

State-of-the-art progress on artificial intelligence and machine learning in accessing molecular coordination and adsorption of corrosion inhibitors

Cite as: Appl. Phys. Rev. **12**, 011302 (2025); doi: [10.1063/5.0228503](https://doi.org/10.1063/5.0228503)

Submitted: 3 August 2024 · Accepted: 2 December 2024 ·

Published Online: 6 January 2025



Taiwo W. Quadri,^{1,a)} Ekemini D. Akpan,¹ Saheed E. Elugoke,^{1,a)} Lukman O. Olasunkanmi,^{2,3} Sheetal,⁴ Ashish Kumar Singh,^{4,5} Balaram Pani,⁶ Jaya Tuteja,⁷ Sudhish Kumar Shukla,⁸ Chandrabhan Verma,^{9,a)} Hassane Lgaz,¹⁰ Valentine Chikaodili Anadebe,^{11,12,13} Rakesh Chandra Barik,^{11,12} Lei Guo,¹⁴ Akram AlFantazi,⁹ Bakang M. Mothudi,^{1,15} and Eno E. Ebenso^{1,a)}

AFFILIATIONS

¹Centre for Materials Science, College of Science, Engineering and Technology, University of South Africa, Johannesburg 1710, South Africa

²Department of Chemistry, Faculty of Science, Obafemi Awolowo University, Ile Ife 220005, Nigeria

³Department of Chemical Sciences, University of Johannesburg, P.O. Box 17011, Doornfontein Campus, Johannesburg 2028, South Africa

⁴Department of Chemistry, Netaji Subhas University of Technology, Dwarka Sector 3, New Delhi 110078, India

⁵Department of Chemistry, Hansraj College, University of Delhi, New Delhi 110007, India

⁶Department of Chemistry, Bhaskaracharya College of Applied Sciences, University of Delhi, Sector 2, Dwarka, New Delhi 110075, India

⁷School of Basic Sciences, Chemistry Division, Galgotias University, Greater Noida 201310, India

⁸Department of Sciences, School of Sciences, Manav Rachna University, Faridabad 121004, India

⁹Department of Chemical and Petroleum Engineering, Khalifa University of Science and Technology, P.O. Box 127788, Abu Dhabi, United Arab Emirates

¹⁰Innovative Durable Building and Infrastructure Research Center, Center for Creative Convergence Education, Hanyang University ERICA, 55 Hanyangdaehak-ro, Sangrok-gu, Ansan-si, Gyeonggi-do 15588, Republic of Korea

¹¹Corrosion and Materials Protection Division, CSIR-Central Electrochemical Research Institute, Karaikudi, Tamil Nadu 630003, India

¹²Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India

¹³Department of Chemical Engineering, Alex Ekwueme Federal University Ndofu Alike, PMB 1010, Abakaliki, Ebonyi State, Nigeria

¹⁴School of Materials and Chemical Engineering, Tongren University, Tongren 554300, China

¹⁵Department of Physics, College of Science, Engineering and Technology, University of South Africa, Johannesburg 1710, South Africa

^{a)}Authors to whom correspondence should be addressed: taiwoquadri27@gmail.com; quadritw@unisa.ac.za; elugose@unisa.ac.za; elugokesaheed@gmail.com; chandraverma.rs.apc@itbhu.ac.in; and ebensee@unisa.ac.za

ABSTRACT

Artificial intelligence (AI) and machine learning (ML) have attracted the interest of the research community in recent years. ML has found applications in various areas, especially where relevant data that could be used for algorithm training and retraining are available. In this review article, ML has been discussed in relation to its applications in corrosion science, especially corrosion monitoring and control. ML tools and techniques, ML structure and modeling methods, and ML applications in corrosion monitoring were thoroughly discussed. Furthermore, detailed applications of ML in corrosion inhibitor design/modeling coupled with associated limitations and future perspectives were reported.

Published under an exclusive license by AIP Publishing. <https://doi.org/10.1063/5.0228503>

TABLE OF CONTENTS

I. INTRODUCTION	1
II. MACHINE LEARNING STRUCTURE AND MODELING METHODS	2
A. Machine learning structure	3
B. Machine learning modeling methods	3
1. Linear regression	5
2. Logistic regression	5
3. Artificial neural networks	5
4. Adaptive neuro-fuzzy inference system (ANFIS)	5
5. Design of experiment (DoE)	7
6. Decision trees	8
7. Evolutionary machine learning algorithms	8
8. Algorithm selection in corrosion research	9
C. Feature selection and analysis of performance metrics in machine learning	10
1. Assessing feature selection techniques	13
2. Feature selection in machine learning for corrosion inhibitors	13
3. Analysis of machine learning performance metrics	13
4. Interpretability of ML models used in corrosion studies	13
D. Bayesian belief network (BBN) in machine learning	13
III. MACHINE LEARNING APPLICATIONS IN MATERIALS DESIGN, PROPERTIES, AND DISCOVERY	14
IV. MACHINE LEARNING FOR CORROSION MONITORING	16
A. Machine learning for atmospheric corrosion monitoring (ACM)	16
B. Machine learning for corrosion monitoring in urban environment	16
C. Machine learning for corrosion monitoring in pipelines	17
D. Machine learning for aqueous corrosion monitoring	18
E. Machine learning in atmospheric corrosion monitoring in marine environment	19
F. Machine learning for pitting corrosion resistance	19
G. Machine learning for microbiologically induced corrosion (MIC) monitoring	20
V. CASE STUDIES OF MACHINE LEARNING IN CORROSION INHIBITION MODELLING	21
A. Artificial neural network (ANN) modeling for corrosion inhibition prediction	22
B. Adaptive neuro-fuzzy inference system (ANFIS) modeling for corrosion inhibition prediction	22
C. Response surface methodology (RSM) for corrosion inhibition prediction	24
D. Random forest (RF) modeling for corrosion inhibition prediction	26
E. Extreme gradient boosting (XGBoost) modeling for corrosion inhibition prediction	27

E. Assorted machine learning techniques for corrosion inhibition prediction	30
VI. REAL-WORLD APPLICATIONS OF MACHINE LEARNING TECHNIQUES FOR CORROSION STUDIES	31
VII. CHALLENGES AND LIMITATIONS OF CORROSION INHIBITION MODELLING	32
VIII. CONCLUSIONS AND PERSPECTIVE	34

I. INTRODUCTION

Machine learning (ML) is a subset of artificial intelligence (AI) that makes use of prerecorded or historical data to predict future results without absolute programming.^{1–3} It operates on intelligent algorithms that are capable of learning with experience.⁴ Traditionally, ML is domain-specific, that is, its responsiveness depends on the field for which it is created and/or being applied. Transfer learning allows a model that had been pre-trained in one domain to be used to fast-track a process in another domain.⁵ ML would not respond to decision-making when applied in a field of study that is different from its primary design. ML is so-called because its algorithm is smart enough to learn the characteristics of the system under investigation by taking in relevant known data and improve its predictive power over time.^{6,7} ML applications require a large set of data about the system for which the algorithm is developed. ML as a technique can perform far more than human capacity. It is capable of analyzing large datasets, identifying patterns or key features in the data that could be used to characterize the system under study.⁷ It essentially becomes a domain expert on the specific topic extremely rapidly.

Modern advancements in computer hardware and software have contributed to the application of intelligence-based techniques in various fields. Applications of ML have been documented in disease analysis and recognition,⁸ biomedical image analysis,⁹ genetics and genomics,¹⁰ cancer prognosis and prediction,¹¹ minerals processing,¹² language processing,¹³ and image recognition.¹⁴ Just like in many other areas, ML has recently found applications in corrosion science. Risk assessment of corrosion in pipeline was reported by Ossai.¹⁵ Simulation of electrochemical profiles of copper corrosion in chloride and sulfide environments was reported by Gong *et al.*¹⁶ Studies that apply ML in various aspects of corrosion research, including corrosion testing, simulation of corrosion profiles, and nondestructive forecast of corrosion in reinforced concretes, have been on the rise.^{17–19}

ML approach features in corrosion inhibition studies usually through quantitative-structure-activity/property relationship (QSAR/QSPR) and artificial neural network (ANN) models.^{20–22} QSAR/QSPR-driven ML approach have been recognized as an efficient, cheap, timely, and reliable means of predicting the potential of new compounds to inhibit corrosion of metals. Correlation between experimental inhibition efficiencies and molecular/electronic descriptors has yielded some success in developing predictive models for corrosion inhibition potentials.²¹ The performance of the models is tested by using internal and external validation approaches. Using a well-performing model, new corrosion inhibitors could be computationally designed prior to laboratory synthesis. In this regard, QSAR/QSPR-based ML promises to reduce the cost of designing new corrosion inhibitors, experimental time, wastage, and by-products. ML can also help to develop the connections between the primary outputs of

quantum mechanical simulations of inhibitors with various aspects of the corrosion inhibition mechanism.^{23–25}

Recently, ML has found diverse uses within corrosion science, such as predicting and monitoring corrosion and designing corrosion resistance materials and inhibitors. The effectiveness of inhibitors, corrosion monitoring, and the impact of microbes on corrosion have all been predicted using ML algorithms. They have been successful in forecasting corrosion rates and microorganisms. They have also forecasted pitting corrosion and the rate of oil pipeline corrosion, proving their usefulness in a range of atmospheric circumstances. The literature suggests that a few previous similar review articles have been published;^{26–30} nevertheless, this review article uniquely provides a focused methodical analysis of ML and AI approaches that are particularly optimized for the coordination and adsorption of corrosion inhibitors. Unlike prevalent reviews focusing on theoretical aspects and perspectives, the present review provides the breadth of ML and AI methodologies employed across various real-world industrial corrosive environments. Through case study assessments, this article examines the effectiveness of various algorithms in addressing complex corrosion-related issues that companies encounter and describes ML's capacity for prediction. This review also provides limitations, challenges, and opportunities for using ML and AI in corrosion science.

II. MACHINE LEARNING STRUCTURE AND MODELING METHODS

A. Machine learning structure

In recent years, ML has emerged as an invaluable approach to construct models from available data, offering the potential to predict the properties of chemical inhibitors that can be synthesized and tested experimentally. The structure of ML typically involves target identification, data preparation, feature selection, model selection and development, and model application.^{31,32}

- Target identification:** The first step in conducting ML modeling is to define the objective in terms of the targeted output. The targeted property or output is expected to be learnable from the pool of available data. A clear identification of the targeted properties is crucial for subsequent ML tool selection. The targeted property or output could be corrosion rate, impedance, inhibition efficiency, etc.
- Data preparation:** Data collection can be sourced from public databases, publications, or generated internally. Examples of such public databases include CORDATA³³ accessible via <https://datacor.shinyapps.io/cordata/>, DATACORTECH³⁴ accessible via <https://datacor.shinyapps.io/datacortech/>, ExCorr³⁵ accessible via <https://excorr.web.app/database>, NACE COR-SUR accessible via <https://app.knovel.com/kn/resources/lkpCSDCORSE/toc>, and machine learning for corrosion database³⁶ accessible via <https://data.mendeley.com/datasets/jfn8yhrphd/1>. Generally, collected data include inputs (variables) and outputs (targeted properties), with inputs encompassing structural attributes, atomic parameters, elemental compositions, and processing conditions.³² The quality and quantity of datasets significantly impact the performance of ML models. The quantity of data required in ML processes typically relies on the chosen ML algorithms. In many cases, ML algorithms favor larger datasets for optimum performance.

Moreover, the quality of the dataset is closely tied to the uncertainty associated with the data.

Due to the incompleteness and noise often found in some original data, it is essential to conduct data pre-processing prior to ML implementation. This ensures the reliability of the models being developed. Data pre-processing generally involves actions such as deleting erroneous or incompatible data and normalizing the original data.^{37,38} Additionally, when sourcing data from publications or databases, thorough checks should be conducted to guarantee accuracy and authenticity. This meticulous approach to data pre-processing contributes significantly to the robustness and effectiveness of ML models.³⁹

- Feature selection:** Effective feature (input variable) selection is fundamental for robust ML models.⁴⁰ In the case of molecular descriptors, they are expected to be related to a wide range of compounds, easily calculable, sensitive to target values, and structurally interpretable.^{38,41,42} Both supervised and unsupervised feature selection methods play a critical role in enhancing model performance. Feature selection is directed at identifying a pertinent subset of features to train the model, with the goal of retaining valuable features while eliminating redundant or irrelevant ones. The resulting effective subset offers benefits such as minimizing training time, mitigating the curse of dimensionality, preventing over-fitting, and simplifying the model, making it both interpretable and universally applicable.⁴³ The commonly employed methods for feature selection encompass both supervised and unsupervised approaches.
- Model selection and development:** Selecting the appropriate ML model is essential for mapping existing data without overfitting or underfitting. The choice of algorithms depends on the learning styles, with options ranging from unsupervised to supervised and semi-supervised learning. Factors such as dataset size, learning rate, and the type of algorithm influence model construction.

ML algorithms can be classified into three main types based on their learning styles, namely, supervised learning, unsupervised, and semi-supervised learning.⁴⁴ In supervised learning, a set of labeled samples are used to train the models. This learning process revolves around the algorithm learning from labeled data, mapping input features to corresponding output labels. Unsupervised learning places emphasis on training unlabeled samples to infer functions that describe hidden structures within the data. The objective is to identify patterns or relationships within the data without predefined output labels. Semi-supervised learning, on the other hand, bridges the gap between supervised and unsupervised learning approaches. This method leverages labeled and unlabeled data, improving model accuracy, particularly in scenarios where labeled training data are limited.⁴⁵ The model utilizes the available labeled data along with additional unlabeled data to enhance its overall performance. Figure 1 presents a thorough overview of AI and ML algorithms together with their numerous subclasses.

Furthermore, understanding these learning styles is essential for selecting the most appropriate ML algorithm based on the characteristics of the available data and the specific objectives of the learning task. The commonly used ML algorithms are depicted in Fig. 2.³²

During the development of ML models, data are typically partitioned into training and testing set based on a specified ratio. The training set is utilized to discern hidden patterns, whereas the testing

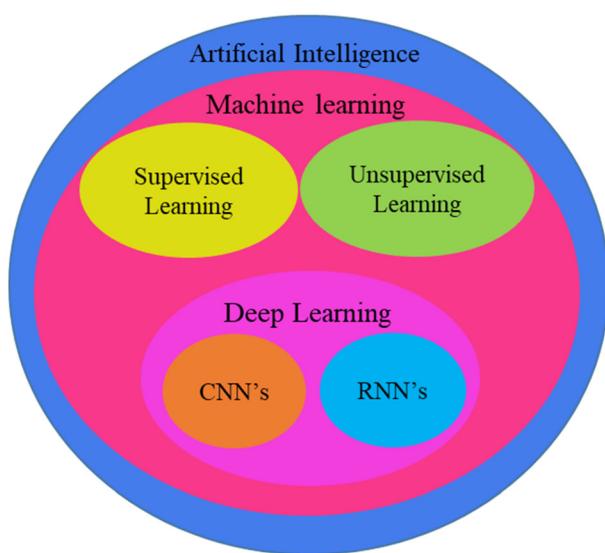
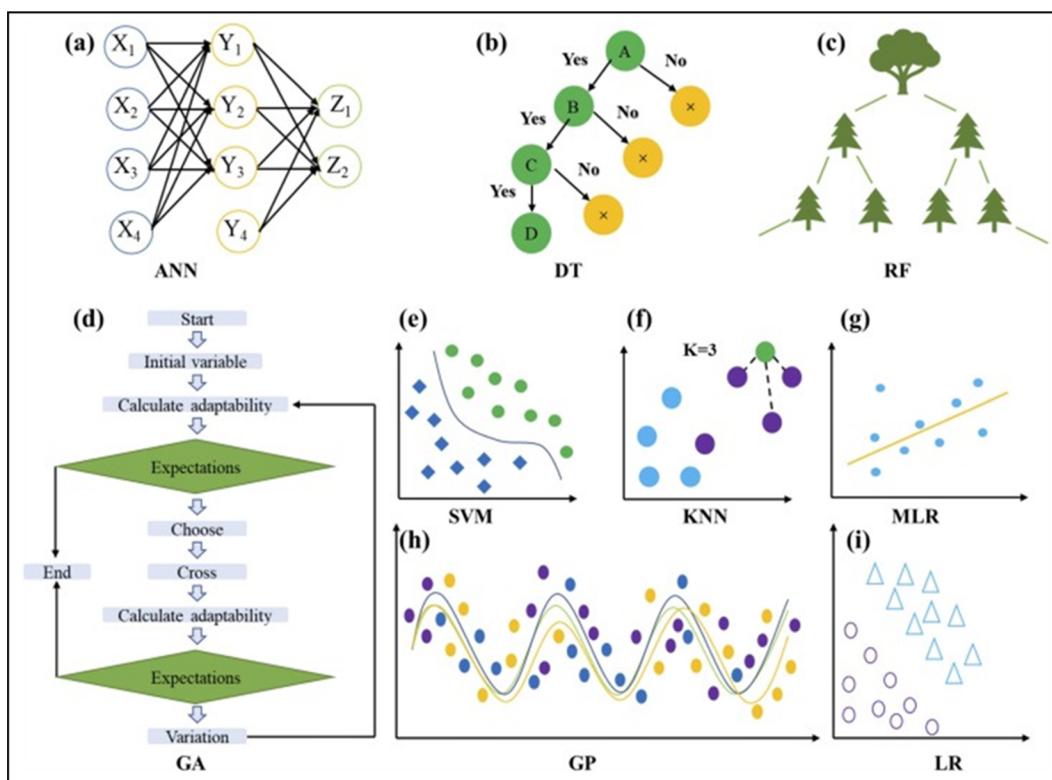


FIG. 1. Overview of AI and ML algorithms with subclasses.

set is used to serve to appraise the model's generalization ability. To assess the reliability of ML models, bootstrapping and cross-validation (CV) techniques are employed.

For regression models, diverse evaluation metrics come into consideration, encompassing the coefficient of determination (R^2), Pearson correlation coefficient (r), and root mean square error (RMSE). Reduced RMSE values generally signify enhanced model performance, while elevated R^2 and r values indicate improved model fitting. The optimal model is discerned through CV, and the testing set is subsequently employed to appraise the model's generalizability by predicting the properties of new materials. Various packages and tools are presently utilized for constructing ML models.^{23,32}

5. Model application: Once the ML model is constructed, its application involves discovering novel inhibitor candidates through methods like high-throughput screening. Virtual samples are designed, and ML models predict their properties, guiding experimental synthesis. Models can also be shared on web servers, enabling convenient online predictions and wider accessibility. ML proves to be a powerful tool for accelerating effective inhibitor design, providing predictive capabilities and insights into inhibitors' properties. Figure 3 presents a typical ML workflow for a systematic and effective application of ML techniques.⁴⁶

FIG. 2. Representations of some widely used machine learning algorithms in corrosion science: (a) Artificial neural network (ANN), (b) decision trees (DT), (c) random forests (RF), (d) genetic algorithm (GA), (e) support vector machine (SVM), (f) K-nearest neighbor (KNN), (g) multiple linear regression (MLR), (h) Gaussian process (GP), (h) linear regression (LR). Reproduced with permission from Liu et al., J. Alloys Compd. 921, 165984 (2022). Copyright 2022 Elsevier.³²

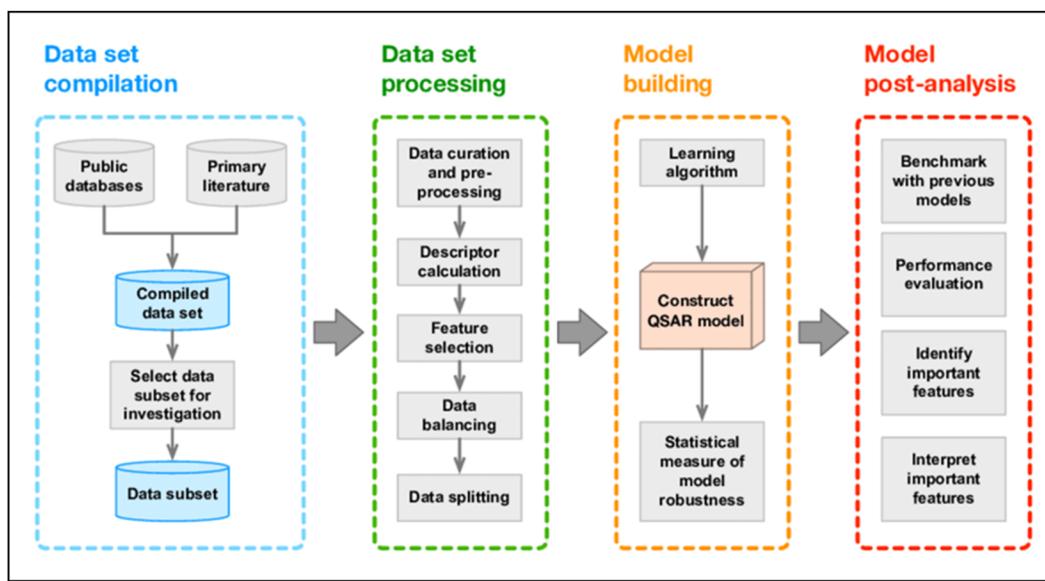


FIG. 3. A typical machine learning workflow. From C. Nantasesamat, *Ecotoxicological QSARs*. Copyright 2020 Springer. Reproduced with permission from Springer.⁴⁶

B. Machine learning modeling methods

1. Linear regression

Linear regression (LR) is a fundamental and widely used technique in the field of ML. This algorithm utilizes supervised learning to predict a continuous outcome variable (dependent variable, Y) by analyzing one or more predictor variables (independent variables, X).^{47,48} It assumes a linear relationship between the variables, meaning that a change in the predictor variables is associated with a constant change in the outcome variable. Its most basic form, otherwise known as simple or univariate linear regression, correlates one X and one Y. In corrosion science, the linear model commonly employed is a modified Hansch-type equation known as the Lukovits–Kálman–Pálinská (LKP) model, which has gained extensive use in correlating calculated and measured inhibition efficiencies.^{49,50} On the other hand, multiple linear regression (MLR) extends to multiple predictor variables as it serves as a predictive tool for a chemical activity (inhibition performance) based on two or more independent predictors (molecular variables). In an earlier review,²³ we extensively covered the use of regression analyses in corrosion inhibition modeling.

2. Logistic regression

Logistic regression shares several similarities with LR. It is often employed to model the probability of an event occurring as a linear function of a set of predictors. This method serves as a statistical technique employed in binary classification scenarios where the outcome variable is restricted to 0 or 1. It can be expanded to handle cases of multivariate classification, such as through the application of polynomial logistic regression. Like MLR, the regression coefficients in logistic regression can elucidate the impact of a molecular descriptor on the predicted outcome. A large coefficient value signifies that the molecular descriptor strongly influences the probability of the outcome, whereas a coefficient with a zero value indicates that the molecular

descriptor has no impact on the probability of the outcome. Additionally, the sign of the coefficients plays a crucial role; a positive coefficient increases the probability of the outcome, while a negative coefficient has the opposite effect. One major limitation of logistic regression is its difficulty in handling high-dimensional data effectively, especially when the number of predictors is large.⁵¹ The application of logistic regression includes predictive modeling for the evaluation of pitting corrosion in C-steel,⁵² atmospheric corrosion rate of low alloy steels,⁵³ and concrete chloride resistance.¹⁸

In addition to the popular linear and logistic regressions, several other LR models have been reported in corrosion inhibition modeling. For instance, in a 2023 study, predictive QSPR models were developed for 13 furan-based corrosion inhibitors using four different LR models, namely, MLR, partial least squares (PLS) regression, principal component analysis (PCA), and principal component regression (PCR).⁵⁴ The study reported that the PCR model yielded the most promising results.

3. Artificial neural networks

Prior to the emergence of neural networks around the 1990s, excitement about existing AI and expert systems had begun to wane due to their practical limitations. The genesis of modern advancements in neural network training can be traced back to the pivotal moment when Rumelhart *et al.* reinvented the backpropagation (BP) algorithm in 1986. This innovation marked a significant breakthrough, especially in the training of feedforward neural networks, enabling them to effectively learn concealed patterns within input–output data.⁵⁵

Neural network analysis represents an AI modeling method, characterized by a sophisticated approach to constructing models that can effectively interpret data through nonlinear functions. The complexity of corrosion, being a highly intricate and nonlinear phenomenon, poses challenges for description using empirical rules or

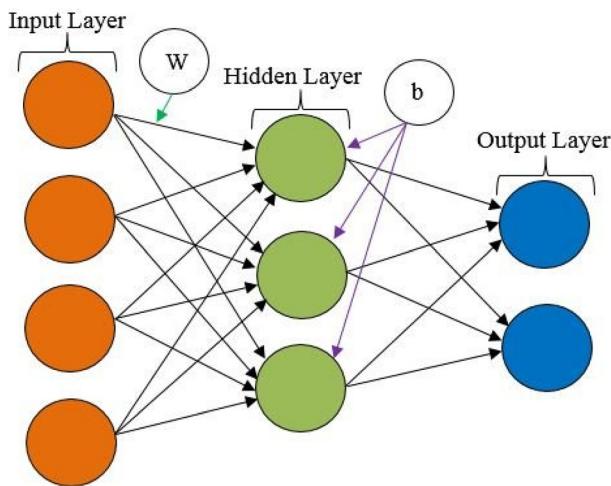


FIG. 4. A simple ANN structure. Reproduced with permission from Amin et al., Materials **15**, 4330 (2022). Copyright 2022 Authors, licensed under a Creative Commons Attribution (CC-BY) license.⁵⁹

conventional analytical methods. This makes ANN particularly well-suited for studying corrosion.^{24,56}

Neurons in a neural network are arranged into layers and connected in a specific pattern, mirroring the interaction observed in natural systems. Each connection between two neurons is assigned a weight coefficient. The prevalent structure for neural networks is the multilayer layout, commonly known as a feedforward network. This architecture typically consists of one or more hidden layers comprising sigmoid neurons, followed by output nodes. These networks can uncover complex relationships between inputs and outputs by stacking layers of neurons with specialized nonlinear activation functions, allowing them to model both linear and nonlinear patterns. Furthermore, the output layer is designed to produce values that exceed the typical range of 1 to +1, enabling the network to generate a wider spectrum of outputs.⁵⁷

The structure of the neural network is depicted in Fig. 4, featuring two main layers: the input layer (variable layer) and the output layer.⁵⁷ The input layer receives the initial input to predict the response variable, with the nodes in this layer expressing the network's prediction. Hidden layers, which can vary in number based on the complexity of the process, play a critical role in processing the input information through a transfer function.⁵⁸

A typical ANN structure includes an input layer, an output layer, and one or more hidden layers. Figure 5 illustrates the topology of a conventional BPNN model. Here, each node comprises a neuron that interacts with neurons in adjacent layers, facilitating information exchange between the input, hidden, and output layers. The input layer processes a stable dataset and applies a nonlinear transformation to the weighted inputs, passing the output to the next layer. If the final output does not meet the desired criteria, an error signal is propagated backward, adjusting the connection weights between nodes in the hidden and input layers, improving its performance through iterative optimization.

Before designing a BPNN, three crucial factors must be specified: the number of nodes, the learning rate, and the activation function. The size of the network is known to have a significant impact on performance, as fewer nodes may result in faster training times and reduced accuracy, while an excessive number of nodes can lead to overfitting and prolonged training durations. The learning rate, which governs the rate of weight adjustment, affects the learning speed, and selecting an optimal rate is crucial to prevent weight fluctuations. Furthermore, the choice of transfer functions in the hidden layer plays a pivotal role in determining the efficiency and accuracy of ANNs.

Determining the pace at which weights are updated impacts how quickly the network learns. Selecting an optimal learning rate is crucial to avoid weight shocks. The learning function in the hidden layer can also significantly impact the accuracy and efficiency of ANNs. Therefore, careful consideration of these three parameters is essential when constructing a BPNN to ensure positive outcomes.⁵⁸ Mathematically, Eq. (2.1) is used to represent the neural network architecture

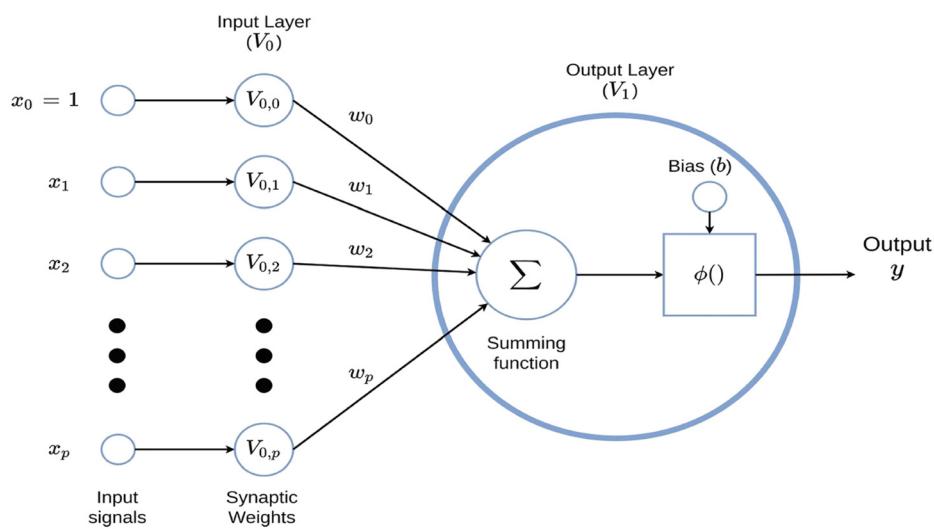


FIG. 5. A backpropagation neural network structure. Reproduced with permission from Montesinos Lopez et al., *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (Springer, 2022). Copyright 2022 Authors licensed under a Creative Commons Attribution (CC-BY) license.⁶⁰

$$y = F \left(\sum_{i=0}^m x_i w_i + b \right). \quad (2.1)$$

In this context, y denotes the output, x_i signifies the input, w_i represents the weight associated with input x_i , m indicates the number of neurons, F denotes the transfer functions, and b represents the bias. The network processes the weighted sum of inputs through nodes using the activation function to produce an output. Logistic (sigmoid) and tan-sigmoid functions stand out as commonly used activation functions in ANN modeling.⁶¹ In contemporary times, there has been an impressive interest in the use of different ANN in the field of corrosion science.^{23,24}

4. Adaptive neuro-fuzzy inference system (ANFIS)

The adaptive neuro-fuzzy inference system (ANFIS) harmoniously combines the learning capabilities of neural networks and the expertise of fuzzy logic to handle uncertainty handling of fuzzy logic, enabling it to dynamically optimize the membership function distribution. This adaptive integration empowers ANFIS for enhanced performance and robustness. Based on the Takagi–Sugeno fuzzy inference system, ANFIS is an ANN algorithm with training capability for approximating nonlinear functions. The fuzzy systems provide a potent tool for modeling and resolving complicated and nonlinear issues that are outside the purview of traditional analytical routes. Recent studies have highlighted the potential of ANFIS to act as an excellent predictive algorithm, particularly in examining the interaction between dependent and independent variables.^{61–64} The versatility of this type of model makes it appropriate for a wide range of applications, including control systems, prediction and forecasting, pattern recognition, and noise cancelation. Its ability to model nonlinear relationships and learn from data makes it particularly useful in fields like finance, healthcare, engineering, and environmental science. Figure 6 presents an overview of ANFIS design with multi-input and single output (MISO) model.

