeF_2-1ThF_4-0.5UF_4, 70LiF-10BeF_2-20UF_4, 62LiF-37BeF_2-1UF_4, 71LiF-16BeF_2-13ThF_4, 58LiF-35BeF_2-7ThF_4, 53LiF-46BeF_2-0.5ThF_4-0.5UF_4, 60LiF-36BeF_2-4UF_4, 62LiF-36.5BeF_2-1ThF_4-0.5UF_4, (NaCl-Na_2SO_4-KCl) + 10%H_2O, FLiNaK, 50wt% Na2SO4 + 50wt%V_2O_5, Na_2SO_4 + V_2O_5, 50%Na_2SO_4 + 50%V_2O_5, NaCl + Na_2SO_4 + KCl, LiCl-Li_2O, and LiCl–KCl. Other range of key conditions included, such as testing temperature

(25–1000 °C), testing duration (0–20,000 h), corrosion rate (0–610 μm/year), feedstock (particle) powder size (11–53 μm), melting point (1260–2700 °C), density (3–9.01 g/cm^3), porosity (1.1–11%), hardness (8–22,050 MPa), tensile strength (69–2500 MPa), elasticity modulus (126–413 GPa), electrical resistivity (0.0000064–1×10^{26} Ω-cm), and thermal conductivity (5.7–60.7 W/m.K).

Regression Models

Application of data analytics approach and regression models for high-temperature corrosion prediction or analysis involves correlating various factors with the corrosion rate or some relevant outcome. In all cases, the approach includes data collection, feature selection, pre-processing of data, exploratory data analysis, regression model selection, uncertainty analysis, followed by model training and its evaluation, interpretation, deployment, and validation.

Regression models help estimate the relationships between a dependent variable and one or more independent variables. Machine learning regression models predict values based on various input data. Overall, the regression models choice should align with the underlying corrosion mechanisms and the data available. Through various regression models [33, 34], we can learn patterns and relationships from the training data which can then be used to make predictions on new data. There are many regression models, however in this research, seven models were used to evaluate their performance: (i) linear regression (LiR), (ii) lasso regression (LaR), (iii) ridge regression (RR), (iv) support vector regression (SVR), (v) random forest regression (RFR), (vi) gradient boosting regression (GBR), and (vii) ada boost regression (ABR). Review of various machine learning and regression models can be cited elsewhere [35], however some rationale on the selection of regression models is provided below.

Linear, Lasso and Ridge Regression

The most straightforward regression approach is linear regression (LiR). It is made up of a dependent variable that depends linearly on the independent variable and two variables. Lasso regression (LaR) linear regression techniques add penalty terms to the linear regression objective function. Lasso supports simple models with limited features, which is ideal for models with large degrees of multi-collinearity. This approach regularises the model to make it general so that it can function for a broad range of data points, which also helps to prevent the problem of overfitting. The ridge regression (RR) model is a method for analysing data from linear regression and multiple regression that exhibits multicollinearity. Ridge regression pre-supposes a linear connection between both the destination values and the individual values, much like simple linear regression does. When the individual values in the set of data have a strong correlation, ridge regression is applied.

Support Vector and Random Forest Regression

Support vector regression (SVR) can effectively achieve a nonlinear classification by utilising the kernel trick, indirectly locating their input data into high feature spaces, with the goal of minimizing the error whereas random forest regression (RFR) is an ensemble ("together") learning approach that functions by building a wide range of decision trees and merging the multiple trees together to get a more accurate and stable prediction. SVR's premise is to locate the better-fit line. The hyperplane with the greatest quantity of points is the better-fit line in SVR, whereas RFR employs a method that integrates estimations from various machine learning models to provide forecasts which are precise than those from an individual model.

Gradient Boost and Ada Boost Regression

Like the RFR model, the gradient boosting regression (GBR) and ada boost or adaptive boosting regression (ABR) are ensemble learning techniques, although they employ several unique calculation methodologies [11]. GBR builds a model in a stage-wise fashion, where each stage corrects the errors of the previous one. ABR combines multiple weak learners to create a strong learner, and is commonly used for classification tasks, but it can also be adapted for regression tasks. The mechanisms of each technique have been described in various literature [36, 37].

Machine Learning Approach

Research on localised corrosion has garnered a lot of attention [24]. For predicting localised corrosion effects, conceptual deterministic approaches are often devised. To comprehend how long-term corrosion of metal behaves, it is crucial to anticipate the rate of corrosion and its progression [38]. As summarised in section above, a set of datasets were collected from public sources and literature studies. Due to limited datasets available in relation to nuclear thermochemical corrosion of materials, the datasets from other high temperature aggressive corrosion of materials (e.g., geothermal, oxidation (wet/dry), solar) were included in the analysis. Such data sets were included from other high temperature aggressive corrosion of materials, as similarities in such types of corrosion and materials degradation can be observed across different application, materials and environments, and several common factors contribute to the degradation process.

Before initiating the analysis and training section, collected datasets were deployed in proper order. Through the data analysis, it helps find the correlation between collected datasets [10, 39]. Thereafter, datasets were divided into training and test parts. A training dataset was utilised by different models (which are imported from the Scikit-learn library in Python programming) to train a corrosion prediction model, and it was evaluated by test and training datasets (representation is depicted in Fig. 2). It is important to note that Scikit-learn is a library that provides

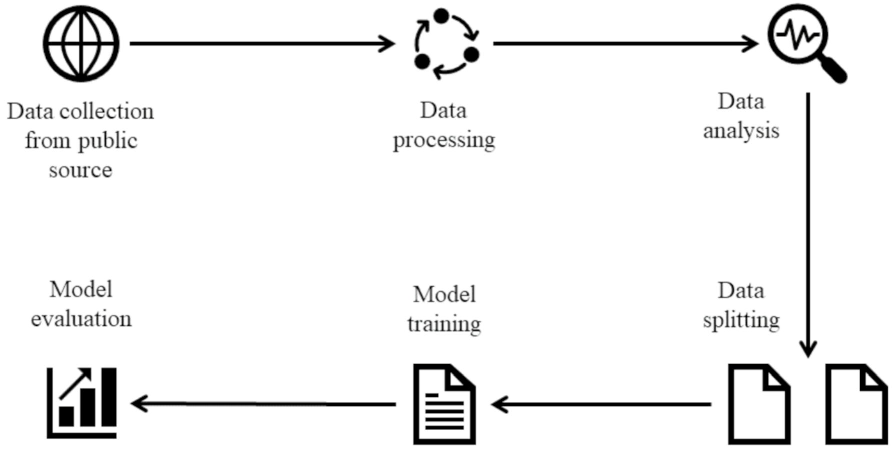


Fig. 2 Working process flowchart for the proposed research

many supervised and unsupervised learning algorithms which is better suited for ML applications with smaller datasets (unlike TensorFlow which is better in deep learning and large-scale datasets).

To investigate the multi-dimensional relationships between materials and corrosion rates, an ML algorithm was implemented using Python 3 package. Model training was performed using the collated datasets gathered from previous studies on the corrosion rates of structural and coating parts. The prepared algorithm was utilised to conduct a preliminary analysis test, which supported checking the correlation between each input parameter and corrosion rate. It also helped to visualise the number of collected corrosion data based on different aggressive corrosion processes with varying temperatures.

