Theoretical Chemistry Accounts (2024) 143:65
https://doi.org/10.1007/s00214-024-03140-x

RESEARCH

Machine learning for pyrimidine corrosion inhibitor small dataset

Wise Herowati[1] · Wahyu Aji Eko Prabowo[1] · Muhamad Akrom[1] · Noor Ageng Setiyanto[1] · Achmad Wahid Kurniawan[1] · Novianto Nur Hidayat[1] · Totok Sutojo[1] · Supriadi Rustad[1]

Received: 25 June 2024 / Accepted: 31 July 2024 / Published online: 9 August 2024

Abstract
Machine learning (ML) approaches have been developed to predict materials' corrosion inhibition efficiency, particularly pyrimidine compounds. Notably, the virtual sample generation (VSG) technique enhances prediction accuracy, a novel approach for handling small datasets in this context. The random forest model, the best-performing nonlinear algorithm, showed substantial accuracy improvement based on the increase in R^2 value from 0.05 to 0.99 and the decrease in RMSE value from 5.60 to 0.42, after applying VSG. These results underscore the efficacy of the VSG technique in boosting the predictive performance of ML models, particularly in scenarios constrained by limited data availability.

Keywords Corrosion · Pyrimidine · Machine Learning · Virtual sample generation

1 Introduction

Corrosion, an omnipresent electrochemical phenomenon, exerts a substantial influence on diverse sectors through the degradation of metallic materials employed in infrastructure and machinery. This degradation not only leads to significant economic losses but also presents safety risks, particularly in sectors such as petrochemicals, construction, and transportation [1–4]. The imperative to tackle this matter has prompted the formulation and execution of diverse strategies for corrosion control.

Corrosion inhibitors have been recognized as a pivotal element in combating the deterioration of metals within these strategies [5–7] These inhibitors operate by creating a protective layer on the surface of the metal, thereby diminishing the corrosion rate and prolonging the material's durability [7–9] The careful choice and efficacy of these inhibitors are crucial, as they have a substantial impact on guaranteeing the longevity and security of metal-based structures and equipment [10–12].

Pyrimidines have demonstrated potential as effective and safe corrosion inhibitors for carbon steel in acidic environments, offering a viable alternative to currently employed toxic chemicals [13, 14]. A comprehensive review has recently been published on using pyrimidines as corrosion inhibitors for carbon steels in acidic environments [15–17]. The results indicate that pyrimidines can act as effective corrosion inhibitors, showing a high level of inhibition efficiency. These studies underscore the potential of pyrimidines in providing efficient corrosion protection, while minimizing environmental and health hazards. Despite these advancements, the conventional methodology employed in experimental research to ascertain and authenticate novel corrosion inhibitors frequently demands substantial resources and time. The expense and labor associated with these experimental methodologies underscore the necessity for more streamlined alternatives [18, 19].

Machine learning (ML), a constituent of artificial intelligence (AI), offers novel approaches to address diverse challenges, such as forecasting corrosion inhibition effectiveness. The utilization of ML in this domain holds great potential due to its capacity to effectively analyze and model extensive datasets. This capability can significantly speed up identifying and assessing novel corrosion inhibitors [20, 21]. Recent developments in ML, particularly in quantitative structure-properties relationships (QSPR), have created

✉ Wahyu Aji Eko Prabowo
prabowo@dsn.dinus.ac.id

✉ Muhamad Akrom
m.akrom@dsn.dinus.ac.id

✉ Supriadi Rustad
srustad@dsn.dinus.ac.id

[1] Research Center for Quantum Computing and Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia

novel opportunities for investigating materials. Through QSPR models, researchers can forecast the characteristics of chemical compounds by analyzing their molecular structures, greatly expediting the screening and detection of potential corrosion inhibitors and enhancing efficiency in this field [22–24].

The application of both linear algorithms, including multiple linear regression (MLR), ridge regression, and lasso regression, as well as nonlinear algorithms, such as support vector machines (SVM), k-nearest neighbors (KNN), and random forests (RF), within ML models, has shown significant achievements in forecasting the effectiveness of diverse compounds [25, 26]. These models have demonstrated their ability to predict corrosion inhibition efficiency (CIE) with considerable accuracy, thus facilitating the discovery of new inhibitors.

Notwithstanding these technological advancements, a notable obstacle in ML is achieving high levels of model accuracy [27–29].The correlation between ML model accuracy and their effectiveness in predicting the CIE of compounds is evident. Inaccuracies in predictions can result in incorrect conclusions and potential hazards when implementing these substances in real-life situations.

A notable challenge in ML for corrosion research is the limited availability of comprehensive datasets. Small or unbalanced datasets can significantly hinder the accuracy of ML models, leading to unreliable predictions. When datasets are limited, models may struggle to capture the underlying patterns and relationships, resulting in overfitting and poor generalization of new data. This limitation underscores the need for techniques that can enhance data quality and quantity to improve model performance.

A major challenge in ML for corrosion research is the limited availability of comprehensive datasets. Small or imbalanced datasets can significantly hamper the accuracy of ML models, leading to unreliable predictions. When datasets are limited, models may struggle to capture underlying patterns and relationships, resulting in overfitting and poor generalization of new data. These limitations underscore the need for techniques that can improve data quality and quantity to improve model performance. Several researchers have addressed the issue of small corrosion datasets using various virtual sample generation (VSG) techniques. VSG is an innovative approach in ML, where synthetic data samples are generated to augment the existing dataset [30–32]. Their results indicate that VSG techniques are capable of improving the accuracy of predictive models [25, 32, and 33]. For example, Akrom et al. [25] utilized VSG to balance their dataset, resulting in improved model robustness and reduced prediction errors. Similarly, Sutojo et al. [32] showed that synthetic data generated through VSG significantly improved the performance of corrosion prediction models by providing additional training samples that better represented the underlying data distribution. These studies underscore the potential of VSG techniques in overcoming data limitations and improving the reliability of ML models in corrosion research.

This study incorporates the VSG technique. This augmentation helps overcome the limitations of small or unbalanced datasets, common issues in corrosion research. By enhancing the dataset with virtually generated samples, the study aims to improve the precision and reliability of ML models. The application of VSG in predicting the CIE of pyrimidine compounds represents a significant stride in combining computational methods with corrosion science, potentially leading to more efficient and effective solutions in the ongoing battle against material degradation. Our research fills the gap by enhancing data quality and model precision, contributing to the development of more accurate and reliable predictions for corrosion inhibition efficiency.

