ARTICLE OPEN

A machine learning approach for corrosion small datasets

Totok Sutojo[1,2], Supriadi Rustad[1,2]✉, Muhamad Akrom[1,3], Abdul Syukur[2], Guruh Fajar Shidik[2] and Hermawan Kresno Dipojono[3]

In this work, we developed a QSAR model using the K-Nearest Neighbor (KNN) algorithm to predict the corrosion inhibition performance of the inhibitor compound. To overcome the small dataset problems, virtual samples are generated and added to the training set using a Virtual Sample Generation (VSG) method. The generalizability of the proposed KNN + VSG model is verified by using six small datasets from references and comparing their prediction performances. The research shows that for the six datasets, the proposed model is able to make predictions with the best accuracy. Adding virtual samples to the training data helps the algorithm recognize feature-target relationship patterns, and therefore increases the number of chemical quantum parameters correlated with corrosion inhibition efficiency. This proposed method strengthens the prospect of ML for developing material designs, especially in the case of small datasets.

npj Materials Degradation (2023)7:18 ; https://doi.org/10.1038/s41529-023-00336-7

INTRODUCTION

Corrosion is an electrochemical process between metal surfaces and a corrosive environment that can cause huge losses in various industrial fields, especially in the oil and gas industries[1,2]. One of the most economical and effective strategies to prevent metal damage due to corrosion is the use of corrosion inhibitors[2–4]. Various organic compounds have been used as corrosion inhibitors to protect metals, for example, nitrogen-containing heterocyclic compounds such as imidazole derivatives, benzimidazoles, and pyridines[5], plant extracts[6–8], and commercial drugs[9,10]. Experimental and theoretical investigations based on density functional theory (DFT) in evaluating various candidates for potential inhibitor compounds require intensive resources, costs, and time. Nowadays, the machine learning (ML) based quantitative structure-activity relationships (QSAR) approach has become a reliable method in revealing the relationship between structural properties of chemical compounds and biological activities. The approach is widely used to explore various inhibitor candidate compounds[11–14] because the electronic properties and chemical reactivity can be quantified against the structure compound chemistry.

Quantum chemical descriptors derived from DFT calculations[15] and molecular dynamics simulations[16] have been implemented in the development of QSAR models using various ML algorithms to evaluate inhibitor performance. An artificial neural network (ANN) model was applied to predict the corrosion inhibition potential of 11 thiophene-derived compounds with 7 quantum chemical descriptors, resulted in a determinant coefficient (R^2) of 0.958[17]. The ANN was also used to predict the efficiency of corrosion inhibition in a chloride solution of 28 amino acids with 12 quantum chemical descriptors, resulted in R^2 of 0.999 and a predictive sum of squares 19.181[18]. Quadri et al.[19] reported that the ANN model implemented on a dataset of 20 pyridazine derivatives gave a root mean square error (RMSE) value of 10.5637. Another QSAR study was also developed to predict 41 pyridine-quinoline-derived compounds with 20 quantum chemical descriptors using a combination of Genetic Algorithm and ANN (GA-ANN) with an average value of RMSE = 8.8%[20]. Liu et al.[21] used an SVM model with 11 descriptors to evaluate 20 benzimidazole derivatives, reported correlation coefficient (r) and RMSE values of 0.9589 and 4.45, respectively. Meanwhile, several other literatures discussing corrosion reported that ANN and support vector machine (SVM) were unreliable when applied to small datasets[22,23]. Beltran-Perez et al.[24] developed an Autoregressive with Exogeneous Inputs ARX model for examining commercial drugs as corrosion inhibitors in steel.

The number of features and samples of the datasets in the studies reported above range from 7 to 20 and 11–69, respectively, this might be classified as small datasets. Research[25] mentioned that small dataset problems refer to the case of small amount of samples, where the number of samples is less than 50 in respect to engineering applications or less than 30 in regard to academic research. Since the availability of high-quality datasets is a key factor in ML for corrosion[26,27], then its related model must have a sufficient number of samples[28]. Small datasets cannot fully reveal all population features due to lack of information[29], and these lead to overfitting[28,29], bias[30,31], decreased accuracy[32,33], poor generalization skills[34], and often make learning algorithms difficult to produce accurate predictions[35]. Solving small dataset problems is important in developing a QSAR model for predicting corrosion resistance with a limited number of data samples.