Some of the collected datasets are required to be converted from string to numerical format, which was done in advance of model training to get the best regression model, and it also helps to easily access those collected datasets for any ML regression models. The overall collected dataset with 16 input parameters was used during the training phase for corrosion rate prediction. The entire dataset was divided at random, with 10% as the testing set and 90% as the training set [24]. The testing set was solely utilised to validate the predicted performance of the improved model; the training set was primarily utilised for training purposes. The seven models—LiR, LaR, RR, SVR, RFR, GBR, and ABR were each independently used to create a corrosion rate prediction model during the training phase. The improved predictive capability was then established as an optimised model. Finally, to confirm the effectiveness of the improved corrosion rate prediction model, a testing set was used. The scikit-learn library [40] was used for all the statistical investigation and information mining tasks.

Furthermore, we directly imported all regression models from the scikit-learn library [40], ensuring no hyperparameter tuning changes were necessary. We imported regression models from the scikit-learn library, as well as metrics for validation. We also used the Matplotlib library [41] for data visualization and the Pandas library [42] for working with the collected datasets.

The random forest model (scheme shown in Fig. 3) was proposed to train the collected data set, as it is a commonly used supervised ML technique which can be employed to address classification and regression issues. Noteworthy, a forest consists of a large number of trees, and the more trees it has, the sturdier it will be (typically combines multiple decision trees output to reach a single result). In contrast, the precision and capacity to solve problems of a random forest algorithm increases with the number of trees in the approach. To increase the dataset's accuracy of prediction, it uses many decision trees on different subsets of the input data. The random forest algorithm's operation steps were as follows; Step 1: a particular data sets random samples are first chosen, Step 2: it then builds a decision tree for each batch of data sets, Step 3: additionally, the decision tree is then averaged during voting, and Step 4: it chooses the final forecast outcome that received most of the votes.

Additionally, a bagging approach was used. Bagging is the process of generating a unique training subset from a sample data set via replacement. The majority of votes determines the outcome (in this research case, it was low, medium, or high corrosion (or corrosion rates) outcome based on the majority of votes received, which is also depicted in Fig. 3 for easier understanding). Based on the aforementioned premise, it can be inferred that random forest employs the Bagging code. Random forest uses a technique called bagging, sometimes referred to as bootstrap aggregation. Any initial random data can be used to start the procedure. It can then be arranged into samples called bootstrap samples. Additionally, each model was then trained separately, producing unique outcomes known as aggregation. The last stage combines all the findings, and the resultant outcome depends on a majority vote. Bagging is the term for this action, which is carried out with an ensemble.

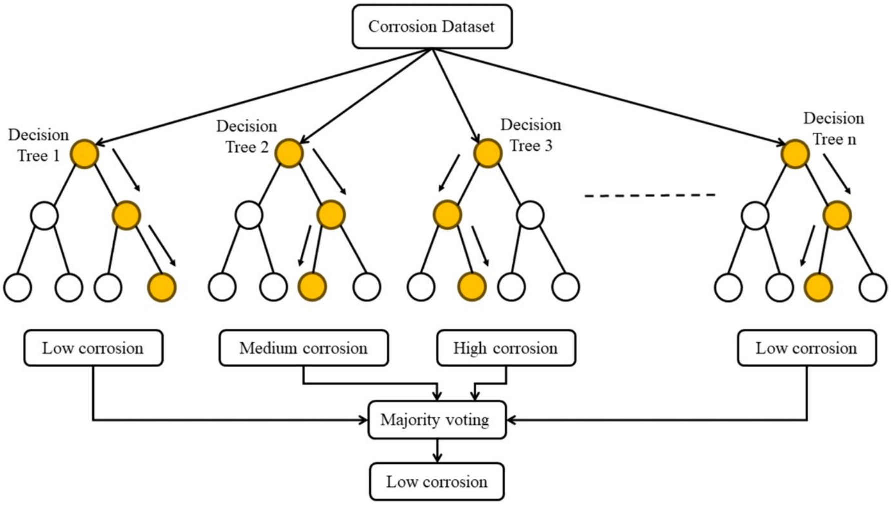


Fig. 3 Simplified illustration of random forest model for corrosion analysis

In the analysis, the data set was crowded with alternatives, for instance, to estimate the corrosion degree of a collection of characteristics based on properties. It then provides a training dataset with details on the type of coating, substrate, temperature, type of thermochemical electrolysis, type of salt, duration of time, etc. In the analysis, the data must then be divided by selecting the shortest chunk first so that it can be divided in as many ways as feasible. Starting by dividing the corrosion level by substrate type, followed by coating material, and so on, could be beneficial. For example, in Fig. 3, let's assume those yellow colour nodes are in decision tree 1, which might hold the first row of data from the collected dataset, and the white colour node could be assumed to be irrelevant data. Similarly, the same principle was followed in decision tree 2 using the second row of data from the collected dataset, and so on. It can anticipate a certain corrosion level with the highest degree of precision, and it can split continuously until a certain node no longer requires it. The level of corrosion rate was predicted based on the major vote count from the decision trees.

Results and Discussion

Exploratory Data Analysis

Exploratory data analysis was the crucial step that involved summarizing the main characteristics of the datasets. It included understanding the structure of datasets, handling missing data (imputation, removal), exploring the statistics for categorical variables, visualization, correlation, data outliers, analysing the distribution, grouping, aggregating to get a high-level understanding, including validating assumptions.

Figure 4 shows the varying corrosion rate with a minimum to maximum temperature range based on different high temperature aggressive corrosion processes (i.e., nuclear, geothermal, oxidation, solar), which are all collected from the public domain. These processes involve the use of chemical reactions at elevated temperatures that could split water molecules to produce hydrogen through electrolysis [2]. As shown in Fig. 4a, the frequency and trend of the plot of the datasets from the literature studies reviewed presents high temperatures aggressive corrosion processes ranging from 250 °C to 2000 °C. Also shown in Fig. 4a, nuclear thermochemical processes have temperatures ranging from 470 °C to 1300 °C with most of the related studies reviewed operating around 500 °C–1000 °C. Solar processes on the other hand had some studies operating at very high temperature of up to 2000 °C, although very limited literature studies presented this is very high temperature values. The spread and centers of each high temperature aggressive corrosion processes with available rate of corrosion data points are represented as black dots with respect to the varying temperature, shown in Fig. 4b. This shows that most of the studies were carried out at temperature values between 500 °C and 1000 °C.

Following the analysis of corrosion rate based on temperature under different high temperature aggressive corrosion processes, additional analysis was carried out to check the range of corrosion rate counts based on the data collected. This included density plotting in different ranges which could also help in identifying