Despite the significant progress in using ML techniques for predicting CIE, current research often struggles with the limitations posed by small or unbalanced datasets, leading to inaccuracies in model predictions. This gap highlights the need for methods to enhance data quality and model precision. Our research addresses this gap by incorporating the VSG technique, which augments the dataset with synthetic samples. By doing so, we aim to improve the accuracy and reliability of ML models in predicting the CIE of pyrimidine compounds. This approach not only enhances the predictive power of the models but also contributes to more efficient and effective identification of potential corrosion inhibitors, thereby advancing the field of corrosion science and offering practical solutions to combat material degradation.

2 Method

2.1 Virtual sample generation (VSG)

VSG is a sophisticated technique employed in machine learning to enhance existing datasets by adding synthetic data points [34–37]. This approach is especially advantageous when working with limited datasets often found in niche areas such as corrosion science [38]. VSG entails generating novel samples with statistical similarity to the original data, while not being exact replicas. This process increases the variety and quantity of the dataset, allowing machine learning models to acquire more generalized patterns and minimize overfitting.

This study uses VSG to tackle the small dataset size of 54 pyrimidine compounds. The method is employed to create extra simulated samples that replicate the qualities of the original compounds, while incorporating slight modifications. The variations are carefully regulated to maintain the realism and chemical plausibility of the virtual samples.

The generation of virtual samples entails the manipulation of the chemical descriptors of the pyrimidine compounds. New compounds are synthesized in a virtual environment by manipulating parameters such as molecular orbital energies, dipole moments, and electronegativity. The virtual compounds undergo the same descriptor calculation process as the original compounds, ensuring dataset consistency.

After being generated, the virtual samples are incorporated into the current dataset, augmenting the sample size. The integration process is executed meticulously to preserve the dataset's equilibrium and integrity. The expanded dataset is subsequently utilized to train machine learning models to achieve more precise and resilient CIE predictions.

KDE's main objective is to create virtual samples to assess how well the ML model predicts the inhibitor chemicals under test would control corrosion. It is assumed that the kernel density estimator $k(x)$ suggested by Rosenblatt [39–41] as Eq. (1). The parameters n and h represent the number of samples and the bandwidth/smoothing factor, respectively, and K is a kernel function that meets the necessary conditions to satisfy the condition $\int K(x)d(x) = 1$. This estimate, which uses the distance x-xi, determines the impact of computing the density of point x at xi.

$$k(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

2.2 Dataset

In this study, the pyrimidine dataset used was adopted from the literature of Alamri et al. [13]. The dataset consists of 54 pyrimidine compounds with 14 descriptors and 1 target. These descriptors are used as features and CIE values as targets. The use of the VSG technique in this study generates the number of training data by 1000 data points addition. Table 1 shows the details of the descriptors used.

The corrosion inhibition process is closely dependent on chemical reactivity processes and inhibitor molecules represented in various chemical reactivity descriptors [42, 43]. In the QSPR context, the expected relationship between the descriptors and the CIE values is that the descriptors will significantly influence the ability of inhibitor compounds to inhibit corrosion. These descriptors, which typically include molecular structure attributes, electronic properties, and steric factors, were selected because they directly impact the interaction between the inhibitor molecules and the metal surface. For instance, molecular descriptors such as HOMO and LUMO energies can indicate the electron-donating and electron-accepting capabilities of the compounds, which are crucial for forming protective layers on the metal surface. Descriptors related to molecular size and shape, such as molecular volume and surface area, influence how well the inhibitor molecules can cover and protect the metal surface. Additionally, descriptors reflecting the hydrophobicity or hydrophilicity of the molecules, such as logP (partition coefficient), affect the solubility and distribution of the inhibitors in the corrosive environment.

Table 1 Descriptors of chemical reactivity of pyrimidine compounds

Descriptor	Description
E_{HOMO} (eV)	Highest occupied molecular orbital energy
E_{LUMO} (eV)	Lowest unoccupied molecular orbital energy
$E_{L\text{-}H}$ (eV)	Energy gap
μ (D)	Dipole moment
IP (eV)	Ionization potential
EA (eV)	Electron affinity
χ (eV)	Absolute electronegativity
η (eV)	Hardness
σ (eV^{-1})	Softness
ΔN	The fraction of electrons shared
ω (eV)	Electrophilicity index
Log P	The logarithm of the partition coefficient
M ($g.mol^{-1}$)	Molecular mass
Vm (cm^3/mol)	Molecular volume

2.3 Machine learning model development

In this study, after dataset preparation, thorough preprocessing procedures, such as missing value checking, outlier handling, normalization, and data partitioning are performed to maintain data integrity and create a uniform framework, which supports efficient analytical processes and ensures that ML models operate efficiently and produce reliable predictions.

The data was normalized using the MinMax-scaling [44] technique to ensure all features were on the same scale. MinMax-scaling transforms features by scaling each feature to the range (0, 1). This technique is important to prevent features with a larger range of values from dominating the model and to ensure each feature contributes equally to the analysis.

The next stage is data division. The dataset is divided into training and testing subsets. This division is done with a certain proportion, such as 80% for training and 20% for testing, to ensure that the model can be evaluated with data that was not seen during training. During the training process, to enhance the precision of our models and strengthen their predictive accuracy, we have incorporated the VSG technique. This approach played a crucial role in increasing the dataset creating a more extensive training environment for the models. During the training process, to improve the precision of our model and strengthen its predictive accuracy, we have

incorporated the KDE-VSG technique. This approach plays a vital role in enhancing the dataset which creates a broader training environment for the model. Additionally, during training, the K-Fold cross-validation technique [45, 46] is used with five folds. This method is crucial in reducing the potential bias and variance issues, which are often encountered in machine learning models.

A wide variety of ML algorithms [47, 48], including linear and nonlinear models, were chosen for this study. This selection aims to assess and compare the efficacy of various algorithmic strategies in predicting corrosion inhibition efficiency. For linear models, these models are chosen for their ability to handle various aspects of linear relationships within the data. The linear models used in this study are linear regression, ARD regression (automatic relevance determination), Bayesian ridge, elastic net, gamma regression, Huber regressor, orthogonal matching pursuit (OMP), passive aggressive regressor, Poisson regressor, RANSAC regressor (random sample consensus), ridge regression, SGD Regressor (stochastic gradient descent), Theil-Sen regressor, tweedie regressor. meanwhile, for nonlinear models, the models used include Adaboost regressor, bagging regressor, gradient boosting regressor, random forest regressor, PLS regression (partial least squares), decision tree regressor, extra tree regressor, kneighbors regressor, XGB regressor (extreme gradient boosting). Following the training phase, each model underwent meticulous optimization and fine-tuning of hyperparameters. This step was pivotal in improving the model's performance and guaranteeing its resilience.