Most ML research for corrosion aims to improve the accuracy of a predictive model. For that reason, Roy et al[26]. used descriptors selection for predicting corrosion resistance in multi-principal element alloys. Their feature selection preprocessing showed both promise and perils of using ML for such a complex chemical phenomenon. We develop an ML-based QSAR model using the K-Nearest Neighbor (KNN) equipped with VSG method to evaluate the corrosion inhibition performance of inhibitor compounds derived from a series of datasets found in previous studies. The KNN method is chosen because, among learning algorithms, it is the simplest and has good performance when working with small datasets[36]. Meanwhile, virtual sample generation (VSG) is chosen because it is the most prominent and popular technique for solving small dataset problems, namely by adding virtual samples to the training data[25]. In this study, all quantum chemical descriptors are considered important features and the addition

[1]Research Center for Materials Informatics, Faculty of Computer Science, Dian Nuswantoro University, Semarang 50131, Indonesia. [2]Doctoral Program of Computer Science, Faculty of Computer Science, Dian Nuswantoro University, Semarang 50131, Indonesia. [3]Advanced Functional Materials Research Group, Bandung Institute of Technology, Bandung 40132, Indonesia. ✉email: srustad@dsn.dinus.ac.id

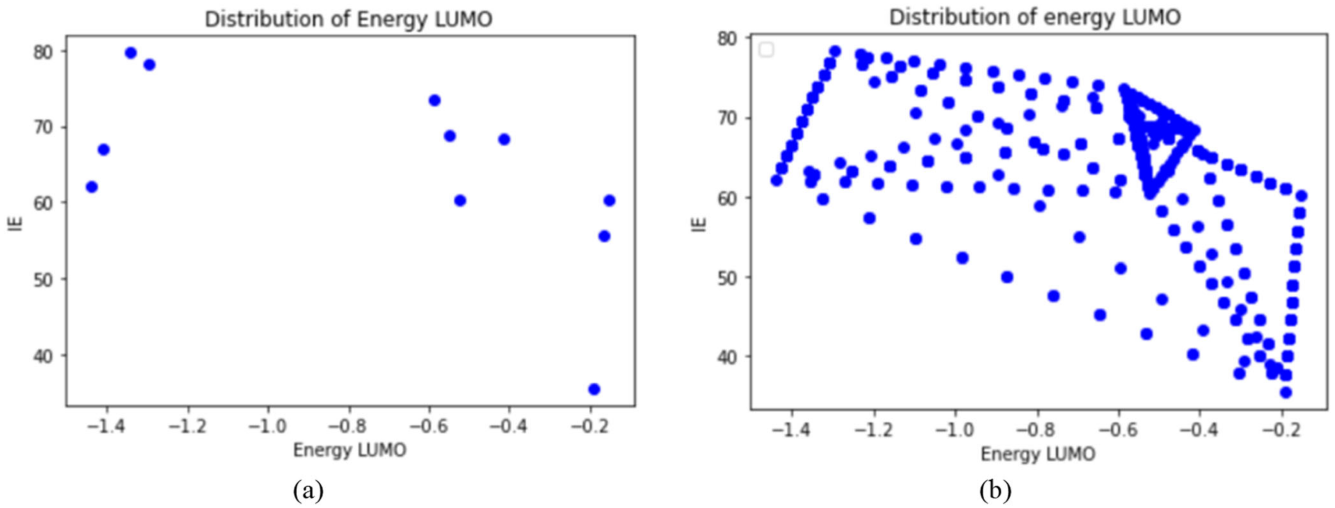

Fig. 1 The effect of virtual samples on the distribution of feature data. Distribution of E_{LUMO} against IE (**a**) before and (**b**) after adding virtual samples for the thiophene and its derivatives.

of virtual samples to the training is intended to increase their relevance and correlation with the target of inhibition efficiencies.

Following the introduction, materials and methods are described in more detail covering datasets, quantum chemical descriptors, KNN algorithm, virtual sample generation method, model development and validation, and assessment metrics. Furthermore, the results of the research are presented along with discussions aimed at knowing the determination of the model and its performance advantages when compared to the previous models. This paper concludes that the KNN + VSG model is proven to outperform the previous model, and that the VSG method is very appropriate to be applied in the field of material sciences of small datasets.

RESULTS AND DISCUSSION

Correlation analysis

Small datasets have at least two characteristics, namely, uneven distribution of observations and the high-value gap between data. These are problems for machine learning algorithms to learn patterns accurately[35]. Virtual samples are generated and added to the training in hopes of addressing the problems. To illustrate how virtual samples help the ML algorithm to learn, in the following, we give an example of adding virtual samples affects the distributions of a descriptor E_{LUMO} against IE and their histograms for the thiophene and its derivatives dataset. Figures 1a, 2a, b show the characteristics of a small dataset with an uneven distribution and a high-value gap between data. The addition of virtual samples makes the distribution more even and the gap between data values becomes smaller as shown in Figs. 1b, 2c, d. With this improved distribution and histogram, the algorithm more easily recognizes patterns even though the number of actual data is small[35].