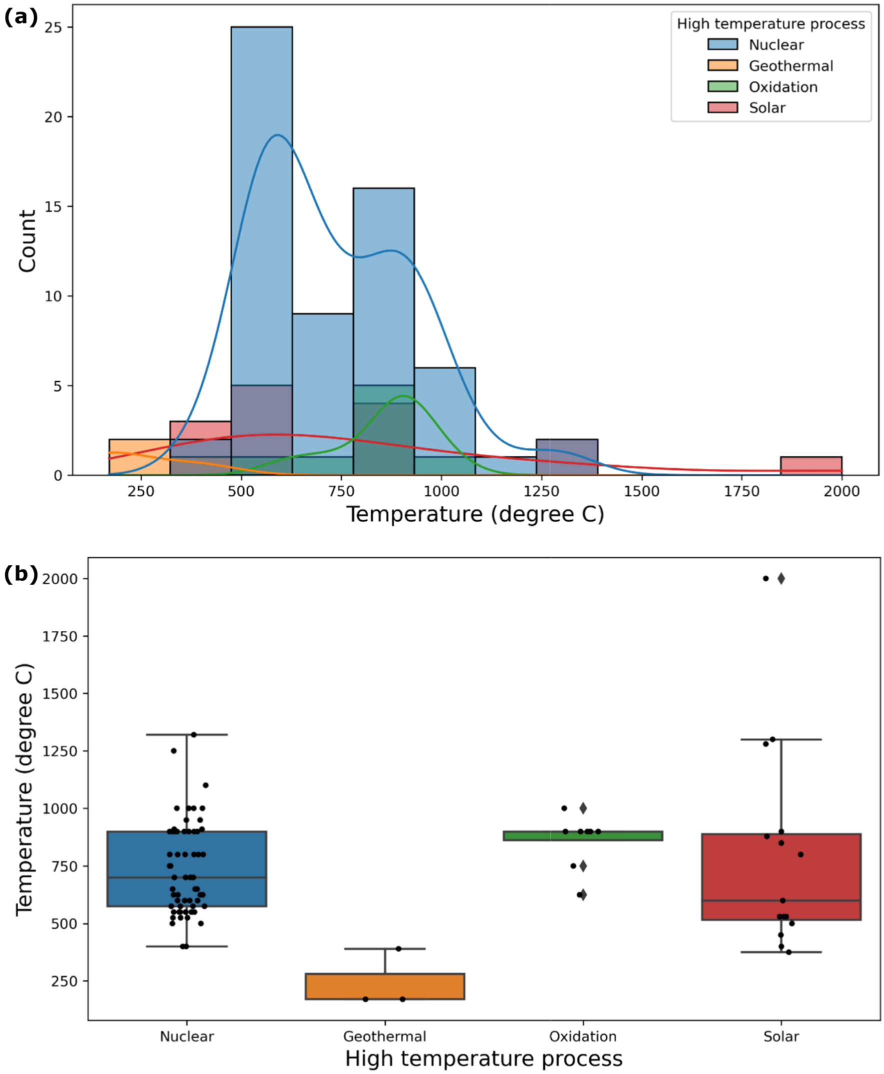


Fig. 4 **a** Temperature distribution plot showing the range of temperature under consideration, and **b** different high temperature aggressive corrosion processes and their respective temperature ranges

density-based thresholds or ranges that could be used for data processing. It could also help in creating new features representing the local data point density around a given observation. However, in the present analysis, the consideration was to use density-driven sampling from regions of higher data density, to improve the model's ability to generalise well in dense regions. As shown in Fig. 5, the collected data has a corrosion rate as high as 610 μm/year. Particularly, it clearly shows that the collected data from the literature study has a huge number of corrosion data points,

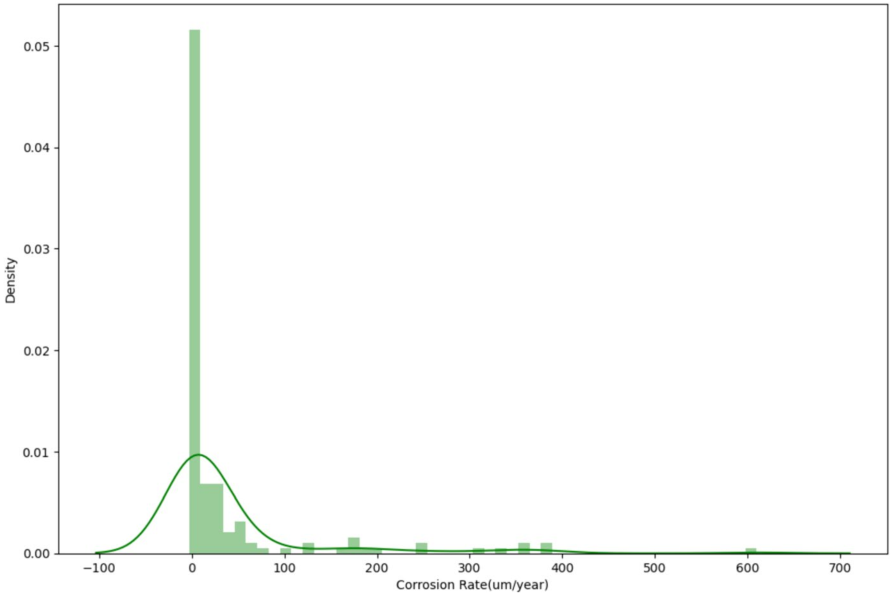


Fig. 5 Distribution plot for range of corrosion rate from the collected dataset

which vary between 0 and 100 μm/year. It is sufficient to train a demonstrable model, which has been discussed in the upcoming section (i.e., predictive modelling of the corrosion rate).

Beforehand, another analysis was carried out to check the correlation between available input parameters and corrosion rate (Fig. 6). In this analysis, a graphical heatmap has been shown in a matrix, represented as colours, which are commonly used for correlation matrices visualisation (i.e., positive, or negative correlations), confusion matrices (i.e., areas of correct and incorrect predictions), and feature importance. In such mapping, the correlation matrix shows how strongly different variables are correlated (i.e., darker colours indicate stronger correlation, (+) correlations means when one variable increases, and the other variable increase, and (−) correlations means when one variable increases, and the other variable tends to decrease). It is important to note that corrosion data are usually measured under several testing environments, and then inconsistencies in testing condition can make the corrosion results difficult for modelling usage and limiting data utilisation [16]. As can be seen from the graphical heatmap analysis (Fig. 6), the features have largely nonlinear relationships, i.e., the input parameters have less correlation with the corrosion rate. Graphical heatmap analysis also helps to extract information which are dominant and redundant factors. It appears to have a limited relationship with the outcome.

Due to the availability of limited data, powder particle size, time and temperature alone have a close correlation to the corrosion rate with positive values. However, the rest of the input parameters have less correlation with the corrosion rate, which are all shown as a negative value in correlation map (except powder

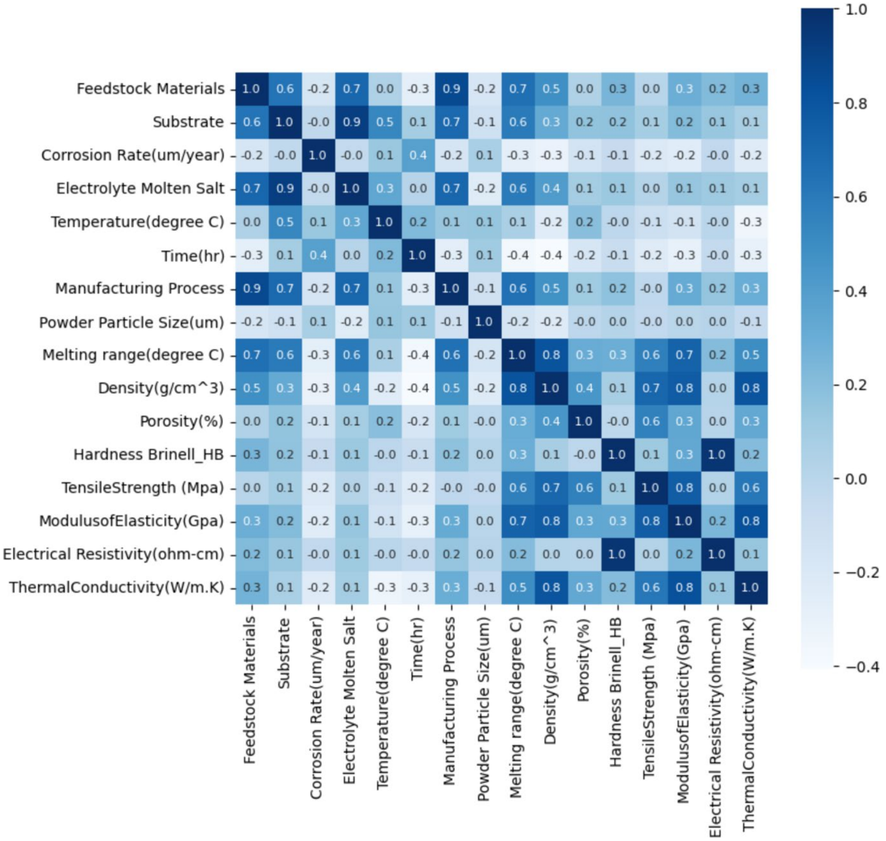


Fig. 6 Correlation map for checking the correlation between available input parameters and corrosion rate