5. Design of experiment (DoE)

The design of experiment (DoE) is imperative in AI and ML prior to optimization and modeling. This stage begins with scrutiny and proper planning of experiments to efficiently explore the variables' space. To choose a collection of conditions for conducting experiments, strategies like factorial designs or central composite designs (CCD) are employed. The next stage involves model development. At this stage, data from the experimental route are used to develop a mathematical model, typically a polynomial equation that predicts the response based on the variables. The final stage is called model validation. At this point, the optimal conditions predicted by the model are tested in additional experiments to verify the predictions of the model.

a. Response surface methodology (RSM). The response surface methodology (RSM) is a set of statistical and mathematical tools that are helpful in modeling and analyzing complex systems where multiple variables affect an interest response. The main objective of RSM is to determine the best response for the variables in order to maximize this response, which might be any measurable quantity influenced by the variables. This approach is very helpful in studies where there are much interactions and complex influences from the variables. The core principle of this model revolves around the design of an approximate model of the response surface essentially a mathematical function that describes the response as a function of predictor variables.^{65,66} By fitting this model to experimental data, researchers can explore the effects of different variables on the response and identify conditions that optimize the response. This statistical tool has gained significant attention in science and engineering discipline due to its robustness and accuracy of the model in analyzing tough engineering problems. The model data also operates on the principles of coded and actual factors to make its prediction.

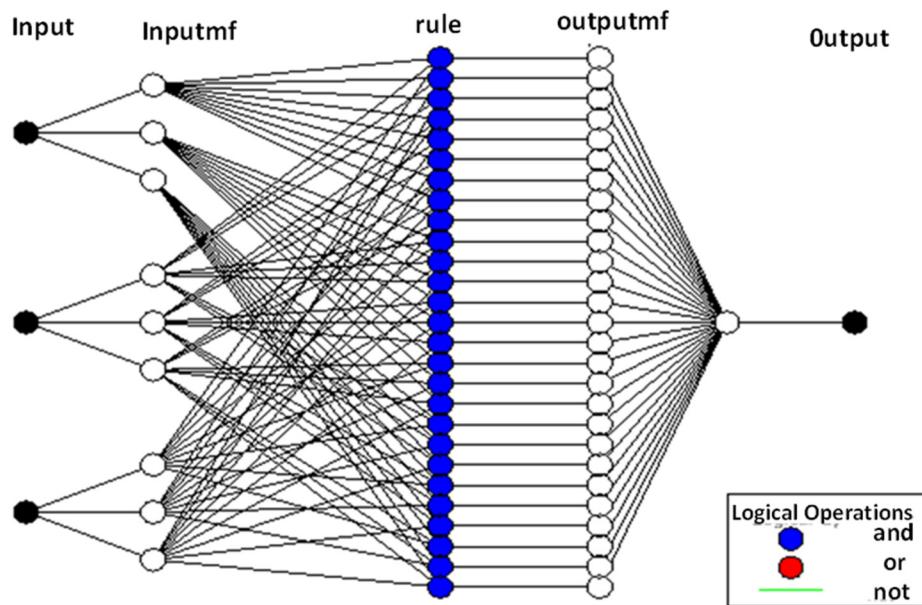


FIG. 6. ANFIS model design. Reproduced with permission from Anadebe *et al.*, *J. Mol. Liq.* **349**, 118482 (2022). Copyright 2022 Elsevier.²⁷⁷

6. Decision trees

Decision trees (DT) represent a fundamental class of algorithms within the domain of ML. A DT is a flow chart-like structure comprising nodes representing tests on attributes, branches corresponding to test outcomes, and leaves representing class labels or regression values as shown in Fig. 7.⁶⁷ The model operates by partitioning the input space into distinct regions, each corresponding to a particular prediction. Starting from the root node, the algorithm assesses an input against the criteria defined at each node, traversing down the branches until reaching a leaf, where a prediction is made.^{68,69} The importance of DT in ML modeling is multi-faceted. DT offer transparency and interpretability as well as enabling a clear understanding of the underlying decision-making process. They can handle both categorical and numerical features, making them applicable to a wide array of tasks including classification, regression, and multi-output problems. Additionally, DT serve as foundational elements for more complex ensemble methods such as random forest (RF), boosting the overall predictive performance.^{70–72}

a. Construction and pruning of decision trees. The construction of DT involves a recursive partitioning of the data space into homogeneous regions, driven by specific criteria that aim to maximize information gain or minimize impurity at each node. Several algorithms have been developed to facilitate this process, each with unique characteristics and applications.^{73–75} These algorithms include the following:

1. **ID3 (Iterative dichotomizer 3):** Utilizes the information gain criterion, which relies on the entropy measure to determine the best attribute for splitting at each node.^{76,77}
2. **C4.5:** An enhancement of the ID3 algorithm, C4.5 employs the gain ratio, which normalizes the information gain using split

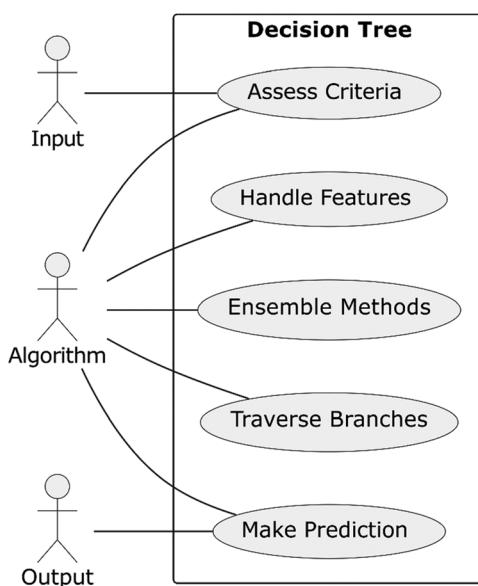


FIG. 7. A flow chart representation of nodes, branches, and leaves for machine learning modeling.

information. This adjustment alleviates the bias toward attributes with many levels.^{75,78}

3. **CART (Classification and regression trees):** Adopts the Gini impurity as a splitting criterion, making it applicable to both classification and regression tasks. The Gini impurity measures the frequency at which a randomly chosen sample would be incorrectly labeled if randomly classified according to the distribution of labels in the subset.^{79–81}

Constructing DT without constraints can lead to overfitting, resulting in a model that excels on training data but fails to generalize well, performing poorly on new, unseen data. Pruning techniques are employed to counteract this by simplifying the tree and improving generalization.

1. **Reduced error pruning:** This approach involves replacing individual subtrees with leaves and assessing whether the change improves or maintains the validation set accuracy.^{82–84}
2. **Cost complexity pruning:** Also known as weakest link pruning, this method introduces a complexity parameter that balances the tree's fit to the data with its size, allowing for a more nuanced tradeoff between over-fitting and under-fitting.^{82,85}

7. Evolutionary machine learning algorithms

Evolutionary machine learning algorithms (EMLAs) represent a class of optimization techniques inspired by the principles of natural evolution (Fig. 8). These algorithms encompass methods that emulate the biological processes of selection, mutation, crossover, and reproduction to explore and exploit the search space of potential solutions.^{86,87} EMLAs operate by maintaining a population of candidate solutions and iteratively applying stochastic operators that reflect natural evolutionary processes.^{88–91} These processes involve the following:

1. **Selection:** This process chooses individuals from the current population based on their fitness, favoring those that perform better in the given task. Various selection methods exist, such as tournament selection and roulette wheel selection, each with distinct characteristics.

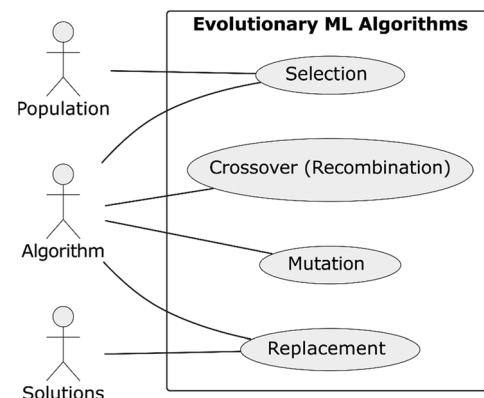


FIG. 8. A visualization of the processes of selection, crossover, mutation, and replacement in optimization techniques.

2. **Crossover (Recombination):** Selected individuals are paired and interchanged to create offspring, promoting the exploration of new regions in the solution space. Crossover methods might include one-point, multi-point, or uniform crossover.
3. **Mutation:** Random changes are introduced to individual solutions, allowing for diversification and prevention of premature convergence to local optima. The mutation might involve flipping bits in binary representations or adding Gaussian noise to real-valued parameters.
4. **Replacement:** Offspring are integrated into the population, and less fit individuals may be replaced, maintaining the population size and gradually driving the search toward optimal or near-optimal solutions.

a. Main types of evolutionary machine learning algorithms.

1. **Genetic algorithm:** Genetic algorithm (GA) constitutes one of the most widely recognized types of EMLA. Mirroring the genetic processes of biological organisms, GA utilizes a chromosome-like representation for solutions and apply genetic operators such as selection, crossover, and mutation to evolve the population. GA has been successfully implemented in various optimization and search problems, exhibiting remarkable adaptability and efficiency.^{89,90}
2. **Genetic programming:** Genetic programming (GP) extends the principles of GAs by evolving not just the parameters but the entire structure of the solution, typically in the form of computer programs or symbolic expressions. This ability to co-evolve both structure and parameters lends GP a unique flexibility, making it suitable for tasks where the functional form of the solution is unknown or complex.^{92–94}

b. Applications of evolutionary machine algorithms in machine learning.

(1) Feature selection

In ML, the selection of relevant features is crucial for model performance and interpretability. EMLAs, particularly GA and GP, have been leveraged for feature selection by encoding the inclusion or exclusion of features as binary strings. The evolutionary search efficiently identifies optimal or near-optimal subsets, reducing dimensionality, and enhancing generalization.^{95,96}

(2) Hyperparameter tuning

Hyperparameters govern the behavior of learning algorithms, and their optimal setting often eludes analytical determination. EMLAs offer a robust solution to this challenge by encoding hyperparameters as individuals within a population and employing evolutionary processes to tune them. This approach navigates the complex and often nonconvex hyperparameter space, finding configurations that improve model accuracy and robustness.^{97,98}

The main types of EMLAs, notably GA and GP, manifest a broad spectrum of applications in contemporary ML. Their application in feature selection and hyperparameter tuning signifies their vital role in enhancing model performance and efficiency. By leveraging the principles of natural evolution, these algorithms bridge the gap between biological systems and computational optimization, marking a significant advancement in the adaptive problem-solving capabilities of ML.

8. Algorithm selection in corrosion research

Selecting the right ML algorithm for corrosion-related tasks requires evaluating problem specifics, data characteristics, and desired outcomes. Below are concise guidelines inspired by current research.

a. Data characteristics. The choice of ML algorithm in corrosion research is heavily influenced by the characteristics of the dataset, particularly volume and feature types. Large, high-quality datasets favor algorithms like ANN and SVM, which are optimized for handling substantial data and complex patterns. However, for smaller datasets, where missing data may be present, DT and RF exhibit robustness and adaptability. Additionally, datasets containing a mix of numerical and categorical features are better suited for DT and RF due to their inherent ability to process such data without extensive preprocessing, making them practical choices in various corrosion studies.⁹⁹

b. Task complexity. The complexity of the corrosion-related task directly informs the algorithm selection. Predictive tasks, such as forecasting corrosion rates or material degradation, benefit from regression models like RF or ANN, which can capture intricate data patterns. In contrast, classification tasks, such as categorizing materials by their corrosion resistance, are well-suited to DT, SVMs, or ensemble methods, which provide effective classification capabilities with manageable computational demands. For image-based analyses, which are becoming prevalent in automated corrosion inspection, convolutional neural networks (CNNs) excel in detecting and segmenting corrosion, enabling precise surface analysis in structural inspections.¹⁰⁰

c. Interpretability vs performance. Corrosion research applications may require either high interpretability or maximum predictive accuracy, with algorithm selection tailored to this balance. DT and LR models are valued for interpretability, as they provide clear decision-making processes essential for research transparency and practical deployment. When predictive performance is prioritized, especially in complex or high-dimensional data settings, more sophisticated models such as neural networks or ensemble methods like RF often yield superior accuracy but at the cost of interpretability. The choice thus depends on the specific requirements of the corrosion-related task.

d. Computational resources. Computational resources play a critical role in algorithm selection for corrosion studies, especially when considering algorithmic complexity. In resource-constrained environments, simpler models such as DT and logistic regression are advantageous due to their lower computational requirements, allowing effective modeling without extensive hardware. Conversely, if resources are abundant, more complex models, including neural networks, become feasible, providing higher predictive power that can justify the increased computational load. Hence, resource availability is a practical determinant in choosing the algorithm for efficient and scalable corrosion research solutions.

e. Specific corrosion applications. Specific ML applications in corrosion research showcase the versatility of algorithms like genetic algorithms and CNNs. For feature selection, genetic algorithms efficiently reduce dimensionality by identifying the most relevant corrosion-influencing variables, optimizing model input for improved

interpretability and performance.¹⁰¹ In automated corrosion detection, CNNs are instrumental in analyzing structural images, facilitating rapid and accurate segmentation of corroded areas, and enhancing the reliability of inspection processes. These specialized applications underline the adaptability of ML algorithms in advancing corrosion studies, from optimizing feature selection to enabling automated damage detection.¹⁰²

C. Feature selection and analysis of performance metrics in machine learning

In ML, feature selection is essential for maximizing model performance while minimizing computational costs. The objective of ML research is to design AI systems that can autonomously acquire knowledge from data, make informed decisions, and draw accurate conclusions without human intervention.¹⁰³ Datasets contain several features or variables that characterize input data in various ML applications. While some features may be more useful than others for a given learning activity, not all features are created equal.^{104,105} Selecting the optimal set of features that has the most influence on the ability of a ML model to make accurate predictions is the aim of feature selection.^{106,107} Features are chosen in supervised, unsupervised, and semi-supervised models, respectively, depending on the type of training data used. As can be seen in Fig. 9, there is a single framework for selecting features in a supervised, unsupervised, or semi-supervised manner.¹⁰⁸

Feature selection is generally driven by several variables:^{103,109,110}

1. **Explainability and interpretability:** Feature selection helps ML models be more explainable by determining which features are most important in making decisions. By zeroing in on a more manageable subset of features, the factors driving model predictions can be better communicated and understood.
 2. **Better model performance:** Overfitting, when the model gets too complicated and performs badly on unknown data, can occur when unnecessary or redundant characteristics are included in the learning process. Models can improve their generalization

and prevent overfitting by only using the most informative features.

3. **The curse of dimensionality:** ML algorithms struggle with high-dimensional datasets that have many features. A phenomenon known as “the curse of dimensionality” describes how finding meaningful patterns or links becomes more challenging as the number of dimensions grows.
 4. **Faster model training, prediction, and deployment:** This can be achieved by minimizing computing costs by reducing the number of features. When working with massive datasets or time-sensitive software, this is crucial.

There are three major categories to consider when classifying feature selection strategies: filter approaches, wrapper approaches, and embedding approaches.^{106,111,112} Methods that evaluate feature subsets using the target learning algorithm are known as wrapper methods, while techniques that embed feature selection within the learning process are known as embedded methods.^{113,114} Table I provides a comparison of the embedded, wrapper, and filter models.¹⁰⁷

To overcome the difficulties of high-dimensional data and intricate learning tasks, sophisticated feature selection algorithms have arisen in recent years, making use of ideas like stability, sparsity, and feature importance.¹¹⁵ The future of feature selection in ML will be determined by these methods, as well as domain-specific knowledge and interpretability needs. Typically, the ML model assesses the effectiveness of the feature selection approach. Hierarchical clustering, K-means, Naive Bayes, BPNN, k-nearest neighbor (KNN), radial basis function neural network (RBF-NN), C4.5, support vector machine (SVM), density-based clustering, and other commonly used ML models are examples.^{116,117} Feature selection is an essential part of the ML process since it improves model efficiency, accuracy, interpretability, and scalability. It is still a thriving field of study, with new advances being made all the time to meet the demands of cutting-edge datasets and learning algorithms.

1. Assessing feature selection techniques

It is essential to assess the efficacy of variable selection approaches before settling on the best approach for an issue. The effect of feature selection on the ML pipeline is measured using a variety of evaluation methodologies and performance indicators.¹¹⁸

- Performance metrics:** To evaluate the efficacy of feature selection strategies, performance metrics offer quantitative measurements. Whether a classification, regression, or clustering task is being evaluated, the metrics chosen will vary. Common indicators of success include regression, clustering, classification accuracy, F1 score, mean squared error, precision and recall, and root mean squared error. Whether a classification, regression, or clustering task is being evaluated, the metrics chosen will vary. Classification tasks benefit more from accuracy, while regression tasks benefit more from MSE or RMSE.^{112,118}
 - Cross-validation techniques:** A common method for evaluating the effectiveness and generality of feature selection algorithms is cross-validation. It reduces the risk of overfitting and gives more reliable estimates of model performance on previously unknown data. Common methods of CV include K-fold, stratified K-fold, and leave-one-out cross-validation (LOOCV).^{119,120} Researchers

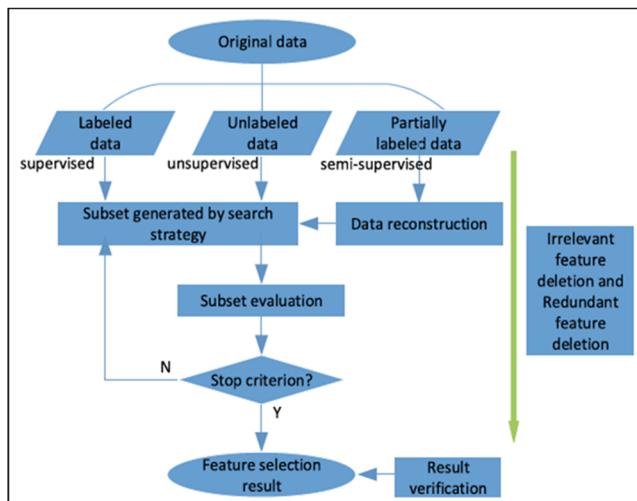


FIG. 9. Methodology for choosing important features. Reproduced with permission from Cai *et al.*, Neurocomputing **300**, 70–79 (2018). Copyright 2018 Elsevier.¹⁰⁸

TABLE I. Analyzing the differences between the wrapper, filter, and embedded approaches. Reproduced with permission from Dhal and Azad, *Appl. Intell.* **52**, 4543–4581 (2022). Copyright 2022 Elsevier.¹⁰⁷

Parameters	Wrapper approach	Embedded method	Filter methods
Procedure			
Assessment	Cross-validation	Cross-validation	Statistical test
Criteria	Feature subset usefulness	Feature subset relevance	Feature subset relevance
Search	Search all possible feature subsets.	The search is governed by the learning process.	By arranging features in a certain hierarchy (either through nested feature subsets or through feature ranking).
Results	<ul style="list-style-type: none"> • Overfitting occurs frequently. • It chooses the most helpful features but is quite time-consuming. 	<ul style="list-style-type: none"> • Overfitting is less likely to occur. • Relatively less complex in terms of time. 	<ul style="list-style-type: none"> • It has strong resistance to overfitting. • The successful selection of beneficial features is not guaranteed.

can use cross-validation to examine the performance of various feature selection approaches and test their stability across numerous training and validation sets. Models that forecast corrosion rates using sensor data, environmental factors, and previous corrosion data can be trained *via* CV. Testing procedures guarantee that the model maintains its accuracy under various circumstances and materials.³⁶ The CV technique makes sure that biased samples or overfitting would not cause the model's performance to be overly optimistic. In corrosion monitoring, CV aids in more accurate performance evaluation because corrosion data are frequently sparse or unbalanced.¹²¹ The method further tracks errors across various data segments to determine how the model improves decision making. This is essential for forecasting corrosion rates under different circumstances, such as temperature, humidity, or exposure time. For instance, Wang *et al.*¹²² described a method for utilizing a ML model to forecast the maximum depth of corrosion in subterranean oil and gas pipelines. By fine-tuning model hyperparameters and performing feature extraction using PCA, the authors' extreme gradient-boosted decision tree (XGBoost) model, which is optimized through grid search cross-validation (GSCV), noticeably improves prediction accuracy. In addition, if the dataset is relatively small, it is imperative to use the stratification method which further ensures that the dataset is represented in each fold.

3. **Robustness and stability measures:** It is crucial to evaluate the robustness and stability of feature selection algorithms to establish their consistency and dependability. The robustness of a feature selection approach is measured by its ability to yield the same results regardless of whether it is used on the entire dataset or on a subset of it. The level of agreement between feature rankings produced by various feature selection algorithms can be measured using stability measures such as stability selection and consensus stability.¹²³

2. Feature selection in machine learning for corrosion inhibitors

The use of ML techniques to foretell corrosion inhibition performance is an exciting new direction. However, proper feature selection is crucial to the efficiency of such models. Feature selection in ML for corrosion inhibitors is discussed in this part, along with the necessary approaches and factors that should be considered when doing so.

a. Importance of feature selection. An accurate and interpretable ML model for corrosion inhibitors relies heavily on a well-selected set of features. The following are some of the most important considerations when choosing features:^{124–126}

1. **Relevant information:** Chemical composition, ambient conditions, and physical qualities all have an impact on corrosion inhibition. Feature selection aids in identifying the most important features that capture critical information about the corrosion process and inhibitor performance.
2. **Dimensionality reduction:** Corrosion data typically have a vast array of features, which might result in the curse of dimensionality. High-dimensional data can degrade model performance by increasing computational complexity and introducing noise. Feature selection strategies minimize the number of features, solving these issues while increasing model efficiency.
3. **Interpretability and insights:** Corrosion inhibition is a complicated process, and feature selection assists in determining the most essential aspects impacting corrosion rates. ML models become more interpretable by picking relevant features, providing insights into the fundamental mechanisms of corrosion inhibition. This information can be used to build effective inhibitors and to inform corrosion prevention techniques.

b. Feature selection techniques for corrosion inhibitors. Several feature selection methods can be used to zero in on important features for making predictions about corrosion inhibitor performance. Features include the following:^{127–132}

1. **Mutual information:** A statistical measurement of the association between two variables, such as the rate of corrosion and the effectiveness of an inhibitor, is known as mutual information. It measures how much we can learn about one variable if we know something else. Corrosion inhibition can be predicted with greater accuracy from features with higher mutual information scores.
2. **Correlation analysis:** Chemical variables, ambient circumstances, and material characteristics are just some of the factors that have been shown to connect with corrosion rates and the efficacy of inhibitors. Correlation analysis can be used to identify features that are both highly predictive and highly associated with the target variable.

3. **Random forest importance:** The relative relevance of features can be determined using RF models. The significance ratings show how each feature helps the model perform predictions. Features with higher significance scores are more likely to accurately predict the effectiveness of a corrosion inhibitor.
4. **Recursive feature elimination (RFE):** RFE is a wrapper approach that iteratively removes features that are not crucial to the model's success. RFE determines which features are most important for forecasting corrosion inhibition by eliminating features one at a time and evaluating model performance.

Corrosion inhibitor feature selection requires attention to domain knowledge, data quality and pre-processing, overfitting, and feature interactions. Accurate and interpretable ML models for corrosion inhibitors rely heavily on careful feature selection. It aids in determining which features are most important, decreases model complexity, improves performance, and elucidates the process of inhibiting corrosion.^{133,134} In order to aid in the creation of efficient corrosion inhibitors and direct preventative measures, ML models can identify the most informative attributes.¹³⁵

It has been demonstrated that feature selection can both improve the interpretability and performance of the corrosion IE predictive models.¹³⁶ According to studies by Li *et al.*¹³⁷ and Schiessler *et al.*,¹³⁸ selecting a suitable set of descriptors for ML models like SVM, kernel ridge regression (KRR), or deep learning models using NN yields better results than using a large number of descriptors. Pham *et al.*¹³⁹ successfully chose the molecular descriptors most pertinent to the inhibition efficiency by applying the permutation feature importance (PFI) technique. A predictive model was trained on a dataset that included several types of organic corrosion inhibitors for carbon steel using these crucial molecular descriptors. According to Hakimain *et al.*, feature selection analysis revealed that the concentrations of hydrogen and sulfide in corrosive environments, along with the total number of alloying elements, were the most significant factors, accounting for up to 77.8% of the corrosion behavior of stainless steel in a variety of environments.¹⁴⁰

Diao *et al.*¹⁴¹ created a corrosion rate prediction models using ML algorithms, by gathering marine corrosion data of low-alloy steels. The RF algorithm was chosen as the modeling algorithm, and input features included both environmental factors and the chemical composition of low-alloy steel. To identify the dominant factors on the corrosion rate, feature reduction techniques such as PCA, Kendall correlation analysis, and gradient boosting decision trees were initially used. In order to transform the chemical composition features into a collection of atomic and physical property features, the authors also suggested two feature creation techniques. Consequently, a model that was no longer restricted to materials with particular chemical compositions was created using the feature creation method. The authors concluded that the R^2 between the actual measured corrosion rate and the corrosion rate predicted by the optimized model was greater than 0.9. Using a small dataset, Qiu *et al.*¹⁴² developed a novel feature selection method called SVM-RFE to predict the rate of atmospheric corrosion by combining RFE and SVM. The experiments showed that this method's top goal was to prove both algorithm efficiency and prediction accuracy. Pourrahimi *et al.*¹⁰⁰ classified steel using DT, RF, and SVM algorithms and described an ML-based prediction model for corrosion behavior of steel in various lactic acid corrosive conditions using a database that defined the corrosion behavior by qualitative labels. The

training and testing accuracies were 97.5% and 92.5%, respectively. Prediction model for pipeline pitting corrosion based on multiple feature selection and residual correction has been reported.¹⁴³ According to the authors, residuals were predicted using the seven features that were obtained through PCA. The prediction result was then combined with the SSA-RELM model's output to determine the predicted depth of pipeline pitting corrosion by combining residual correction and multiple feature selection. In comparison with the SSA-RELM model, the study showed that the suggested model decreased mean squared error, mean absolute percentage error, and mean absolute error by 66.80%, 42.71%, and 42.64%, respectively.

By creating virtual samples, Rustad *et al.*¹⁴⁴ presented the Gaussian mixture model virtual sample generation (GMM-VSG) technique to improve feature correlation for ML on anti-corrosion materials. Based on the RMSE and R^2 values, the experimental results showed that the virtual samples generated by the GMM-VSG increased the number of training samples, improved the correlation between features and targets, and increased the ML prediction accuracy. They also showed that the virtual samples extrapolated the original dataset. The four most crucial elements for forecasting the corrosion weight gain of uranium and uranium alloys were chosen using feature selection techniques, such as PCA and extra trees algorithm.¹³³ Consequently, a good corrosion prediction for uranium and uranium alloys was obtained with the corrosion weight gain prediction model having 96.8% prediction accuracy. A slime mold optimization SVM detection method and a RF algorithm with simplified features were proposed to address the issue of detection accuracy of internal wall damage in water pipelines.¹⁴⁵ According to the authors, the feature set with the strongest feature expression ability was chosen for classification and recognition after the contribution of the fused features was examined using the feature simplified RF algorithm. According to experimental results, the SMA-SVM algorithm based on FS-RF feature selection had a higher classification accuracy than other algorithms found in the literature.

3. Analysis of machine learning performance metrics

The success of ML models is assessed using ML performance metrics. They allow one to appraise the predictive accuracy of a model by considering how well a model can foresee the right output given a certain input.¹⁴⁶ There are numerous performance measures available, each tailored to a certain sort of ML task. Model selection, hyperparameter tuning, and the general enhancement of ML models can all be aided by using performance metrics.¹⁴⁷

Regression metrics and classification metrics are the two primary categories of ML performance measurements. Models that forecast a continuous output, such as the cost of a home or the volume of sales, are evaluated using regression metrics. Models that forecast a discrete output, such as whether a patient has an illness or not, are evaluated using classification metrics.^{148,149} Popular measures of classification and regression problem performance include:^{148–152}

1. **Precision and recall:** In binary classification tasks, precision and recall are crucial parameters, particularly when classes have unequal representation. Precision evaluates the proportion of accurate predictions made, while recall estimates the proportion of correct predictions made across all positive predictions.

Precision and recall must be balanced because a rise in one can result in a fall in the other.

2. **Accuracy:** One of the easiest performance indicators to understand is accuracy, which is the ratio of correctly identified examples to all occurrences in the dataset. Accuracy is beneficial for balanced datasets, but it could be deceiving for unbalanced ones. In these situations, additional performance metrics that take into consideration class distribution may eclipse it.
3. **ROC-AUC:** A performance indicator called receiver operating characteristic area under the curve (ROC-AUC) evaluates how well a model can distinguish between classes in binary classification situations. On a ROC curve, the area under the curve (AUC) is calculated by plotting the true positive rate against the false positive rate. Better model discrimination is indicated by a higher ROC-AUC score.
4. **F1 Score:** The F1 score offers a single metric that includes the precision and recall measures. It is the harmonic mean of these two metrics. It is frequently employed when recall and precision are equally important, particularly in situations where there are unbalanced datasets.
5. **Mean squared error (MSE) and mean absolute error (MAE):** Performance indicators for regression tasks include MAE and MSE. MSE represents the average squared difference, whereas MAE appraises the average absolute difference between anticipated and actual values. Both measurements assist in assessing the precision and goodness-of-fit of regression models.
6. **R-squared (R^2):** In regression models, the R^2 statistic is used to quantify how much of the variance in the variable that is dependent (the target) can be explained by the variables that are independent (features). A higher R^2 value indicates that the model fits the data more accurately.