The effect of adding virtual samples is further analyzed using correlation calculation to see the strength of the relationship between the descriptors and the corrosion inhibition efficiency (IE). Some literatures[37–39] used these calculations of correlation for their feature selection method to get an increase in the model accuracy. The features that correlate with the target are selected to get the best model[38,40]. This study uses the Spearman correlation coefficient as in Eq. (5) with a 95% confidence interval or a 5% significance level to describe how sensitive the IE responses to the increase or decrease of monotonous descriptors. The sign 1 means both the descriptor and IE samples are correlated (p value <0.05). The sign 0 means the two samples are not correlated (p value ≥0.05). Tables 1–7 present a complete picture of the effect of adding virtual samples to the learning of the KNN algorithm on the correlation between each descriptor and the IE target for all datasets.

It can be seen in Table 1, based on the p value of 0.05, there are only four descriptors that are correlated with the IE target, namely Hammet constant, E_{LUMO}, energy gap, and molecular volume, while three other descriptors (dipole moment, E_{HOMO}, molecular surface area) are not correlated (sign = 0). This situation reduces the ability of ML algorithms to recognize patterns. With the addition of virtual samples, there is an improvement in the status of the three descriptors so that all descriptors are correlated with IE (sign = 1). The addition of virtual samples improves the correlation between descriptors and the target for the Thiophene and Its Derivatives dataset. This is the status that every ML algorithm expects in recognizing patterns. In Table 2, for the Benzimidazole Derivatives dataset, there is only 1 descriptor (E_{HOMO}) not correlated with IE. The use of virtual samples improves the status of this descriptor so that the whole descriptors of the dataset are correlated with IE, making it easier for the ML algorithm to capture the pattern.

Table 3 shows that for the Amino Acids, as many as 8 features are not correlated with IE, namely: E_{HOMO}, adsorption energy, total energy, dipole x, dipole y, dipole z, molecular surface area, and molecular volume. In such circumstances, the performance of the ML algorithm is not optimal because the number of uncorrelated features is more than that of their correlated counterparts. The addition of virtual samples reduces the level of feature uncorrelatedness to only 1 feature, namely molecular volume. On the Pyridines and Quinolones dataset that consists of 20 descriptors (Table 4), the addition of virtual samples improves the status of descriptor's correlation. Learning without virtual samples results in 15 correlated and the remaining five are not correlated. With the addition of virtual samples, only two descriptors are not correlated, namely dipole moment and electronegativity.

The most striking observation on the benefits of virtual samples in ML performance is found in the Commercial Drugs and Pyridazine Derivatives datasets, as presented in Tables 5, 6. In Commercial Drugs, from 10 descriptors, there is no single descriptor has correlation with IE, but the addition of virtual samples increases the number of correlated descriptors into 8 with 2 stay uncorrelated, namely pKa and E_{LUMO}. In the Pyridazine Derivatives dataset (Table 6), the correlation between attributes