particle size, time, and temperature). With other few variables, highly correlated features (e.g., correlation coefficient $\geq$ 0.7) include feedstock materials, substrates, electrolyte molten salt, manufacturing process, melting range, density, hardness, tensile strength, modulus of elasticity, electrical resistivity, and thermal conductivity. Some of the correlations are strong (e.g., electrical resistivity, hardness), though not have obvious and direct physical meaning. However, the materials composition and structure (which influence hardness) can influence both properties independently. For example, certain materials may have high electrical resistivity due to their electronic structure, while also exhibiting high hardness due to a strong atomic structure and could be part of future investigations. Too many input features could reduce the generalisation ability of the model [16]. However, the potential analysis could include assessing the effect of confounding variables (not directly visible in the datasets), nonlinear relationships, effect of outliers, errors, or missing values, creating new features, assess the statistical significance, or may be visualise it differently. However, in the current analysis, a suitable regression model was trained using available data and compared with other regression models, all discussed in the following section.

Predictive Modelling of the Corrosion Rate

After observing the data through exploratory analysis, data point density observation and graphical heatmap visualisation, this section presents the corrosion estimation outcomes for the seven regression models. For a flawless prediction model, the predictive performance was represented as functions of the quantified values, which matched the observed data exactly. It can be seen that the data points fall on the diagonal lines, and closer the data points are to the diagonal, representing the more precise predicted outcomes.

Abundance of noise and overfitting issues occurred frequently when using linear regression model. Similarly, SVR does not have the capability to perform well with noisy data. Due to these, both regression models demonstrated a poor match for the corrosion. Additionally, RR and lasso regression both aided in reducing overfitting issues by picking characteristics of lower relevance and reducing the size of big coefficients. Thus, the RR and lasso regressions' fitting effects on the two target qualities were substantial and effective, and this also made these two regression models perform similarly. The ABR model performed well in the training dataset. However, it has poor performance in the test dataset due to its progressively learning-boosting technique. As shown in Figs. 7, and 8, respectively, the GBR and RFR models are interconnected strategies that serve to strengthen the model's capability and provide well-fitting outcomes. Additionally, these models have the capability to handle multiple types of data, like categorical, textual, and numerical. Therefore, a good result was obtained from these models. Nevertheless, RFR (Fig. 8) performed well compared to other models due to its precision-improving capability by reducing the overfitting issues in the decision tree.

Moreover, each trained model has not predicted the corrosion rate completely. It is imperative to remember that there is a finite dataset used to train each model. Therefore, we must make sure that this is considered while interpreting the findings. Thus, it shows variation in the x- and y-axes based on each model's robustness.

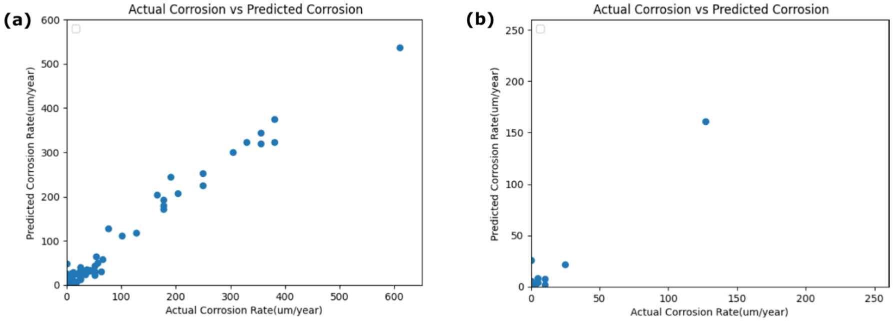


Fig. 7 Prediction accuracy of GBR model: **a** train datasets, and **b** test datasets

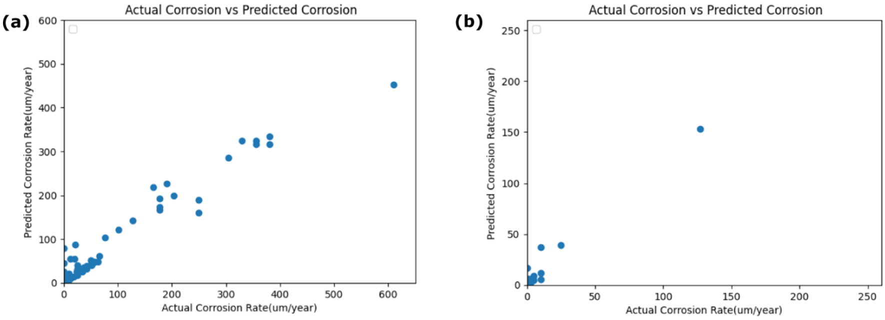


Fig. 8 Prediction accuracy of RFR model: **a** train datasets, and **b** test datasets

However, each model performed badly on the training sets (except ABR, GBR, and RFR), possibly because of overfitting while adequately fitting the training dataset [43]. The results of the poorly performed regression models are presented in Appendix A as supplementary material.

In this research work, we randomly divided 90% of the data into a training set and 10% into a test set from the collected dataset. We used this separated data to train the optimal model and cross-validate each model. This research work employed two metrics: the coefficient of determination $\left(R^2\right)$ and the mean absolute error (MAE) to cross-validate the prediction accuracy and performance of each model.

For evaluating the predictive performance of model's, the coefficient of determination $\left(R^2\right)$, which analyses two sets of data using a value between real and predicted, were both used [11]. It provides an indication of how well the model predicts the variability in the data (for example, $R^2 = 0$, where the model does not predict any variability in the dependent variable, whereas $R^2 = 1$, where the model perfectly predicts variability in the dependent variable. Additionally, mean absolute error (MAE) also used, which can more intuitively reflect the error. The following equations are used to formulate them:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(f_i - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - y'\right)^2} \tag{1}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|f_i - y_i\right| \tag{2}$$

where y' is the average target value across all test samples, f_i is the predicted value, y_i is the target value, and n signifies the number of test samples. The MAE is used to determine the average absolute variation between the predicted and target values. A lower MAE value indicates the most accurate model. Most evaluations of regression model efficacy use these metrics (R and MAE). Each improved model's R^2 and MAE calculation results for the training and testing sets are shown in

Table 1. The outcomes demonstrated that the RF regression's fitting impact had the highest level of accuracy in comparison [43].