The optimal performance metric for an ML task depends on the nature of the problem at hand and the desired outcomes of the project. Analyzing ML performance indicators is a crucial step in determining how well a model performs across different tasks. The choice of acceptable metrics depends on the nature of the problem and the features of the data, and each statistic offers a distinct perspective on model performance.¹⁵³ While precision, accuracy, F1 score, and recall are useful for classification, handling imbalanced datasets is where ROC-AUC is most pertinent. However, for regression tasks, MAE, MSE, and R^2 offer insightful information.¹⁵²

4. Interpretability of ML models used in corrosion studies

Many ML models are incomprehensible, and users are unaware of the underlying inference of the predictions, even though the predictions are highly accurate. In other words, the ML model's prediction process is like a mysterious black box that is hard to comprehend, especially for those who are not familiar with computer programs.¹⁵⁴ Improved trust, validation, and usefulness in real-world applications are made possible by interpretability in ML models for corrosion studies. The applicability and dependability of these tools can be improved by combining feature importance analysis, physics-informed constraints, and interpretable models specifically designed for corrosion research.¹⁵⁵ Engineers and scientists studying corrosion can provide important information about how materials behave and interact with

their surroundings. Therefore, rather than relying exclusively on complex data patterns for insights, interpretable models enable this knowledge to be integrated and validated. Additionally, interpretability aids in comparing the models to existing experimental and theoretical knowledge. The model's predictions are more reliable if its logic is consistent with established corrosion mechanisms.¹⁵⁶ Some of the methods used in the interpretability of ML models used in corrosion studies include:

1. **Partial dependence plots (PDP):** PDP are considered the primary foundation for conclusions in pertinent research and are one of the most popular techniques for estimating the marginal impact of independent variables on the expected result of a ML model.¹⁵⁷ A PDP, for instance, can help corrosion researchers understand how a ML model reacts to changes in the environment by illustrating how rising concentrations of corroders impact corrosion rates. In order to predict and interpret concrete corrosion caused by carbon dioxide using ML, Wang *et al.*¹⁵⁸ used PDP and the SHAP algorithm to assess the independent effects of "carbonation time," "water," "aggregate," and each raw material, as well as the interactive effects between features and oxide components. According to the authors, this analysis helped to understand the significance of the feature, investigate possible synergies, and ultimately identify the ideal value range, all of which support its useful applications and offer insightful information for studying the corrosion of steel bars in concrete.
2. **SHapley Additive exPlanations (SHAP):** The SHAP technique allows for the determination of the features with the greatest influence on model predictions. In order to better understand the predictive logic and feature contribution, SHAP values were utilized to visually interpret the model both locally and globally in Song *et al.*'s interpretable ML for maximum corrosion depth and influence factor analysis.¹⁵⁶ Additionally, the interactions and impact of the features on the corrosion depth were also effectively explained by the accumulated local effect (ALE). Electrochemical corrosion mechanisms, PDP, and the SHAP algorithm were used to analyze the interpretability of the physics-guided neural networks (PGNN) predictive model with an improved particle swarm optimization (IPSO) algorithm.¹⁵⁹ Four indicators, namely, mean absolute percentage error (MAPE), MAE, RMSE, and R^2 , were used to evaluate the model's accuracy, and its uniformity with corrosion mechanisms was used to gauge its reliability. The results showed that the PGNN-IPSO model can improve prediction accuracy by avoiding learning incorrect patterns that go against physical principles. Some other authors have reported that the use of the SHAP interpretability technique more accurately captures the trend of how various feature interactions affect corrosion rate and to visually analyze the output of ML models thereby arriving at consequential conclusions.^{160–164}

D. Bayesian belief network (BBN) in machine learning

Bayesian belief network (BBN) is recognized as a belief network or probabilistic model which is basically entitled for graphical representations of probabilistic relationships among a varied set of variables. Judia Pearl in late 19th century was the first known to use the term "Bayesian network" and it marked valuable additions to probabilistic

theory.¹⁶⁵ In a Bayesian network (BN), variables are depicted as nodes, and their connections are being represented as directed edges or arcs linking the nodes. The edges show any dependencies or conditional relationships between the nodes, which can represent observable or concealed variables. Each node in the network has a probability distribution attached to it that expresses how uncertain or certain each node is about the value of a certain variable given the values of its parent nodes. Figure 10 symbolizes A, B, C, and D as random variables represented by nodes where a directed arrow connects node B to node A, indicating that node A is the parent of node B. Furthermore, node C is shown to be independent of node A, indicating independence between the two variables. Generalized use of BN is confined for uncertain reasoning, making predictions and data learning, etc. Moreover, BN has been reported to provide more correlated data in cases where uncertain or incomplete knowledge implies.¹⁶⁶

A BBN is defined as a composition of two major components: (a) directed acyclic graph (DAG) and (b) set of conditional probability tables (CPT). DAG illustrates the reliance among variables where each node signifies a random variable and edges explain the probabilistic dependencies among varied variables, whereas CPT defines the conditional probabilities distributions corresponded to each variable given its parent.¹⁶⁸ BBN can be constructed for artificial intelligence.¹⁶⁹

- Problem identification:** Factors pertinent to area of concern are identified and every variable needs to be discrete or handled accordingly.
- Defining the dependencies:** Probabilistic or causal links between the variables have been defined, and this could be accomplished through domain knowledge, professional judgment, or data analysis.
- Construction of direct acrylic graph (DAG):** Further by using the variables as nodes and the dependencies as edges, create a directed acyclic graph where the edges' directions reveal any causal or probabilistic links.
- Assigning probabilistic distributions:** Each variable has been assigned with a probability distribution based on the parent's variable and majorly involves the construction of conditional probability tables. CPT for each variable corresponds to the probability distribution provided the values of its parent variables.
- Learning the parameters:** Bayesian inferences can further be utilized to estimate the network parameters.

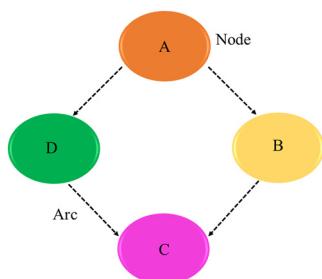


FIG. 10. Representation of nodes and arcs in Bayesian belief network. From Subaira et al. *Translating Healthcare Through Intelligent Computational Methods*. Copyright 2023 Springer. Reproduced with permission from Springer.¹⁶⁷

- Performing inference:** Construction of a network and estimation of parameters are followed by probabilistic inferences to make predictions. Inferences include calculation of posterior probabilities of variables given observed evidence using certain techniques like belief propagation and variable elimination.
- Evaluation and improvement:** Assess the effectiveness of the BBN, improve the model as needed, and repeat the process if new information or understanding becomes available.

The prime advantage of BN focuses on resolution of uncertain situations using probabilistic approaches. We can alter our ideas and make predictions in light of fresh information or observations by defining the conditional dependencies between variables. Numerous applications, including natural language processing, risk assessment, gene expression analysis, and medical diagnostics, can benefit from the use of BN.¹⁶⁹ Calculating relevant probabilities given observable evidence is the process of inference in BN. Variable elimination, belief propagation, and Markov chain Monte Carlo (MCMC) sampling techniques are just a few examples of algorithms that can be used to do this. A critical job is to use data to learn the parameters and topology of BN. In this, the conditional probability distributions are estimated, and the interdependence between the variables is identified. It is possible to learn BN using a variety of techniques, including constraint-based methods, score-based methods, and hybrid techniques that incorporate both types.¹⁷⁰

Overall, BN offers a strong framework for describing and deducing from uncertain knowledge and has been widely used in many different fields. However, the evolution of BBN often searches for proficiency with probability theory and graphical models. Though, there are software libraries and tools that offer utility for building and carrying out inferences in BBN such as python.¹⁷¹

III. MACHINE LEARNING APPLICATIONS IN MATERIALS DESIGN, PROPERTIES, AND DISCOVERY

The screenings of high-performance chemical species for different industrial and biological applications using ML approaches are scientifically critical to developing many technologically relevant fields.¹⁷²⁻¹⁷⁴ Regrettably, time-consuming and ineffective recurring practical and theoretical evaluations are frequently required since meaningful advancement often necessitates a blend of chemical intuition and serendipity. Finding new materials takes a long time; from beginning study to first usage, it usually takes 10 to 20 years.¹⁷² The seven independent stages of new materials research are “discovery, development, property optimization, system design and integration, certification, manufacturing, and deployment,” as shown in Fig. 11.¹⁷² Various institutions, either engineering or scientific teams, may perform the different stages. Two established techniques frequently used in materials research are computational modeling and experimental measurement. However, employing these methods to speed up material discovery and design is challenging due to the inherent constraints of practical settings and theoretical underpinnings.

It is anticipated that materials design informed by computation will speed up and reduce the cost of developing new materials.^{175,176} Computational modeling takes less time than actual measurement and is better for providing genuine experiments because it gives complete control over the pertinent variables.¹⁷⁷ However, computational simulation also faces several difficulties, such as the fact that it is highly dependent on the microstructures of the materials concerned, that it

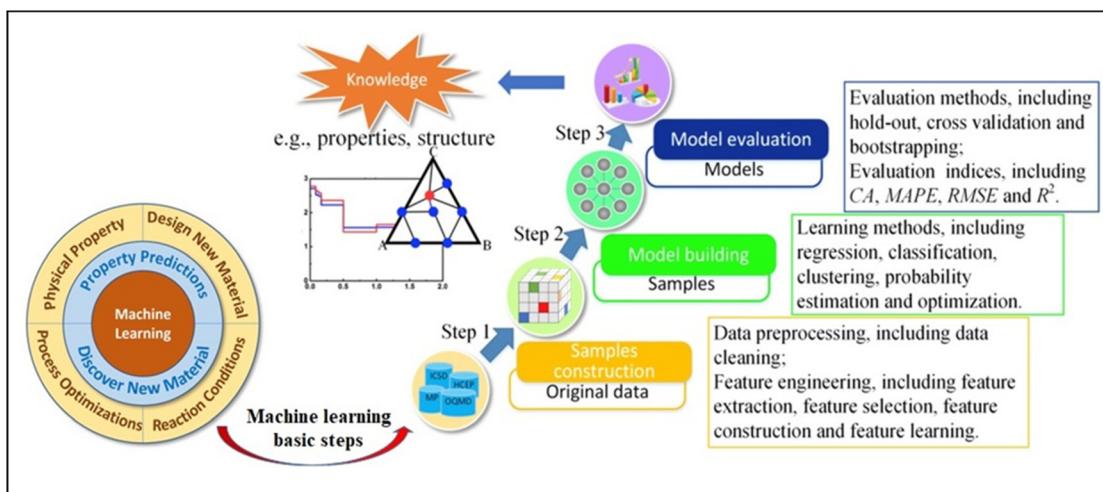


FIG. 11. Application of ML in new materials discovery and property predication and basic steps involved in ML. Reproduced with permission from Liu et al., J. Materomics 3, 159–177 (2017). Copyright 2017 Elsevier.¹⁷²

necessitates outstanding performance computing hardware, typically in the form of enormous computing clusters, to ensure the computational simulators can run, and that no explicit utilization can be obtained from previous calculations when an entirely novel system is examined.^{178–180} Experimental investigation and computational modeling tend to be unable to handle rapidly developing challenges due to the complex needs involved in understanding materials' basic physical and chemical characteristics and advancing their industrial applications. For example, exploring the glass transition temperature (GTT) using experimental observations is extremely difficult and ineffective due to the extensive temperature over which the transition occurs.¹⁸¹ However, since the GTT is affected by external and internal variables, including pressure, structure, and constitutive and torsional properties, computer algorithms cannot accurately mimic it.¹⁸² As a result, material scientists and engineers have made several efforts to improve upon these two widely used approaches.

ML uses algorithms to help computers learn from empirical information by simulating the linear or nonlinear correlations between material qualities and related factors.^{172,183} This technique is effective for identifying patterns in high-dimensional data. ML algorithms have recently been effective at overcoming the challenges of modeling the interactions between material qualities and intricate physical parameters.^{172,183} Notably, the beneficial uses of ML, including aiding in identifying materials based on unsuccessful experiments and searching for effective molecular organic light-emitting diodes, are considered an inventive mode of material advancement.^{184,185} Modern approaches to material study can balance realistic experimental demands with a low error rate, fully utilize the vast amount of data at hand, and expedite the materials study procedure. In industry and academia, emphasis has been drawn to several novel machine techniques, such as deep learning for in-depth data processing. Developing analytical models can be automated using ML. Thanks to automated learning algorithms that continuously learn from data, computers can discover hidden insights despite being expressly instructed to hunt for them.^{172,186} Regression, classification, and other high-dimensional data-related problems show suitable ML applications.^{172,186} ML, which aims to extract information

and gain knowledge from enormous databases, learns from prior computations to provide trustworthy, repeatable judgments and results and has thus been crucial in many fields. The earliest known uses of ML in the field of materials science date back to the 1990s, when ceramic-matrix composites' corrosion behavior and tensile and compressive strength were predicted using ML techniques like ANNs and symbol methods.^{172,186–188} ML has since been applied to several materials' science issues, including discovering novel materials and predicting material properties. The three stages of ML system are typically sample creation, model construction, and model assessment.

The first step gathers the initial data through computer simulations and experiment observations.¹⁷² Data cleaning should be done before creating a sample from the initial information because these data are frequently unreliable, noisy, and incomplete. Additionally, several dependent factors, some unrelated to the decision qualities, can alter any sample obtained. It works as a "black box," connecting input and output data with a specific set of linear or nonlinear algorithms.^{189,190} The values of coefficients with which a particular mapping function most closely resembles the target function are discovered via ML using samples of the target function. There are three main categories of ML usages for materials research and design: material properties forecasting, novel material discovery, and numerous other uses (Fig. 12).¹⁷² Experimental testing and physical modeling are conventional methods to explore various material properties to ascertain the attributes of an unknown substance or material. Both macroscopic and microscopic scales can be used to characterize and store the features of a material, such as its electrical, thermal, and lattice characteristics. Conventional techniques constrain the assessment of material attributes. The experimental setting and intricate operations are used in both experiments and simulations. Consequently, computational calculations cannot simulate complicated interactions between features and associated factors, and specific linkages may be unknown.^{191,192} Additionally, the assessment tests are frequently carried out at the very end of the material selection process, so if the results are inadequate, the engaged time and resources could be better spent. ML offers an additional method to reduce the time and assistance needed for

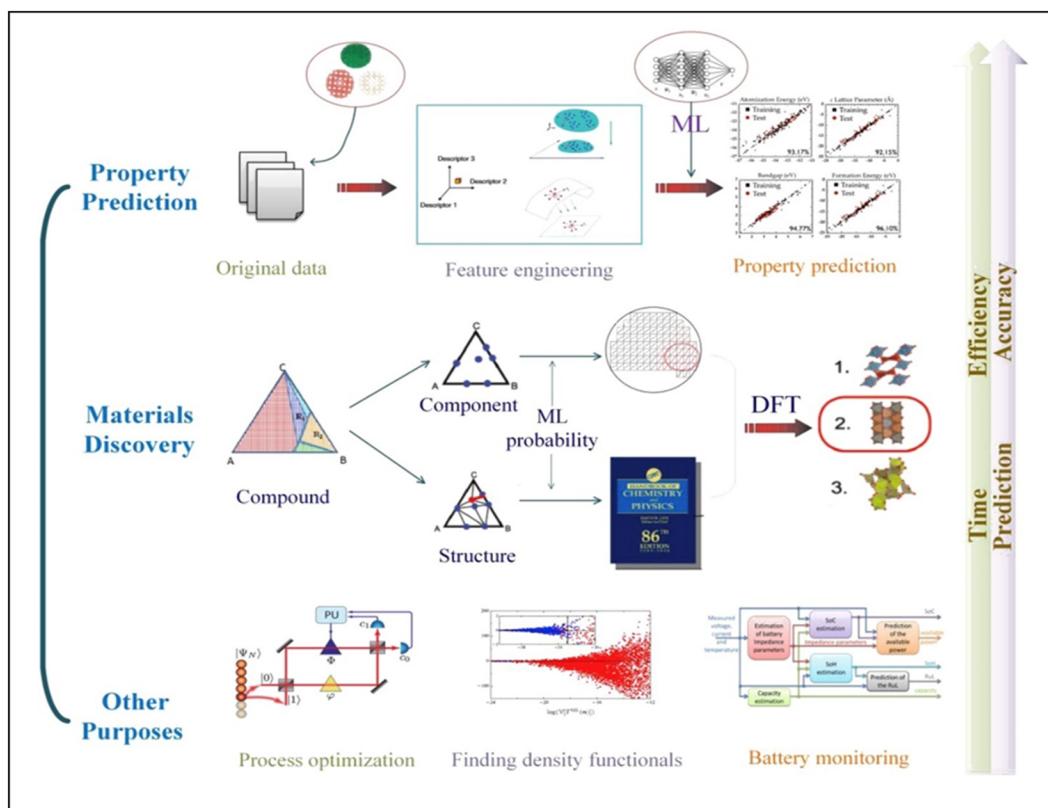


FIG. 12. An outlook on the use of machine learning in material science. Reproduced with permission from Liu *et al.*, *J. Materiomics* **3**, 159–177 (2017). Copyright 2017 Elsevier.¹⁷²

11 January 2025 14:34:59

decision-making.^{172,186–188} The main idea behind utilizing ML for property forecasting is to examine and identify the nonlinear correlations between the qualities and associated components using existing data.

To discover the novel material, compositional and structural expertise are required. One of the most critical problems in materials science is predicting the crystal structure of novel substances because developing new materials necessitates using relevant expertise.^{193,194} Conventional techniques, such as compositional and structural screening, involve many computations and tests and frequently demand more time and money than is available. Additionally, the complexity of the energy surfaces in the combinatorial space makes it difficult to anticipate the structures from first principles because it comprises all conceivable atomic combinations in three dimensions.^{172,186–188} Several cutting-edge techniques, including random sampling, metadynamics, and optimum hopping, improve the usefulness of traditional crystal structure forecasting methods. However, evaluating these approaches is highly resource-intensive and incurs high computational costs.^{195,196} The ML techniques can be broadly divided into composition prediction techniques and methods of structural forecasting.^{172–174} The expected materials' thermodynamic stability needs to be considered. A thermodynamically stable substance does not separate into several phases or compounds. Energy assessment is one of the measures used to assess the stability of novel materials as an

outcome. ML is also used to address various material science-related issues.^{197–199} Traditional approaches need too much time, are challenging to operate, and have high computing costs to overcome these issues. Therefore, scientists combine ML with the established approach to address the pertinent problem in material science.

IV. MACHINE LEARNING FOR CORROSION MONITORING

A. Machine learning for atmospheric corrosion monitoring (ACM)

In order to prevent accidental damage from old buildings or aging of infra-structure and most importantly to maintain safety and structural integrity, prevention of corrosion is an important aspect.^{200,201} Corrosion refers to the deterioration of materials, such as metal, due to chemical reactions with the environment. It can significantly weaken the structural components of a building and lead to costly repairs or even structural failure if not addressed promptly. Corrosion detection or its monitoring techniques can be done via two routes: destructive and nondestructive routes. Where destructive testing is done by loading specimen under test until it is broken on other way around, nondestructive testing is to evaluate the material in its original condition without affecting its serviceability. ML is a nondestructive technique and can be applied to atmospheric corrosion monitoring (ACM) to analyze environmental data, predict corrosion rates,

and provide real-time insights. It is a useful tool for ACM as it can analyze large amounts of data from multiple sources, including environmental conditions, corrosion rates, and material properties. With this information, ML algorithms can make predictions and identify patterns that may be difficult for humans to recognize.

If we investigate the monthly/quarterly/yearly data observed from corrosion coupons, we will notice that the data obtained from ACM studies contains detailed information about the corrosion kinetics impacted by environmental factors, be it pollutants, pH, humidity, temperature, and many other. Thus, it is difficult to process the data by applying physics base corrosion law and to derive a direct or indirect relationship of corrosion with environmental factors or to predict the corrosion free life of materials. Thus, automation is utmost required to ease the understanding of environmental factors and to predict the life of materials.

For example, ML in ACM is the use of ANNs to predict the deterioration rate of metals in a specified environment. ANNs can be trained on historical data and then used to make predictions about future corrosion rates based on current environmental conditions. This can be particularly useful in industrial settings where corrosion can cause significant damage to equipment and infrastructure. Another example is the use of ML algorithms to identify corrosion hotspots in large structures, such as bridges or pipelines. By analyzing data from sensors placed throughout the structure, these algorithms can identify areas that are more susceptible to corrosion and alert maintenance teams to act before significant damage occurs. Overall, ML can be a valuable tool for ACM as it can provide insights into complex systems and help prevent costly damage. However, it is crucial to ensure that the algorithms are properly trained and validated to ensure accuracy and reliability.

In vehicles and bridges, sensors are being employed to track the corrosion and to understand how environmental factors influences the rate of corrosion.^{202,203} For many years, electrochemical methods like electrochemical noise and electrochemical impedance spectroscopy have been used for ACM. For instance, to monitor the damage of substrate in terms of weight change, its acoustic properties and electrical resistance nonelectrochemical sensors are generally used.^{204–208} On contrary to that, electrochemical sensors functions by creating a closed loop on surface of sensor and scan the corrosion activity by observing impedance,^{209,210} electrochemical noise,^{211,212} and linear polarization resistance.²¹³

B. Machine learning for corrosion monitoring in urban environment

Owing to commercialization in urban environment, corrosion in urban environments is a significant concern due to the presence of various factors that can accelerate the corrosion process. Urban areas often exhibit higher pollution levels, increased humidity, and exposure to aggressive agents, such as industrial emissions, vehicle emissions, and deicing salts. Urban areas are characterized by higher levels of air pollution resulting from industrial activities, transportation, and urbanization. Pollutants such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon dioxide (CO_2), and particulate matter can contribute to the corrosion of metals and structures. These pollutants can react with moisture in the air to form corrosive acids, accelerating the corrosion process.

Industrial activities in urban areas can release corrosive gases and chemicals into the atmosphere. Acidic gases like sulfur dioxide and hydrogen sulfide, as well as corrosive chemicals used in manufacturing processes, can lead to accelerated corrosion of metals and infrastructure. High traffic volumes in urban areas contribute to increased vehicle emissions, which contain pollutants like nitrogen oxides (NO_x) and volatile organic compounds (VOCs). These pollutants can react with atmospheric moisture to form corrosive acids, impacting vehicles, bridges, and other metal structures exposed to traffic-related pollution. In urban environments, deicing salts are often used to combat icy conditions during winter. These salts, such as sodium chloride or calcium chloride, can come into contact with metal surfaces, leading to corrosion. Salt corrosion is particularly common in urban areas with colder climates where deicing salts are heavily used.

Urban areas tend to experience the urban heat island effect, where temperatures are higher compared to surrounding rural areas. Increased temperatures can accelerate corrosion processes, particularly for metals exposed to heat and moisture. Urban environments can exhibit microclimate variations due to factors such as building shadows, airflow patterns, and variations in pollution levels. These microclimate variations can lead to localized corrosion, where certain areas experience higher corrosion rates than others, depending on their exposure to environmental conditions.

In 2013, NACE International IMPACT report estimates the approximate global cost for corrosion to be 2.5 trillion US dollar, which represents the 3.4% of global gross domestic product (GDP).²¹⁴ ML techniques can be applied to study and address corrosion-related challenges in urban environments in many ways. For instance, ML algorithms can analyze historical corrosion data, along with various environmental and infrastructure parameters, to develop predictive models. These models can forecast corrosion rates or identify areas with a higher risk of corrosion in urban settings. By understanding the factors contributing to corrosion, proactive maintenance and corrosion prevention strategies can be implemented. The same can be used to analyze vast amounts of environmental data collected from sensors and monitoring stations in urban areas. This includes factors such as air quality, temperature, humidity, rainfall, and pollution levels. Here are few references that discuss the application of ML in the context of corrosion in urban environments.

In a study published in Corrosion Science in 2019, ML was used to predict the corrosion rate of C-steel in a simulated urban atmosphere.²¹⁵ They trained a SVM algorithm on a dataset of environmental and material variables, such as temperature, humidity, and surface roughness. The SVM was able to accurately predict corrosion rates based on these variables. The authors concluded that this approach could be used for real-time corrosion monitoring in urban environments. In another study published in the Journal Sensors in 2019, researchers used ML to predict the deterioration rate of copper in a simulated urban environment.²¹⁶ They used RF algorithms to analyze data on environmental factors such as temperature, humidity, and pollutant concentrations and summarized that the RF model was able to accurately predict the deterioration rate of copper based on the environmental factors. In 2020, researchers used ML to analyze data from a network of sensors placed in an outdoor environment. They employed a DT algorithm to identify patterns in the corrosion data obtained. The outcome revealed that the DT algorithm was able to accurately identify regions with high corrosion rates, suggesting that it

could be a useful tool for real-time monitoring of corrosion in outdoor settings.²¹⁷

C. Machine learning for corrosion monitoring in pipelines

Corrosion monitoring in pipelines is a critical process used to assess the integrity and condition of the pipeline infrastructure. It involves the continuous or periodic measurement of corrosion-related parameters to identify and mitigate corrosion issues before they lead to pipeline failures or leaks. Some common techniques used to monitor the corrosion in pipelines are as follows:

- Coupons and probes:** Corrosion coupons are metal samples that are inserted into the pipeline and periodically removed for analysis. They provide an accurate measurement of the deterioration rate over a specific time. Similarly, corrosion probes are devices that are permanently installed in the pipeline to measure the corrosion rate continuously.
- Electrical resistance (ER) probes:** ER probes are installed in pipelines to measure the changes in electrical resistance caused by corrosion. They can provide real-time data on the corrosion rate and the location of corrosion along the pipeline.
- Ultrasonic thickness measurement (UTM):** UTM is a nondestructive testing technique used to measure the remaining wall thickness of the pipeline. By comparing the current thickness with the initial thickness, the rate of corrosion can be determined.
- Acoustic emission (AE):** This technique entails the detection of high-frequency acoustic signals emitted by the pipeline when corrosion or other damage occurs. It can provide information about the location and severity of corrosion activity.
- Close interval potential surveys (CIPS) and direct current voltage gradient (DCVG):** These techniques involve measuring the voltage and potential differences along the pipeline to identify areas of potential corrosion or coating damage.
- Real-time corrosion monitoring systems:** These systems use a combination of sensors, data loggers, and communication technologies to continuously monitor parameters such as temperature, pressure, and corrosion potential. The data collected are transmitted to a central control room for real-time analysis and prompt action.

It is important to note that the selection of the corrosion monitoring technique depends on factors such as the type of pipeline, its operating conditions, and the budget allocated for monitoring activities. Pipeline operators often employ multiple techniques to glean a thorough understanding of the corrosion behavior and take appropriate preventive or corrective measures to ensure the pipeline's integrity.

In a study published in Materials and Corrosion in 2021, ML was employed to identify corrosion hotspots in a pipeline. They used a KNN algorithm to analyze data from corrosion sensors placed along the pipeline.²¹⁸ The KNN was able to identify areas of the pipeline that were more susceptible to corrosion and alert maintenance teams to act. The authors suggest that this approach could be used for real-time corrosion monitoring in industrial settings. In another study published in the Journal of Materials Science in 2020, researchers used ML to analyze data from a network of sensors placed in an industrial environment.²¹⁹ They used a self-organizing map (SOM) algorithm to identify

patterns in the corrosion data. The outcome revealed that the SOM algorithm could accurately identify regions with high corrosion rates, suggesting that it could be a useful tool for real-time monitoring of corrosion in industrial settings.

In a study published in the journal Measurement in 2020, researchers used ML to identify potential corrosion sites on a steel surface using a CNN to analyze images of the steel surface and to identify regions with potential corrosion. The outcome showed that the CNN was capable of accurately identifying potential corrosion sites on the steel surface.²²⁰ In a study published in Corrosion Science in 2019, researchers used ML to forecast the dissolution rate of mild steel in an industrial setting. They collected data on environmental variables such as temperature, humidity, and air pollution levels, as well as on the dissolution rate of the metal. They used an RF algorithm to make predictions based on the data. The results showed that the RF algorithm was able to accurately predict the corrosion rate based on the environmental variables.²²¹

In a study published in the Journal of Loss Prevention in the Process Industries in 2020, researchers used ML to optimize a corrosion prevention strategy for a chemical plant.²²² They collected data on environmental variables and corrosion rates and used a DT algorithm to identify the most effective corrosion prevention strategy based on the data. The outcome obtained showed that the ML approach was able to identify an optimal corrosion prevention strategy that reduced the corrosion rate by 20%. In another study published in the Journal of the Electrochemical Society in 2019, researchers used ML to identify corrosion hotspots in a pipeline system.²¹⁸ They collected data from sensors placed throughout the pipeline system and used a clustering algorithm to identify regions with similar corrosion patterns. The obtained outcome showed that the ML approach was able to identify corrosion hotspots with high accuracy.

Another interesting and nonintrusive corrosion monitoring technique is sensor technology. Corrosion monitoring sensors are devices designed to measure and monitor the corrosive effects of the atmosphere on various materials, particularly metals. These sensors provide valuable data for assessing the corrosion rate, identifying corrosion hotspots, and evaluating the effectiveness of corrosion protection measures. These sensors utilize ultrasonic radiation to measure the corrosion. The methodology is very useful for corrosion monitoring in pipeline. These sensors help to maintain the plant safety without intrusive access points which are required in traditional corrosion monitoring. Most importantly with the help of monitoring pipe wall thickness, the job of pipeline operator can be made convenient for their decision making in day-to-day production.

Electrochemical sensors, electrical resistance sensors, acoustic sensors, corrosion potential sensors, corrosion coupons, and wireless sensor networks are some of the common sensors employed to attain the valuable data for assessing the corrosion rate, identifying corrosion hotspots, and evaluating the effectiveness of corrosion protection measures. Although, all said methods are capable enough to measure the degree of corrosion but failed to provide real-time monitoring which encourages researchers to develop sensors which can work on real time monitoring to reduce the chemicals required to deal with corrosion after being identified.

Considering this, Nining *et al.*²²³ in 2020 developed the atmospheric corrosion sensor based on strain management (ACSSM) instead of electrochemical method. Previously also, ACSSM techniques

have been used for determining the reduction in thickness owing to corrosion by elastic strain.^{224–226} The calculations for thickness reduction produced by corrosion were done by theoretical formulas and applying finite element method simulations. It was seen that strain management circuits fabricated with active–dummy circuits produce better results than elastic strain management.

The literature survey of last 10 years suggested that many optical fiber techniques have been applied in the field of ACSSM, namely, distributed fiber sensors,^{227,228} long period grating sensors,^{229,230} etc. The application of fiber Bragg grating measurement in strain management monitoring has been seen in metal bridges,²³¹ metal in concrete structure,²³² in landslide displacement,²³³ and in welding joints²³⁴ and in aircraft materials²³⁵ too.

D. Machine learning for aqueous corrosion monitoring

Aqueous corrosion is an electrochemical process that occurs when a metal sample is unprotected in an aqueous environment. The metal surface becomes oxidized, and the electrons released in this reaction are used to reduce another substance in the environment. This can lead to the formation of rust, pits, or other forms of corrosion damage. There can be a variety of aqueous corrosion, and a few examples may include the following:

1. Corrosion occurrence when the entire metal surface is unprotected in an aqueous environment.
2. Corrosion occurrence when two distinct metal samples interact with each other in an aqueous environment. The more active metal (the one that is more easily oxidized) will corrode more rapidly than the less active metal.
3. Corrosion occurrence when a metal sample is exposed to an aqueous environment in a crevice or crack. The electrolyte can be trapped in the crevice, which creates a localized environment that is more conducive to corrosion.
4. Corrosion occurrence that results in the creation of pits on the surface of the metal. The pits can be very small, but they can grow over time and cause significant damage to the metal. Aqueous corrosion can have a significant impact on the performance and lifespan of metal structures and equipment. It can also lead to safety hazards, such as the failure of pipelines or other critical infrastructure. By understanding the causes and mechanisms of aqueous corrosion, it is possible to take steps to prevent or minimize this problem.

Aqueous corrosion monitoring adopts the use of a set of equipment and methods or a corrosion meter to offer online or offline corrosion rate information expressed in mpy (millimeters per year). There are varieties of ways to monitor aqueous corrosion, out of which some common methods are weight loss (WL) measurements, ER measurements, potentiodynamic polarization (PDP) measurements, and scanning electrochemical microscopy (SECM) method. The adopted method basically depends on the specific application and the requirements of the user. For example, WL measurements are a simple and inexpensive method, but they can be time-consuming and may not be sensitive enough to detect small changes in the corrosion rate.²³⁶ ER measurements are a more sensitive method, but they can be more complex to perform. PDP measurements are a very sensitive method, but they require specialized equipment and expertise.

Aqueous corrosion monitoring is a valuable tool for preventing and controlling corrosion. By understanding the different methods of aqueous corrosion monitoring, it is possible to choose the most appropriate method for a particular application. ML is an emerging and nondestructive aqueous corrosion monitoring technique. The ML algorithms that can be applicable to ACM include SVMs. SVMs belong to a supervised ML algorithm capable of classifying data into two or more categories. They are often used for aqueous corrosion monitoring to classify data into different types of corrosion, such as general corrosion, pitting corrosion, and galvanic corrosion. Another is RF that represents a type of ensemble learning algorithm that can be used to classify data into multiple categories. It is often used for aqueous corrosion monitoring to identify the factors that contribute to corrosion, such as the metal type, medium, and the presence of contaminants. Also, ANNs belong to deep learning algorithms with capacity to learn existing complicated patterns in data. They are often used for aqueous corrosion monitoring to predict the corrosion rate of a metal or to identify areas of localized corrosion.