The effectiveness of each model was evaluated using essential performance metrics: the coefficient of determination (R^2), root mean square error (RMSE), mean square error (MSE), and mean absolute error (MAD). These metrics were calculated for both the training and test datasets. These metrics were selected based on their capacity to indicate the model's predictive capabilities precisely. An ideal model would exhibit an R^2 value close to one and a lower RMSE, indicating high accuracy and reliability in predicting the corrosion inhibition efficiency of pyrimidine compounds. Eq. (2–5) are the formulation of metrics, where n represents the number of observations or samples, and x and y are variables representing the independent and dependent data features, respectively. Y_i is the observation value, $\widehat{Y}_i$ is the predictive value.

$$R^2 = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}} \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2\right)} \tag{3}$$

$$\mathrm{MSE} = \frac{1}{n}\left(\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2\right) \tag{4}$$

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i\prime - \overline{Y}_i\right| \tag{5}$$

3 Results and discussion

Tables 2 and 3 showcase model performances before and after the addition of virtual samples for several linear and nonlinear models.

Table 2 presents the performance of 14 linear regression models for predicting the effectiveness of corrosion inhibitors, both before and after the application of VSG. The results indicate that VSG generally enhances model performance, evidenced by a decrease in RMSE values and an increase in R^2 values. This improvement can be attributed to the expanded and more balanced dataset provided by VSG, which helps the models better capture the underlying patterns in the data. Linear regression showed a significant improvement post-VSG. The substantial increase in the R^2 value from − 0.62 to 0.31 indicates that the linear regression model benefited from the augmented dataset, which provided a better representation of the corrosion inhibitors' performance. This improvement suggests that VSG effectively mitigated the overfitting issues typically associated with small datasets by providing a more comprehensive dataset for training. Both ARD Regression and Bayesian Ridge Regression displayed similar improvements. These

Table 2 Linear model performances based on testing set

No	Models	Before VSG		After VSG	
		RMSE	R^2	RMSE	R^2
1	Linear Regression	7.32	− 0.62	4.40	0.31
2	ARD Regression	5.83	− 0.03	4.42	0,29
3	Bayesian Ridge	5.83	− 0.03	4.67	0.29
4	Elastic Net	5.84	− 0.03	4.67	0.20
5	Gamma Regression	5.82	− 0.03	4.60	0.23
6	Huber Regressor	7.41	− 0.66	4.50	0.26
7	Orthogonal Matching Pursuit	5.85	− 0.04	4.72	0.19
8	Passive Aggressive Regressor	6.56	− 0.30	5.70	-0.14
9	Poisson Regressor	5.84	− 0.03	4.58	0.30
10	RANSAC Regressor	7.57	− 0.73	5.93	-0.30
11	Rigde	5.83	− 0.03	4.53	0.21
12	SGD Regressor	6.48	− 0.27	4.33	0.26
13	Theil-Sen Regressor	–	–	4.50	0.24
14	Tweedie Regressor	5.83	− 0.03	4.51	0.20

Table 3 Nonlinear model performances based on testing set

No	Models	Before VSG		After VSG	
		RMSE	R^2	RMSE	R^2
1	AdaBoost regressor	5.67	0.03	2.61	0.76
2	Bagging regressor	5.78	−0.01	0.44	0.99
3	Gradient boosting regressor	5.66	0.03	1.05	0.96
4	Random forest regressor	5.60	0.05	0.42	0.99
5	PLS regression	5.97	−0.08	4.54	0.25
6	Decision tree regressor	7.35	−0.63	0.97	0.97
7	Extra tree regressor	–	–	0.70	0.98
8	KNeighbors regressor	5.86	−0.04	1.92	0.86
9	XGB regressor	5.89	−0.05	0.49	0.99

algorithms apply regularization, which helps in managing overfitting and enhancing generalization. The improvement in R^2 values post-VSG signifies that the additional synthetic samples provided by VSG allowed these models to better learn the relationships within the data, reducing prediction errors and improving accuracy. Both elastic net and ridge regression models incorporate regularization techniques, which help to handle multicollinearity and prevent overfitting. The performance improvements post-VSG reflect the effectiveness of these regularization methods in leveraging the enhanced dataset to produce more accurate predictions.

Despite the general trend of improvement, some models like the passive aggressive regressor and RANSAC regressor continued to exhibit negative R^2 values even after VSG. These models are designed to handle specific types of data anomalies (e.g., Passive aggressive regressor for online learning and RANSAC for robust fitting). The persistence of negative R^2 values suggests that these models may not be well-suited to this particular dataset or that further tuning is necessary. The limited improvement implies that, while VSG aids in data augmentation, the fundamental assumptions and characteristics of these models might not align well with the dataset's nature.

The application of VSG generally improves the predictive performance of linear regression models by addressing the limitations of small datasets. The significant enhancements observed in most models post-VSG underscore the potential of synthetic data augmentation in overcoming data constraints and improving model accuracy. However, the continued poor performance of certain models indicates that further model-specific adjustments or alternative modeling approaches might be required for optimal results. Overall, these findings highlight the value of VSG in enhancing the predictive capabilities of machine learning models in the field of corrosion inhibition.

Table 3 provides a detailed comparison of the performance of various nonlinear regression models before and after the application of VSG. The results illustrate the impact of VSG on improving the models' predictive accuracy by providing additional synthetic data samples.