Machine learning algorithms are better at making predictions than traditional statistical approaches because of their strong generalisation capabilities. Nevertheless, the research's training set included a variety of substrates and extra variables, and we believe that the volume of data it included was inadequate to accurately capture the influence that these variables had on corrosion rate. As a result, the model's ability to predict outcomes for the testing set was substantially less accurate than it was predicted for the training set. The prediction accuracies for the RFR and GBR models were all rather good [11]. But in general, to increase accuracy, the Random Forest model can identify the most essential features from a dataset. It is also quick and precise, processing big datasets with smaller parameters. Thus, these features make the RF model perform better than other regression models.

Corrosion Rate Modelling Complexity and Model Limitations

Corrosion rate modelling is a critical tool in the design of coating and structural parts prone to degradation due to corrosive environment. Materials and corrosion engineers often rely on models which can predict corrosion rates to select appropriate construction materials for construction. Experimental investigations to determine corrosion rates of materials under various operational conditions can be costly. Therefore, development of ML models, including simulations, mechanistic, empirical (or semi-empirical), and mathematical models, etc., could play an important role in estimating the corrosion rates for various systems.

Application of ML models demonstrates the promise in predicting corrosion rates, but they come with some limitations. ML model limitations for evaluating corrosion rates have been well argued [17, 44]. The limitations are due to inherent corrosion processes complexity, challenges with the data, and the ML models. As demonstrated through this work, corrosion data can be inadequate, particularly for high temperature corrosion environmental conditions (nuclear, geothermal, oxidation (dry/wet), solar applications) which can have very different corrosion mechanisms affecting the materials, there can be inconsistencies in data collection methods,

Table 1 Coefficient of determination (R^2) and mean absolute error (MAE) values of each model

Regression model	R^2 value		MAE value	
	Train datasets	Test datasets	Train datasets	Test datasets
Linear regression (LiR)	0.001	−0.981	56.926	40.519
Support vector regression (SVR)	-0.151	−0.063	40.722	12.116
Ridge regression (RR)	0.205	−2.248	46.216	39.071
Lasso regression (LaR)	0.204	−2.251	45.931	38.319
Ada boost regression (ABR)	0.856	−0.389	28.455	31.827
Gradient boosting regression (GBR)	0.977	0.725	7.763	9.733
Random forest regression (RF)	0.944	0.864	9.388	6.451

and there can be incomplete datasets. Features identified may not be relevant for the models and data processing may result in poor model performance. Prediction and comparison of the corrosion rates using different models under nearly identical conditions for a sample can be complex. Analysis shows that not all corrosion rate models demonstrated a high generalisation ability [17]. Also, as seen in current analysis (sections above), the features have largely nonlinear relationships, i.e., the input parameters have less correlation with the corrosion rate.

We collected several corrosion datasets from the public domain, each related to different conditions. All of these datasets are combined based on the same parameters (such as feedstock (coating) materials, substrates (structural part), manufacturing (coating deposition) process, corrosive environment (electrolyte or salt), temperature, testing duration, feedstock (particle) powder size, melting temperature, density, porosity, hardness, tensile strength, elasticity modulus, electrical resistivity, and thermal conductivity), which is used as an input and corrosion rate (µm/year) as an output. We deployed the collected datasets into two parts: input and output, as previously mentioned. Using the properly deployed dataset for training each regression model, the RFR model accurately predicted the corrosion rate behavior in both the train and test datasets, as illustrated in Fig. 8. The inclusion of various corrosion types allowed the model to account for a wide range of conditions during the training phase. Specifically, it can aid in predicting the behavior of corrosion rates when it receives unseen datasets related to various conditions. However, at times, it may not be able to accurately predict corrosion rates based on various unexplored corrosion condition-related variables. Such an issue can be resolved by incorporating the unexplored corrosion datasets into the training phase.

The amount of data including completeness within rows representing observations and columns representing features would have a significant impact on the model's ability to predict regression since the ML model's primary mechanism of operation was to acquire the dataset's underlying information. To increase the associated forecast accuracy, additional pertinent data must be added. The database might be further enhanced, it was thought, to increase the method's precision as well as reliability [11]. Literature trends show that there is a need to incorporate basic descriptors based on domain knowledge which could simplify the representation and improve the interpretability of ML modelling. Such models should have an in-depth knowledge of fundamentals of materials and environment, in particular mechanisms related to thermodynamics, reaction kinetics and mass transfer, heat transfer and behavior of degradation products on the rate of corrosion [44]. It has also been proposed to select temporal variables which positively affect the overall model's performance [17].

Opportunities

As demonstrated through present work, inclusion of different corrosion environments in the training data for ML models provides an opportunity to enhance the model's ability to generalise across various conditions. However, inclusion of data can impact ML model performance to predict corrosion rates, as the inclusion introduces increased complexity and potential data imbalance. It can be part of further analysis, but this offers an opportunity to employ a combination of strategies, starting with leveraging domain expertise, hybrid modelling, and feature engineering, which could potentially mitigate such challenges, and could lead to reliable corrosion prediction models.

In the present work, to begin with, the outcomes of the correlation and multicollinearity evaluations assisted in determining which characteristics were most important to the corrosion rate. The RFR model then assessed the order of importance of the dominant characteristics and showed how they affected corrosion rate in an intuitive way. ML was able to deliver extra data in numerous formats compared to traditional analytical techniques. Consequently, depending on the training data, the suggested RFR model could be utilised to provide long-term predictions of corrosion rates for several substrate types during high temperature aggressive corrosion processes under varied salt solutions. For data with different variables, such as substrates, feedstock materials, manufacturing process, electrolyte molten salt, temperature, melting point, powder particle size, porosity, tensile strength, modulus of elasticity, thermal conductivity, electrical resistivity, and corrosion rate data, etc., the ML model demonstrated improved regression capability. This corrosion rate prediction model's efficacy has led to its recognition as a useful tool for ongoing corrosion research [11].

Data mining benefits greatly from ML, a common large data processing technique. Generally, corrosion data is typically sparse, ML's benefits and requirements are not generally clear. The amount and clarity of the data are typically regarded as the precision assurance of ML modelling and analysis [45]. There are several corrosion test terms, it might be challenging to immediately summaries and utilise corrosion data from the publications [45–47]. It is becoming difficult to get sufficient data on corrosion, particularly during high temperature aggressive corrosion processes. Nevertheless, the substrates and other features might be accurately represented, if required data are acquired during high temperature aggressive corrosion of materials. ML just offers some exploratory analysis findings for the hazy phenomena and unexplored rules. Technical experiments are also required to be conducted to verify and analyse the true process.

This model performs substantially better than other models in modelling the limited data series [10]. Unfortunately, because of the limited amount of corrosion data, overfitting can quickly result from utilising excessive input characteristics. Moreover, a model's capacity for generalisation may be hampered by an excessive number of input characteristics. Hence, the analysis precision and application efficacy of a model are significantly influenced by suitable feature selection, which aids in extracting dominant information and removing unnecessary components in the original data [16]. Nevertheless, an overfitting issue readily occurs when these approaches are applied to the corrosion rate series data, indicating that the model's historical data fitting precision is huge but its ability to forecast upcoming data is poor [10].

To solve these problems, it is highly recommended to prepare a miniature high temperature aggressive corrosion of materials (e.g., nuclear, geothermal, oxidation, solar applications) experimental setup with the required number of substrates and applied feedstock coating materials and salt substances, including other variables. Thus, it could help to generate several datasets. However, it requires a substantial level of funding to proceed with this approach. Nevertheless, there is another possibility to generate numerous datasets than the last approach using computational simulation approach.