ML has been shown to be effective for aqueous corrosion monitoring in several studies. For instance, Jiménez-Come *et al.* demonstrated the superior performance of SVMs compared to Naïves-Bayes techniques in the evaluation of the impact of environmental variables on pitting corrosion behavior of stainless steels. The SVMs yielded precision and accuracy values of 0.96 and 0.87, respectively.²³⁷ Another study by Xiong *et al.* used 16 different ML algorithms to predict the corrosion rate of 7XXX Al alloys with prediction accuracy of 89%.²³⁸

E. Machine learning in atmospheric corrosion monitoring in marine environment

The transportation industry, mainly the shipping of large consignments, is generally done via sea, and there is significant damage of ships by corrosion in marine environment, and the efficiency of mechanical parts in marine environment is significantly reduced²³⁹ via hull structural failure.²⁴⁰ Studies have indicated that approximately 90% of the failure of ship structure is due to the corrosion.²⁴¹ Thus, corrosion detection and estimating the impact of corrosion become very important to avoid marine accidents or oil spilling. The prior detection of corrosion can significantly minimize the accidents and can save the life of people as well as environment by avoiding structural failure.^{242,243} In this new era, technology has been updated to such a great extent that artificial technologies, ML, can be utilized to appraise the corrosion rate and its phenomena.^{244,245}

ML has proven to be a valuable tool in ACM, including in marine environments. By leveraging the power of data analysis and pattern recognition, ML algorithms can analyze complex datasets and provide insights into corrosion processes, enabling proactive maintenance and corrosion control strategies. Here are a few relevant references that explore the application of ML in ACM in marine environments. In a study by Yan *et al.*, 6 ML algorithms were employed to study the correlation between environmental variables, material properties and the corrosion rate of low-alloy steels. The RF algorithm was able to optimally predict the corrosion rates based on the 9 dominant factors.²⁴⁶ In one more study published in the journal Corrosion Science in 2009, ML was employed to predict the corrosion rate of C-steel in a marine environment.²⁴⁷ They used a support vector regression (SVR) algorithm to make predictions based on the data.

Another important aspect is to monitor the corrosion of steel in a marine environment. Li *et al.*, in 2023 published a study where they used ML to foretell the corrosion rate of steel in different marine environments.¹⁶⁴ They collected data on environmental variables, material composition and exposure time as well as the corrosion rate of the metal. They used a XGBoost algorithm to make predictions based on the data and obtained an accuracy of 93%, 93%, and 96% for splash, submerged and tidal zones, respectively.

Deep learning, a subset of ML, involves training neural networks with multiple layers to automatically learn patterns and representations from data. In the context of ACM, deep learning models can be used for tasks such as corrosion rate prediction, classification of corrosion severity, or anomaly detection. Deep learning models, such as CNNs or RNNs, can analyze complex patterns in multi-dimensional data and make accurate predictions.^{102,248}

On one side, ML focuses on learning algorithms and the pattern learning focuses on classification or identifying the types. Pattern learning techniques, such as clustering algorithms, can be used to identify groups or clusters of similar corrosion behavior. By grouping similar data points together, patterns can emerge, highlighting areas where corrosion is more prevalent or where certain environmental conditions lead to accelerated corrosion. These patterns can provide insights into the factors influencing corrosion and help in understanding corrosion mechanisms. Overall, pattern learning and deep learning techniques can enhance the field of ACM by uncovering hidden patterns, predicting corrosion behavior, and facilitating proactive maintenance strategies. However, it is imperative to note that these techniques require high-quality data and careful model development to ensure accurate and reliable results.

F. Machine learning for pitting corrosion resistance

ML undeniably can be a tool utilized to enhance and analyze the pitting corrosion resistance. Pitting corrosion generally damages metallic parts owing to formation of small pits and holes over surfaces

and primarily affects the metal surface more specifically by making small areas depassivated, and this depassivation makes metals prone to electrochemical attacks. Certain factors including exposure to corrosive environments, dissolution of protective layers, impurities, etc., can be considered liable for pitting corrosion. Figure 13 presents an overview of some countries' network pitting corrosion modeling based on ML.

Pitting corrosion is majorly a cause of concern owing to its deep penetration to metallic structures causing structural failure and deep perforation. Complex and localized nature make it challenging to predict pitting corrosion. Though, by leveraging ML, it is convenient to design models that can assist in understanding and predicting pitting corrosion behavior. In 2023, Akhlaghi *et al.* predicted the accurate pit depth of pitting corrosive attacks in buried transmission pipelines. Real-world datasets have been collected in this context, and the utilized ML model was seen to resolve the overfitting problem exhibited by support vector regression-firefly algorithm (SVR-FFA), therefore enhancing the accuracy of the dataset. Generalization and generalization–memorization based approaches were utilized to design deep learning. Here, generalization and generalization–memorization models have been compared with others like SVR, SVR-GA and SVR-FFA, etc., and a noticeable comparison has been located with overfitting issues, thus outperforming the compared models.²⁴⁹ In 2022, Sasidhar *et al.* modeled a deep learning framework to unveil the effect of varying compositional and environmental factors over pitting corrosion resistance of diverse passivating alloys.

1. Pitting potential was enhanced with decreased chloride concentration, whereas when chloride was kept constant, certain alloys like Ni-Cr and Al-Cr experienced augmented nitrogen as a contributing factor to improved pitting corrosion.
2. Ni-Co alloys affirmed the carbon concentration rise, a major factor for maximized pitting potential.
3. Additionally, Mo was inferred to exhibit pitting corrosion resistance in Ni-Cr-based alloys and stainless steel, though its effectiveness was seen to increase in acidic situations.

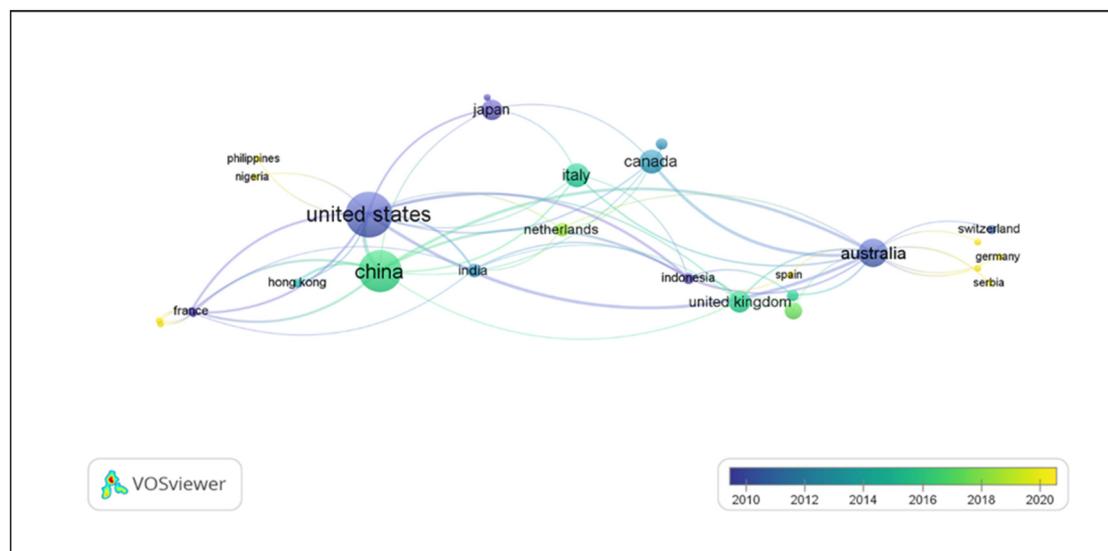


FIG. 13. Variation of countries network of pitting corrosion and machine learning.

4. Chromium content optimization data for Ni-Cr and Fe-Cr based alloys revealed the effective critical wt.% for both (i.e., active transitions from negative value to positive have been marked at mentioned concentrations 15% and 12%).²⁵⁰ In 2021, Qu *et al.* proposed the use of ML model for the judgment of pipeline pitting corrosion. To examine the relationship between the likelihood of pitting and input features such material factors, solution factors, and environmental factors, ML algorithms and feature engineering correlation methods were applied as shown in Fig. 14. The main variables to determine the probability of occurrence of pitting corrosion include CO_2 , temperature (T), Cl^- , Ca^{2+} , and pH have been considered prominent factors to judge the occurrence of pitting corrosion.²⁵¹

In 2020, Hoang *et al.* utilized ML in modeling a method for automatic recollection of metallic surfaces diagnosed with pitting corrosion. Linear population size reduction (LSHADE) along with SVM was modeled including the model training followed by a sample collection of 213 images and utilized the collected data to develop an intelligent approach to programmed detection of pitting corrosion. The latterly developed approach was found to be 91.8% accurate and 0.91 precise.²⁵² In 2014, Alkanhal *et al.* practiced the use of wavelet and fractal analysis to evaluate the effect of test time overgrowth rate of pitting corrosion by assessing the images of corroded samples. Results affirmed the advancement of tubular structure in terms of corrosion product along with corrosive pit. Furthermore, certain parameters, such as energy loss, Shannon entropy, fractal slope, and dimension, experienced an increase with time.²⁵³ In order to solve the nonlinear problem of evaluating the danger of steel rebar pitting corrosion, the study conducted by Chou *et al.* examined the applicability of four single AI models (ANN, SVM/SVR, CART and LR, four meta ensembles (voting, bagging, stacking, and tiering), and a hybrid metaheuristic regression).²⁵⁴

G. Machine learning for microbiologically induced corrosion (MIC) monitoring

Microorganisms such as bacteria, archaea, and fungi often colonize and stimulate corrosion. Microbial-induced corrosion (MIC), being termed as microbiologically influenced corrosion, is a weighty concern in diverse industries including infrastructure, marine, and oil and gas and a range of metallic counterparts such as copper, steel, and aluminum have been seen affected from it. Microorganisms have been

seen contributing to corrosion via different mechanisms; production of corrosive metabolites, biofilms formation, or variation in the nearby environment of metallic counterparts.²⁵⁵ Additionally, production of certain metabolites like ammonia, hydrogen sulfide, etc., by microorganisms advocates the phenomenon of corrosion. Localized microenvironments that favor corrosion can be produced by biofilms, which are microbial colonies enmeshed in a matrix of extracellular polymeric substances (EPS). A variety of tactics are used to prevent and mitigate MIC. These can include the adoption of protective coatings, the utilization of corrosion-resistant materials, and the application antimicrobial treatments to inhibit microbial growth, and ambient conditions that are optimized to reduce microbial activity.²⁵⁶

MIC, being a complex process, utilizes the interactions between microorganisms and metals, leading to the collapse of structures, and associated metallic counterparts. By exploiting ML techniques, researchers can gain valuable insights into MIC mechanisms, prediction of corrosion rates, and develop effective corrosion combating strategies.

Figure 15 shows some effective ways ML is being utilized in microbial-induced corrosion study including data analysis and pattern recognition, predictive modeling, biofilm analysis, and sensor data analysis.

In 2021, Sabour *et al.* predicted microbial-induced corrosion in cement and sulfur concrete by exploiting ANFIS, GP, and multi-expression programming (MEP) techniques. These models were trained and tested using 40 diverse concrete samples collected and exposed to varying microbial conditions. MEP outperformed the other two models in predicting corrosion in both cases with remarkable correlation coefficient (93.3% and 99.5% for SC and CC) and least error.²⁵⁷ In 2022, Wang *et al.* introduced XGBoost model for concrete corrosion behavior over a dataset of 379 specimens. The XGBoost model was equipped with Bayesian optimization that calculated corrosion as a result of temperature variation, H_2S concentration, and RH observed in the sewage environment. Case study was randomly segmented into 8:2 for test and training datasets, and the R^2 values for both were obtained as 0.85 and 0.87.²⁵⁸

In 2021, Rahman *et al.* used image processing and deep learning models to establish an estimation for the size attributes of bacterial cells in MIC. Authors focused the study on sulfate reducing bacteria (SRB) grown on the steel surfaces. A deep CNN was trained with scanning electron microscopic images of biofilms to perform the detection of varying shapes including pores, surface, bacterial cells, and other

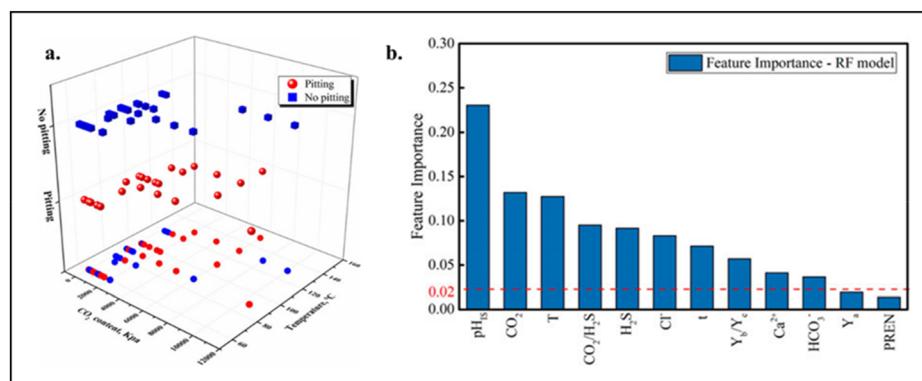


FIG. 14. (a) Effect of temperature (T) and CO_2 concentration over pitting prediction, (b) RF model-based feature priority order for 12 features. Reproduced from Qu *et al.*, Front. Mater. 8, 733813 (2021). Copyright 2021 Authors, licensed under a Creative Commons Attribution (CC-BY) License.²⁵¹

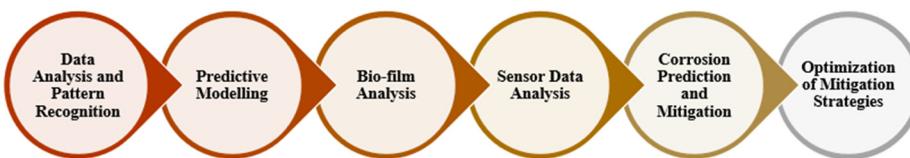


FIG. 15. Machine learning in the prediction of microbiological induced corrosion.

corrosive products. This model study was being involved as a major tool for quantitative analysis of bacterial cells over metallic surfaces when exposed to microbial corrosion.²⁵⁹

In 2021, Bommanapally *et al.* developed a self-supervised learning strategy for evaluating microbially induced corrosive products. The datasets were collected based on scanning electron microscopic images of biofilms, and the whole network was established on simple Siamese architecture which resulted in utilization of unlabeled data for the prediction of corrosion in biofilms. A 62% accuracy was recorded using a simple Siamese architectural network which thereafter predicted the corrosion in affected biofilms.²⁶⁰

In 2020, Adumene *et al.* exploited an integrated BN-Markov process model for anticipation of microbial-induced corrosion, pit depths, and its consequential effects on offshore systems. The MIC rate was calculated probabilistically using the BN network, considering the interdependence and dynamic nonlinearity of the main input components. The consequences of the nonlinear interactions of numerous significant components were assessed, and their level of influence was investigated. The failure chance, critical failure year, and temporal development of the MIC pit depth distribution are all predicted using the projected MIC rate as transition intensity in the Markov process. The created model was found flexible and captured the changing effects of MIC.²⁶¹ Another probabilistic Bayesian model for MIC rate determination was worked by Dawuda *et al.*, which was constituted of 45 nodes comprising biofilm and microbiological conditions, operating and fluid conditions, and metal surfaces. Two case studies (hot water piping and liquid water pipeline) were preferred for the testing of the designed model, which projected the accurate corrosion rate from 0.6 to 0.9 mm/yr, respectively. Additionally, MIC was found most sensitive to surface roughness, material's PREN value, and cathodic reactions.²⁶² In 2004, Hewayde *et al.* investigated the viability of two ANNs to forecast the degradation and compressive strength of concrete under varying concentrations of sulfuric acid. Datasets included concrete mixtures as input variables and weight loss exhibited at various concentrations of sulfuric acid and compressive strength as output variables. RMSE values of 0.013 magnitude along with 8.45% of absolute error corresponded to the successful association of ANNs in the prediction of concrete degradation.²⁶³

V. CASE STUDIES OF MACHINE LEARNING IN CORROSION INHIBITION MODELLING

As information technology and industrial revolution advances, predictive models for corrosion inhibition properties of materials have also improved significantly. Designing new corrosion inhibitors has been hugely influenced by advancements in computer technology, software development, and data management. Using available data for known compounds, ML has been used to predict the corrosion inhibition efficiencies of unknown or novel compounds. This approach does not only reduce the amount of time required to design new efficient

material but also reduces cost and redundant experimentation. Some of the cases of ML applications in corrosion inhibitor design are highlighted here.

A. Artificial neural network (ANN) modeling for corrosion inhibition prediction

In many chemical industries, corrosion is a serious problem that results in financial losses, harm to the environment, and safety issues. To minimize the cost of repair, corrosion inhibitors are intermittently employed to reduce corrosion. However, because there are so many operating conditions affecting the system, evaluating the efficacy of corrosion inhibitors for specific applications is still a challenging issue. Predicting the effectiveness of corrosion inhibitors for specific applications remains a complex task due to the diverse range of influencing factors.^{64,264} Thus, in the context of corrosion inhibition prediction, the use of ANN models is a desirable development. ANN models are powerful optimization and predicting software that can predict the optimal performance of an inhibitor on the metal surface.^{264–266} ANN model operates on the principles of model design, training, model evaluation, and interpretation.²⁶⁷ Due to its robustness and accuracy, many corrosion engineers have used ANN models to advance their research.

Quadri *et al.* reported the ML and ANN predictive models for the corrosion inhibition efficiencies of pyrimidine-based compounds using 40 datasets for model development. The study also involved model performance evaluation metrics and validation.²⁶⁸ In a similar study, Alamri and Alhazmi utilized 54 datasets of pyrimidine corrosion inhibitors for C-steel in HCl to develop ML models for the prediction of inhibition efficiencies. In their study, PLS regression and RF methods were used to select seven weighing descriptors including, electrophilicity, molecular volume, logarithm of the partition coefficient, electron affinity, energy of the lowest unoccupied molecular orbital, molecular mass, and electronegativity as the major contributors to the corrosion inhibition performances of the pyrimidines.¹²⁶ Similar modeling developments for pyrimidine-based corrosion inhibitors have been reported in the literature.^{269–272}

El-Idrissi *et al.* developed QSAR-based ML models from a dataset of 25 benzimidazoles for the prediction of corrosion inhibition efficiencies of benzimidazoles for mild steel in HCl.²⁷³ Our recent study on ML predictive models for inhibition properties of ionic liquids utilized 30 datasets of ionic liquids and the model was based on MLR and multi-layer perceptron neural network (MLPNN) methods applied to six chemical descriptors.²⁷⁴ ML predictive model has also been applied to amino acids based corrosion inhibitors as reported by Akrom *et al.*²⁷⁵ The study involved the use of 47 datasets of known amino acids to develop 20 prospective ML models out of which three models gave satisfactory performance, and the models were used to predict the inhibition efficiencies of three new amino acids.

Sanni *et al.*⁶¹ compared ANN and ANFIS to scrutinize the inhibitory capacity of agricultural waste for stainless steel plate. They considered 1–10 neurons with different training models. The model performance between ANN and ANFIS were studied using statistical parameters. Evidently, the ANFIS model with 20 clusters has much credence over the ANN model, although it was concluded that both models predicted the inhibition efficiency of the agricultural waste with high degree of accuracy. Akbarzadeh *et al.*²⁷⁶ studied corrosion resistance of sol-gel nanocomposite coating via ANN modeling and experimental validation. The proposed ANN model operates based on single hidden-layer perceptron. Next, they further used correlation coefficient and a schematic comparison of the simulated and experimental data in the Nyquist diagrams to assess the optimized model's accuracy. Evidently, there is a strong relationship between the empirical and the ANN estimated data. Also, Anadebe and coworkers²⁷⁷ predicted the inhibition activity of salbutamol derivative using hybrid models based on ANN and ANFIS. These models were further integrated with GA. The model design and plots for process are displayed in Fig. 16. The validation process and other statistical factors served as

the assessment criteria. They suggested that to attain low deviations, zero error, and good R^2 , the network must be trained correctly after several iterations. Based on the values obtained using statistical indices, the ANFIS-GA outperformed the ANN-GA, though both models exhibited good prediction. A recent study has also documented the predictive capacity of CNN model for the inhibition behavior of Timoho leaf extract on stainless steel in H_2SO_4 .²⁷⁸

In another study by Udunwa *et al.*,²⁷⁹ ANN and ANFIS models were used to appraise the inhibition activity of imidazole based ionic liquid on Al alloy. The predicted values 88.48 and 89.06% were recorded for ANN and ANFIS, respectively. Based on the model response obtained, the ANFIS model best suits the experimental data. Also, Abeng and Anadebe⁶³ further scrutinized the performance of ANN and ANFIS model in predicting the corrosion inhibition of doxorubicin drug on mild steel. The experimental dataset used for the ML was based on the gravimetric test where different variables exist. The proposed ANN model has three inputs, five hidden layers, one output layer, and a single response as shown in Fig. 17(a). The regression plots for the training, validation and testing phases are shown in

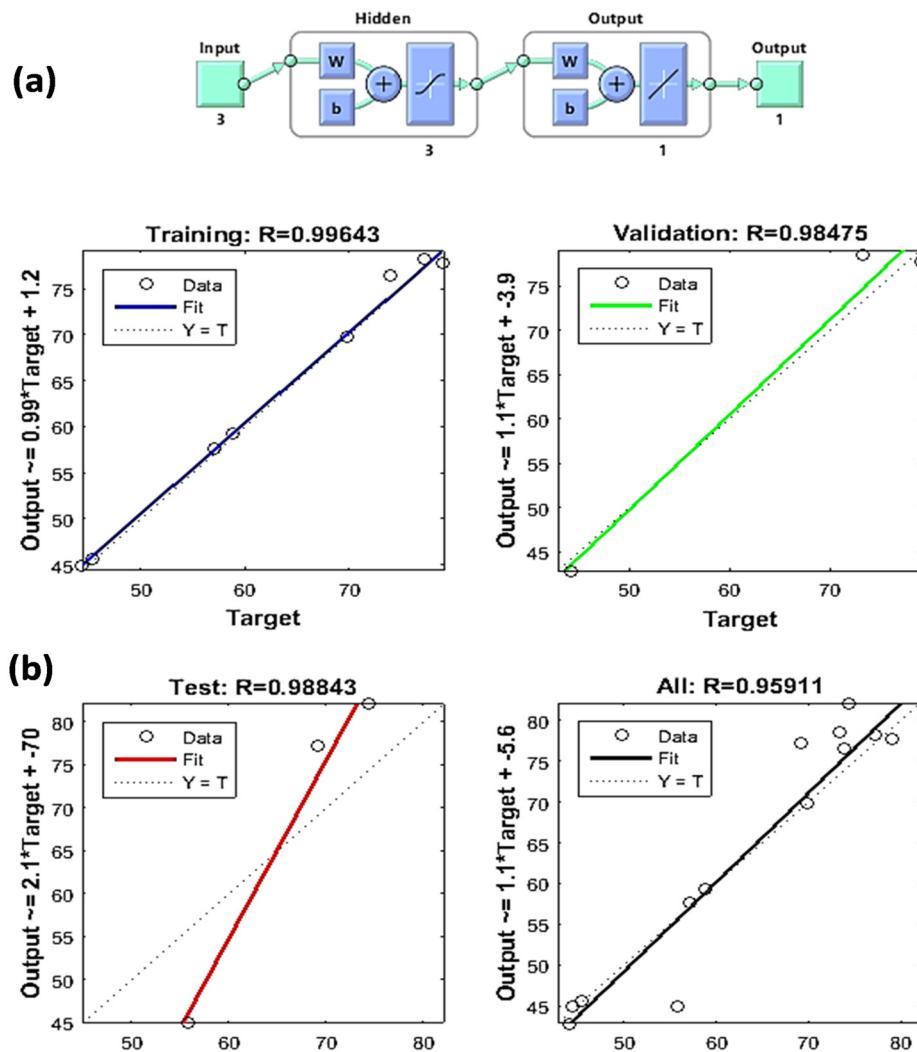


FIG. 16. (a) ANN model, (b) regression plots. Reproduced with permission from Anadebe *et al.*, J. Mol. Liq. 349, 118482 (2022). Copyright 2022 Elsevier.²⁷⁷

Fig. 17(b). Interestingly, the FCM-clustered ANFIS model showed superior performance over the ANN model. In summary, the ANN model showed good data prediction though, one of the limitations is analyzing of few experimental data points.

B. Adaptive neuro-fuzzy inference system (ANFIS) modeling for corrosion inhibition prediction

An attempt has been made to combine the fuzzy logic and ANN into a single framework called “adaptive neuro-fuzzy inference system” (ANFIS). This hybrid model combines the learning and connectionist structure of neural networks with the human-like reasoning style of fuzzy systems to provide a potent tool for modeling and solving complex, nonlinear problems. One of the uniqueness of this model is its ability to handle small and large datasets. This has further necessitated its use in predicting corrosion inhibition of a molecule. Abeng and Ewona²⁸⁰ used RSM and ANFIS model to predict the efficacy of *Thaumatomoccus daniellii* extract on C-steel. From their findings, ANFIS model showed superiority of ($R^2 = 0.994$) over ANN model (0.989). Both models were able to analyze the interactive effects impacting the response (output). Anadebe *et al.*²⁸¹ used ANFIS model

to predict the inhibition efficiency of Ce-MOF in sweet corrosive environment. The ANFIS model was designed based on multi-input and single output response. The ANFIS model as depicted in Fig. 18 analyzes the interactive effects of the system with 3D plots. Three major colors exist, where the light-yellow regions denote the optimum points. The experimental and the ANFIS predicted data were further plotted, and a linear graph was obtained showing good correlation between the two parameters. Also, Emembolu and coworkers⁶² further used RSM-GA and ANFIS-GA models to predict *Aspilia Africana* leaf as corrosion inhibitor for Al alloy. The ANFIS-GA predicted value was 80% at optimum inhibitor dose while that of RSM-GA was 77.3%. The high value of ANFIS-GA is an indication of robustness of the model over RSM which is mainly regarded as a statistical tool. Sanni *et al.*²⁸² predicted the anticorrosion effect of agro-waste on stainless steel in acidizing liquid using ANN and ANFIS models. The proposed models predicted the agro-waste inhibition effect, and the statistical measures were recorded. Based on the statistical measures acquired, ANFIS was found to be the most accurate predictive model compared to the ANN model. From the findings, they concluded that the most important factors for system prediction are the effect of exposure time and inhibitor concentration. Furthermore, Anadebe *et al.*²⁸³ used ANFIS-MISO

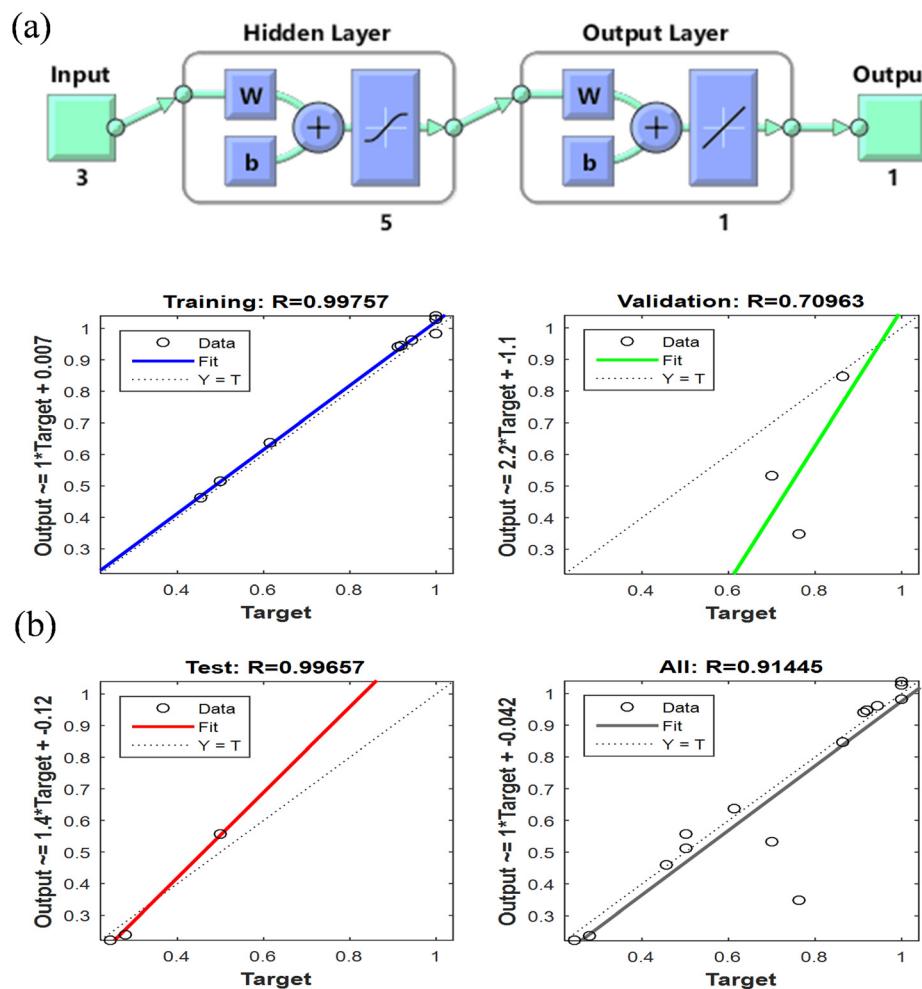


FIG. 17. (a) ANN model (b) regression plots. Reproduced with permission from F. E. Abeng and V. C. Anadebe, Comput. Theor. Chem. 2023, 114334. Copyright 2023 Elsevier.⁶³

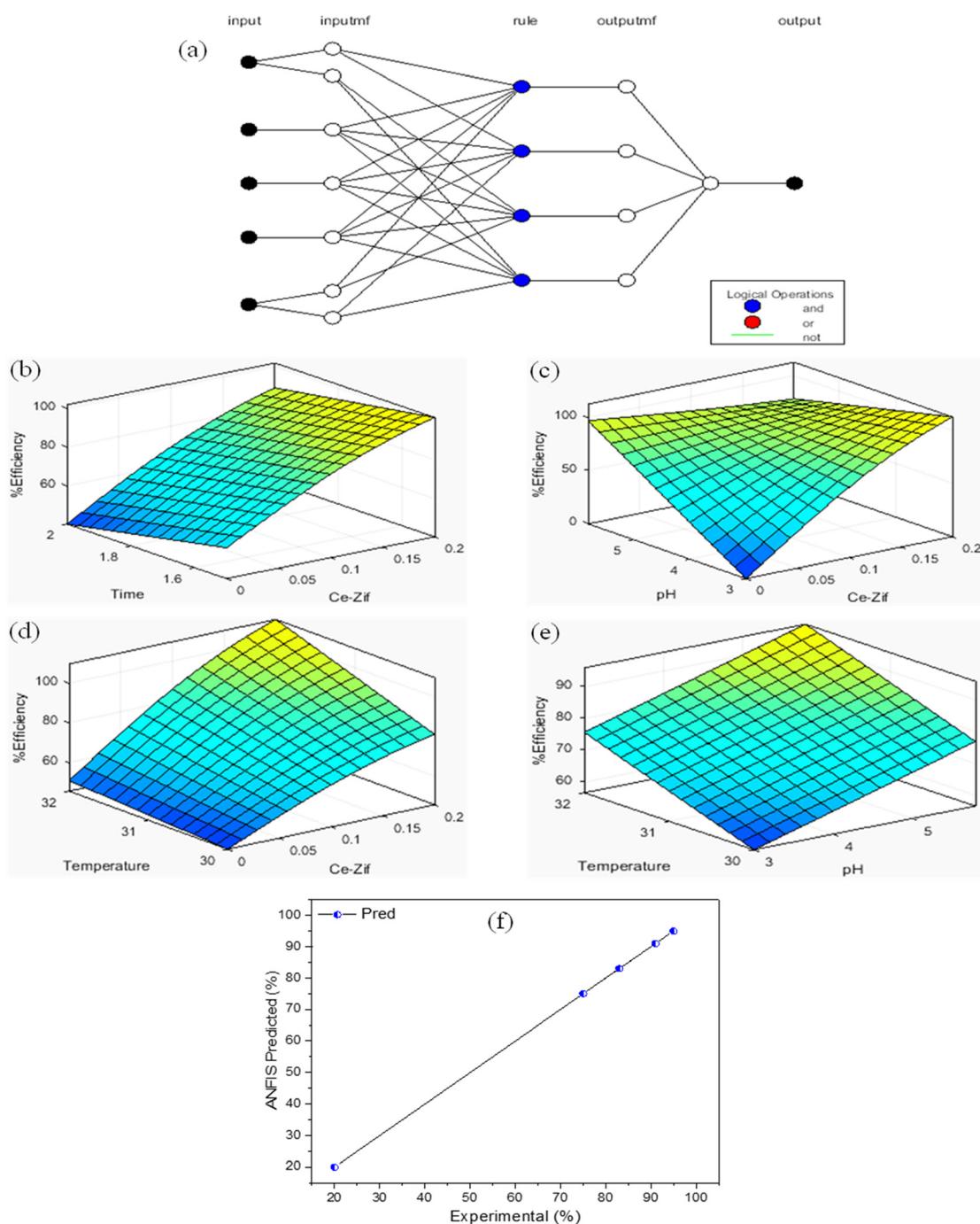


FIG. 18. (a) Model, (b)–(e) 3D plots (f) ANFIS model predicted vs experimental data. Reproduced with permission from Anadebe et al., Process Saf. Environ. Prot. **168**, 499–512 (2022). Copyright 2022 Elsevier.²⁸¹

model to predict the inhibition effect of dexamethasone drug on mild steel. Three independent and one output variables were considered based on gravimetric dataset. The interactive effects of these variables were illustrated as shown in Fig. 19. Evidently, there is a mutual

relationship among the variables, which resulted in a high R^2 of 0.993. This section summarizes the robustness of ANFIS model over other models.