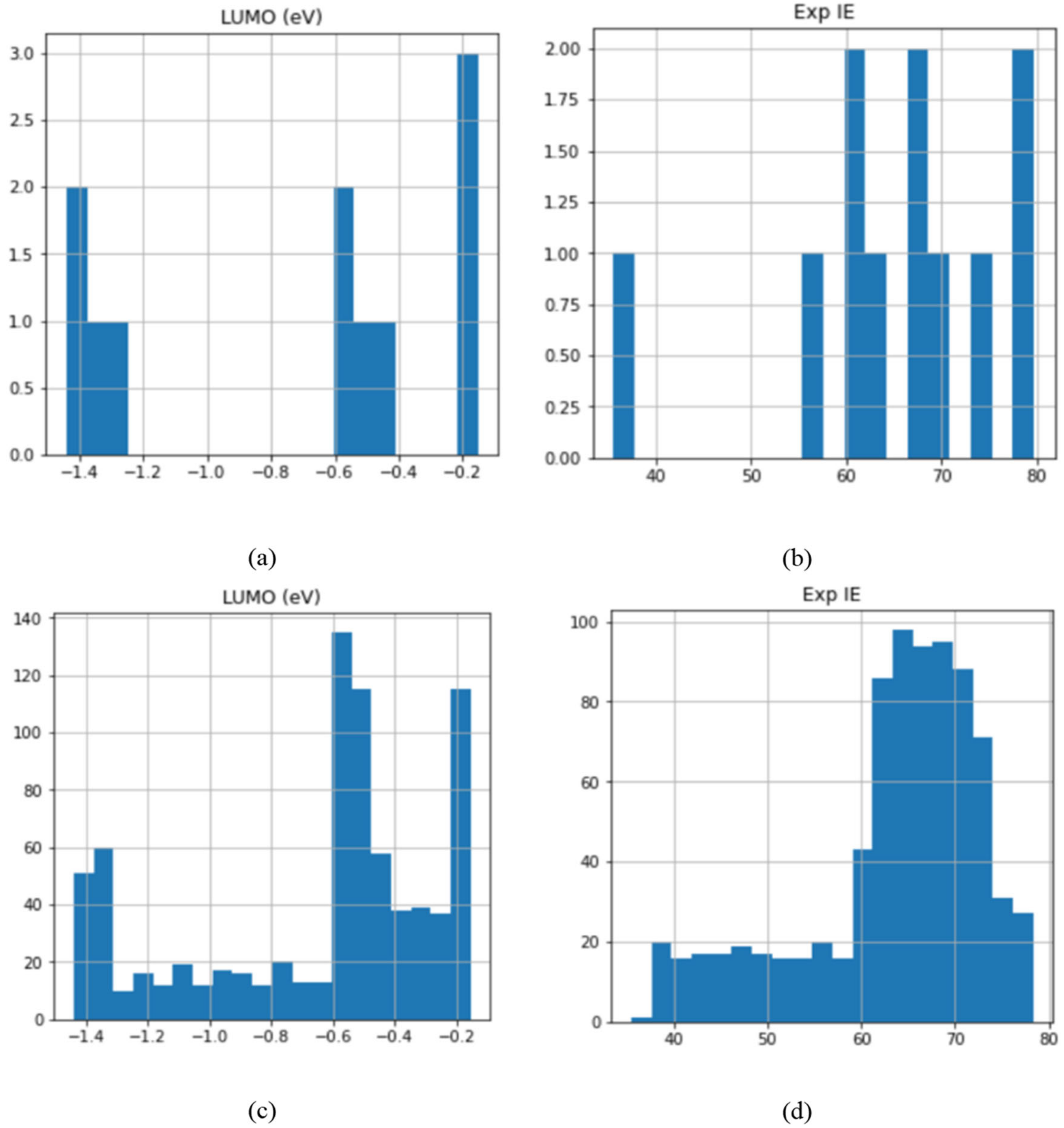

Fig. 2 The effect of virtual samples on histograms of feature data. Histograms of (**a**) E_{LUMO} and (**b**) IE before, and (**c**) E_{LUMO} and (**d**) IE after adding virtual samples for the thiophene and its derivatives.

Table 1. Feature-target correlation for thiophene derivatives.

Descriptor	Without virtual samples			With virtual samples		
	r	*p* value	Sign	*r*	*p* value	Sign
Hammett constant	0.65	0.03	1	0.61	0.00	1
Dipole moment	0.60	0.05	0	0.62	0.00	1
HOMO	−0.45	0.17	0	−0.26	0.00	1
LUMO	−0.61	0.04	1	−0.63	0.00	1
Gap energy	−0.68	0.02	1	−0.54	0.00	1
Molecular surface area	0.57	0.06	0	0.66	0.00	1
Molecular volume	0.79	0.00	1	0.75	0.00	1

shows that there is only 1 descriptor, namely the total energy with the status of correlated before adding the virtual samples. The presence of virtual samples changes the status of correlation from one into nine descriptors, only two, namely dipole moment and electronegativity, are still uncorrelated.

Overall, the experiment shows that virtual samples affect the status of descriptors from uncorrelated to mostly correlated. This correlation (between the descriptor and IE) is needed to get the best model so that it improves the performance of the model[37,41]. The experiment shows that the use of virtual samples in the KNN algorithm (Table 7), significantly decreases the RMSE values from 12.122 to 1.639, 12.89 to 3.88, 45.711 to 3.888; 20.424 to 2.707%; 7.113 to 3.858; and 12.848 to 1.135 each for the thiophene and its derivatives, benzimidazole derivatives, amino acids and their related compounds, pyridines and quinolones, commercial drugs, and pyridazine derivatives datasets. This clearly indicates that the

Table 2. Feature-target correlation for benzimidazole derivatives.

	Without virtual samples			With virtual samples		
Descriptor	*r*	*p* value	Sign	*r*	*p* value	Sign
HOMO	0.39	0.09	0	0.36	0.00	1
LUMO	0.61	0.00	1	0.44	0.00	1
Polarizability	0.47	0.04	1	0.62	0.00	1
Total negative natural charges	−0.92	0.00	1	−0.91	0.00	1
Molecular volume	0.54	0.01	1	0.66	0.00	1
Ionization potential	−0.82	0.00	1	−0.84	0.00	1
Electron affinity	−0.63	0.00	1	−0.51	0.00	1
Electrophilicity	−0.61	0.01	1	−0.45	0.00	1
Fraction of electron transferred	0.78	0.00	1	0.79	0.00	1
ΛNICS(1)B	0.73	0.00	1	0.75	0.00	1
ΛNICS(1)I	0.74	0.00	1	0.7	0.00	1

Table 3. Feature-target correlation for amino acids.