AdaBoost works by combining multiple weak learners to create a strong learner. The improvement in performance post-VSG indicates that the increased dataset allowed AdaBoost to better capture the nuances in the data, thereby improving the overall accuracy and predictive capability of the ensemble model. Bagging reduces variance by training multiple instances of a base estimator on different subsets of the data. The significant improvement after VSG suggests that the augmented dataset enabled the model to better generalize, capturing the underlying data distribution more effectively and reducing prediction error dramatically. Gradient boosting works by sequentially building an ensemble of models that correct the errors of the previous models. The improvement in performance after VSG indicates that the enriched dataset provided better insights for the model to learn from, leading to a significant reduction in RMSE and a corresponding increase in R^2. Random Forest, an ensemble of decision trees, benefits from the reduction of variance and the ability to handle a larger variety of patterns. The dramatic improvement in performance after VSG highlights how the additional data samples helped the random forest model to better capture the complexities of the dataset, resulting in near-perfect predictions. PLS Regression is a linear model that tries to find the fundamental relations between two matrices. The moderate improvement after VSG suggests that, while the additional data helped, PLS Regression may still face limitations in capturing nonlinear relationships compared to other nonlinear models. Decision trees can easily overfit on small datasets but perform better with more comprehensive datasets. The significant improvement post-VSG indicates that the additional synthetic samples helped the model generalize better and avoid overfitting. Extra trees, like random forests, reduce variance through multiple random splits. The improvement after VSG indicates that the augmented dataset helped the model to better identify patterns and relationships within the data, leading to a significant enhancement in predictive accuracy. KNeighbors regressor predicts the target by averaging the targets of the k-nearest neighbors. The improvement post-VSG suggests that the increased dataset provided a better distribution of samples, allowing for more accurate neighbor-based predictions. XGBoost is an efficient and scalable implementation of gradient boosting. The substantial improvement after VSG highlights the model's ability to leverage the augmented dataset to enhance predictive performance, achieving near-perfect accuracy significantly.

The results clearly show that the application of VSG substantially improves the performance of nonlinear regression models. The additional synthetic data samples provided by VSG help to overcome the limitations of small or unbalanced datasets, enabling the models to better capture the

underlying patterns and relationships within the data. This leads to significant reductions in RMSE and corresponding increases in R^2 values, indicating enhanced predictive accuracy and reliability. The findings underscore the value of VSG in augmenting datasets and improving the performance of ML models in predicting the effectiveness of corrosion inhibitors.

Table 4 presents the performance of the three most optimal linear and nonlinear models before and after the application of VSG. For linear model, linear regression, a simple yet powerful model, showed significant improvement after VSG. The increase in R^2 value from − 0.62 to 0.31 indicates that the augmented dataset provided by VSG helped the model better capture the underlying linear relationships in the data, reducing the prediction error substantially. The Huber regressor, which is robust to outliers, displayed notable improvements post-VSG. The model's ability to handle outliers effectively, combined with the enriched dataset, contributed to better generalization and increased predictive accuracy, as reflected by the improved R^2 value. Theil-Sen regressor, known for its robustness to multicollinearity and outliers, also showed improved performance after VSG. The synthetic samples provided by VSG helped the model to capture more accurate trends in the data, leading to a significant reduction in RMSE and an increase in R^2 value.

For nonlinear models, random forest regressor, an ensemble learning method, showed exceptional improvement after VSG. The model benefits from the diversity of trees and the ability to reduce overfitting, which, combined with the enhanced dataset, resulted in near-perfect predictions as evidenced by the significant increase in R^2 to 0.99. The bagging regressor, which creates multiple versions of a model and aggregates their predictions, showed remarkable performance improvement post-VSG. The augmented dataset provided a more comprehensive training set for each base model, significantly reducing variance and improving predictive accuracy. Decision tree regressor, which splits the data into subsets based on feature values, benefited greatly from VSG. The additional synthetic data samples allowed the model to avoid overfitting and capture more accurate splits, resulting in a drastic reduction in RMSE and a substantial increase in R^2 value. The application of VSG significantly enhanced the performance of both linear and nonlinear models by addressing the limitations of small and unbalanced datasets. The enriched dataset allowed the models to better capture the underlying patterns and relationships, resulting in substantial improvements in predictive accuracy. The findings underscore the value of synthetic data augmentation in improving the robustness and reliability of ML models in predicting the effectiveness of corrosion inhibitors.

Table 4 The three most optimal linear and nonlinear models based on testing set

Model	Before VSG		After VSG	
	RMSE	R^2	RMSE	R^2
Linear				
Linear regression	7.32	− 0.62	4.40	0.31
Huber regressor	7.41	− 0.66	4.50	0.26
Theil-Sen regressor	7.18	− 0.56	4.50	0.24
Nonlinear				
Random forest	5.60	0.05	0.42	0.99
Bagging regressor	5.78	− 0.01	0.44	0.99
Decision tree	7.35	− 0.63	0.97	0.97

The findings of the study indicate that the nonlinear model is more relevant to the evaluated pyrimidine compound dataset. Although the application of VSG also improves the accuracy of the linear model, its ability to capture patterns in the tested dataset is still less than optimal. In contrast, the nonlinear model tends to be relevant to this dataset, as evidenced by its superior prediction accuracy. This claim is supported by the scatterplot of predicted data points in Fig. 1, where the nonlinear model shows a distribution of data points that is close to the predicted line, compared to the pattern produced by the linear model. Figure 1 illustrates the scatterplot of actual values versus predicted values for the three most optimal linear and nonlinear models after applying VSG. The random forest model is the superior model, corresponding to a high level of accuracy with an R^2 of 0.99 and an RMSE of 0.42. The enriched dataset allows the random forest model to generalize better, resulting in highly accurate predictions. This model shows remarkable improvement after the application of VSG. Almost, all predicted values are parallel to the predicted line, indicating very high accuracy. Data augmentation from VSG enabled the model to generalize better and capture complex patterns in the data.

Overall, both linear and nonlinear models benefit from the augmented dataset, allowing them to better capture the underlying patterns and relationships in the data, although the application of VSG improved the accuracy of the linear model, the nonlinear model proved to be more relevant, and effective in capturing the complex patterns in the tested pyrimidine compound dataset. Data augmentation with VSG provided a significant improvement in the predictive capability of the nonlinear model, showing great potential in corrosion research applications. This improvement was evident in the closer alignment of predicted values with actual values, reduced RMSE, and increased R^2 values. These findings underscore the effectiveness of VSG in enhancing the predictive capability of models, particularly nonlinear ones in corrosion inhibition research.