C. Response surface methodology (RSM) for corrosion inhibition prediction

A statistical or mathematical model called RSM can be helpful for modeling and analyzing complex systems where multiple variables affect an interest response. By determining the ideal points for the variables, the main goal of RSM is to maximize this response, which can be any quantifiable quantity influenced by the variables. This methodology is very useful for trials where there are many interactions between the variables and a complex influence from the variables.²⁸⁴ The process can be modeled using a variety of response surface technique models, such as first-order, second-order, and three-level factorial models. The most popular response surface models are second-order models, such as central composite design (CCD) and Box-Behnken design (BBD), because of their great flexibility and increased efficiency for process optimization.^{285,286} However, many researchers preferred BBD over CCD because it has fewer tests, is more economical and practical, and requires less time.²⁸⁷ In the quest for proper design of experiment and understanding of the interactive effects affecting corrosion inhibition process, many researchers have used this mathematical model. Uduwua *et al.*²⁸⁸ used RSM to optimize and forecast the inhibition activity of ionic liquid based on methylimidazolium chloride for Al alloy in acid. In their study, 20 experimental runs were considered with three independent factors and one

response output. Furthermore, the interactive effects of the overall process were studied using 3D RSM plots as shown in Fig. 20. The 3D graphs clearly indicate that relationship exists among the variables for optimization and prediction. The predicted value was 83.30%, which was close to the experimental value of 84.21%.

Uduwua *et al.*²⁸⁹ also studied the applicability of RSM over ANN-GA prediction for *Gongronema latifolium* extract as corrosion inhibitor for Al alloy. Both models were used for optimization and modeling the interactive effects. The process optimization was based on WL dataset. The RSM had prediction of 81.3% at maximum inhibitor dose. Furthermore, Anadebe and coworkers²⁹⁰ used the RSM model to maximize the pigeon pea leaf extract's inhibitory effectiveness for mild steel as shown in Fig. 21.

Prior to the modeling, the authors designed the experiment considering 30 experimental runs where different variables exist. The predicted vs actual plot evidenced a linear graph on the 45° line with data points dispersed randomly, suggesting that the model fits the experimental dataset. They further used a quadratic model to illustrate the connection between the factors of inhibition efficiency and inhibition variables. Based on the analysis of variance obtained (ANOVA), the model was significant with F-value of 39.13. Also, other statistical parameters aligned credibly with the experimental data.

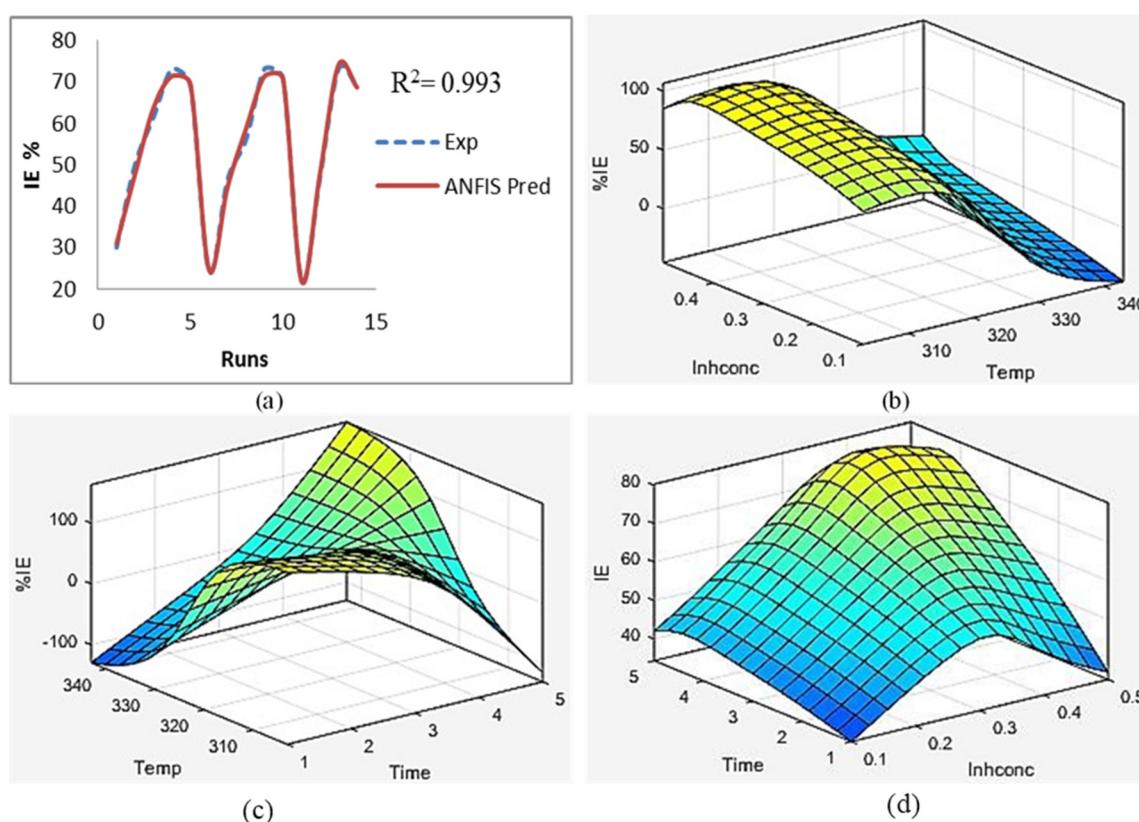


FIG. 19. (a) Experimental vs ANFIS data, (b)–(d) 3D plots for independent and output response. Reproduced with permission from Anadebe *et al.*, J. Taiwan Inst. Chem. Eng. 115, 251–265 (2020). Copyright 2020 Elsevier.²⁸³

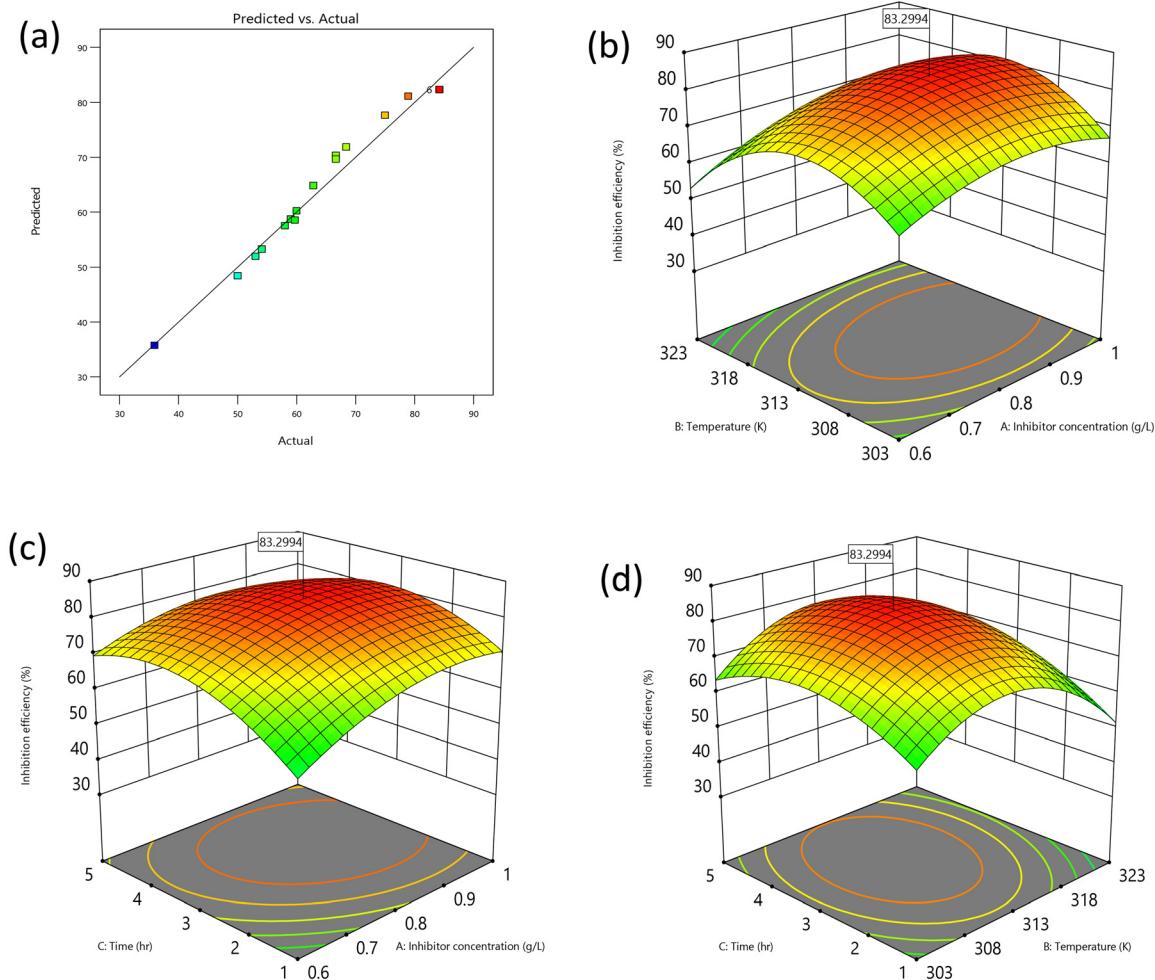


FIG. 20. (a) Predicted vs exp. (b)–(d) 3D RSM plots for overall interactive effects. Reproduced with permission from Uduwana *et al.*, *J. Mol. Liq.* **364**, 120019 (2022). Copyright 2022 Elsevier.²⁸⁸

11 January 2025 14:34:59

D. Random forest (RF) modeling for corrosion inhibition prediction

This cluster model comprises DT.²⁹¹ The DT model operates on the principles of tree-like structure to make response. With the experimental data, the tree model is developed based on the branching parameters affecting the system.²⁹² Thereafter, the nodes are identified which assist in making new input data. The model further combines CART. Based on the uniqueness of RF, many corrosion engineers have used it to make predictions on corrosion behavior and inhibition efficiency. Alamri and Alhazmi¹²⁶ developed a data-driven model based on RF and PLS to design and predict pyrimidine corrosion inhibitors. In developing the model, 54 datasets from the literature were used. They used PLS to select seven descriptors and design the model, while the RF was used to evaluate the nonlinear behavior of the data in order to obtain accurate inhibition prediction. For proper model assessment, important statistical parameters were used to validate the model. Interestingly, the RF gave much credence with MSE of 32.602 compared to the MSE of 64.641 obtained from the PLS model (Fig. 22). Pei

*et al.*²¹⁶ predicted the atmospheric corrosion of Fe/Cu corrosion sensor using three ML models, namely, ANN, RF, and SVM. First, they monitored the atmospheric corrosion of carbon steel by galvanic-type corrosion sensor for a period of 34 days. By the virtue of the models proposed, the impact of relative humidity, rainfall, and temperature was predicted to be the major corrosion contributing factor than air-borne particles. Based on the modeling and validation, the RF model exhibited the highest degree of accuracy compared to ANN and SVM models as shown in Fig. 23. In another interesting work, the RF algorithm was used to model the time-dependent corrosion rate of carbon steel in the presence of inhibitors.²⁹³ The main independent variable considered by the RF model was time. The experimental findings revealed that the corrosion rate is dependent on the dose schedule and final rates are governed by the environment's severity. Given the above, the RF algorithm gave the best prediction on the time-dependent corrosion rate with MSE in assortment 0.005–0.093. Liu *et al.*²⁹⁴ used RF model to evaluate the self-healing of epoxy coating containing ZIF-8@Ca microfillers. In this excellent work, the ML

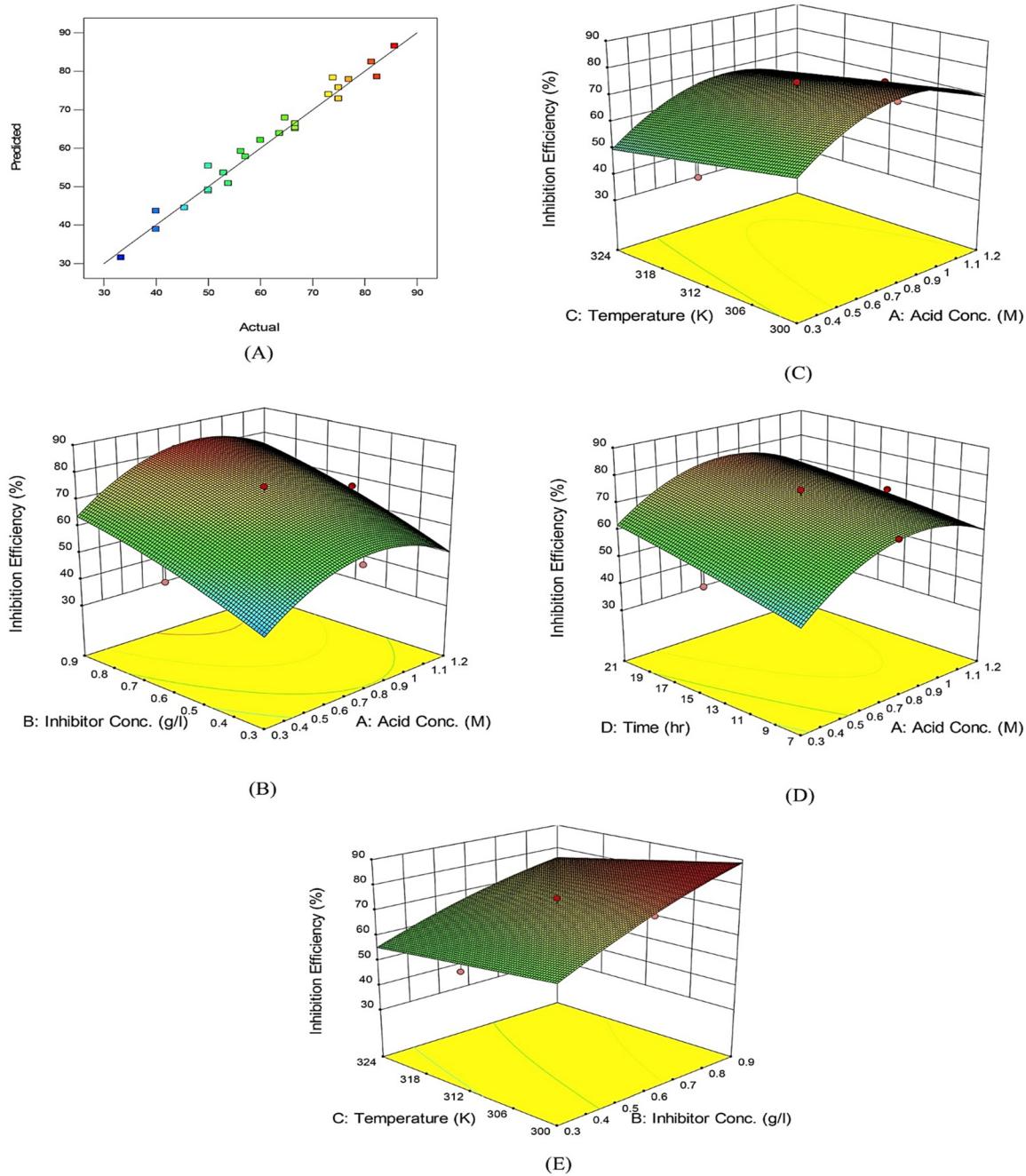


FIG. 21. (a) Predicted vs actual, (b) 3D optimization plots for the interactive effects. Reproduced with permission from Anadebe *et al.*, Mater. Chem. Phys. **233**, 120–132 (2019). Copyright 2019 Elsevier.²⁹⁰

11 January 2025 14:34:59

model was divided into two stage wise processes. First, five distinctive models were evaluated, and the RF was selected for learning process. Following series of five cycles of active learning process, statistical parameters were used to appraise the RF model. The RF model demonstrated good prediction accuracy with R^2 of 0.709, MAPE of 0.081,

and RMSE of 0.685. Second, the optimum coating design was determined via Bayesian optimization route.

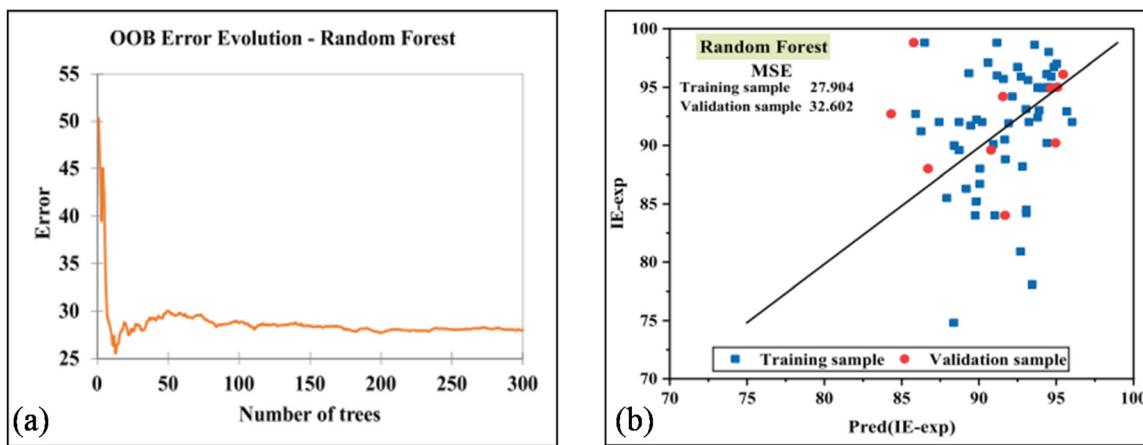


FIG. 22. (a) Error vs number of trees for during optimization procedure of Random forest model and (b) experimental vs calculated IE % efficiencies for training and validated data obtained by Random forest model. Reproduced with permission from Alamri and Alhazmi, J. Saudi Chem. Soc. **26**, 101536 (2022). Copyright 2022 Elsevier.¹²⁶

E. Extreme gradient boosting (XGBoost) modeling for corrosion inhibition prediction

The XGBoost model has received significant attention in the area of AI and ML for solving complex engineering problems.^{295,296} The model operates on the principle of combining multiple models and training them to enhance their accuracy. In view of obtaining optimal values for hyperparameters, Bayesian optimization process have been proven to be more suitable for XGBoost. To achieve this fit, it is imperative to develop an objective function parameter that evaluates the model's performance and uses the data generated to effectively inspect

the optimal configurations.¹²¹ Therefore, considering the compatibility of Bayesian optimization and XGBoost, researchers in diverse fields have adopted this model for optimization and modeling. Wang *et al.*²⁹⁷ used a data-driven optimization route based on Spearman correlation analysis and XGBoost model to predict atmospheric corrosion of low alloy steel. In their study, XGBoost denotes the values of β , and the correlation parameter represent whether β values are positive or negative. From the XGBoost prediction, the β values for the environmental factors were found to be as follows: RH is 0.375, NaCl concentration is 0.334, SO₂ concentration is 0.196, and temperature is 0.095. Qiao *et al.*²⁹⁸ predicted the pitting corrosion of stainless steel in

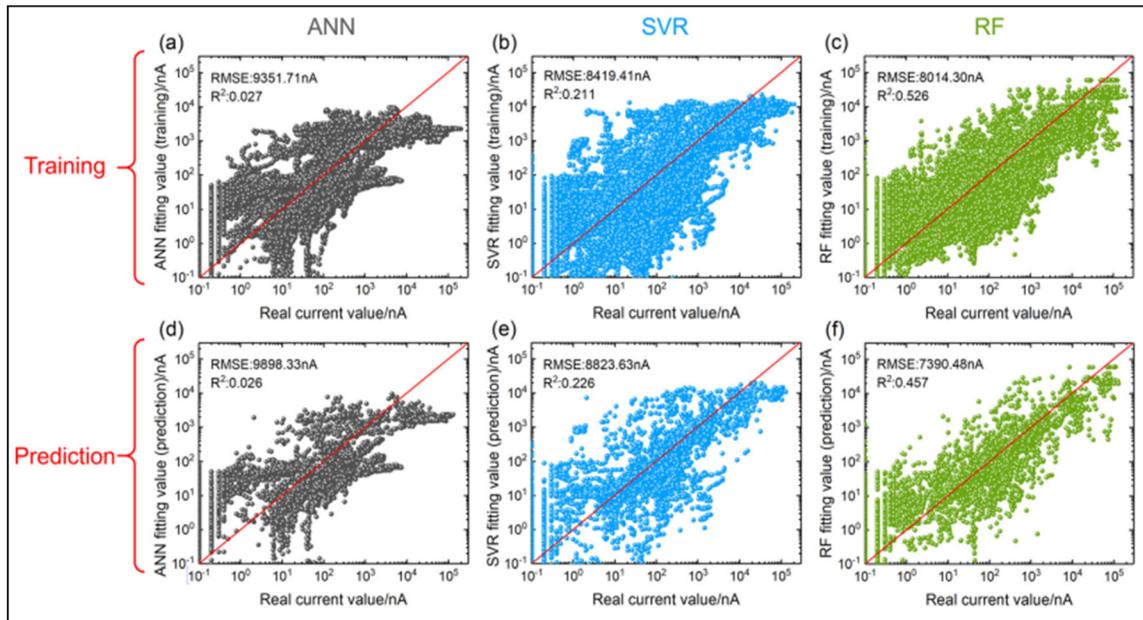


FIG. 23. The fitting results for the training samples by (a) ANN, (b) SVR, and (c) RF models, and for the testing samples by (d) ANN (e) SVR, and (f) RF models. Reproduced with permission from Pei *et al.*, Corros. Sci. **170**, 108697 (2020). Copyright 2020 Elsevier.²¹⁶

chloride-induced environment. The XGBoost predicted the pitting potential of stainless steel. Figure 24 represents the experimental and predicted values. To be specific, the XGBoost model in Fig. 24(a) demonstrated significant and ordered data points from the training point. This convergence resulted in a high value of 0.996 by the XGBoost model. On the contrary, a more dispersed distribution of the data points can be seen in the scatter plots of the training sets for the remaining final five algorithms [Figs. 24(b)–24(f)]. In summary, with experimental and predicted values firmly concentrated around both sides of the diagonal, the XGBoost model clearly outperforms the other models in terms of scatterplot for the test. Based on the statistical factors considered, R^2 and RMSE were 0.877 and 132.51, respectively. Ye *et al.*²⁹⁹ developed a thermal protective coating and further predicted the coating using ML based on SVM and XGBoost model. In their study, proper scrutiny of the output value prediction of varying models was achieved based on 20% test dataset. As shown in Fig. 25, two distinctive colors are visible (blue and orange). The blue denotes the training dataset while the orange describes the test dataset, respectively. Based on the four datasets studied, the XGBoost model had the highest R^2 value, the lowest MSE, and the least bias in the predicted values. The excellent predictive capabilities of GBoost and XGBoost models for organic corrosion inhibitors have been documented in several recently published studies.^{269,271,300–306}

E. Assorted machine learning techniques for corrosion inhibition prediction

It is interesting to note that ML algorithms are transforming corrosion science by enabling more accurate, efficient, and data-driven approaches to predict corrosion rates, assess inhibitor performance, and optimize materials selection. Advanced ML techniques, including GA, GFA, SVM, and quantum machine learning (QML), offer diverse capabilities for analyzing complex, nonlinear relationships typical in corrosion processes.^{264,307} For example, SVM is known for their robust classification capabilities, helping to distinguish between high- and low-risk corrosion scenarios by analyzing complex environmental and material data, while GA is widely used for optimization in corrosion science, as they emulate natural selection to fine-tune feature combinations and hyperparameters in ML models. This approach is valuable in corrosion science for optimizing predictive models to deliver enhanced accuracy across diverse environments and corrosion inhibitors. Quantum neural networks (QNNs) and QML represent cutting-edge developments that leverage quantum computing principles to address the multi-factorial, nonlinear nature of corrosion data. QML offers unprecedented computational efficiency, enabling rapid analysis of massive datasets, which is essential for understanding the vast parameter space in corrosion science. QNN, a specialized form of QML, can be particularly transformative by applying quantum-enhanced neural network architectures to model complex corrosion mechanisms with higher precision and speed. These quantum techniques, although still emerging, presents an exciting frontier for advancing the predictive

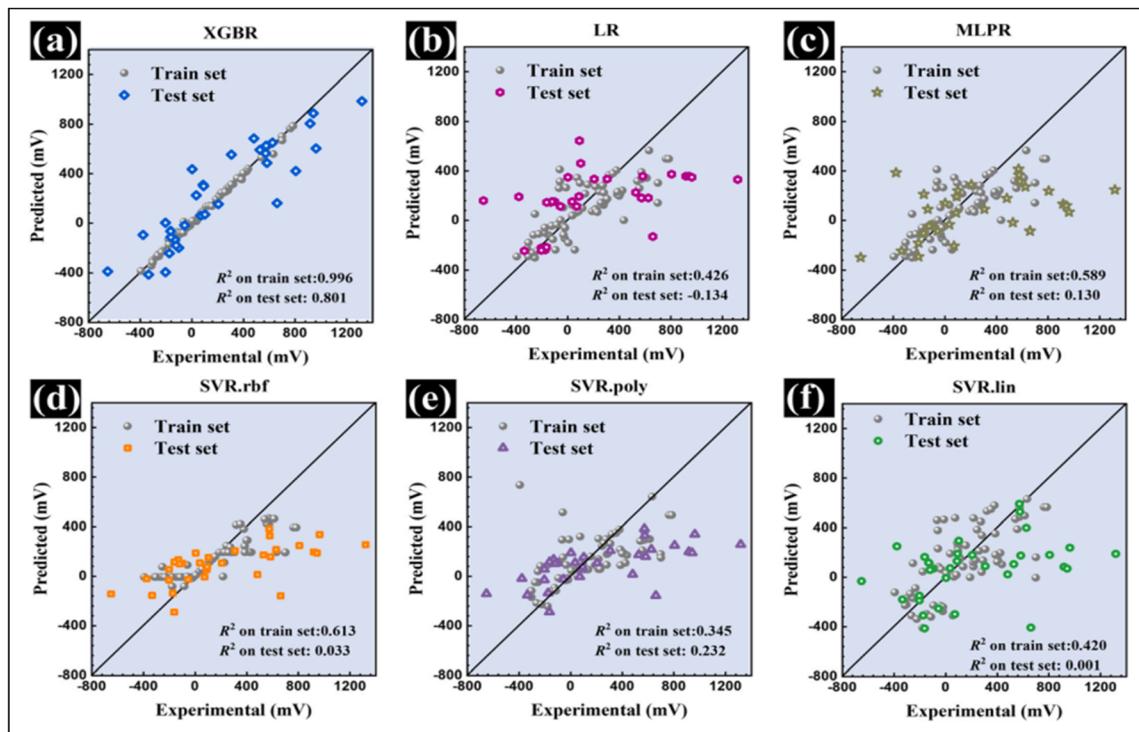


FIG. 24. Experimental-predicted plots of the models trained by six algorithms (XGBR, LR, MLP.R, SVR.rbf, SVR.poly, SVR.lin). Reproduced with permission from Qiao *et al.*, *Colloids Surf., A* **676**, 132274 (2023). Copyright 2023 Elsevier.²⁹⁸

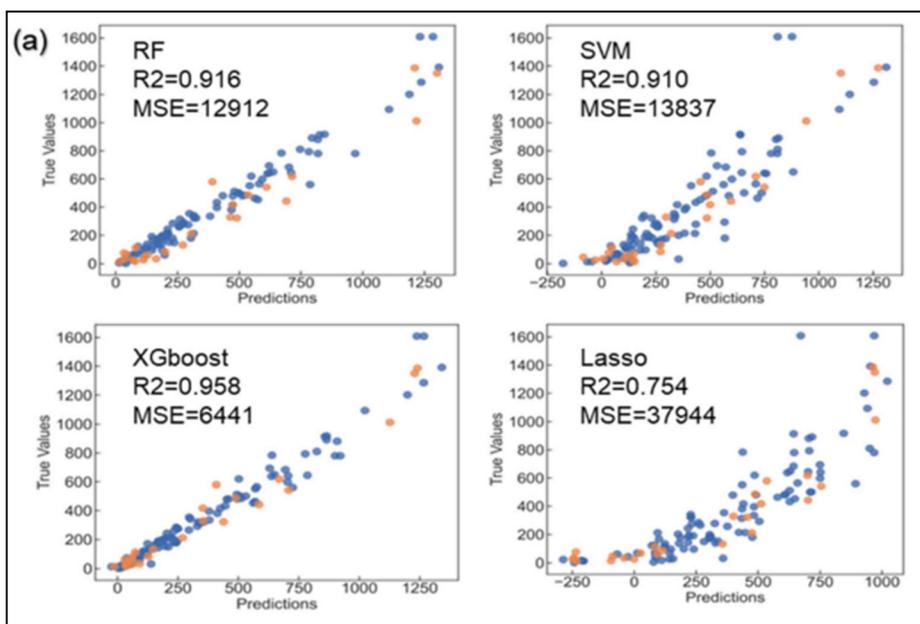


FIG. 25. Prediction results of the output value from different machine learning models. RF, SVM, XGBoost, and Lasso. Reproduced with permission from Ye et al., Corros. Sci. **238**, 112388 (2024). Copyright 2024 Elsevier.²⁹⁹

power and scalability of ML in corrosion science, enabling breakthroughs in materials durability and corrosion prevention.^{35,269}

In a study conducted by Liu *et al.*,¹²⁴ ML models such as MLR, SVM, and RF were employed to correlate the inhibition data of 20 benzimidazoles with obtained descriptors. Multi-collinearity-based clustering analysis and Gini importance evaluation were utilized to eliminate multi-collinearity and identify 47 nonlinear chemical variables, respectively. The SVM model achieved the best outcome of R^2 and RMSE of 0.96 and 4.45, respectively, with top 11 descriptors. The authors concluded by designing six novel benzimidazoles, which yielded excellent inhibition performance ranging from 75% to 96%. Using the same dataset, another study that considered neglected and improved parameters such as aromaticity adopted an SVM approach to model inhibition data. The study attained a good model with R^2 and RMSE of 0.96% and 6.79%, respectively. The inhibition efficacies of the six designed molecules based on the yielded varied from 89% to 98%.³⁰⁸ Some early works exploring the use of GFA for corrosion inhibition prediction of different organic inhibitors were reported by Khaled *et al.*^{309–312} In a recent study, Elsamman *et al.*³¹³ documented the performance of GFA to model the inhibition data of seven bispyrazole (PYRs) compounds tested for their anticorrosive properties against stainless steel in HCl using selected quantum variables. Evaluating the model using a lack-of-fit (LOF) score, the authors established a good agreement between empirical and predicted data. A new inhibitor (BPYR-8) was further employed to validate the GFA model using extensive experimental methods and the obtained inhibition efficiency demonstrated a perfect alignment.