Conclusions

In this work, machine learning approach were to estimate corrosion rates of materials when operated at high temperatures conditions (e.g., nuclear, geothermal, oxidation (dry/wet), solar applications) but geared towards nuclear thermochemical cycles. In all cases, the approach includes data collection, feature selection, pre-processing of data, exploratory data analysis, regression model selection, uncertainty analysis, followed by model training and its evaluation, interpretation, deployment, and validation. To estimate the rate of corrosion using a random forest regression (RFR), a corrosion growth model was presented utilising the situation of a substrate with feedstock applied product. Analysis demonstrated that RFR model is highly precise compared to other models. The suggested method's accuracy percentage using coefficient of determination is 94.4% in training datasets, and 86.4% in test datasets. These results suggest that machine learning techniques might be helpful tools for corrosion research since they offer an effective way to utilise corrosion data. Based on this outcome, the model's applicability was expanded, and machine learning was utilised to confirm the precision and viability of forecasting aggressive corrosion of materials during high temperature. The approach offers opportunities to improve corrosion predictive capabilities and develop materials and corrosion protection technologies for critical infrastructure applications.

We can say that this is First-Of-A-Kind research where we developed database (data classification, algorithm, relationship between materials and information) for high temperature aggressive corrosion of materials. As a recommendation, the database can be used as part of further research in relation to electrolyser material selection, alongside the modular construction of high temperature thermochemical electrolyser for hydrogen production.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11085-024-10312-4.

Acknowledgements Authors primarily acknowledges high temperature thermochemical electrolysis related funding by Henry Royce Institute via Materials Challenge Accelerator Programme Grant No. MCAP034 (METALYSIS), and UK's National Nuclear Laboratory (NNL) via gamechanger Grant No. GC 596 (THERMOSIS). Additionally, the authors acknowledge high temperature steam electrolysis related funding by the UKRI EPSRC via Grants No. EP/W033178/1 (METASIS), and ScotGov's Emerging Energy Transition Fund (EETF) via Grant No. EETF/HIS/APP/007 (Hy-One).

Author Contributions Ramkumar Muthukrishnan done investigation (lead); methodology (lead); software (lead); data curation (lead); formal analysis (lead); validation (lead); visualisation (lead); writing—original draft (lead). Yakubu Balogun performed investigation (lead); methodology (lead); software (lead); data curation (lead); formal analysis (lead); validation (lead); visualisation (lead); writing—review and editing (supporting). Vinooth Rajendran done investigation (lead); methodology (lead); data curation (lead); validation (lead). Anil Prathuru and Mamdud Hossain helped in supervision (supporting); writing—review and editing (supporting). Nadimul Haque Faisal contributed to conceptualization (lead); funding acquisition (lead); investigation (supporting); methodology (supporting); project administration (lead); supervision (lead); validation (supporting); writing—original, review and editing (supporting). All authors read and approved the final manuscript.

Funding This study was supported by Henry Royce Institute, MCAP034, MCAP034, MCAP034, National Nuclear Laboratory, GC596, GC596, GC596, Engineering and Physical Sciences Research Council, EP/W033178/1, EP/W033178/1, EP/W033178/1, Scottish Government, EETF/HIS/APP/007

Data Availability In this research work, we collected and extracted several datasets from published research articles. The Excel file (data with reference.xlsx) appropriately names the references for those datasets. Additionally, we kept the source article for the collected dataset in a separate folder under the reference name. Those folders can be found in the 'data with reference folder'. The datasets used and/or analysed during the current study with reference and code are available in this GitHub repository: https://github.com/rk3839/Predicting-Corrosion-Rate-using-ML/tree/master.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. R. Roper, M. Harkema, P. Sabharwall, C. Riddle, B. Chisholm, B. Day, and P. Marotta, *Annals of Nuclear Energy* **169**, 2022 108924. https://doi.org/10.1016/j.anucene.2021.108924.
2. N. H. Faisal, A. Prathuru, R. Ahmed, V. Rajendran, M. Hossain, V. Venkatachalapathy, N. K. Katiyar, J. Li, Y. Liu, Q. Cai, B. A. Horri, D. Thanganadar, G. S. Sodhi, K. Patchigolla, C. Fernandez, S. Joshi, S. Govindarajan, V. Kurushina, S. Katikaneni, and S. Goel, *ChemNanoMat* **8**, 2022 e202200384. https://doi.org/10.1002/cnma.202200384.
3. Nuclear Decommissiong Authority (NDA) Factsheet, 2014: Operating a nuclear power reactor. https://ukinventory.nda.gov.uk/wp-content/uploads/2014/01/Fact-sheet-operating-a-nuclear-power-reactor.pdf
4. Carty, R. H., Mazumder, M., Schreider, J. D., Panborn, J. B., 1981. Thermochemical hydrogen production, Gas Research Institute for the Institute of Gas Technology, GRI Report 80–0023, vol. 1, Chicago, IL 60616.
5. K. F. Knoche, P. Schuster, and T. Ritterbex, *International Journal of Hydrogen Energy* **9**, 1984 (473). https://doi.org/10.1016/0360-3199(84)90099-5.
6. Z. Ping, W. Laijun, C. Songzhe, and X. Jingming, *Renewable and Sustainable Energy Reviews* **81**, 2018 (1802). https://doi.org/10.1016/j.rser.2017.05.275.
7. G. F. Naterer, S. Suppiah, M. A. Rosen, K. Gabriel, I. Dincer, O. A. Jianu, Z. Wang, E. B. Easton, B. M. Ikeda, G. Rizvi, and I. Pioro, *International Journal of Hydrogen Energy* **42**, 2017 (15708). https://doi.org/10.1016/j.ijhydene.2017.03.133.
8. A. Farsi, I. Dincer, and G. F. Naterer, *Journal of Cleaner Production* **276**, 2020 123833. https://doi.org/10.1016/j.jclepro.2020.123833.