	Without virtual samples			With virtual samples		
Descriptor	*r*	*p* value	Sign	*r*	*p* value	Sign
HOMO	0.28	0.15	0	0.2	0.00	1
LUMO	−0.39	0.04	1	−0.18	0.00	1
Gap energy	−0.76	0.00	1	−0.49	0.00	1
Binding energy	0.99	0.00	1	0.95	0.00	1
Adsorption energy	−0.06	0.77	0	0.04	0.05	1
Total energy	0.07	0.73	0	0.1	0.00	1
Dipole moment	−0.98	0.00	1	−0.85	0.00	1
Dipole x	−0.04	0.86	0	0.12	0.00	1
Dipole y	0.17	0.4	0	0.06	0.01	1
Dipole z	−0.11	0.57	0	−0.05	0.02	1
Molecular surface area	0.07	0.73	0	0.05	0.04	1
Molecular volume	0.04	0.86	0	0.04	0.08	0

Table 4. Feature-target correlation for pyrimidine and quinoline derivatives.

Descriptor	Without virtual samples			With virtual samples		
	r	*p* value	Sign	*r*	*p* value	Sign
Dipole moment	0.02	0.92	0	0.02	0.20	0
Polarizability	0.49	0.00	1	0.39	0.00	1
HOMO	0.22	0.17	0	0.13	0.00	1
LUMO	−0.6	0.00	1	−0.56	0.00	1
Gap energy	−0.65	0.00	1	−0.63	0.00	1
Ionization potential	−0.41	0.01	1	−0.36	0.00	1
Electron affinity	0.59	0.00	1	0.57	0.00	1
Electronegativity	0.04	0.79	0	0.02	0.22	0
Gobal hardness	−0.61	0.00	1	−0.6	0.00	1
Global softness	0.61	0.00	1	0.6	0.00	1
Electrophilicity	0.55	0.00	1	0.5	0.00	1
Electron donor capacity	0.48	0.00	1	0.41	0.00	1
Electron acceptor capacity	0.58	0.00	1	0.57	0.00	1
Fraction of electrons transferred	0.43	0.01	1	0.39	0.00	1
NBO atomic charge	0.53	0.00	1	0.39	0.00	1
Hydrophobicity	0.35	0.03	1	0.21	0.00	1
Solvent accessible surface area	0.2	0.21	0	0.1	0.00	1
Molecular surface area	0.25	0.11	0	0.16	0.00	1
Molecular volume	0.42	0.01	1	0.29	0.00	1
Adsorption energy	−0.68	0.00	1	−0.66	0.00	1

Table 5. Feature-target correlation for commercial drugs.

	Without virtual samples			With virtual samples		
Descriptor	*r*	*p* value	Sign	*r*	*p* value	Sign
Molecular weight	0.01	0.95	0	0.04	0.00	1
Acid dissociation constant	−0.07	0.57	0	−0.01	0.31	0
Octanol-water partition coefficient	0.11	0.4	0	0.17	0.00	1
Water solubility	−0.04	0.76	0	−0.08	0.00	1
Polar surface area	−0.07	0.58	0	−0.06	0.00	1
Polarizability	0.06	0.64	0	0.1	0.00	1
HOMO	0.21	0.09	0	0.21	0.00	1
LUMO	0.00	0.99	0	−0.01	0.45	0
Electronegativity	−0.13	0.33	0	−0.09	0.00	1
Fraction of electrons transferred	0.22	0.08	0	0.25	0.00	1

addition of virtual samples improves the performance of the KNN model in making predictions of corrosion inhibition efficiencies.

Comparisons to different models

This section compares the exploratory power of different models, and at the moment, the effect of material representation is not discussed yet. Comparisons are made purely regarding ML performance parameters. The generalizability of our proposed KNN + VSG model is verified with the six datasets in Table 8, and its performance was compared with the performance of the ANN, GA-ANN, SVM, and ARX models from related previous references. The RMSE metric is used to assess the prediction error for each model, while R^2 is used to assess the suitability of the model predictions with observations. The experimental results of corrosion inhibitor prediction for the six datasets are presented in Table 7, where the RMSE_CV is the prediction error tested during cross-validation training, and the RMSE is that tested using a testing set, while NA stands for not available information. The RMSE values marked with an asterisk (*) are prediction errors that are tested using the entire datasets. As the RMSE_CV and RMSE represent cross-validation training and testing errors, respectively, their values become a measure of whether overfitting occurs.