In addition, in the context of evaluating ML models, the R^2 value is a measure of how well the model's predictions

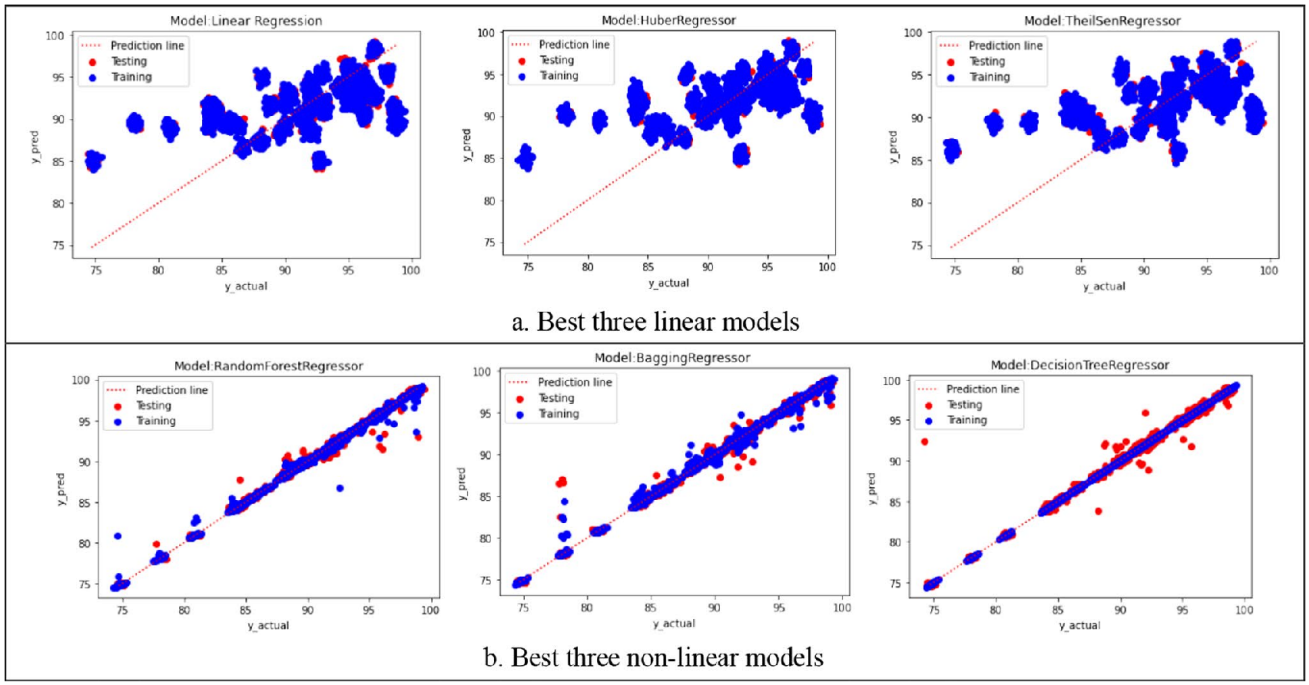

Fig. 1 Scatter plot of best three **a** linear and **b** nonlinear models

match the actual data. An R^2 value of 1 indicates perfect predictions, while an R^2 value of 0 suggests that the model is no better than a simple mean of the observed data. Negative R^2 values, as seen in some models before the application of VSG, indicate that the model's predictions are worse than the mean prediction, which can occur due to several reasons related to the characteristics of the dataset and the performance of the inhibitors against corrosion. Before VSG, the dataset was too small or unbalanced, lacking sufficient representation of the variability in the performance of corrosion inhibitors. This limitation makes it challenging for models to learn meaningful patterns, leading to poor prediction accuracy and consequently negative R^2 values. Corrosion inhibition is influenced by various factors including the chemical structure of the inhibitors, environmental conditions, and the type of metal. The intricate interactions and dependencies among these factors may not be adequately captured by the limited dataset, resulting in models that fail to generalize well and produce suboptimal predictions. With small datasets, several model can easily overfit, capturing noise rather than the underlying trends. Overfitting leads to excellent performance on the training data but poor performance on unseen data, which contributes to negative R^2 values. The performance of corrosion inhibitors can vary significantly under different conditions. If the dataset does not capture this variability adequately, models may struggle to predict the CIE accurately, leading to negative R^2 values. In summary, Negative R^2 values before applying VSG indicate the limitations of the original dataset and the complexity of accurately predicting corrosion inhibition efficiency. The application of VSG has proven to be an effective solution, significantly improving the dataset's quality and balance, leading to substantial improvements in model performance. This justifies the use of VSG techniques to overcome data limitations and enhance the predictive accuracy of machine learning models in corrosion research.

From Table 5 and Fig. 2, the random forest model showed high performance on the training dataset but significantly lower performance on the test dataset. This indicates overfitting, where the model learns the training data too well but fails to generalize to new, unseen data. Similar to random forest, the bagging model performed well on the training dataset but poorly on the test dataset, again indicating overfitting. The decision tree model exhibited the most significant overfitting, with very high performance on the training data and poor performance on the test data. After applying VSG, the random forest model improved significantly on the test dataset, achieving a good fit with high R^2 and low RMSE values, indicating it can generalize well. The Bagging model also showed substantial improvement, indicating that VSG helped mitigate the overfitting issue, resulting in a good fit. The decision tree model, although less accurate than random forest and Bagging, demonstrated a significant improvement with VSG, indicating better generalization compared to the initial overfitted model.

Before applying VSG, the models exhibited a tendency towards overfitting, where they performed exceptionally well on the training data but poorly on the test data. This

Table 5 Best three model performances before and after the addition of VSG

Model	Before VSG							
	Training				Testing			
	R^2	RMSE	MSE	MAD	R^2	RMSE	MSE	MAD
Random Forest	0.98	0.90	0.81	0.64	0.05	5.6	31.36	3.96
Bagging	0.98	0.88	0.77	0.62	−0.01	5.78	33.41	4.09
Decision Tree	0.99	1.1	1.21	0.78	−0.63	7.35	54.02	5.20
Model	Before VSG							
	Training				Testing			
	R^2	RMSE	MSE	MAD	R^2	RMSE	MSE	MAD
Random Forest	0.99	0.43	0.18	0.30	0.99	0.42	0.18	0.30
Bagging	0.99	0.44	0.19	0.31	0.99	0.44	0.19	0.31
Decision Tree	0.98	0.98	0.96	0.69	0.97	0.97	0.94	0.69