In three studies published by Akrom *et al.* in 2024,^{270,314,315} the performance of QNN was compared to classical neural networks in the development of predictive model for pyrimidine, pyridine-quinoxaline, and quinoxaline inhibitors, respectively. Interestingly, QNN that integrates special properties of quantum computing such as entanglement and superposition with those of ML during calculations offered

superior predictive performance. Figure 26 presents the proposed quantum neural network framework. Employing existing quinoxaline dataset, Akrom *et al.*³¹⁵ constructed variational quantum circuit-based QML. Exploratory data analysis and PCA were employed in data pre-processing, and QML was used to generate the model. The obtained outcomes of 0.987, 0.92, 0.97, and 1.10 for R^2 , MAE, RMSE, and MAD, respectively, revealed a better predictive capacity than comparative studies. Newly designed quinoxaline compounds yielded inhibition efficiency ranging from 89% to 95%, which favorably compares with the performed experimental results.

VI. REAL-WORLD APPLICATIONS OF MACHINE LEARNING TECHNIQUES FOR CORROSION STUDIES

ML techniques are increasingly applied in real-world corrosion research to improve prediction accuracy, optimize inhibitor usage, and extend material lifespan. In industries like oil and gas, manufacturing, and maritime, ML models help to analyze complex, multi-dimensional corrosion data, enabling the prediction of corrosion rates and the effectiveness of inhibitors under various conditions. Algorithms like DT, RF, and SVM are commonly used to model the impact of environmental variables, such as temperature, humidity, and chemical concentrations, on corrosion outcomes. Additionally, ML is applied in selecting the most effective inhibitors for specific materials and environments, reducing costs associated with overuse and minimizing environmental impact. By integrating real-time data from sensors and historical datasets, ML-driven corrosion prediction supports proactive maintenance, reducing downtime, and enhancing the safety and durability of critical infrastructure.^{28,29,35,316} This section briefly captures some applications of machine learning in real-world scenarios.

1. **Oil and gas industry:** In pipeline maintenance, ML algorithms have been applied to predict corrosion rates and identify potential failure points, leading to targeted inhibitor use. For instance,

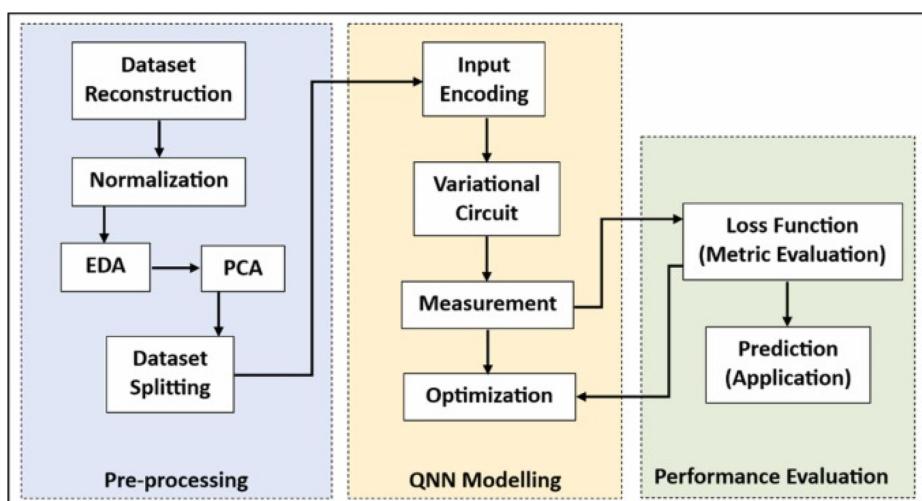


FIG. 26. The proposed quantum neural network workflow. Reproduced with permission from Akrom *et al.*, Mater. Today Commun. **39**, 108758 (2024). Copyright 2022 Elsevier.²⁷⁰

a study by Xu *et al.*³¹⁷ leveraged on an innovative hybrid ML model to accurately predict the corrosion rate of natural gas pipelines in Southwest China. Using decomposing algorithm for data preprocessing, an improved PSO algorithm for optimization, and the complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and SVM for prediction, authors evaluated the performance of the model using prediction effectiveness, seven typical measurement rules, Taylor diagram, and Gray relational analysis (GRA). The outstanding performance of the model in terms of R^2 (0.925) and MAPE (5.73%) outperformed other reported conventional state-of-the-art models. In another study, Hu³¹⁸ demonstrated the superior performance of Extra Trees Regression (ETR) to other five ML models (AdaBoost, CatBoost, GBR, XGBoost, and RF) using experimental corrosion data obtained from an oil and gas pipeline in Eastern Hainan, China. With obtained R^2 , MAE, MSE, and RMSE of 0.93, 0.035, 0.002, and 0.048, respectively, SHAP and ALE were used to visually explain the proposed model to understand the impact of input variables and understand how they affect the corrosion rate. The SHAP values revealed that pressure, temperature and partial pressure of CO_2 largely contributed to the internal corrosion rate of the pipelines.

2. **Marine industry:** ML models like DT and ANN are used to evaluate corrosion in marine environments, where saltwater exposure accelerates degradation. In a recent study by Cai *et al.*,³¹⁹ spatial-temporal model including ANN was employed to predict the corrosion rate of three different industrial steel types in marine environment. Various influential conditions, such as temperature, pH, velocity, oxidation-reduction potential, and dissolved O_2 , obtained from World Ocean Atlas 2018 (WOA18) were integrated into the model development. The ANN model (R^2 of 0.91–0.94 and RMSE of 0.47–1.32) presented a better predictive outcome than MLR and the general Eyring model with salinity, dissolved O_2 and temperature showing positive impacts on the corrosion rate.
3. **Aerospace industry:** Studies show that ML has been used to develop predictive models for the performance of organic inhibitors tested against aerospace Al alloys such as AA2024 and

AA7075. For instance, Winkler *et al.* adopted ML models such as MLR and Bayesian regularized neural networks (BRANN) to develop quantitative, robust, and predictive models for numerous organic inhibitors tested against aerospace AA2024 and AA7075. Thousands of descriptors were obtained using three different approaches, which were trimmed to 11 and 31 depending on the pH and tested Al alloy.^{320,321} Building on this work, Galvao *et al.*¹²⁸ adopted various ML tools (KNN, DT, ANN, SVM) and ensemble methods comprising RF and bagging to analyze the inhibition data of 102 organic compounds for aerospace Al alloy. Among other selected features, the authors further examined the significance of incorporating Gibbs energies and dimerization enthalpies in the model. Statistical parameters such as balanced accuracy, specificity, and sensitivity revealed the most successful performance of RF model.

4. **Infrastructure and civil engineering:** ML models are used to assess and predict corrosion in steel-reinforced concrete, a common issue in bridges and buildings. Various advanced ML models have been utilized to predict corrosion rates for steel rebar and reinforced concrete structures.^{99,322,323} A case study by Munawar *et al.*³²⁴ applied deep hierarchical CNN model based on 16 convolution layers and cycle generative adversarial network (CycleGAN) using image data collected from two locations in Australia: the Sky Rail areas and The Bolte Bridge. Using CycleGAN, the authors successfully achieved automatic corrosion detection and the minimization of infrastructural defects with satisfactory statistical criteria.

VII. CHALLENGES AND LIMITATIONS OF CORROSION INHIBITION MODELLING

The first step at achieving a reliable ML model for predicting corrosion trends is to acquire high-quality data. Mostly, raw corrosion data possess irregularities such as gaps within rows, insufficiency, and the presence of erroneous values. Missing values (gaps in data) are addressed with data preprocessing steps by either removing the data containing gaps (where possible) or filling the gaps with probable values via interpolation. It is worth noting that interpolation can also be

administered when small datasets are to be transformed into larger datasets to solve the problem of insufficient datasets and overfitting associated with corrosion inhibition modeling. Overfitting is inevitable for ANN models developed with small datasets due to the uneven distribution of observations. Unfortunately, corrosion inhibition data adopted for the design of ANN models often have high dimensions and ridiculously small data points. To circumvent the implications of the limited corrosion inhibition data bedeviling corrosion inhibition predictive models, random datasets are generated using dedicated models such as the VSG.^{125,303,325} Essentially, VSG techniques are used to create synthetic datasets that are homogeneous with the original datasets. Additionally, interpolations by synthetic data generation help in preventing overfitting of ML models and enable generalization of the model. However, most of the VSG techniques including Monte Carlo, Exponential Transform, and Bootstrap simulations depend on assumptions that may produce corrosion data that considerably differ from the actual data. Other VSG networks such as the GA, GAN, variational autoencoders (VAE), and nearest neighbor interpolation (NNI) have also been considered for improved synthetic data generation, even though they come with limitations ranging from lack of precision in estimations to restrictions based on acceptability criteria and applications.¹⁴⁴ For instance, GAN, which was originally designed for image generation, has been expanded to cater to tabular data generation through the development of the conditional tabular GAN (CTGAN). Furthermore, recent improvements on these VSG techniques have been instrumental to addressing the challenges associated with their application, which mostly emanated from single-distribution assumptions. For instance, the GMM-VSG addressed the issue of precision by accommodating complex data distributions. The need for data pre-processing processes such as data decomposition and the identification of the appropriate algorithm for pre-modeling exercises could be quite laborious, especially when complex data are involved. In the real sense, complex hybrid models that take care of the data decomposition, parameters optimization, and prediction are required for reliable predictions within the permissible error margin. It is worth noting that a reliable optimization algorithm should be devoid of premature convergence so that the model output could be free from elevated variance, when tested with external inputs. On the other hand, top-notch data decomposition gets rid of insignificant noise during the prediction process. Although from the literature,³¹⁷ the results of the evaluation of the data decomposition–optimization algorithm–prediction hybrid models comprising other models such as dual ensemble models confirm that there is a need for such ternary hybrid model, the development of such system for predicting the inhibition efficiency or corrosion rate of novel inhibitors may be somewhat laborious and time-intensive.

Statistical methods such as Z-score and min-max normalization procedures are also essential in data preprocessing to present datasets on a uniform scale, particularly in image recognition-based corrosion prediction. Multiplication with coefficients or negative logarithm normalization (log scaling) approach is also applicable to data normalization. In instances where the linear normalization technique offers low precision, log scaling method may be helpful.^{16,326} Essentially, the selection of the appropriate normalization approach sometimes requires a *priori* knowledge of the features. For instance, Gong *et al.*¹⁶

used the log-scale approach for normalization because of the logarithmic relationship between the features in the electrochemical impedance dataset. Furthermore, normalization processes such as Z-score normalization are very efficient in identifying outliers and erroneous data, thus helping to avoid working with noisy data, especially in image-based corrosion prediction. Noisy data can also be normalized by physically inspecting the dataset and removing irrelevant entries. However, it is more appropriate to apply statistical methods to identify erroneous data to be eliminated. This is important because, in some instances, important structural descriptors are excluded as part of the variables for the actual model development during the data standardization process. The fact that the excluded descriptors are sometimes excluded before testing implies that the predictions did not consider the contribution of such excluded variables to the outcome of the prediction. This development suggests that domain knowledge and statistical tools are essential to guide the design and execution of ML algorithms for corrosion studies.

Feature screening is also an important step in preparing data for modeling because it shows the importance of the features. Statistical correlation analysis, and advanced statistical methods such as factor analysis and principal component analysis (PCA) are useful for feature selection. The PCA is preferred because it aids the reduction of the dimensionality of the data. However, when numerous features are involved in corrosion studies, the results of PCA may lack interpretability, making alternative means of feature selection inevitable. One of the alternatives is the two-stage feature down selection method for feature selection.³²⁷

ML algorithms have provided answers to numerous research questions, and valid answers to corrosion research problems are not left out. ML algorithms such as ANN, RF, KNN, SVM, GBR, and DT assist in making predictions that aid decision making in the selection of corrosion inhibitors and optimum experimental conditions. Technically, the ML algorithms have proven to be important for predicting the failure or success of oil pipelines and some other economically significant installations that are prone to corrosion.¹⁵ While the predictions of the ML are reliable within the permissible error limit, developing ML models for monitoring corrosion inhibition or identifying potential inhibitors comes with its challenges. The adoption of predictive ML models for identifying potential corrosion inhibitors often depends on the correlation between selected predictors and inhibition efficiency using QSPR and ML algorithms. Since many algorithms are tested to select the best predictor, it is usually difficult to identify the best model when limited performance metrics are deployed. This has played out in a number of publications where authors relied on the RMSE, R², and MSE as performance metrics, and the statistical analysis shows that the models are not statistically different.³²⁸ While this seems like a challenge, further evaluation of the models by estimating the prediction effectiveness or GRA of the models could assist in reaching a reasonable conclusion.

The ANN is one of the most popular nonlinear ML algorithms for corrosion inhibition modeling. The ANN, unlike the KNN algorithm, requires a large amount of data to make reliable predictions devoid of the undesirable overfitting. Additionally, the primary deep learning models such as CNN and RNN have given predictions with a level of accuracy that suggests that further improvement is required. To improve the accuracy of predictions in corrosion inhibition modeling, deep learning models are either combined with other ML

algorithms such as RF, SVM, and CNN, or further developed for improved speed and accuracy. Notably, the YOLO (you only look once) algorithm is a great improvement on the traditional deep learning techniques in terms of accuracy. Also, to avoid the gradient vanishing constraints in RNN, the gated recurrent unit (GRU) and the long short-term memory (LSTM) network built on RNN are utilized.³²⁹ Particularly, LSTM has been very effective in predictions based on temporal windows where data are treated as temporal sequences, thus enabling long-term learning with great accuracy. This type of model is recommended for predicting long-term evolution of corrosion rate in pipelines.

Overall, a significant challenge for ML models in corrosion inhibition research lies in translating computational predictions into practical, real-world applications. Although ML models provide valuable insights into corrosion inhibition efficiency, their predictions are often based on simplified conditions that do not capture the full complexity of real-world environments. In practical settings, variations in temperature, pressure, and multi-component interactions introduce factors that can alter corrosion processes in ways not accounted for by current ML models. These discrepancies make it difficult to directly apply ML-based predictions without further adjustments or recalibration for real-world conditions. To address these limitations, there is a need to explore hybrid approaches that integrate ML with traditional physical and chemical modeling. While promising, these approaches are computationally intensive and require extensive experimental validation, posing additional resource challenges. There is a need for a collaborative approach between computational and experimental researchers, ensuring that ML models are iteratively refined and validated against experimental data. Strengthening the alignment between computational predictions and real-world experimental results will enhance the reliability and applicability of ML models for corrosion inhibition, ultimately providing a more accurate reflection of their effectiveness in practical applications.

VIII. CONCLUSIONS AND PERSPECTIVE

The application of ML algorithms predicts the inhibition efficiency of potential inhibitors, corrosion monitoring under various atmospheric conditions, and the influence of microbes on corrosion inhibition. The various ML algorithms deployed for these purposes have satisfactorily predicted the inhibition efficiency of novel inhibitors, corrosion rate over time, and the contribution of diverse microorganisms to corrosion inhibition. The corrosion rate of oil pipelines, the contribution of microbes, and other factors to pitting corrosion have equally been predicted with ML algorithms. Although SVM, ANN, BN, RF, DT, and the ensemble algorithms have provided platforms for corrosion occurrence or inhibition forecast, it is essential to develop a generalized law for corrosion prediction from the results of previous ML simulations to improve the design of future hybrid algorithms for better output. Also, it is essential for ML algorithm developers to provide full details about the training, testing, and validation steps to provide better understanding of the modeling steps. Furthermore, predictions bordering on identifying potential inhibitors should be substantiated with experimental validation of such prediction because the predicted inhibition efficiency for potential inhibitors is hardly investigated. While the use of multi-dimensional data for the design of ML algorithms is counterproductive, predictions made without some of the eliminated descriptors may only be theoretically acceptable. Therefore, developing ML algorithms that could accommodate several

descriptors without compromising the quality of the prediction is very important.

ACKNOWLEDGMENTS

All the authors acknowledge their institutions for providing the platform to conduct this research work. T.W.Q., E.D.A., and S.E. acknowledge the University of South Africa for their postdoctoral fellowships. Council of Scientific and Industrial Research (CSIR), India, and The World Academy of Sciences (TWAS), Italy, are gratefully acknowledged by V.C.A. for the CSIR-TWAS Postgraduate Fellowship (Award No. 22/FF/CSIRTWAS/2019). V.C.A. and R.C.B. wish to thank CSIR-CECRI (CECRI/PESVC/Pubs/2024-132) for their support.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Taiwo W. Quadri: Conceptualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Chandrabhan Verma:** Validation (equal); Visualization (equal); Writing – original draft (equal). **Hassane Igaz:** Writing – original draft (equal); Writing – review & editing (equal). **Valentine Chikaodili Anadebe:** Writing – original draft (equal); Writing – review & editing (equal). **Rakesh Chandra Barik:** Formal analysis (equal); Writing – original draft (equal). **Lei Guo:** Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Akram AlFantazi:** Formal analysis (equal); Visualization (equal); Writing – original draft (equal). **Bakang Moses Mothudi:** Formal analysis (equal); Visualization (equal). **Eno E. Ebenso:** Conceptualization (equal); Project administration (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Ekemini D. Akpan:** Formal analysis (equal); Writing – original draft (equal). **Saheed Eluwale Elugoke:** Visualization (equal); Writing – original draft (equal). **Lukman O. Olasunkanmi:** Formal analysis (equal); Writing – original draft (equal); Writing – review & editing (equal). **Sheetal Sheetal:** Writing – original draft (equal); Writing – review & editing (equal). **Ashish Kumar Singh:** Formal analysis (equal); Validation (equal); Writing – original draft (equal). **Balaram Pani:** Validation (equal); Writing – original draft (equal). **Jaya Tuteja:** Visualization (equal); Writing – original draft (equal). **Sudhish Kumar Shukla:** Formal analysis (equal); Writing – original draft (equal).

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

- ¹H. Nozari and M. E. Sadeghi, "Artificial intelligence and machine learning for real-world problems (A survey)," *Int. J. Innovation Eng.* **1**, 38–47 (2021).
- ²D. Jakhar and I. Kaur, "Artificial intelligence, machine learning and deep learning: Definitions and differences," *Clin. Exp. Dermatol.* **45**, 131–132 (2020).
- ³T. Panch, P. Szolovits, and R. Atun, "Artificial intelligence, machine learning and health systems," *J. Global Health* **8**, 020303 (2018).

- ⁴I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Comput. Sci.* **2**, 160 (2021).
- ⁵K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data* **3**, 1–40 (2016).
- ⁶M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," *Cognit. Rob.* **3**, 54–70 (2023).
- ⁷A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access* **5**, 7776–7797 (2017).
- ⁸M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," *Healthcare* **10**, 541 (2022).
- ⁹M. Strzelecki and P. Badura, "Machine learning for biomedical application," *Appl. Sci.* **12**, 2022 (2022).
- ¹⁰M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nat. Rev. Genet.* **16**, 321–332 (2015).
- ¹¹K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.* **13**, 8–17 (2015).
- ¹²J. T. McCoy and L. Auret, "Machine learning applications in minerals processing: A review," *Miner. Eng.* **132**, 95–109 (2019).
- ¹³A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. Devylder, M. Walter, and S. Berrouiguet, "Machine learning and natural language processing in mental health: Systematic review," *J. Med. Internet Res.* **23**, e15708 (2021).
- ¹⁴H. Fujiyoshi, T. Hirakawa, and T. Yamashita, "Deep learning-based image recognition for autonomous driving," *IATSS Res.* **43**, 244–252 (2019).
- ¹⁵C. I. Ossai, "A data-driven machine learning approach for corrosion risk assessment—A comparative study," *Big Data Cognit. Comput.* **3**, 28 (2019).
- ¹⁶X. Gong, C. Dong, J. Xu, L. Wang, and X. Li, "Machine learning assistance for electrochemical curve simulation of corrosion and its application," *Mater. Corros.* **71**, 474–484 (2020).
- ¹⁷C. Völker, S. Kruschwitz, and G. Ebelt, "A machine learning-based data fusion approach for improved corrosion testing," *Surv. Geophys.* **41**, 531–548 (2020).
- ¹⁸M. Hosseinzadeh, S. S. Mousavi, A. Hosseinzadeh, and M. Dehestani, "An efficient machine learning approach for predicting concrete chloride resistance using a comprehensive dataset," *Sci. Rep.* **13**, 15024 (2023).
- ¹⁹R. K. Tipu, V. Batra, Suman, V. R. Panchal, and K. S. Pandya, "Predictive modelling of surface chloride concentration in marine concrete structures: A comparative analysis of machine learning approaches," *Asian J. Civ. Eng.* **25**, 1443–1465 (2024).
- ²⁰D. E. Millán-Ocampo, A. Parrales-Bahena, J. G. González-Rodríguez, S. Silva-Martínez, J. Porcayo-Calderón, and J. A. Hernández-Pérez, "Modelling of behavior for inhibition corrosion of bronze using artificial neural network (ANN)," *Entropy* **20**, 409 (2018).
- ²¹T. W. Quadri, L. O. Olasunkanmi, E. D. Akpan, O. E. Fayemi, H.-S. Lee, H. Igaz, C. Verma, L. Guo, S. Kaya, and E. E. Ebenso, "Development of QSAR-based (MLR/ANN) predictive models for effective design of pyridazine corrosion inhibitors," *Mater. Today Commun.* **30**, 103163 (2022).
- ²²T. W. Quadri, L. O. Olasunkanmi, O. E. Fayemi, H. Igaz, O. Dagdag, E.-S. M. Sherif, A. A. Alrashdi, E. D. Akpan, H.-S. Lee, and E. E. Ebenso, "Computational insights into quinoxaline-based corrosion inhibitors of steel in HCl: Quantum chemical analysis and QSPR-ANN studies," *Arabian J. Chem.* **15**, 103870 (2022).
- ²³T. W. Quadri, L. O. Olasunkanmi, O. E. Fayemi, E. D. Akpan, C. Verma, E.-S. M. Sherif, K. F. Khaled, and E. E. Ebenso, "Quantitative structure activity relationship and artificial neural network as vital tools in predicting coordination capabilities of organic compounds with metal surface: A review," *Coord. Chem. Rev.* **446**, 214101 (2021).
- ²⁴E. E. Ebenso, C. Verma, L. O. Olasunkanmi, E. D. Akpan, D. K. Verma, H. Igaz, L. Guo, S. Kaya, and M. A. Quraishi, "Molecular modelling of compounds used for corrosion inhibition studies: A review," *Phys. Chem. Chem. Phys.* **23**, 19987–20027 (2021).
- ²⁵C. Verma, D. S. Chauhan, R. Aslam, P. Banerjee, J. Aslam, T. W. Quadri, S. Zehra, D. K. Verma, M. A. Quraishi, and S. Dubey, "Principles and theories of green chemistry for corrosion science and engineering: Design and application," *Green Chem.* **26**, 4270–4357 (2024).
- ²⁶D. A. Winkler, "Predicting the performance of organic corrosion inhibitors," *Metals* **7**, 553 (2017).
- ²⁷A. E. Hughes, D. A. Winkler, J. Carr, P. Lee, Y. Yang, M. Laleh, and M. Y. Tan, "Corrosion inhibition, inhibitor environments, and the role of machine learning," *Corros. Mater. Degrad.* **3**, 672–693 (2022).
- ²⁸M. M. H. Imran, S. Jamaludin, and A. F. M. Ayob, "A critical review of machine learning algorithms in maritime, offshore, and oil and gas corrosion research: A comprehensive analysis of ANN and RF models," *Ocean Eng.* **295**, 116796 (2024).
- ²⁹M. M. H. Imran, S. Jamaludin, A. F. M. Ayob, A. A. I. M. Ali, S. Z. A. S. Ahmad, M. F. A. Akbar, M. I. R. Suhrab, N. Zainal, S. M. Norzeli, and S. B. Mohamed, "Application of artificial intelligence in marine corrosion prediction and detection," *J. Mar. Sci. Eng.* **11**, 256 (2023).
- ³⁰S. Li, C. Li, and F. Wang, "Computational experiments of metal corrosion studies: A review," *Mater. Today Chem.* **37**, 101986 (2024).
- ³¹C. Nantasesamat, C. Isarankura-Na-Ayudhya, T. Thanakorn Naenna, and V. Prachayavasittikul, "A practical overview of quantitative structure-activity relationship," *EXCLI J.* **8**, 74–88 (2009).
- ³²X. Liu, P. Xu, J. Zhao, W. Lu, M. Li, and G. Wang, "Material machine learning for alloys: Applications, challenges and perspectives," *J. Alloys Compd.* **921**, 165984 (2022).
- ³³T. L. Galvão, I. Ferreira, A. Kuznetsova, G. Novell-Leruth, C. Song, C. Feiler, S. V. Lamaka, C. Rocha, F. Maia, and M. L. Zheludkevich, "CORDATA: An open data management web application to select corrosion inhibitors," *npj Mater. Degrad.* **6**, 48 (2022).
- ³⁴T. L. Galvão, I. Ferreira, F. Maia, J. R. Gomes, and J. Tedim, "DATACORTECH: Artificial intelligence platform for the virtual screen of aluminum corrosion inhibitors," *npj Mater. Degrad.* **8**, 70 (2024).
- ³⁵D. A. Winkler, A. E. Hughes, C. Özkan, A. Mol, T. Würger, C. Feiler, D. Zhang, and S. Lamaka, "Impact of inhibition mechanisms, automation, and computational models on the discovery of organic corrosion inhibitors," *Prog. Mater. Sci.* **149**, 101392 (2025).
- ³⁶L. B. Coelho, D. Zhang, Y. Van Ingelgem, D. Steckelmacher, A. Nowé, and H. Terryn, "Reviewing machine learning of corrosion prediction in a data-oriented perspective," *npj Mater. Degrad.* **6**, 8 (2022).
- ³⁷C. Nantasesamat, C. Isarankura-Na-Ayudhya, and V. Prachayavasittikul, "Advances in computational methods to predict the biological activity of compounds," *Expert Opin. Drug Discovery* **5**, 633–654 (2010).
- ³⁸K. Roy, S. Kar, and R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment* (Academic Press, 2015).
- ³⁹A. Golbraikh, X. S. Wang, H. Zhu, and A. Tropsha, "Predictive QSAR modeling: Methods and applications in drug discovery and chemical risk assessment," in *Handbook of Computational Chemistry* (Springer, Netherlands, 2012), pp. 1309–1342.
- ⁴⁰B. F. Begam and J. S. Kumar, "Computer assisted QSAR/QSPR approaches—A review," *Indian J. Sci. Technol.* **9**, 1–8 (2016).
- ⁴¹T. Puzyń, J. Leszczyński, and M. T. Cronin, *Recent Advances in QSAR Studies: Methods and Applications* (Springer Science & Business Media, 2010).
- ⁴²K. Roy, S. Kar, and R. N. Das, *A Primer on QSAR/QSPR Modeling: Fundamental Concepts* (Springer, 2015).
- ⁴³C. W. Champ and D. K. Shepherd, *Encyclopedia of Statistics in Quality and Reliability* (John Wiley & Sons, 2007).
- ⁴⁴B. Yang, Y. Lv, Z. Zhang, X. He, and Y. Zhu, "Anticorrosion mechanism of natural acidic amino acids on steel in chloride solution: Experimental, theoretical and machine learning approaches," *J. Build. Eng.* **79**, 107801 (2023).
- ⁴⁵T. Hai, A. El-Shafay, R. D. Thanoon, K. Sharma, F. M. Alhomayani, and A. S. M. Metwally, "Development of machine learning techniques in corrosion inhibition evaluation of 5-methyl-1-H-benzotriazole on N80 steel in acidic media," *Mater. Today Commun.* **36**, 106778 (2023).
- ⁴⁶C. Nantasesamat, "Best practices for constructing reproducible QSAR models," in *Ecotoxicological QSARs* (Springer, 2020), pp. 55–75.
- ⁴⁷K. A. Marill, "Advanced statistics: Linear regression, part I: Simple linear regression," *Acad. Emerg. Med.* **11**, 87–93 (2004).
- ⁴⁸K. A. Marill, "Advanced statistics: Linear regression, part II: Multiple linear regression," *Acad. Emerg. Med.* **11**, 94–102 (2004).
- ⁴⁹I. Lukovits, K. Pálfi, I. Bakó, and E. Kálmán, "LKP model of the inhibition mechanism of thiourea compounds," *Corrosion* **53**, 915–919 (1997).