9. Bryson-Jones, H. and Bollet, Y. 2023. New Royce Landscape Report: Materials for Nuclear Enabled Hydrogen, A landscape report exploring the technologies, recommended research, and wider enablers for the development of a nuclear enabled hydrogen sector. https://www.royce.ac.uk/news/new-royce-landscape-report-materials-for-nuclear-enabled-hydrogen/ (Accessed March 2024).
10. Y. Zhi, D. Fu, T. Yang, D. Zhang, X. Li, and Z. Pei, *Anti-Corrosion Methods and Materials* **66**, (4), 2019 (403). https://doi.org/10.1108/ACMM-11-2017-1858.
11. L. Yan, Y. Diao, Z. Lang, and K. Gao, *Science and Technology of Advanced Materials* **21**, 2020 (359). https://doi.org/10.1080/14686996.2020.1746196.
12. G. L. Hart, T. Mueller, C. Toher, and S. Curtarolo, *Nature Reviews Materials* **6**, 2021 (730). https://doi.org/10.1038/s41578-021-00340-w.
13. L. Zhu, J. Zhou, and Z. Sun, *The Journal of Physical Chemistry Letters* **13**, 2022 (3965). https://doi.org/10.1021/acs.jpclett.2c00576.
14. J. Cai, R. A. Cottis, and S. B. Lyon, *Corrosion Science* **41**, 1999 (2001). https://doi.org/10.1016/S0010-938X(99)00024-4.
15. T. Parthiban, R. Ravi, G. T. Parthiban, S. Srinivasan, K. R. Ramakrishnan, and M. Raghavan, *Corrosion Science* **47**, 2005 (1625). https://doi.org/10.1016/j.corsci.2004.08.011.
16. Y. Diao, L. Yan, and K. Gao, *Materials & Design* 2021. https://doi.org/10.1016/j.matdes.2020.109326.
17. L. B. Coelho, D. Zhang, Y. Van Ingelgem, D. Steckelmacher, A. Nowé, and H. Terryn, *npj Materials Degradation* **6**, 2022 (8). https://doi.org/10.1038/s41529-022-00218-4.
18. M. Aghaaminiha, R. Mehrani, M. Colahan, B. Brown, M. Singer, S. Nesic, S. M. Vargas, and S. Sharma, *Corrosion Science* **193**, 2021 109904. https://doi.org/10.1016/j.corsci.2021.109904.
19. X. Jiang, Y. Yan, and Y. Su, *npj Materials Degradation* **6**, 2022 (92). https://doi.org/10.1038/s41529-022-00307-4.
20. A. Shewalkar, D. Nyavanandi, and S. A. Ludwig, *Journal of Artificial Intelligence and Soft Computing Research* **9**, 2019 (235). https://doi.org/10.2478/jaiscr-2019-0006.
21. T. Lookman, P. V. Balachandran, D. Xue, and R. Yuan, *npj Computational Materials* **5**, 2019 (21). https://doi.org/10.1038/s41524-019-0153-8.
22. W. Wang, X. Jiang, S. Tian, P. Liu, D. Dang, Y. Su, T. Lookman, and J. Xie, *NPJ Computational Materials* **8**, 2022 (9). https://doi.org/10.1038/s41524-021-00687-2.
23. P. Liu, H. Huang, X. Jiang, Y. Zhang, T. Omori, T. Lookman, and Y. Su, *Acta Materialia* **235**, 2022 118101. https://doi.org/10.1016/j.actamat.2022.118101.
24. X. Jiang, B. Jia, G. Zhang, C. Zhang, X. Wang, R. Zhang, H. Yin, X. Qu, Y. Song, L. Su, and Z. Mi, *Scripta Materialia* **186**, 2020 (272). https://doi.org/10.1016/j.scriptamat.2020.03.064.
25. C. Shen, C. Wang, X. Wei, Y. Li, S. van der Zwaag, and W. Xu, *Acta Materialia* **179**, 2019 (201). https://doi.org/10.1016/j.actamat.2019.08.033.
26. C. Nyby, X. Guo, J. E. Saal, S. C. Chien, A. Y. Gerard, H. Ke, T. Li, P. Lu, C. Oberdorfer, S. Sahu, and S. Li, *Scientific Data* **8**, 2021 (58). https://doi.org/10.1038/s41597-021-00840-y.
27. C. D. Taylor and B. M. Tossey, *npj Materials Degradation* **5**, 2021 (38). https://doi.org/10.1038/s41529-021-00184-3.
28. A. Roy, M. F. N. Taufique, H. Khakurel, R. Devanathan, D. D. Johnson, and G. Balasubramanian, *Npj Materials Degradation* **6**, 2022 (9). https://doi.org/10.1038/s41529-021-00208-y.
29. G. Koch, 1 - Cost of corrosion, in Woodhead Publishing Series in Energy, Trends in Oil and Gas Corrosion Research and Technologies, eds.: A.M. El-Sherik, (Woodhead Publishing, 2017), pp. 3–30. https://doi.org/10.1016/B978-0-08-101105-8.00001-2
30. Global Information, Anti-Corrosion Coatings - Market Share Analysis, Industry Trends & Statistics, Growth Forecasts (2024 - 2029). https://www.giiresearch.com/report/moi1432779-anti-corrosion-coatings-market-share-analysis.html (accessed March 2024)
31. Y. Peng and M. H. Nagata, *Chaos, Solitons & Fractals* **139**, 2020 110055. https://doi.org/10.1016/j.chaos.2020.110055.
32. Faisal, N. H., Rajendran, V., Prathuru, A., Hossain, M., Muthukrishnan, R., Balogun, Y., Pancholi, K., Hussain, T., Lokachari, S., Horri, B. A., Bankhead, M., 2024. Thermal spray coatings for molten salt facing structural parts and enabling opportunities for thermochemical cycle electrolysis. *Engineering Reports* (Accepted, 22 May 2024).
33. B. Yildiz, J. I. Bilbao, and A. B. Sproul, *Renewable and Sustainable Energy Reviews* **73**, 2017 (1104). https://doi.org/10.1016/j.rser.2017.02.023.
34. S. Kumar and V. Bhatnagar, *Journal of Intelligent Systems and Computing* **3**, 2022 (40).

35. M.O.K. Mendonça, S.L. Netto, P.S.R. Diniz, S. Theodoridis, Chapter 13 - Machine learning: Review and trends, ed. P.S.R. Diniz, Signal Processing and Machine Learning Theory (Academic Press, 2024), pp. 869–959 https://doi.org/10.1016/B978-0-32-391772-8.00019-3
36. A. Agrawal, P. D. Deshpande, A. Cecen, G. P. Basavarsu, A. N. Choudhary, and S. R. Kalidindi, *Integrating Materials and Manufacturing Innovation* **3**, 2014 (90). https://doi.org/10.1186/2193-9772-3-8.
37. D. Shin, Y. Yamamoto, M. P. Brady, S. Lee, and J. A. Haynes, *Acta Materialia* **168**, 2019 (321). https://doi.org/10.1016/j.actamat.2019.02.017.
38. M. Kamrunnahar and M. Urquidi-Macdonald, *Corrosion Science* **52**, 2010 (669). https://doi.org/10.1016/j.corsci.2009.10.024.
39. Z. Pei, D. Zhang, Y. Zhi, T. Yang, L. Jin, D. Fu, X. Cheng, H. A. Terryn, J. M. Mol, and X. Li, *Corrosion Science* **170**, 2020 108697. https://doi.org/10.1016/j.corsci.2020.108697.
40. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Journal of Machine Learning Research* **12**, 2011 (2825).
41. J. D. Hunter, *Computing in Science & Engineering* **9**, 2007 (90). https://doi.org/10.1109/MCSE.2007.55.
42. W. McKinney, 2010. Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference, Austin, 28 June-3 July 2010, p. 56. http://conference.scipy.org.s3-website-us-east-1.amazonaws.com/proceedings/scipy2010/pdfs/mckinney.pdf
43. Z. Lu, S. Si, K. He, Y. Ren, S. Li, S. Zhang, Y. Fu, Q. Jia, H. B. Jiang, H. Song, and M. Hao, *Advances in Materials Science and Engineering* **9597155**, 2022 (1). https://doi.org/10.1155/2022/9597155.
44. S. A. Mazari, L. Ghalib, A. Sattar, M. M. Bozdar, A. Qayoom, I. Ahmed, A. Muhammad, R. Abro, A. Abdulkareem, S. Nizamuddin, H. Baloch, and N. M. Mubarak, *International Journal of Greenhouse Gas Control* **96**, 2020 103010. https://doi.org/10.1016/j.ijggc.2020.103010.
45. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, *Nature* **559**, 2018 (547). https://doi.org/10.1038/s41586-018-0337-2.
46. Knotkova, D., Kreislova, K. and Sheldon Jr, S.D., 2012. International atmospheric exposure program: summary of results. ASTM Data Series, 71.
47. B. Chico, D. De la Fuente, I. Díaz, J. Simancas, and M. Morcillo, *Materials* **10**, 2017 (601). https://doi.org/10.3390/ma10060601.