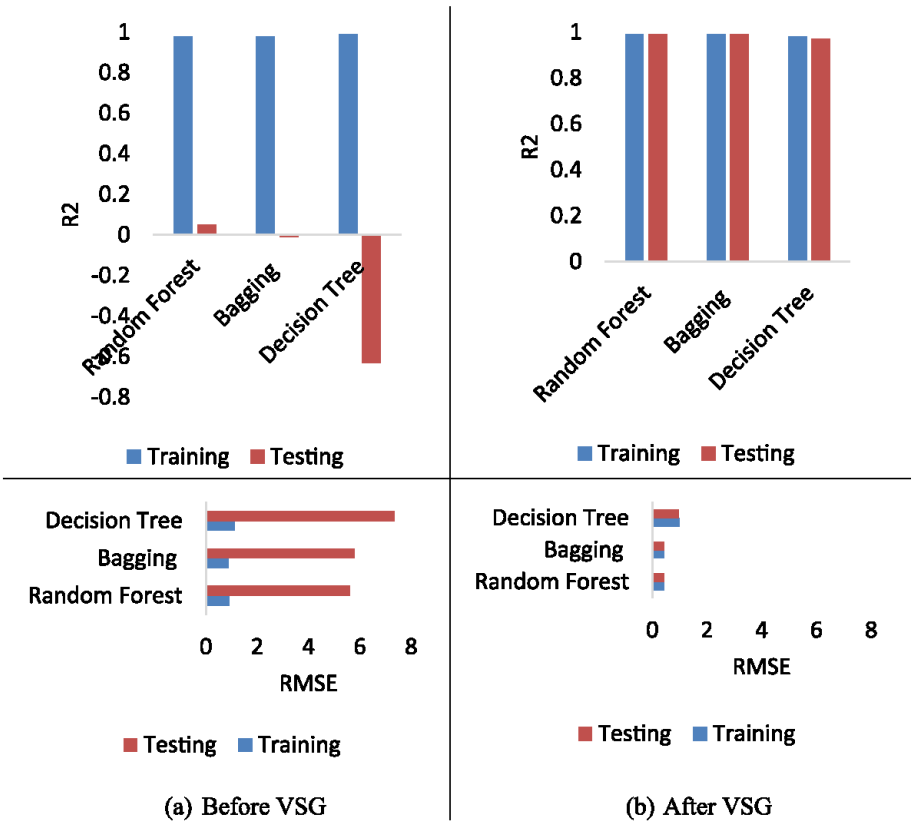

Fig. 2 Training and testing metrics between **a** before and **b** after the addition of VSG for best three models

was evident from the high R^2 and low RMSE values for the training sets contrasted with the low R^2 and high RMSE values for the test sets. After implementing VSG, the models achieved a better balance, with improved R^2 and RMSE values for the test datasets. This demonstrates that VSG effectively enhances the models' ability to generalize from training data to unseen test data, leading to a good fit and more reliable predictions in real-world scenarios.

Table 6 demonstrates the performance of the random forest model on three different datasets, pyridazine, quinoxaline, and plant extract, both before and after the application of the VSG technique. The comparison highlights a significant improvement in model performance across all evaluated metrics. The application of VSG leads to remarkable improvements in the random forest model's performance across all datasets. The *R*2 values

Table 6 The random forest performance before and after the application of VSG for testing set

Dataset	Without VSG				With VSG			
	R^2	RMSE	MSE	MAD	R^2	RMSE	MSE	MAD
Pyridazine [49]	0.52	1.12	1.25	0.79	0.99	0.20	0.04	0.14
Quinoxaline [50]	0.40	1.56	2.43	1.10	0.99	0.18	0.03	0.13
Plant extract [25]	0.51	1.38	1.90	0.98	0.99	0.15	0.02	0.11

consistently improve to 0.99, indicating accurate predictive capability. Both RMSE and MSE show substantial reductions, reflecting lower prediction errors. Similarly, MAD values drop significantly, indicating increased precision and accuracy of predictions. This suggests that VSG is highly effective in enhancing the performance of random forest models by optimizing variable selection and reducing noise in the data.

Table 7 shows a comparison of the performance of the model in this study with other relevant works for same dataset. It shows that the random forest model in this study has a very good ability to explain data variability, with an almost perfect R^2 value, with a low RMSE value indicating a very low prediction error, and a low MSE value indicating a very good model performance with a low mean squared error. In contrast, the model in previous work [13], reported an MSE value of 32.60 indicating that the current model with the effect of adding VSG can improve the accuracy of the random forest model. MAD provides a direct indication of how much the model's predictions typically deviate from the actual values. A relatively small MAD value indicates that the model has good predictive performance because its average prediction error is smaller. A MAD value of 0.30 indicates that, on average, the model's predictions deviate by about 0.3 from the actual values, which is relatively small. The results of this study indicate that the random forest model applied with VSG significantly increased its accuracy, indicating a substantial increase in performance. This emphasizes the importance of implementing VSG in improving the accuracy and predictive ability of the model, especially in the corrosion research domain. Thus, this study not only shows a significant increase in model performance but also highlights the relevance and effectiveness of nonlinear approaches, especially random forest, in handling the tested pyrimidine compound dataset.

4 Conclusion

This study comprehensively analyzes pyrimidine compounds as corrosion inhibitors, utilizing advanced ML techniques to predict CIE. The application of virtual samples through VSG has been shown to significantly enhance the accuracy of predictive models. Among these, nonlinear models, especially the random forest model, emerged as particularly effective, outperforming linear models based on RMSE and R^2 metrics. The conclusive evidence from the study places the random forest model at the forefront, with an R^2 of 0.99 and an RMSE of 0.42, highlighting its high accuracy in modeling the complex relationships within the chemical descriptors of pyrimidine compounds. The observed results underscore the efficacy of VSG in bolstering the predictive performance of machine learning models in the context of corrosion inhibitor efficiency. Such advancements in model accuracy, particularly in fields with limited datasets, highlight the potential of synthetic data augmentation as a robust approach for enhancing machine learning applications in scientific research.

Table 7 Comparison with another model for the pyrimidine dataset

Model	This work				Other work [13]		
	R^2	RMSE	MSE	MAD	R^2	RMSE	MSE
RF	0.99	0.42	0.18	0.30	–	–	32.60

Acknowledgements The authors would like to express their appreciation to the LPPM Universitas Dian Nuswantoro for their significant assistance in funding this research through the "Penelitian Internal" project, with reference number 049/A.38-04/UDN-09/V/2023. The computational work for this study was conducted at the Research Center for Materials Informatics, located in the Faculty of Computer Science at Dian Nuswantoro University in Semarang, Indonesia.