- ⁵⁰I. Lukovits, A. Shaban, and E. Kálman, "Corrosion inhibitors: Quantitative structure–activity relationships," *Russ. J. Electrochim.* **39**, 177–181 (2003).
- ⁵¹K. Varmuza, M. Dehmer, and D. Bonchev, *Statistical Modelling of Molecular Descriptors in QSAR/QSPR* (Wiley Online Library, 2012).
- ⁵²J. Leifer, P. Zapp, and J. Mickalonis, "Predictive models for determination of pitting corrosion versus inhibitor concentrations and temperature for radioactive sludge in carbon steel waste tanks," *Corrosion* **55**, 31–37 (1999).
- ⁵³Y. Zhi, D. Fu, D. Zhang, T. Yang, and X. Li, "Prediction and knowledge mining of outdoor atmospheric corrosion rates of low alloy steels based on the random forests approach," *Metals* **9**, 383 (2019).
- ⁵⁴S. Hadisaputra, A. D. Irham, A. A. Purwoko, E. Junaidi, and A. Hakim, "Development of QSPR models for furan derivatives as corrosion inhibitors for mild steel," *Int. J. Electrochim. Sci.* **18**, 100207 (2023).
- ⁵⁵V. Venkatasubramanian, "The promise of artificial intelligence in chemical engineering: Is it here, finally?" *AIChE J.* **65**, 466–478 (2019).
- ⁵⁶K. F. Khaled and N. A. Al-Mobarak, "A predictive model for corrosion inhibition of mild steel by thiophene and its derivatives using artificial neural network," *Int. J. Electrochim. Sci.* **7**, 1045–1059 (2012).
- ⁵⁷K. Khaled and A. Sherik, "Using neural networks for corrosion inhibition efficiency prediction during corrosion of steel in chloride solutions," *Int. J. Electrochim. Sci.* **8**, 9918–9935 (2013).
- ⁵⁸B. Jain and R. Rawat, "QSAR and ANN-based molecular modeling," in *Computational Modelling and Simulations for Designing of Corrosion Inhibitors* (Elsevier, 2023), pp. 183–199.
- ⁵⁹M. N. Amin, M. Iqbal, M. Ashfaq, B. A. Salami, K. Khan, M. I. Faraz, A. A. Alabdullah, and F. E. Jalal, "Prediction of strength and CBR characteristics of chemically stabilized coal gangue: ANN and random forest tree approach," *Materials* **15**, 4330 (2022).
- ⁶⁰O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Fundamentals of artificial neural networks and deep learning," in *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (Springer, 2022), pp. 379–425.
- ⁶¹O. Sanni, O. Adeleke, K. Ukoba, J. Ren, and T.-C. Jen, "Application of machine learning models to investigate the performance of stainless steel type 904 with agricultural waste," *J. Mater. Res. Technol.* **20**, 4487–4499 (2022).
- ⁶²L. N. Emembolu, P. E. Ohale, C. E. Onu, and N. J. Ohale, "Comparison of RSM and ANFIS modeling techniques in corrosion inhibition studies of *Aspilia africana* leaf extract on mild steel and aluminium metal in acidic medium," *Appl. Surf. Sci. Adv.* **11**, 100316 (2022).
- ⁶³F. E. Abeng and V. C. Anadebe, "Combined electrochemical, DFT/MD-simulation and hybrid machine learning based on ANN-ANFIS models for prediction of doxorubicin drug as corrosion inhibitor for mild steel in 0.5 M H₂SO₄ solution," *Comput. Theor. Chem.* **1229**, 114334 (2023).
- ⁶⁴V. C. Anadebe, V. I. Chukwuike, M. A. Chidiebere, and R. C. Barik, "Synthesis, characterization, and evaluation of Co-MOF Based ZIF-67 for CO₂ corrosion inhibition of X65 steel: Insights from electrochemical studies and a machine learning algorithm," *J. Phys. Chem. C* **127**, 9871 (2023).
- ⁶⁵O. S. Amodu, M. O. Odunlami, J. T. Akintola, T. V. Ojumu, and O. S. Ayanda, "Artificial neural network and response surface methodology for optimization of corrosion inhibition of mild steel in 1 M HCl by Musa paradisiaca peel extract," *Helijon* **8**, e11955 (2022).
- ⁶⁶S. E. Kaskah, G. Ehrenhaft, J. Gollnick, and C. B. Fischer, "Prediction model for optimal efficiency of the green corrosion inhibitor oleoylsarcosine: Optimization by statistical testing of the relevant influencing factors," *Eng.* **4**, 635–649 (2023).
- ⁶⁷W. Y. Loh, "Classification and regression trees," *WIREs Data Min. Knowl. Discovery* **1**, 14–23 (2011).
- ⁶⁸W.-Y. Loh, Q. Zhang, W. Zhang, and P. Zhou, "Missing data, imputation and regression trees," *Stat. Sin.* **30**, 1697–1722 (2020).
- ⁶⁹G. De'ath and K. E. Fabricius, "Classification and regression trees: A powerful yet simple technique for ecological data analysis," *Ecology* **81**, 3178–3192 (2000).
- ⁷⁰D. S. Aguilera-Torres, J. A. Galaviz-Aguilar, and J. R. Cárdenas-Valdez, "Reliable comparison for power amplifiers nonlinear behavioral modeling based on regression trees and random forest," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2022), pp. 1527–1530.
- ⁷¹M. R. Segal, *Machine Learning Benchmarks and Random Forest Regression* (Kluwer Academic Publishers, 2004).
- ⁷²V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geol. Rev.* **71**, 804–818 (2015).
- ⁷³J. Mingers, "An empirical comparison of pruning methods for decision tree induction," *Mach. Learn.* **4**, 227–243 (1989).
- ⁷⁴E. Frank, *Pruning Decision Trees and Lists* (The University of Waikato, 2000).
- ⁷⁵F. Esposito, D. Malerba, G. Semeraro, and J. Kay, "A comparative analysis of methods for pruning decision trees," *IEEE Trans. Pattern Anal. Mach. Intell.* **19**, 476–491 (1997).
- ⁷⁶L. G. Sison and E. K. Chong, "Fuzzy modeling by induction and pruning of decision trees," in *Proceedings of 1994 9th IEEE International Symposium on Intelligent Control* (IEEE, 1994), pp. 166–171.
- ⁷⁷J. Chen, X. Wang, and J. Zhai, "Pruning decision tree using genetic algorithms," in *2009 International Conference on Artificial Intelligence and Computational Intelligence* (IEEE, 2009), pp. 244–248.
- ⁷⁸L. Rokach and O. Maimon, *Decision Trees, Data Mining and Knowledge Discovery Handbook* (Springer, 2005), pp. 165–192.
- ⁷⁹R. Timofeev, *Classification and Regression Trees (CART) Theory and Applications* (Humboldt University, Berlin, 2004), pp. 54.
- ⁸⁰J. M. Luna, E. D. Gennatas, L. H. Ungar, E. Eaton, E. S. Diffenderfer, S. T. Jensen, C. B. Simone, J. H. Friedman, T. D. Solberg, and G. Valdes, "Building more accurate decision trees with the additive tree," *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19887–19893 (2019).
- ⁸¹V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees: An overview and their use in medicine," *J. Med. Syst.* **26**, 445–463 (2002).
- ⁸²W. N. H. W. Mohamed, M. N. M. Salleh, and A. H. Omar, "A comparative study of reduced error pruning method in decision tree algorithms," in *2012 IEEE International Conference on Control System, Computing and Engineering* (IEEE, 2012), pp. 392–397.
- ⁸³R. Pandya and J. Pandya, "C5.0 algorithm to improved decision tree with feature selection and reduced error pruning," *Int. J. Comput. Appl.* **117**, 18–21 (2015).
- ⁸⁴T. Elomaa and M. Kaariainen, "An analysis of reduced error pruning," *J. Artif. Intell. Res.* **15**, 163–187 (2001).
- ⁸⁵B. R. Kiran and J. Serra, "Cost-complexity pruning of random forests," in *Mathematical Morphology and Its Applications to Signal and Image Processing: 13th International Symposium, ISMM 2017, Fontainebleau, France, May 15–17, 2017* (Springer, 2017), pp. 222–232.
- ⁸⁶J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence* (MIT Press, 1992).
- ⁸⁷J. R. Sampson, "Adaptation in natural and artificial systems (John H. Holland)," *Soc. Ind. Appl. Math. Rev.* **18**, 529 (1976).
- ⁸⁸L. Booker, *Perspectives on Adaptation in Natural and Artificial Systems* (Santa Fe Institute Studies on the Sciences of Complexity) (Oxford University Press, 2005).
- ⁸⁹M. Dianati, I. Song, and M. Treiber, *An Introduction to Genetic Algorithms and Evolution Strategies* (University of Waterloo, Ontario N2L 3G1, Canada, 2002).
- ⁹⁰J. Carr, "An introduction to genetic algorithms," *Senior Project* **1**, 40 (2014).
- ⁹¹J.-B. Mouret, *Evolutionary Adaptation in Natural and Artificial Systems* (Université Pierre et Marie Curie, 2015).
- ⁹²S. Winkler, M. Affenzeller, and S. Wagner, "Advanced genetic programming based machine learning," *J. Math. Model. Algorithms* **6**, 455–480 (2007).
- ⁹³H. Iba, Y. Hasegawa, and T. K. Paul, *Applied Genetic Programming and Machine Learning* (CRC Press, 2009).
- ⁹⁴W. B. Langdon and R. Poli, *Foundations of Genetic Programming* (Springer Science & Business Media, 2013).
- ⁹⁵A. S. Abdullah, C. Ramya, V. Priyadharsini, C. Reshma, and S. Selvakumar, "A survey on evolutionary techniques for feature selection," in *2017 Conference on Emerging Devices and Smart Systems (ICEDSS)* (IEEE, 2017), pp. 58–62.
- ⁹⁶N. Zhang, A. Gupta, Z. Chen, and Y.-S. Ong, "Evolutionary machine learning with minions: A case study in feature selection," *IEEE Trans. Evol. Comput.* **26**, 130–144 (2022).

- ⁹⁷K. Lakra and A. Chug, "Improving software maintainability prediction using hyperparameter tuning of baseline machine learning algorithms," *Applications of Artificial Intelligence and Machine Learning: Select Proceedings of ICAAIIML 2020* (Springer, 2021), pp. 679–692.
- ⁹⁸P. A. Adedeji, O. O. Olatunji, N. Madushele, and T.-C. Jen, "Evolutionary-based hyperparameter tuning in machine learning models for condition monitoring in wind turbines—A survey," in *2021 IEEE 12th International Conference on Mechanical and Intelligent Manufacturing Technologies (ICMIMT)* (IEEE, 2021), pp. 254–258.
- ⁹⁹Z. Dong, L. Ding, Z. Meng, K. Xu, Y. Mao, X. Chen, H. Ye, and A. Poursaei, "Machine learning-based corrosion rate prediction of steel embedded in soil," *Sci. Rep.* **14**, 18194 (2024).
- ¹⁰⁰S. Pourrahimi, S. Hakimian, A.-H. Bouzid, and L. A. Hof, "On the use of machine learning algorithms to predict the corrosion behavior of stainless steels in lactic acid," *Metals* **13**, 1459 (2023).
- ¹⁰¹M. Hussain, T. Zhang, I. Jamil, A. A. Soomro, and I. Hussain, "Application of machine learning approaches to prediction of corrosion defects in energy pipelines," in *Advances in Corrosion Modelling* (Springer, 2024), pp. 127–166.
- ¹⁰²A. R. M. Forkan, Y.-B. Kang, P. P. Jayaraman, K. Liao, R. Kaul, G. Morgan, R. Ranjan, and S. Sinha, "CorrDetector: A framework for structural corrosion detection from drone images using ensemble deep learning," *Expert Syst. Appl.* **193**, 116461 (2022).
- ¹⁰³I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
- ¹⁰⁴M. M. Taye, "Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions," *Computers* **12**, 91 (2023).
- ¹⁰⁵Y. Saéys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics* **23**, 2507–2517 (2007).
- ¹⁰⁶R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.* **97**, 273–324 (1997).
- ¹⁰⁷P. Dhal and C. Azad, "A comprehensive survey on feature selection in the various fields of machine learning," *Appl. Intell.* **52**, 4543–4581 (2022).
- ¹⁰⁸J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing* **300**, 70–79 (2018).
- ¹⁰⁹M. A. Hall and L. A. Smith, "Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper," in *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference* (Association for the Advancement of Artificial Intelligence, 1999), pp. 235–239.
- ¹¹⁰V. Bolón-Canedo, N. Sánchez-Marín, and A. J. K. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowl. Inf. Syst.* **34**, 483–519 (2013).
- ¹¹¹B. Venkatesh and J. Anuradha, "A review of feature selection and its methods," *Cybern. Informat. Technol.* **19**, 3–26 (2019).
- ¹¹²G. Chandrashekhar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.* **40**, 16–28 (2014).
- ¹¹³N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Front. Bioinform.* **2**, 927312 (2022).
- ¹¹⁴Y. Lyu, Y. Feng, and K. Sakurai, "A survey on feature selection techniques based on filtering methods for cyber attack detection," *Information* **14**, 191 (2023).
- ¹¹⁵J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.* **50**, 94 (2018).
- ¹¹⁶D. S. Huang, "Radial basis probabilistic neural networks: Model and application," *Int. J. Pattern Recognit. Artif. Intell.* **13**, 1083–1101 (1999).
- ¹¹⁷J.-R. Zhang, J. Zhang, T.-M. Lok, and M. R. Lyu, "A hybrid particle swarm optimization-back-propagation algorithm for feedforward neural network training," *Appl. Math. Comput.* **185**, 1026–1037 (2007).
- ¹¹⁸S. D. Asir Antony Gnana, B. S. Appavu Alias, and L. E. Jebamalar, "Performance metrics for feature selection, clustering and classification algorithms," *Asian J. Res. Soc. Sci. Human* **6**, 107–122 (2016).
- ¹¹⁹L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: A review with examples from ecology," *Ecol. Monogr.* **93**, e1557 (2023).
- ¹²⁰J. Kalaiappan, A. R. Bagepalli, S. Almal, R. Mishra, Y.-C. Hu, and K. Srinivasan, "Impact of cross-validation on machine learning models for early detection of intrauterine fetal demise," *Diagnostics* **13**, 1692 (2023).
- ¹²¹D. Ruiz, A. Casas, C. A. Escobar, A. Perez, and V. Gonzalez, "Advanced machine learning techniques for corrosion rate estimation and prediction in industrial cooling water pipelines," *Sensors* **24**, 3564 (2024).
- ¹²²N. Wang, L. Song, H. Fang, B. Li, and F. Wang, "Multi-parameter maximum corrosion depth prediction model for buried pipelines based on GSCV-XGBoost," *IEEE Access* **11**, 116452–116471 (2023).
- ¹²³R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *J. Appl. Sci. Technol. Trends* **1**, 56–70 (2020).
- ¹²⁴Y. Liu, Y. Guo, W. Wu, Y. Xiong, C. Sun, L. Yuan, and M. Li, "A machine learning-based QSAR model for benzimidazole derivatives as corrosion inhibitors by incorporating comprehensive feature selection," *Interdiscip. Sci. Comput. Life Sci.* **11**, 738–747 (2019).
- ¹²⁵T. Sutojo, S. Rustad, M. Akrom, A. Syukur, G. F. Shidik, and H. K. Dipojono, "A machine learning approach for corrosion small datasets," *npj Mater. Degrad.* **7**, 18 (2023).
- ¹²⁶A. H. Alamri and N. Alhazmi, "Development of data driven machine learning models for the prediction and design of pyrimidine corrosion inhibitors," *J. Saudi Chem. Soc.* **26**, 101536 (2022).
- ¹²⁷M. Fernandez, M. Breedon, I. S. Cole, and A. S. Barnard, "Modeling corrosion inhibition efficacy of small organic molecules as non-toxic chromate alternatives using comparative molecular surface analysis (CoMSA)," *Chemosphere* **160**, 80–88 (2016).
- ¹²⁸T. L. P. Galvão, G. Novell-Leruth, A. Kuznetsova, J. Tedim, and J. R. B. Gomes, "Elucidating structure–property relationships in aluminum alloy corrosion inhibitors by machine learning," *J. Phys. Chem. C* **124**, 5624–5635 (2020).
- ¹²⁹L. Yan, Y. Diao, Z. Lang, and K. Gao, "Corrosion rate prediction and influencing factors evaluation of low-alloy steels in marine atmosphere using machine learning approach," *Sci. Technol. Adv. Mater.* **21**, 359–370 (2020).
- ¹³⁰Y. Cai, Y. Xu, Y. Zhao, and X. Ma, "Atmospheric corrosion prediction: A review," *Corros. Rev.* **38**, 299–321 (2020).
- ¹³¹C. Aldrich, "Process variable importance analysis by use of random forests in a shapley regression framework," *Minerals* **10**, 420 (2020).
- ¹³²S. Kilmen and O. Bulut, "Scale abbreviation with recursive feature elimination and genetic algorithms: An illustration with the test emotions questionnaire," *Information* **14**, 63 (2023).
- ¹³³X. Wang, W. Zhang, W. Zhang, and Y. Ai, "A machine learning method for predicting corrosion weight gain of uranium and uranium alloys," *Materials* **16**, 631 (2023).
- ¹³⁴C. Özkan, L. Sahlmann, C. Feiler, M. Zheludkevich, S. Lamaka, P. Sewlikar, A. Kooijman, P. Taheri, and A. Mol, "Laying the experimental foundation for corrosion inhibitor discovery through machine learning," *npj Mater. Degrad.* **8**, 21 (2024).
- ¹³⁵J. Wang, P. Xu, X. Ji, M. Li, and W. Lu, "Feature selection in machine learning for perovskite materials design and discovery," *Materials* **16**, 3134 (2023).
- ¹³⁶H. Lu, H. Peng, Z.-D. Xu, J. C. Matthews, N. Wang, and T. Iseley, "A feature selection-based intelligent framework for predicting maximum depth of corroded pipeline defects," *J. Perform. Constr. Facil.* **36**, 04022044 (2022).
- ¹³⁷X. Li, B. Vaghfinazari, T. Würger, S. V. Lamaka, M. L. Zheludkevich, and C. Feiler, "Predicting corrosion inhibition efficiencies of small organic molecules using data-driven techniques," *npj Mater. Degrad.* **7**, 64 (2023).
- ¹³⁸E. J. Schiessler, T. Würger, S. V. Lamaka, R. H. Meißner, C. J. Cyron, M. L. Zheludkevich, C. Feiler, and R. C. Aydin, "Predicting the inhibition efficiencies of magnesium dissolution modulators using sparse machine learning models," *npj Comput. Mater.* **7**, 193 (2021).
- ¹³⁹T. H. Pham, P. K. Le, and D. N. Son, "A data-driven QSPR model for screening organic corrosion inhibitors for carbon steel using machine learning techniques," *RSC Adv.* **14**, 11157–11168 (2024).
- ¹⁴⁰S. Hakimian, S. Pourrahimi, A.-H. Bouzid, and L. A. Hof, "Application of machine learning for the classification of corrosion behavior in different environments for material selection of stainless steels," *Comput. Mater. Sci.* **228**, 112352 (2023).
- ¹⁴¹Y. Diao, L. Yan, and K. Gao, "Improvement of the machine learning-based corrosion rate prediction model through the optimization of input features," *Mater. Des.* **198**, 109326 (2021).

- ¹⁴²X. Qiu, D. Fu, Z. Fu, K. Riha, and R. Burget, "The method for material corrosion modelling and feature selection with SVM-RFE," in *2011 34th International Conference on Telecommunications and Signal Processing (TSP) Budapest, Hungary* (IEEE, 2011), pp. 443–447.
- ¹⁴³Z. Zhu, Q. Zheng, H. Liu, J. Zhang, T. Wu, and X. Qu, "Prediction model for pipeline pitting corrosion based on multiple feature selection and residual correction," *J. Mar. Sci. Appl.* (published online).
- ¹⁴⁴S. Rustad, M. Akrom, T. Sutojo, and H. K. Dipojono, "A feature restoration for machine learning on anti-corrosion materials," *Case Stud. Chem. Environ. Eng.* **10**, 100902 (2024).
- ¹⁴⁵Q. Zhao, L. Li, L. Zhang, and M. Zhao, "Recognition of corrosion state of water pipe inner wall based on SMA-SVM under RF feature selection," *Coatings* **13**, 26 (2023).
- ¹⁴⁶M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.* **45**, 427–437 (2009).
- ¹⁴⁷A. Botchkarev, "A new typology design of performance metrics to measure errors in machine learning regression algorithms," *Interdiscip. J. Inf., Knowl., Manage.* **14**, 45–79 (2019).
- ¹⁴⁸F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Mach. Learn.* **42**, 203–231 (2001).
- ¹⁴⁹L. Breiman, *Classification and Regression Trees* (Routledge, 2017).
- ¹⁵⁰V. Plevris, G. Solarzano, N. Bakas, and M. B. Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," in *ECCOMAS Congress 2022—8th European Congress on Computational Methods in Applied Sciences and Engineering* (European Community on Computational Methods in Applied Sciences, 2022).
- ¹⁵¹S. R. Arumugam, S. Gowr, Abimala, Balakrishna, and O. Manoj, "Performance evaluation of machine learning and deep learning techniques," in *Convergence of Deep Learning in Cyber-IoT Systems and Security* edited by R. Chakraborty, A. Ghosh, J. K. Mandal, and S. Balamurugan (Wiley, 2022), pp. 21–65.
- ¹⁵²M. Steurer, R. J. Hill, and N. Pfeifer, "Metrics for evaluating the performance of machine learning based automated valuation models," *J. Prop. Res.* **38**, 99–129 (2021).
- ¹⁵³M. Lee, Y. Yu, Y. Cheon, S. Baek, Y. Kim, K. Kim, H. Jung, D. Lim, H. Byun, C. Lee, and J. Jeong, "Machine learning-based prediction of controlled variables of apc systems using time-series data in the petrochemical industry," *Processes* **11**, 2091 (2023).
- ¹⁵⁴I. Ekanayake, D. Meddage, and U. Rathnayake, "A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP)," *Case Stud. Constr. Mater.* **16**, e01059 (2022).
- ¹⁵⁵Z. Wang, T. Zhou, and K. Sundmacher, "Interpretable machine learning for accelerating the discovery of metal-organic frameworks for ethane/ethylene separation," *Chem. Eng. J.* **444**, 136651 (2022).
- ¹⁵⁶Y. Song, Q. Wang, X. Zhang, L. Dong, S. Bai, D. Zeng, Z. Zhang, H. Zhang, and Y. Xi, "Interpretable machine learning for maximum corrosion depth and influence factor analysis," *npj Mater. Degrad.* **7**, 9 (2023).
- ¹⁵⁷B. Liang, H. Liu, E. L. Cressey, C. Xu, L. Shi, L. Wang, J. Dai, Z. Wang, and J. Wang, "Uncertainty of partial dependence relationship between climate and vegetation growth calculated by machine learning models," *Remote Sens.* **15**, 2920 (2023).
- ¹⁵⁸J. Wang, Z. Zhang, X. Liu, Y. Shao, X. Liu, and H. Wang, "Prediction and interpretation of concrete corrosion induced by carbon dioxide using machine learning," *Corros. Sci.* **233**, 112100 (2024).
- ¹⁵⁹C. H. E. N. Qian, H. Wei, Z. Changhui, G. U. A. N. Aocheng, Z. Chen, and Y. Xiaopeng, "Physics-guided prediction of corrosion rate inside gathering and transportation pipelines and explanatory analysis," *J. Chin. Soc. Corros. Prot.* (published online).
- ¹⁶⁰M. Liu and W. Li, "Prediction and analysis of corrosion rate of 3C steel using interpretable machine learning methods," *Mater. Today Commun.* **35**, 106408 (2023).
- ¹⁶¹M. E. A. Ben Seghier, O. A. Mohamed, and H. Ouamer, "Machine learning-based Shapley additive explanations approach for corroded pipeline failure mode identification," *Structures* **65**, 106653 (2024).
- ¹⁶²A. Moses, D. Chen, P. Wan, and S. Wang, "Prediction of electrochemical corrosion behavior of magnesium alloy using machine learning methods," *Mater. Today Commun.* **37**, 107285 (2023).
- ¹⁶³J. Kuang and Z. Long, "Prediction model for corrosion rate of low-alloy steels under atmospheric conditions using machine learning algorithms," *Int. J. Miner. Metall. Mater.* **31**, 337–350 (2024).
- ¹⁶⁴Z. Li, Z. Long, S. Lei, X. Liu, L. Yang, W. Zhang, and T. Zhang, "Evaluating the corrosion resistance of marine steels under different exposure environments via machine learning," *Phys. Scr.* **98**, 015402 (2022).
- ¹⁶⁵P. Barbrook-Johnson and A. S. Penn, *Bayesian Belief Networks, Systems Mapping: How to Build and Use Causal Models of Systems* (Springer, 2022), pp. 97–112.
- ¹⁶⁶L. M. Gladence, M. Karthi, and V. M. Anu, "A statistical comparison of logistic regression and different Bayes classification methods for machine learning," *ARPN J. Eng. Appl. Sci.* **10**, 5947–5953 (2015).
- ¹⁶⁷A. Subaira, T. Primya, C. Vinothini, K. Karpagavadi, S. Saranya, and C. Ram Kumar, "Emerging techniques and algorithms used in soft computation," in *Translating Healthcare Through Intelligent Computational Methods* (Springer, 2023), pp. 245–263.
- ¹⁶⁸N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.* **29**, 131–163 (1997).
- ¹⁶⁹Building Bayesian Networks, in *Modeling and Reasoning with Bayesian Networks*, edited by A. Darwiche (Cambridge University Press, Cambridge, 2009), pp. 76–125.
- ¹⁷⁰P. Larranaga, H. Karshenas, C. Bielza, and R. Santana, "A review on evolutionary algorithms in Bayesian network learning and inference tasks," *Inf. Sci.* **233**, 109–125 (2013).
- ¹⁷¹B. G. Marcot and T. D. Penman, "Advances in Bayesian network modelling: Integration of modelling technologies," *Environ. Modell. software* **111**, 386–393 (2019).
- ¹⁷²Y. Liu, T. Zhao, W. Ju, and S. Shi, "Materials discovery and design using machine learning," *J. Materomics* **3**, 159–177 (2017).
- ¹⁷³Z. Lu, "Computational discovery of energy materials in the era of big data and machine learning: A critical review," *Mater. Rep.: Energy* **1**, 100047 (2021).
- ¹⁷⁴K. Rajan, "Materials informatics," *Mater. Today* **8**, 38–45 (2005).
- ¹⁷⁵C. Tian, T. Li, J. Bustillos, S. Bhattacharya, T. Turnham, J. Yeo, and A. Moridi, "Data-driven approaches toward smarter additive manufacturing," *Adv. Intell. Syst.* **3**, 2100014 (2021).
- ¹⁷⁶A. Paul, M. Mozaffar, Z. Yang, W.-k Liao, A. Choudhary, J. Cao, and A. Agrawal, "A real-time iterative machine learning approach for temperature profile prediction in additive manufacturing processes," in *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (IEEE, 2019), pp. 541–550.
- ¹⁷⁷D. C. Miller and N. J. Salkind, *Handbook of Research Design and Social Measurement* (Sage, 2002).
- ¹⁷⁸J. H. Panchal, S. R. Kalidindi, and D. L. McDowell, "Key computational modeling issues in integrated computational materials engineering," *Comput.-Aided. Des.* **45**, 4–25 (2013).
- ¹⁷⁹J. Marian, C. S. Becquart, C. Domain, S. L. Dudarev, M. R. Gilbert, R. J. Kurtz, D. R. Mason, K. Nordlund, A. E. Sand, and L. L. Snead, "Recent advances in modeling and simulation of the exposure and response of tungsten to fusion energy conditions," *Nucl. Fusion* **57**, 092008 (2017).
- ¹⁸⁰J. Leite, V. Slowik, and H. Mihashi, "Computer simulation of fracture processes of concrete using mesolevel models of lattice structures," *Cem. Concr. Res.* **34**, 1025–1033 (2004).
- ¹⁸¹C. Camacho-Zuniga and F. Ruiz-Trevino, "A new group contribution scheme to estimate the glass transition temperature for polymers and diluents," *Ind. Eng. Chem. Res.* **42**, 1530–1534 (2003).
- ¹⁸²X.-L. Yu, B. Yi, and X.-Y. Wang, "Prediction of the glass transition temperatures for polymers with artificial neural network," *J. Theor. Comput. Chem.* **07**, 953–963 (2008).
- ¹⁸³A. Sharma, T. Mukhopadhyay, S. M. Rangappa, S. Siengchin, and V. Kushvaha, "Advances in computational intelligence of polymer composite materials: Machine learning assisted modeling, analysis and design," *Arch. Comput. Methods Eng.* **29**, 3341–3385 (2022).
- ¹⁸⁴Q. Tao, P. Xu, M. Li, and W. Lu, "Machine learning for perovskite materials design and discovery," *npj Comput. Mater.* **7**, 23 (2021).
- ¹⁸⁵Y. Wu, J. Guo, R. Sun, and J. Min, "Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells," *npj Comput. Mater.* **6**, 120 (2020).

- ¹⁸⁶C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electron. Markets* **31**, 685–695 (2021).
- ¹⁸⁷H. Rao and A. Mukherjee, "Artificial neural networks for predicting the macro-mechanical behaviour of ceramic-matrix composites," *Comput. Mater. Sci.* **5**, 307–322 (1996).
- ¹⁸⁸L. Chonghe, G. Jin, Q. Pei, C. Ruiliang, and C. Nianyi, "Some regularities of melting points of AB-type intermetallic compounds," *J. Phys. Chem. Solids* **57**, 1797–1802 (1996).
- ¹⁸⁹Z. Zhang, M. W. Beck, D. A. Winkler, B. Huang, W. Sibanda, and H. Goyal, "Opening the black box of neural networks: Methods for interpreting neural network models in clinical applications," *Ann. Transl. Med.* **6**, 216 (2018).
- ¹⁹⁰F.-Y. Tzeng and K.-L. Ma, *Opening the Black Box-Data Driven Visualization of Neural Networks* (IEEE, 2005).
- ¹⁹¹A. F. Smith and G. O. Roberts, "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods," *J. R. Stat. Soc.: Ser. B* **55**, 3–23 (1993).
- ¹⁹²J. J. Hopfield and D. W. Tank, "Computing with neural circuits: A model," *Science* **233**, 625–633 (1986).
- ¹⁹³C. R. Groom and F. H. Allen, "The Cambridge structural database in retrospect and prospect," *Angew. Chem. Int. Ed.* **53**, 662–671 (2014).
- ¹⁹⁴M. Ohring, *Materials Science of Thin Films: Deposition and Structure* (Elsevier, 2001).
- ¹⁹⁵Y. Fu, H. Jiang, N. Xiao, L. Tian, and F. Liu, "Aa-dedupe: An application-aware source deduplication approach for cloud backup services in the personal computing environment," in *2011 IEEE International Conference on Cluster Computing* (IEEE, 2011), pp. 112–120.
- ¹⁹⁶A. Shakarami, M. Ghobaei-Arani, and A. Shahidinejad, "A survey on the computation offloading approaches in mobile edge computing: A machine learning-based perspective," *Comput. Networks* **182**, 107496 (2020).
- ¹⁹⁷M. Ceriotti, C. Clementi, and O. Anatole von Lilienfeld, "Introduction: Machine learning at the atomic scale," *Chem. Rev.* **121**, 9719–9721 (2021).
- ¹⁹⁸M. Meuwly, "Machine learning for chemical reactions," *Chem. Rev.* **121**, 10218–10239 (2021).
- ¹⁹⁹J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, and A. Tkatchenko, "Combining machine learning and computational chemistry for predictive insights into chemical systems," *Chem. Rev.* **121**, 9816–9872 (2021).
- ²⁰⁰X. Li, D. Zhang, Z. Liu, Z. Li, C. Du, and C. Dong, "Materials science: Share corrosion data," *Nature* **527**, 441–442 (2015).
- ²⁰¹T. Grøntoft, A. Verney-Carron, and J. Tidblad, "Cleaning costs for European sheltered white painted steel and modern glass surfaces due to air pollution since the year 2000," *Atmosphere* **10**, 167 (2019).
- ²⁰²Y. Shi, D. Fu, X. Zhou, T. Yang, Y. Zhi, Z. Pei, D. Zhang, and L. Shao, "Data mining to online galvanic current of zinc/copper Internet atmospheric corrosion monitor," *Corros. Sci.* **133**, 443–450 (2018).
- ²⁰³A. Nishikata, Q. Zhu, and E. Tada, "Long-term monitoring of atmospheric corrosion at weathering steel bridges by an electrochemical impedance method," *Corros. Sci.* **87**, 80–88 (2014).
- ²⁰⁴B. Kamsu-Foguem, "Knowledge-based support in non-destructive testing for health monitoring of aircraft structures," *Adv. Eng. Inf.* **26**, 859–869 (2012).
- ²⁰⁵Z. Li, D. Fu, Y. Li, G. Wang, J. Meng, D. Zhang, Z. Yang, G. Ding, and J. Zhao, "Application of an electrical resistance sensor-based automated corrosion monitor in the study of atmospheric corrosion," *Materials* **12**, 1065 (2019).
- ²⁰⁶C. Kleber, U. Hilfrich, and M. Schreiner, "In situ QCM and TM-AFM investigations of the early stages of degradation of silver and copper surfaces," *Appl. Surf. Sci.* **253**, 3712–3721 (2007).
- ²⁰⁷F. Hernandez-Valle, A. Clough, and R. S. Edwards, "Stress corrosion cracking detection using non-contact ultrasonic techniques," *Corros. Sci.* **78**, 335–342 (2014).
- ²⁰⁸A. Legat, "Monitoring of steel corrosion in concrete by electrode arrays and electrical resistance probes," *Electrochim. Acta* **52**, 7590–7598 (2007).
- ²⁰⁹S. Wan, J. Hou, Z.-F. Zhang, X.-x Zhang, and Z.-H. Dong, "Monitoring of atmospheric corrosion and dewing process by interlacing copper electrode sensor," *Corros. Sci.* **150**, 246–257 (2019).
- ²¹⁰C. Thee, L. Hao, J. Dong, X. Mu, X. Wei, X. Li, and W. Ke, "Atmospheric corrosion monitoring of a weathering steel under an electrolyte film in cyclic wet-dry condition," *Corros. Sci.* **78**, 130–137 (2014).
- ²¹¹E. Budevski, W. Obretenov, W. Bostanov, G. Staikov, J. Doneit, K. Jüttner, and W. Lorenz, "Noise analysis in metal deposition—expectations and limits," *Electrochim. Acta* **34**, 1023–1029 (1989).
- ²¹²D. Xia, S. Song, W. Jin, J. Li, Z. Gao, J. Wang, and W. Hu, "Atmospheric corrosion monitoring of field-exposed Q235B and T91 steels in Zhoushan offshore environment using electrochemical probes," *J. Wuhan Univ. Technol.* **32**, 1433–1440 (2017).
- ²¹³D. Law, S. Millard, and J. Bungey, "Linear polarisation resistance measurements using a potentiostatically controlled guard ring," *NDT E Int.* **33**, 15–21 (2000).
- ²¹⁴G. Koch, J. Varney, N. Thompson, O. Moghissi, M. Gould, and J. Payer, *International Measures of Prevention, Application, and Economics of Corrosion Technologies Study* (NACE International, 2016), p. 216.
- ²¹⁵M. Terrados-Cristos, F. Ortega-Fernández, G. Alonso-Iglesias, M. Díaz-Piloneta, and A. Fernández-Iglesias, "Corrosion prediction of weathered galvanized structures using machine learning techniques," *Materials* **14**, 3906 (2021).
- ²¹⁶Z. Pei, D. Zhang, Y. Zhi, T. Yang, L. Jin, D. Fu, X. Cheng, H. A. Terryn, J. M. Mol, and X. Li, "Towards understanding and prediction of atmospheric corrosion of an Fe/Cu corrosion sensor via machine learning," *Corros. Sci.* **170**, 108697 (2020).
- ²¹⁷Q. Li, X. Xia, Z. Pei, X. Cheng, D. Zhang, K. Xiao, J. Wu, and X. Li, "Long-term corrosion monitoring of carbon steels and environmental correlation analysis via the random forest method," *npj Mater. Degrad.* **6**(1), 1 (2022).
- ²¹⁸Q. Ma, G. Tian, Y. Zeng, R. Li, H. Song, Z. Wang, B. Gao, and K. Zeng, "Pipeline in-line inspection method, instrumentation and data management," *Sensors* **21**, 3862 (2021).
- ²¹⁹K. Popova and T. Prošek, "Corrosion monitoring in atmospheric conditions: A review," *Metals* **12**, 171 (2022).
- ²²⁰S. Wang, X. Xia, L. Ye, and B. Yang, "Automatic detection and classification of steel surface defect using deep convolutional neural networks," *Metals* **11**, 388 (2021).
- ²²¹E. D. Kenny, R. S. Paredes, L. A. de Lacerda, Y. C. Sica, G. P. de Souza, and J. Lázaris, "Artificial neural network corrosion modeling for metals in an equatorial climate," *Corros. Sci.* **51**, 2266–2278 (2009).
- ²²²O. Mythreyi, M. R. Srinivas, T. Amit Kumar, and R. Jayaganthan, "Machine-learning-based prediction of corrosion behavior in additively manufactured Inconel 718," *Data* **6**, 80 (2021).
- ²²³N. Purwasih, H. Shinohaki, S. Okazaki, H. Kihira, Y. Kuriyama, and N. Kasai, "Atmospheric corrosion sensor based on strain measurement with active-dummy fiber Bragg grating sensors," *Metals* **10**, 1076 (2020).
- ²²⁴N. Kasai, M. Hiroki, T. Yamada, H. Kihira, K. Matsuoka, Y. Kuriyama, and S. Okazaki, "Atmospheric corrosion sensor based on strain measurement," *Meas. Sci. Technol.* **28**, 015106 (2017).
- ²²⁵N. Purwasih, N. Kasai, S. Okazaki, and H. Kihira, "Development of amplifier circuit by active-dummy method for atmospheric corrosion monitoring in steel based on strain measurement," *Metals* **8**, 5 (2017).
- ²²⁶N. Purwasih, N. Kasai, S. Okazaki, H. Kihira, and Y. Kuriyama, "Atmospheric corrosion sensor based on strain measurement with an active dummy circuit method in experiment with corrosion products," *Metals* **9**, 579 (2019).
- ²²⁷S. Liehr, P. Lenke, M. Wendt, K. Krebber, M. Seeger, E. Thiele, H. Metschies, B. Gebreselassie, and J. C. Munich, "Polymer optical fiber sensors for distributed strain measurement and application in structural health monitoring," *IEEE Sens. J.* **9**, 1330–1338 (2009).
- ²²⁸D. A. Drake, R. W. Sullivan, and J. C. Wilson, "Distributed strain sensing from different optical fiber configurations," *Inventions* **3**, 67 (2018).
- ²²⁹G. Chen, H. Xiao, Y. Huang, Z. Zhou, and Y. Zhang, "A novel long-period fiber grating sensor for large strain measurement," *Proc. SPIE* **7292**, 360–371 (2009).
- ²³⁰J.-N. Wang and J.-L. Tang, "Feasibility of fiber Bragg grating and long-period fiber grating sensors under different environmental conditions," *Sensors* **10**, 10105–10127 (2010).
- ²³¹C. Rodrigues, F. Cavadas, C. Félix, and J. Figueiras, "FBG based strain monitoring in the rehabilitation of a centenary metallic bridge," *Eng. Struct.* **44**, 281–290 (2012).
- ²³²S. Yehia, T. Landolsi, M. Hassan, and M. Hallal, "Monitoring of strain induced by heat of hydration, cyclic and dynamic loads in concrete structures using fiber-optics sensors," *Measurement* **52**, 33–46 (2014).