For the KNN model, bad suitability of the model predictions is observed for its very low determination coefficient values below 0.50. For the KNN + VSG model, it can be seen that RMSE < RMSE_CV values are true for all datasets. It means that the KNN + VSG prediction of all datasets are in good-fitting[42]. The use of virtual samples in small datasets facilitates ML algorithms and effectively improves the quality of prediction fitting. In general, for all datasets, adding virtual samples significantly decreases RMSE and increases R^2 values, and therefore improves the performance of KNN model[43]. The range of RMSE values of 7.113−45.711 decreases to 1.135−3.888 after adding virtual samples, and the coefficient of determination increases from a maximum of 0.50 to a minimum of 0.96 after the implementation of the VSG.

Based on the RMSE and R^2 values, the KNN + VSG model performs better than all other models do for the same datasets. The model performance improvement is so obvious on the RMSE values from 5.836 to 1.639, 6.79 to 3.88, 4.45 to 3.88, 5.157 to 3.888, 8.831 to 2.707, 4.870 to 3.858, 10.5637 to 1.1350, each for the thiophene and its derivatives, benzimidazole derivatives, amino acids and their compounds, pyridines and quinolones, commercial drugs, and pyridazines derivatives datasets. The improvement is also supported by the relatively high determination coefficient of the proposed model ranging from 0.96 to 0.99, as compared to that of the previous models, ranging from −0.65 to 0.99.

The superiority of the proposed model performance can be observed through the visual representation of prediction data. Figure 3a–f presents the KNN + VSG model prediction data for thiophene and its derivatives, benzimidazole derivatives, amino acids and their compounds, pyridines and quinolones, commercial drugs, and pyridazines derivatives datasets, respectively. The same data from other models are also presented for comparison, except for the last dataset where the associated data is not available. It can be seen that compared to the previous model for the same dataset, the prediction data of the KNN + VSG model is relatively closer to the prediction line, indicating that this model is superior to the previous models[44].

In particular, we discuss the case of implementing the ARX model on the Commercial Drugs dataset because the suitability is visually very poor (Fig. 3e). This can happen because the data pattern from the Commercial Drugs dataset which is predicted using the linear ARX model[24] is nonlinear. This linear ARX model approach causes the prediction of IE to be unrealistic, especially for the following five data, sulfadiazine (106.31%), metacyclic (111.72%), glycine (124.03%), ethosuximide (158.88%), and hexetidine (259.25%). The KNN + VSG model can overcome the problems faced by the ARX model, such as poor suitability and unrealistic IE predictions being realistic, namely 82.51, 82.96, 93.19, 96.71, and 93.93% for ethosuximide, hexetidine, metacyclic, glycine, and sulfadiazine.

In summary, the KNN + VSG model is used to predict the efficiency of corrosion inhibition on the surface of the material based on computational and/or experimental data, which are classified as small datasets. In general, the addition of virtual samples helps increase the number of quantum chemical descriptors that correlate with IE as a target. Consequently, it improves the performance of the model with much lower errors and high determination coefficient values. Predictive performance assessment using RMSE and R^2 shows that the proposed model performs better than the ANN[17], SVM[16], SVM[21], ANN[18], GA-ANN, ARX[24], and ANN[19]

Table 6. Feature-target correlation for pyridazine derivatives.

	Without virtual samples			With virtual samples		
Descriptor	*r*	*p* value	Sign	*r*	*p* value	Sign
Total energy	−0.69	0.00	1	−0.68	0.00	1
HOMO	0.38	0.10	0	0.38	0.00	1
LUMO	−0.41	0.08	0	−0.23	0.00	1
Gap energy	−0.39	0.08	0	−0.27	0.00	1
Dipole moment	−0.12	0.60	0	0	0.99	0
Ionization potential	0.38	0.10	0	−0.38	0.00	1
Electron affinity	0.41	0.08	0	0.23	0.00	1
Electronegativity	0.13	0.57	0	0.04	0.17	0
Global hardness	−0.39	0.08	0	−0.27	0.00	1
Global softness	0.39	0.09	0	0.30	0.00	1
Fraction of electrons transferred	0.37	0.11	0	0.36	0.00	1

Table 7. Performance comparison between models for each dataset.