Author contribution WH, WAEP, MA were involved in writing—original draft, methodology, conceptualization, analysis, data collection and construction, Performed machine learning calculation. NAS, AWK, NNH, TS, SRwere involved in review, supervision, validation.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Gardner L (2019) Stability and design of stainless steel structures – Review and outlook, thin-walled structures. Thin-Walled Struct 141:208–216. https://doi.org/10.1016/j.tws.2019.04.019
2. Wanli W, Chen R, Yang Z, He Z, Zhou Y, Lv F (2021) Corrosion resistance of 45 carbon steel enhanced by laser graphene-based coating. Diam Relat Mater 116:108370. https://doi.org/10.1016/j.diamond.2021.108370
3. Raja VB, Palanikumar K, Renish RR, Babu AG, Varma J, Gopal P (2021) Corrosion resistance of corten steel–A review. Mater Today Proc 1(46):3572–3577
4. Xin H, Iulia Tarus L, Cheng MV, Persem N, Lorich L (2021) Experiments and numerical simulation of wire and arc additive manufactured steel materials. Structures 34:1393–1402. https://doi.org/10.1016/j.istruc.2021.08.055
5. Mythreyi OV, Rohith Srinivaas M, Kumar TA, Jayaganthan R (2021) Machine-learning-based prediction of corrosion behavior in additively manufactured inconel 718. Data 6(8):80. https://doi.org/10.3390/data6080080
6. Skrifvars BJ, Backman R, Hupa M, Salmenoja K, Vakkilainen E (2008) Corrosion of superheater steel materials under alkali salt deposits Part 1: the effect of salt deposit composition and temperature. Corros Sci 50(5):1274–1282
7. Akrom M, Saputro AG, Maulana AL, Ramelan A, Nuruddin A, Rustad S, Dipojono HK (2023) DFT and microkinetic investigation of oxygen reduction reaction on corrosion inhibition mechanism of iron surface by *Syzygium Aromaticum* extract. Appl Surf Sci 1(615):156319
8. Akrom M (2024) Green corrosion inhibitors for iron alloys: a comprehensive review of integrating data-driven forecasting, density functional theory simulations, and experimental investigation. J Mult Mater Inf 1(1):22–37. https://doi.org/10.62411/jimat.v1i1.10495
9. Budi S et al (2024) Implementation of polynomial functions to improve the accuracy of machine learning models in predicting the corrosion inhibition efficiency of pyridine-quinoline compounds as corrosion inhibitors. KnE Eng. https://doi.org/10.18502/keg.v6i1.15351
10. Mu'azu ND et al (2023) Inhibition of low carbon steel corrosion by a cationic gemini surfactant in 10wt.% H2SO4 and 15 wt% HCl under static condition and hydrodynamic flow. SouAfrican J Chem Eng 43:232–244. https://doi.org/10.1016/j.sajce.2022.10.006
11. Quadri TW et al (2022) Predicting protection capacities of pyrimidine-based corrosion inhibitors for mild steel/HCl interface using linear and nonlinear QSPR models. J Mol Model 28(9):1–23. https://doi.org/10.1007/s00894-022-05245-1
12. Singh R, Prasad D, Safi Z, Wazzan N, Guo L (2022) De-scaling, experimental, DFT, and MD-simulation studies of unwanted growing plant as natural corrosion inhibitor for SS-410 in acid medium. Colloid Surf Physicochem Eng Asp 649:129333. https://doi.org/10.1016/j.colsurfa.2022.129333
13. Alamri AH, Alhazmi N (2022) Development of data driven machine learning models for the prediction and design of pyrimidine corrosion inhibitors. J Saudi Chem Soc 26(6):101536. https://doi.org/10.1016/j.jscs.2022.101536
14. Akrom M, Rustad S, Dipojono HK (2024) Variational quantum circuit-based quantum machine learning approach for predicting corrosion inhibition efficiency of pyridine-quinoline compounds. Mater Today Quantum 2:100007. https://doi.org/10.1016/j.mtquan.2024.100007
15. Rasheeda K, Alva VDP, Krishnaprasad PA, Samshuddin S (2018) Pyrimidine derivatives as potential corrosion inhibitors for steel in acid medium—an overview. Int J Corros Scale Inhibit 7(1):48–61. https://doi.org/10.17675/2305-6894-2018-7-1-5
16. Akrom M, Rustad S, Dipojono HK (2024) Prediction of anti-corrosion performance of new triazole derivatives via machine learning. Comput Theor Chem 1(1236):114599
17. Herowati W et al (2024) Prediction of corrosion inhibition efficiency based on machine learning for pyrimidine compounds: a comparative study of linear and non-linear algorithms. KnE Eng 7:68–77. https://doi.org/10.18502/keg.v6i1.15350
18. Quraishi MA, Chauhan DS, Saji VS (2021) Heterocyclic biomolecules as green corrosion inhibitors. J Mol Liq 341:117265. https://doi.org/10.1016/j.molliq.2021.117265
19. Obot IB, Macdonald DD, Gasem ZM (2015) Density functional theory (DFT) as a powerful tool for designing new organic corrosion inhibitors: part 1: an overview. Corros Sci 99:1–30. https://doi.org/10.1016/j.corsci.2015.01.037
20. Takagi T (2023) How beneficial or threatening is artificial intelligence? Chem-Bio Inf J 23:7–13. https://doi.org/10.1273/cbij.23.7
21. Akrom M, Rustad S, Dipojono HK (2024) SMILES-based machine learning enables the prediction of corrosion inhibition capacity. MRS Commun. https://doi.org/10.1557/s43579-024-00551-6
22. Belghiti ME, Benhiba F, Benzbiria N, Lai CH, Echihi S, Salah M, Zeroual A, Karzazi Y, Tounsi A, Abbiche K, Belaaouad S (2022) Performance of triazole derivatives as potential corrosion inhibitors for mild steel in a strong phosphoric acid medium: Combining experimental and computational (DFT, MDs & QSAR) approaches. J Mol Struct 15(1256):132515
23. Galvão TLP, Novell-Leruth G, Kuznetsova A, Tedim J, Gomes JRB (2020) Elucidating structure-property relationships in aluminum alloy corrosion inhibitors by machine learning. J Phys Chem C 124(10):5624–5635. https://doi.org/10.1021/acs.jpcc.9b09538
24. Lemaoui T, Hammoudi NE, Alnashef IM, Balsamo M, Erto A, Ernst B, Benguerba Y (2020) Quantitative structure properties relationship for deep eutectic solvents using Sσ-profile as molecular descriptors. J Mol Liq 309:113165
25. Akrom M, Rustad S, Dipojono HK (2024) A machine learning approach to predict the efficiency of corrosion inhibition by natural product-based organic inhibitors. Phys Scr 99(3):036006. https://doi.org/10.1088/1402-4896/ad28a9
26. Akrom M, Rustad S, Dipojono HK (2024) Development of quantum machine learning to evaluate the corrosion inhibition capability of pyrimidine compounds. Mater Today Commun 39:108758. https://doi.org/10.1016/j.mtcomm.2024.108758