- ²³³H-l Liu, Z-w Zhu, Y. Zheng, B. Liu, and F. Xiao, "Experimental study on an FBG strain sensor," *Opt. Fiber Technol.* **40**, 144–151 (2018).
- ²³⁴H. Murayama, K. Ohara, N. Kanata, K. Kageyama, and H. Igawa, "Strain monitoring and defect detection in welded joints by using fiber-optic distributed sensors with high spatial resolution," *E-J. Adv. Maint.* **2**, 191–199 (2011).
- ²³⁵M. C. Emmons, S. Karnani, S. Trono, K. P. Mohanchandra, W. L. Richards, and G. P. Carman, "Strain measurement validation of embedded fiber Bragg gratings," *Int. J. Optomechatronics* **4**, 22–33 (2010).
- ²³⁶T. W. Quadri, E. D. Akpan, L. O. Olasunkanni, O. E. Fayemi, and E. E. Ebenso, "Fundamentals of corrosion chemistry," in *Environmentally Sustainable Corrosion Inhibitors* (Elsevier, 2022), pp. 25–45.
- ²³⁷M. J. Jiménez-Come, I. J. Turias, J. J. Ruiz-Aguilar, and F. J. Trujillo, "A comprehensive approach based on SVM to model pitting corrosion behaviour of EN 1.4404 stainless steel," *Mater. Corros.* **65**, 1024–1032 (2014).
- ²³⁸X. Xiong, N. Zhang, J. Yang, T. Chen, and T. Niu, "Machine learning-assisted prediction of corrosion behavior of 7XXX aluminum alloys," *Metals* **14**, 401 (2024).
- ²³⁹T. Walker, O. Adebambo, M. Feijoo, E. Elhaimer, T. Hossain, S. Edwards, C. Morrison, J. Romo, N. Sharma, S. Taylor, and S. Zomorodi, "Environmental effects of marine transportation," in *World Seas: An Environmental Evaluation* (Elsevier, 2021), pp. 505–530.
- ²⁴⁰W. Fan and H. Lin, "Current status and prospects of reliability analysis of hull structures under corrosion and fatigue damage," *Zhongguo Jianchuan Yanjiu* **12**, 52–63 (2017).
- ²⁴¹M. Khayatazad, L. De Pue, and W. De Waele, "Detection of corrosion on steel structures using automated image processing," *Dev. Built Environ.* **3**, 100022 (2020).
- ²⁴²D. Wu and J. R. Birge, "Risk intelligence in big data era: A review and introduction to special issue," *IEEE Trans. Cybern.* **46**, 1718–1720 (2016).
- ²⁴³V. S. Agarwala, P. L. Reed, and S. Ahmad, "Corrosion detection and monitoring-A review," in *CORROSION 2000* (One Petro, 2000), p. NACE-00271.
- ²⁴⁴G. Schmitt, M. Schütze, G. F. Hays, and W. Burns, "Global needs for knowledge dissemination, research, and development in materials deterioration and corrosion control," *World Corros. Org.* **38**, 14 (2009).
- ²⁴⁵A. Sharon and D. Itzhak, "The environmental behaviour of SAE 304L stainless steel sintered in the presence of a copper base additive," *Mater. Sci. Eng., A* **224**, 177–186 (1997).
- ²⁴⁶R. Vashi and H. Kadiya, "Corrosion study of metals in marine environment," *J. Chem.* **6**, 1240–1246 (2009).
- ²⁴⁷Y. Wen, C. Cai, X. Liu, J. Pei, X. Zhu, and T. Xiao, "Corrosion rate prediction of 3C steel under different seawater environment by using support vector regression," *Corros. Sci.* **51**, 349–355 (2009).
- ²⁴⁸N.-L. Tran, T.-H. Nguyen, V.-T. Phan, and D.-D. Nguyen, "A machine learning-based model for predicting atmospheric corrosion rate of carbon steel," *Adv. Mater. Sci. Eng.* **2021**, 1–25.
- ²⁴⁹B. Akhlagh, H. Mesghali, M. Ehteshami, J. Mohammadpour, F. Salehi, and R. Abbassi, "Predictive deep learning for pitting corrosion modeling in buried transmission pipelines," *Process Saf. Environ. Prot.* **174**, 320–327 (2023).
- ²⁵⁰K. N. Sasidhar, N. H. Siboni, J. R. Mianroodi, M. Rohwerder, J. Neugebauer, and D. Raabe, "Deep learning framework for uncovering compositional and environmental contributions to pitting resistance in passivating alloys," *npj Mater. Degrad.* **6**, 71 (2022).
- ²⁵¹Z. Qu, D. Tang, Z. Wang, X. Li, H. Chen, and Y. Lv, "Pitting judgment model based on machine learning and feature optimization methods," *Front. Mater.* **8**, 733813 (2021).
- ²⁵²N.-D. Hoang, "Image processing-based pitting corrosion detection using metaheuristic optimized multilevel image thresholding and machine-learning approaches," *Math. Probl. Eng.* **2020**, 6765274.
- ²⁵³T. A. Alkanhal, "Image processing techniques applied for pitting corrosion analysis," *Entropy* **5**, 2 (2014).
- ²⁵⁴J.-S. Chou, N.-T. Ngo, and W. K. Chong, "The use of artificial intelligence combiners for modeling steel pitting risk and corrosion rate," *Eng. Appl. Artif. Intell.* **65**, 471–483 (2017).
- ²⁵⁵M. Wu, T. Wang, K. Wu, and L. Kan, "Microbiologically induced corrosion of concrete in sewer structures: A review of the mechanisms and phenomena," *Constr. Build. Mater.* **239**, 117813 (2020).
- ²⁵⁶A. S. H. Makhlouf and M. A. Botello, "Failure of the metallic structures due to microbiologically induced corrosion and the techniques for protection," in *Handbook Materials Failure Analysis* (Elsevier, 2018), pp. 1–18.
- ²⁵⁷M. R. Sabour, G. A. Dezvareh, and K. P. Niavol, "Application of artificial intelligence methods in modeling corrosion of cement and sulfur concrete in sewer systems," *Environ. Process.* **8**, 1601–1618 (2021).
- ²⁵⁸Y. Wang, F. Su, Y. Guo, H. Yang, Z. Ye, and L. Wang, "Predicting the microbiologically induced concrete corrosion in sewer based on XGBoost algorithm," *Case Stud. Constr. Mater.* **17**, e01649 (2022).
- ²⁵⁹M. H. Rahman, J. Duckworth, S. Ragi, P. Chundi, V. R. Gadhamshetty, and G. Chilkkoor, "Deep learning approach to extract geometric features of bacterial cells in biofilms," in *Advances in Data Science and Information Engineering: Proceedings from IC DATA 2020 and IKE 2020* (Springer, 2021), pp. 359–368.
- ²⁶⁰V. Bommanapally, M. Ashaduzzman, M. Malshe, P. Chundi, and M. Subramanian, "Self-supervised learning approach to detect corrosion products in biofilm images," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE, 2021), pp. 3555–3561.
- ²⁶¹S. Adumene, S. Adedigba, F. Khan, and S. Zendehboudi, "An integrated dynamic failure assessment model for offshore components under microbiologically influenced corrosion," *Ocean Eng.* **218**, 108082 (2020).
- ²⁶²A.-W. Dawuda, M. Taleb-berrouane, and F. Khan, "A probabilistic model to estimate microbiologically influenced corrosion rate," *Process Saf. Environ. Prot.* **148**, 908–926 (2021).
- ²⁶³E. Hewayne, M. Nehdi*, E. Allouche, and G. Nakhla, "Neural network prediction of concrete degradation by sulphuric acid attack," *Struct. Infrastruct. Eng.* **3**, 17–27 (2007).
- ²⁶⁴I. Baskin and Y. Ein-Eli, "Chemoinformatics for corrosion science: Data-driven modeling of corrosion inhibition by organic molecules," *Mol. Inf.* **43**, e202400082 (2024).
- ²⁶⁵A. A. Khadom, M. S. Mahdi, and H. B. Mahood, "Mathematical regression and artificial neural network for prediction of corrosion inhibition process of steel in acidic media," *J. Bio-Triblo-Corros.* **6**, 92 (2020).
- ²⁶⁶V. C. Anadebe, V. I. Chukwu, C. C. Okoye, L. Guo, and R. C. Barik, "Artificial intelligence and machine learning in corrosion research," in *Handbook of Research on Corrosion Sciences and Engineering* (IGI Global, 2023), pp. 1–23.
- ²⁶⁷E. Betiku, O. R. Omilakin, S. O. Ajala, A. A. Okeleye, A. E. Taiwo, and B. O. Solomon, "Mathematical modeling and process parameters optimization studies by artificial neural network and response surface methodology: A case of non-edible neem (*Azadirachta indica*) seed oil biodiesel synthesis," *Energy* **72**, 266–273 (2014).
- ²⁶⁸T. W. Quadri, L. O. Olasunkanni, O. E. Fayemi, H. Lgaz, O. Dagdag, E.-S. M. Sherif, E. D. Akpan, H.-S. Lee, and E. E. Ebenso, "Predicting protection capacities of pyrimidine-based corrosion inhibitors for mild steel/HCl interface using linear and nonlinear QSAR models," *J. Mol. Model.* **28**, 1–23 (2022).
- ²⁶⁹M. Akrom, S. Rustad, A. G. Saputro, and H. K. Diponjono, "Data-driven investigation to model the corrosion inhibition efficiency of pyrimidine-pyrazole hybrid corrosion inhibitors," *Comput. Theor. Chem.* **1229**, 114307 (2023).
- ²⁷⁰M. Akrom, S. Rustad, and H. K. Diponjono, "Development of quantum machine learning to evaluate the corrosion inhibition capability of pyrimidine compounds," *Mater. Today Commun.* **39**, 108758 (2024).
- ²⁷¹M. Akrom, S. Rustad, A. G. Saputro, A. Ramelan, F. Fathurrahman, and H. K. Diponjono, "A combination of machine learning model and density functional theory method to predict corrosion inhibition performance of new diazine derivative compounds," *Mater. Today Commun.* **35**, 106402 (2023).
- ²⁷²W. Herowati, W. A. E. Prabowo, M. Akrom, T. Sutojo, N. A. Setiyanto, A. W. Kurniawan, N. N. Hidayat, and S. Rustad, "Prediction of corrosion inhibition efficiency based on machine learning for pyrimidine compounds: A comparative study of linear and non-linear algorithms," *KnE Eng.* **6**, 68–77 (2024).
- ²⁷³H. El-Idrissi, A. Diane, M. Driouch, M. Lahyaoui, N. Saffaj, R. Mamouni, B. Ihssane, T. Saffaj, A. Haoudi, and A. Mazzah, "Insights into corrosion inhibition development through QSAR and machine learning: Application to benzimidazole derivatives," *Int. J. Corros. Scale Inhib.* **12**, 2101–2128 (2023).
- ²⁷⁴T. W. Quadri, L. O. Olasunkanni, O. E. Fayemi, E. D. Akpan, H.-S. Lee, H. Lgaz, C. Verma, L. Guo, S. Kaya, and E. E. Ebenso, "Multilayer perceptron neural network-based QSAR models for the assessment and prediction of

- corrosion inhibition performances of ionic liquids," *Comput. Mater. Sci.* **214**, 111753 (2022).
- ²⁷⁵M. Akrom, S. Rustad, and H. K. Dipojono, "Machine learning investigation to predict corrosion inhibition capacity of new amino acid compounds as corrosion inhibitors," *Results Chem.* **6**, 101126 (2023).
- ²⁷⁶S. Akbarzadeh, K. Akbarzadeh, M. Ramezanizadeh, R. Naderi, M. Mahdavian, and M.-G. Olivier, "Corrosion resistance enhancement of a sol-gel coating by incorporation of modified carbon nanotubes: Artificial neural network (ANN) modeling and experimental explorations," *Prog. Org. Coat.* **174**, 107296 (2023).
- ²⁷⁷V. Anadebe, P. Nnaji, O. Onukwuli, N. Okafor, F. Abeng, V. Chukwuike, C. Okoye, I. Udoh, M. Chidiebere, and L. Guo, "Multidimensional insight into the corrosion inhibition of salbutamol drug molecule on mild steel in oilfield acidizing fluid: Experimental and computer aided modeling approach," *J. Mol. Liq.* **349**, 118482 (2022).
- ²⁷⁸F. Gapsari, F. Utaminigrum, C. W. Lai, K. Anam, A. M. Sulaiman, M. F. Haidar, T. S. Julian, and E. E. Ebenso, "A convolutional neural network-VGG16 method for corrosion inhibition of 304SS in sulfuric acid solution by timio leaf extract," *J. Mater. Res. Technol.* **30**, 1116–1127 (2024).
- ²⁷⁹D. I. Uduwana, O. D. Onukwuli, S. C. Nwanonenyi, and C. B. Ezekannaghha, "Novel imidazole based ionic liquid as anti-corrosion additive for aluminum alloy: Combined experimental, DFT/MD simulation and soft computing approach," *Appl. Surf. Sci. Adv.* **19**, 100578 (2024).
- ²⁸⁰F. E. Abeng and I. O. Ewona, "Insights into the corrosion resistance of *Thaumatomoccus daniellii* on carbon steel in simulated sea water: Experimental and machine learning prediction," *Appl. Surf. Sci. Adv.* **18**, 100457 (2023).
- ²⁸¹V. C. Anadebe, V. I. Chukwuike, S. Ramanathan, and R. C. Barik, "Cerium-based metal organic framework (Ce-MOF) as corrosion inhibitor for API 5L X65 steel in CO₂-saturated brine solution: XPS, DFT/MD-simulation, and machine learning model prediction," *Process Saf. Environ. Prot.* **168**, 499–512 (2022).
- ²⁸²O. Sanni, O. Adeleke, K. Ukoba, J. Ren, and T.-C. Jen, "Prediction of inhibition performance of agro-waste extract in simulated acidizing media via machine learning," *Fuel* **356**, 129527 (2024).
- ²⁸³V. C. Anadebe, O. D. Onukwuli, F. E. Abeng, N. A. Okafor, J. O. Ezeugo, and C. C. Okoye, "Electrochemical-kinetics, MD-simulation and multi-input single-output (MISO) modeling using adaptive neuro-fuzzy inference system (ANFIS) prediction for dexamethasone drug as eco-friendly corrosion inhibitor for mild steel in 2 M HCl electrolyte," *J. Taiwan Inst. Chem. Eng.* **115**, 251–265 (2020).
- ²⁸⁴A. Elebo, S. Uba, P. A. Ekwumemgbo, and V. O. Ajibola, "Unravelling adsorption potential of expired clindamycin as corrosion inhibitor at mild Steel/HCl interface via response surface methodology, artificial neural network, electrochemical, and computational strategies," *Curr. Res. Green Sustainable Chem.* **8**, 100402 (2024).
- ²⁸⁵P. C. Nnaji, V. C. Anadebe, C. Agu, I. G. Ezemagu, J. C. Edeh, A. A. Ohanehi, O. D. Onukwuli, and E. E. Eluno, "Statistical computation and artificial neural algorithm modeling for the treatment of dye wastewater using *Mucuna sloanei* as coagulant and study of the generated sludge," *Results Eng.* **19**, 101216 (2023).
- ²⁸⁶O. Onukwuli, P. Nnaji, M. Menkiti, V. Anadebe, E. Oke, C. Ude, C. Ude, and N. Okafor, "Dual-purpose optimization of dye-polluted wastewater decontamination using bio-coagulants from multiple processing techniques via neural intelligence algorithm and response surface methodology," *J. Taiwan Inst. Chem. Eng.* **125**, 372–386 (2021).
- ²⁸⁷M. Salar, M. R. Nikoo, A. Al-Mamun, G. R. Rakhshandehroo, and M. G. Mooselu, "Optimizing Fenton-like process, homogeneous at neutral pH for ciprofloxacin degradation: Comparing RSM-CCD and ANN-GA," *J. Environ. Manage.* **317**, 115469 (2022).
- ²⁸⁸D. I. Uduwana, O. D. Onukwuli, and V. C. Anadebe, "Synthesis and evaluation of 1-butyl-3-methylimidazolium chloride based ionic liquid for acid corrosion inhibition of aluminum alloy: Empirical, DFT/MD-simulation and RSM modeling," *J. Mol. Liq.* **364**, 120019 (2022).
- ²⁸⁹D. I. Uduwana, O. D. Onukwuli, M. C. Menkiti, S. C. Nwanonenyi, C. B. Ezekannaghha, and C. O. Aniagor, "Experimental, computational, and theoretical studies on the corrosion inhibition potential of green *Gongronema latifolium* extract as corrosion inhibitor for aluminum alloy in HCl solutions," *J. Mol. Struct.* **1302**, 137508 (2024).
- ²⁹⁰V. Anadebe, O. Onukwuli, M. Omotomiwa, and N. Okafor, "Experimental, theoretical modeling and optimization of inhibition efficiency of pigeon pea leaf extract as anti-corrosion agent of mild steel in acid environment," *Mater. Chem. Phys.* **233**, 120–132 (2019).
- ²⁹¹S. J. Rigatti, "Random forest," *J. Insur. Med.* **47**, 31–39 (2017).
- ²⁹²Y. Guo, M. Sun, W. Zhang, and L. Wang, "Machine learning in enhancing corrosion resistance of magnesium alloys: A comprehensive review," *Metals* **13**, 1790 (2023).
- ²⁹³M. Aghaaminiha, R. Mehrani, M. Colahan, B. Brown, M. Singer, S. Nesic, S. M. Vargas, and S. Sharma, "Machine learning modeling of time-dependent corrosion rates of carbon steel in presence of corrosion inhibitors," *Corros. Sci.* **193**, 109904 (2021).
- ²⁹⁴T. Liu, Z. Chen, J. Yang, L. Ma, A. Mol, and D. Zhang, "Machine learning assisted discovery of high-efficiency self-healing epoxy coating for corrosion protection," *npj Mater. Degrad.* **8**, 11 (2024).
- ²⁹⁵A. Ogunleye and Q.-G. Wang, "XGBoost model for chronic kidney disease diagnosis," *IEEE/ACM Trans. Comput. Biol. Bioinf.* **17**, 2131–2140 (2020).
- ²⁹⁶A. I. A. Osman, A. N. Ahmed, M. F. Chow, Y. F. Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," *Ain Shams Eng. J.* **12**, 1545–1556 (2021).
- ²⁹⁷B. Wang, Y. Li, X. Cheng, D. Zhang, C. Liu, X. Wang, X. Yong, and X. Li, "Data-driven optimization model customization for atmospheric corrosion on low-alloy steel: Incorporating the dynamic evolution of the surface rust layer," *Corros. Sci.* **221**, 111349 (2023).
- ²⁹⁸C. Qiao, H. Luo, X. Wang, H. Cheng, D. Bi, and X. Li, "Machine learning-based prediction of pitting corrosion resistance in stainless steels exposed to chloride environments," *Colloids Surf., A* **676**, 132274 (2023).
- ²⁹⁹Z. Ye, K. Li, G. Chen, Y. Zou, S. Wang, Z. Wang, S. Yu, Y. Wang, D. Jia, and Y. Zhou, "Accelerated design and fabrication of thermal protection coating via high-throughput experiments and machine learning," *Corros. Sci.* **238**, 112388 (2024).
- ³⁰⁰A. N. Safitri, G. A. Trisnapradika, A. W. Kurniawan, W. A. E. Prabowo, and M. Akrom, "Development of a machine learning model to predict the corrosion inhibition ability of benzimidazole compounds," *J. Multiscale Mater. Inf.* **1**, 16–21 (2024).
- ³⁰¹M. Akrom, S. Rustad, and H. K. Dipojono, "SMILES-based machine learning enables the prediction of corrosion inhibition capacity," *MRS Commun.* **14**, 379–379 (2024).
- ³⁰²N. A. Setiyanto, H. A. Azies, U. Sudibyo, A. Pertiwi, S. Budi, and M. Akrom, "A machine learning model for evaluation of the corrosion inhibition capacity of quinoxaline compounds," *J. Multiscale Mater. Inf.* **1**, 10–15 (2024).
- ³⁰³M. Akrom, S. Rustad, and H. K. Dipojono, "A machine learning approach to predict the efficiency of corrosion inhibition by natural product-based organic inhibitors," *Phys. Scr.* **99**, 036006 (2024).
- ³⁰⁴D. A. Rachman and M. Akrom, "Ensemble learning model in predicting corrosion inhibition capability of pyridazine compounds," *J. Multiscale Mater. Inf.* **1**, 38–43 (2024).
- ³⁰⁵M. Akrom, T. Sutojo, A. Pertiwi, S. Rustad, and H. K. Dipojono, "Investigation of best QSPR-based machine learning model to predict corrosion inhibition performance of pyridine-quinoline compounds," *J. Phys. Conf. Ser.* **2673**, 012014 (2023).
- ³⁰⁶M. Akrom, S. Rustad, and H. K. Dipojono, "Prediction of anti-corrosion performance of new triazole derivatives via machine learning," *Comput. Theor. Chem.* **1236**, 114599 (2024).
- ³⁰⁷J. Fang, X. Cheng, H. Gai, S. Lin, and H. Lou, "Development of machine learning algorithms for predicting internal corrosion of crude oil and natural gas pipelines," *Comput. Chem. Eng.* **177**, 108358 (2023).
- ³⁰⁸L. Li, X. Zhang, S. Gong, H. Zhao, Y. Bai, Q. Li, and L. Ji, "The discussion of descriptors for the QSAR model and molecular dynamics simulation of benzimidazole derivatives as corrosion inhibitors," *Corros. Sci.* **99**, 76–88 (2015).
- ³⁰⁹K. Khaled, "Modeling corrosion inhibition of iron in acid medium by genetic function approximation method: A QSAR model," *Corros. Sci.* **53**, 3457–3465 (2011).
- ³¹⁰K. F. Khaled and N. S. Abdel-Shafie, "Quantitative structure and activity relationship modeling study of corrosion inhibitors: Genetic function

- approximation and molecular dynamics simulation methods," *Int. J. Electrochem. Sci.* **6**, 4077–4094 (2011).
- ³¹¹K. F. Khaled and A. M. El-Sherik, "Using molecular dynamics simulations and genetic function approximation to model corrosion inhibition of iron in chloride solutions," *Int. J. Electrochem. Sci.* **8**, 10022–10043 (2013).
- ³¹²K. F. Khaled and N. Al-Nofai, "N. Abdel-Shafi, QSAR of corrosion inhibitors by genetic function approximation, neural network and molecular dynamics simulation methods," *J. Mater. Environ. Sci.* **7**, 2121–2136 (2016).
- ³¹³A. Elsammam, K. Khaled, S. A. Halim, and N. Abdelshafi, "A critical view of the QSAR model for the prediction of a new bispyrazole derivative BPYR-P as a corrosion inhibitor for 304 SS in a 1.0 M HCl solution," *J. Mol. Struct.* **1297**, 136728 (2024).
- ³¹⁴M. Akrom, S. Rustad, and H. K. Dipojono, "Variational quantum circuit-based quantum machine learning approach for predicting corrosion inhibition efficiency of pyridine-quinoline compounds," *Mater. Today Quantum* **2**, 100007 (2024).
- ³¹⁵M. Akrom, S. Rustad, H. K. Dipojono, and R. Maezono, "A comprehensive approach utilizing quantum machine learning in the study of corrosion inhibition on quinoxaline compounds," *Artif. Intell. Chem.* **2**, 100073 (2024).
- ³¹⁶L. Xu, Y. Wang, L. Mo, Y. Tang, F. Wang, and C. Li, "The research progress and prospect of data mining methods on corrosion prediction of oil and gas pipelines," *Eng. Fail. Anal.* **144**, 106951 (2023).
- ³¹⁷L. Xu, J. Yu, Z. Zhu, J. Man, P. Yu, C. Li, X. Wang, and Y. Zhao, "Research and application for corrosion rate prediction of natural gas pipelines based on a novel hybrid machine learning approach," *Coatings* **13**, 856 (2023).
- ³¹⁸J. Hu, "Prediction of the internal corrosion rate for oil and gas pipelines and influence factor analysis with interpretable ensemble learning," *Int. J. Pressure Vessels Piping* **212**, 105329 (2024).
- ³¹⁹Y. Cai, Y. Xu, Y. Zhao, K. Zhou, and X. Ma, "A spatial-temporal approach for corrosion prediction in time-varying marine environment," *J. Loss Prev. Process Ind.* **66**, 104161 (2020).
- ³²⁰D. Winkler, M. Breedon, P. White, A. Hughes, E. Sapper, and I. Cole, "Using high throughput experimental data and *in silico* models to discover alternatives to toxic chromate corrosion inhibitors," *Corros. Sci.* **106**, 229–235 (2016).
- ³²¹D. A. Winkler, M. Breedon, A. E. Hughes, F. R. Burden, A. S. Barnard, T. G. Harvey, and I. Cole, "Towards chromate-free corrosion inhibitors: Structure-property models for organic alternatives," *Green Chem.* **16**, 3349–3357 (2014).
- ³²²K. Khan, M. Iqbal, F. E. Jalal, M. N. Amin, M. W. Alam, and A. Bardhan, "Hybrid ANN models for durability of GFRP rebars in alkaline concrete environment using three swarm-based optimization algorithms," *Constr. Build. Mater.* **352**, 128862 (2022).
- ³²³A. Rajender, A. K. Samanta, and A. Paral, "Comparative study of corrosion-based service life prediction of reinforced concrete structures using traditional and machine learning approach," *Int. J. Struct. Integr.* (published online).
- ³²⁴H. S. Munawar, F. Ullah, D. Shahzad, A. Heravi, S. Qayyum, and J. Akram, "Civil infrastructure damage and corrosion detection: An application of machine learning," *Buildings* **12**, 156 (2022).
- ³²⁵R. S. Iyer, N. S. Iyer, and A. Joseph, "Harnessing machine learning and virtual sample generation for corrosion studies of 2-alkyl benzimidazole scaffold small dataset with an experimental validation," *J. Mol. Struct.* **1306**, 137767 (2024).
- ³²⁶T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2009).
- ³²⁷A. Roy, M. F. N. Taufique, H. Khakurel, R. Devanathan, D. D. Johnson, and G. Balasubramanian, "Machine-learning-guided descriptor selection for predicting corrosion resistance in multi-principal element alloys," *npj Mater. Degrad.* **6**, 9 (2022).
- ³²⁸G. Vishwakarma, A. Sonpal, and J. Hachmann, "Metrics for benchmarking and uncertainty quantification: Quality, applicability, and best practices for machine learning in chemistry," *Trends Chem.* **3**, 146–156 (2021).
- ³²⁹D. Ruiz, A. Casas, C. A. Escobar, A. Perez, and V. Gonzalez, "Advanced machine learning techniques for corrosion rate estimation and prediction in industrial cooling water pipelines," *Sensors* **24**, 3564 (2024).