Datasets	Model	RMSE_CV	R^2	RMSE
Thiophene	KNN	8.9	−0.07	12.122
	KNN + VSG	2.127	0.98	1.639
	ANN[17]	NA	0.96	5.836*
Benzimidazole	KNN	14.388	−0.27	12.89
	KNN + VSG	4.366	0.98	3.88
	SVM[16]	NA	0.92	6.79
	SVM[21]	NA	0.92	4.45
Amino acids	KNN	29.564	0.45	45.711
	KNN + VSG	15.095	0.99	3.888
	ANN[18]	NA	0.99	5.157*
Pyridines and quinolones	KNN	28.010	0.50	20.424
	KNN + VSG	8.661	0.99	2.707
	GA-ANN[20]	16.700	0.80	8.831
Commercial Drugs	KNN	5.738	0.22	7.113
	KNN + VSG	4.3	0.96	3.858
	ARX[24]	NA	−0.65	4.870
Pyridazine	KNN	6.871	0.11	12.848
	KNN + VSG	1.854	0.99	1.135
	ANN[19]	NA	NA	10.5637

The RMSE values marked with an asterisk (*) are prediction errors that are tested using the entire datasets.

Table 8. List of 6 small datasets on corrosion inhibition efficiencies.

No	Dataset Name	Number of samples	Number of descriptors	Model	Data sources and users
1	Thiophene and its derivatives	11	7	ANN	[17]
2	Benzimidazole derivatives	20	11	SVM	[16,21,37]
3	Amino acids	28	12	ANN	[18]
4	Pyridines and quinolones	41	20	GA-ANN	[20]
5	Commercial drugs	69	10	ARX	[24]
6	Pyridazine derivatives dataset	20	11	ANN	[19]