27. Akrom M, Rustad S, Dipojono HK (2023) Machine learning investigation to predict corrosion inhibition capacity of new amino acid compounds as corrosion inhibitors. Result Chem 6:101126. https://doi.org/10.1016/j.rechem.2023.101126
28. Akrom M, Rustad S, Saputro AG, Dipojono HK (2023) Data-driven investigation to model the corrosion inhibition efficiency of pyrimidine-pyrazole hybrid corrosion inhibitors. Comput Theor Chem 1229:114307. https://doi.org/10.1016/J.COMPTC.2023.114307
29. Akrom M, Sutojo T, Pertiwi A, Rustad S, Dipojono HK (2023) Investigation of best QSPR-based machine learning model to predict corrosion inhibition performance of pyridine-quinoline compounds. J Phys Conf Series 2673(1):012014. https://doi.org/10.1088/1742-6596/2673/1/012014
30. Pengcheng X, Ji X, Li M, Wencong L (2023) Virtual sample generation in machine learning assisted materials design and discovery. J Mater Inf 3:16. https://doi.org/10.20517/jmi.2023.18
31. Akrom M, Rustad S, Saputro AG, Ramelan A, Fathurrahman F, Dipojono HK (2023) 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. https://doi.org/10.1016/J.MTCOMM.2023.106402
32. Sutojo T, Rustad S, Akrom M, Syukur A, Shidik GF, Dipojono H (2023) A machine learning approach for corrosion small datasets. npj Materials Degradation 7:18. https://doi.org/10.1038/s41529-023-00336-7
33. Iyer RS, Iyer NS, Joseph A (2024) Harnessing machine learning and virtual sample generation for corrosion studies of 2-alkyl benzimidazole scaffold small dataset with an experimental validation. J Mol Struct 15(1306):137767
34. Li DC, Wen IH (2014) A genetic algorithm-based virtual sample generation technique to improve small data set learning. Neurocomputing 143:222–230. https://doi.org/10.1016/j.neucom.2014.06.004
35. Cui C, Tang J, Xia H, Qiao J, Yu W (2023) Virtual sample generation method based on generative adversarial fuzzy neural network. Neural Comput Appl 35(9):6979–7001. https://doi.org/10.1007/s00521-022-08104-5
36. Yang J, Yu X, Xie ZQ, Zhang JP (2011) A novel virtual sample generation method based on Gaussian distribution. Knowl Based Syst 24(6):740–748. https://doi.org/10.1016/j.knosys.2010.12.010
37. Zhang XH, Xu Y, He YL, Zhu QX (2021) Novel manifold learning based virtual sample generation for optimizing soft sensor with small data. ISA Trans 109:229–241. https://doi.org/10.1016/j.isatra.2020.10.006
38. Sutojo T, Rustad S, Akrom M, Syukur A, Shidik GF, Dipojono HK (2023) A machine learning approach for corrosion small datasets. Npj Mater Degrad 7:18. https://doi.org/10.1038/s41529-023-00336-7
39. Kamalov F (2020) Kernel density estimation based sampling for imbalanced class distribution. Inf Sci N Y 512:1192–1201. https://doi.org/10.1016/J.INS.2019.10.017
40. Zhu Q.Z., Wang Z. H., He Y. L., and Xu Y., "A monte carlo and kernel density estimation based virtual sample generation method for small data modeling problem," In: *Proceedings - 2020* Chinese Automation Congress, CAC 2020, Institute of Electrical and Electronics Engineers Inc, pp. 1123–1128. (2020) https://doi.org/10.1109/CAC51589.2020.9326486.
41. J. Kim and C. D. Scott, "Robust Kernel Density Estimation," 2012. [Online]. Available: www.eecs.umich.edu/
42. Akrom M (2022) Investigation of natural extracts as green corrosion inhibitors in steel using density functional theory. J Teori dan Aplikasi Fisika. 31:89–102
43. Budi S, Akrom M, Trisnapradika GA, Sutojo T, Prabowo WA (2023) Optimization of polynomial functions on the NuSVR algorithm based on machine learning: case studies on regression datasets. Sci J Inf 10(2):151–158
44. Pedregosa F et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12(85):2825–2830
45. Browne MW (2000) Cross-validation methods. journal of Mathematical Psychology. J Math Psychol 44:108–132
46. Friedl H, Stampfer E (2001) Cross-validation. In: Abdel H, -Shaarawi E, Piegorsch W (eds) Encyclopedia of Environmetrics. Wiley, USA. https://doi.org/10.1002/9780470057339.vac062
47. Pugliese R, Regondi S, Marini R (2021) Machine learning-based approach: global trends, research directions, and regulatory standpoints. Data Sci Manag 4:19–29. https://doi.org/10.1016/j.dsm.2021.12.002
48. Sarker IH (2021) Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2(3):1–21. https://doi.org/10.1007/s42979-021-00592-x
49. Quadri TW, Olasunkanmi LO, Akpan ED, Fayemi OE, Lee HS, Lgaz H, Verma C, Guo L, Kaya S, Ebenso EE (2022) Development of QSAR-based (MLR/ANN) predictive models for effective design of pyridazine corrosion inhibitors. Mater Today Commun 1(30):103163
50. Quadri TW, Olasunkanmi LO, Fayemi OE, Lgaz H, Dagdag O, Sherif ES, Alrashdi AA, Akpan ED, Lee HS, Ebenso EE (2022) Computational insights into quinoxaline-based corrosion inhibitors of steel in HCl: quantum chemical analysis and QSPR-ANN studies. Arabian J Chem 15(7):103870