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

Appendix A. Supplementary material

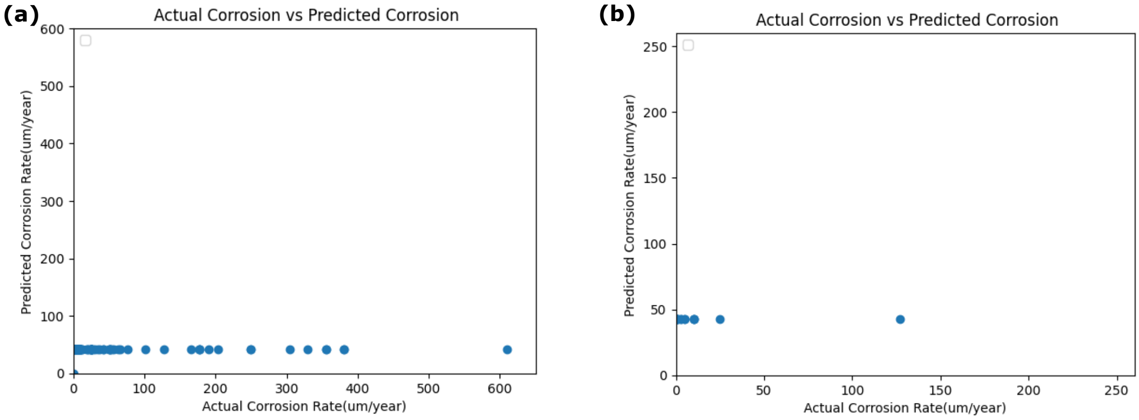


Fig. A1. Prediction accuracy of LR model: (a) train datasets, and (b) test datasets.

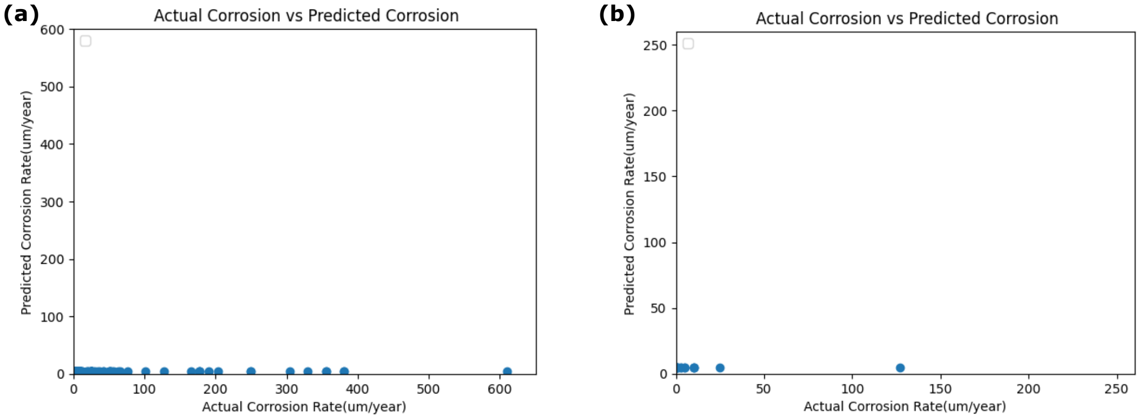


Fig. A2. Prediction accuracy of SVR model: (a) train datasets, and (b) test datasets.

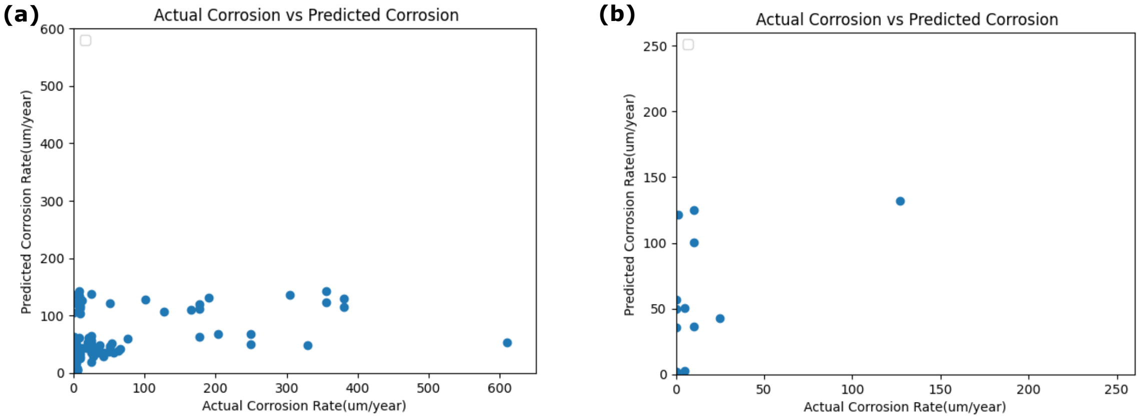


Fig. A3. Prediction accuracy of RR model: (a) train datasets, and (b) test datasets.

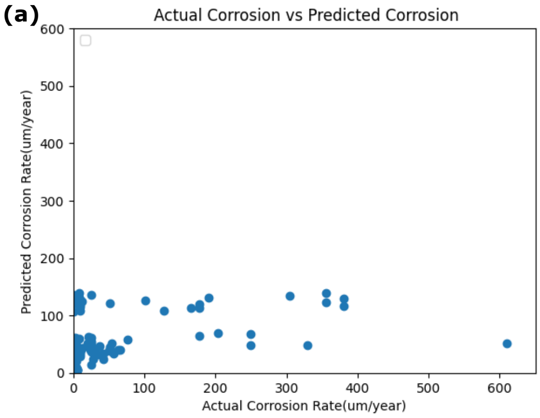


Fig. A4. Prediction accuracy of LaR model: (a) train datasets, and (b) test datasets.

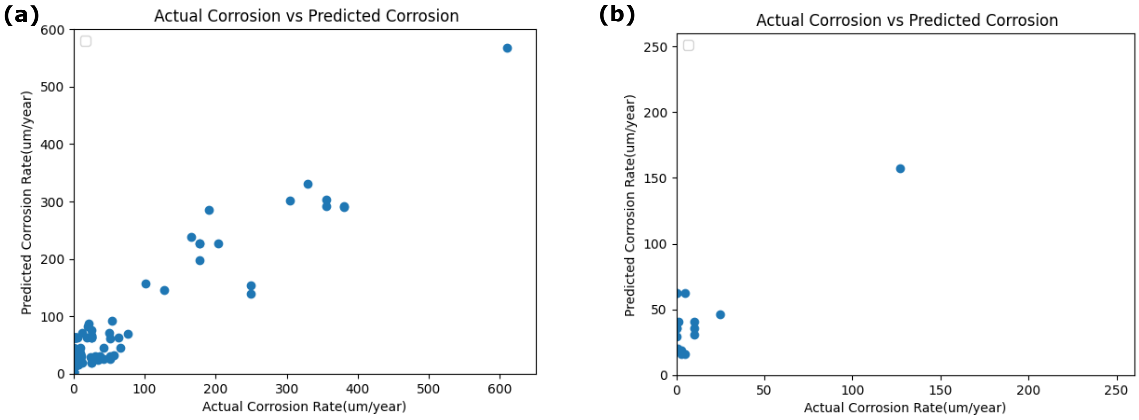


Fig. A5. Prediction accuracy of ABR model: (a) train datasets, and (b) test datasets.