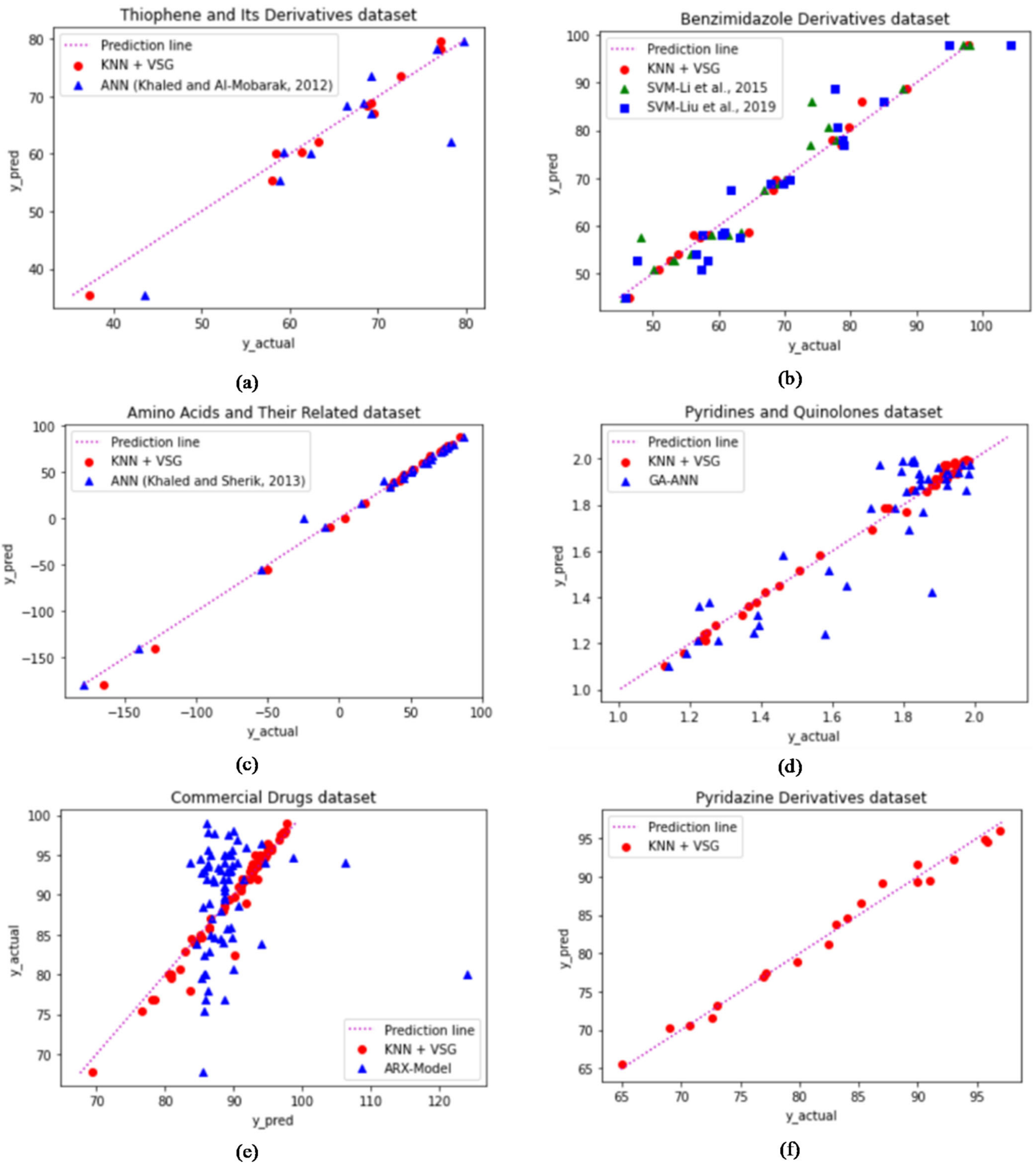

Fig. 3 The effect of virtual samples on the performance of machine learning model. Prediction data of KNN + VSG and that of (**a**) ANN for the thiophene, (**b**) SVM for the benzimidazole, (**c**) ANN for amino acids, (**d**) GA-ANN for pyridine and quinoline, (**e**) ARX for commercial drugs datasets, and (**f**) Prediction data of KNN + VSG for pyridazine dataset.

models do when working on the same dataset as used in this research. Model development is open for improvement either in an algorithm or preprocessing, and this research shows that VSG preprocessing is appropriate for solving small dataset problems in corrosion research.

METHODS

Datasets, samples, and descriptors

This study uses six small datasets from previous studies about corrosion inhibition efficiencies. Their name, number of samples, number of descriptors, associated models, and users are listed in

Table 8. Almost all datasets have samples less than 50, except Commercial Drugs, with 69 samples. It's not too much of a stretch to classify them as small datasets. For all the datasets, the number of dimensions (descriptors) are lower than that of samples, ranging from 7 to 20. Different ML algorithms had been implemented into the datasets to make predictions of inhibition efficiencies, namely ANN-based, SVM, and ARX.

Thiophene and its derivatives dataset has 11 samples with seven quantum chemical descriptors, namely: Hammett constant, dipole moment, HOMO, LUMO, gap energy, molecular surface area, and molecular volume[17]. The benzimidazole derivative dataset was adopted from the work[45] and described by ref. [16]. This dataset consists of 20 samples and 11 descriptors: HOMO, LUMO, polarizability, molecular volume, ionization potential, electron affinity, electrophilicity, total negative natural charges of all non-hydrogen atoms, the fraction of electrons transferred, ΛNICS(1)B, and ΛNICS(1)I. The amino acids dataset consists of 28 samples and 12 descriptors, taken directly from ref. [18]. The descriptors associated with this dataset are HOMO, LUMO, gap energy, binding energy, adsorption energy, total energy, dipole moment, dipole x, dipole y, dipole z, molecular surface area, and molecular volume. The dataset of 41 Pyridines and Quinolines is taken from ref. [20]. This dataset consists of 20 quantum chemical descriptors are HOMO, LUMO, gap energy, ionization potential, electron affinity, electronegativity, global hardness, global softness, dipole moment, polarizability, electrophilicity, electron donor capacity, electron acceptor capacity, the fraction of electrons transferred, NBO atomic charge, adsorption energy, hydrophobicity, molecular volume, molecular surface area, and solvent accessible surface area. The Commercial drugs dataset was obtained from ref. [24]. This dataset consists of ten descriptors: molecular weight, acid dissociation constant, octanol-water partition coefficient, water solubility, polar surface area, polarizability, HOMO, LUMO, electrophilicity, and the fraction of electrons transferred. The pyridazine derivatives dataset is taken from[19] with 11 quantum chemical descriptors, namely total energy, gap energy, HOMO, LUMO, dipole moment, ionization potential, electron affinity, electronegativity, global hardness, global softness, and the fraction of transferred electrons.

Corrosion inhibition is highly dependent on the chemical reactivity of the inhibitor molecule represented in various quantum chemical descriptors[20]. Hammett constant indicates the negativity of an atom. A negative value indicates that the atom acts as an electron donor, whereas a positive value indicates that the atom acts as an electron acceptor. The dipole moment describes the ability of molecules to interact with the metal surface dipole. Gap energy shows the level of inhibitor molecular binding ability to be adsorbed by the metal surface. Molecular surface area and molecular volume are parameters to measure the ability of molecules to prevent access to corrosive agents on metal surfaces. Since physisorption and chemisorption are related to the electronic interaction with the metal surface, the polarization of the charge around the molecule is very influential. Molecular polarizability has something to do with the distribution of electron density around the molecule, and the ability of the molecule to distort the electron density. Ionization potential is defined as the amount of energy needed to release one outer electron of the atom used to measure the reactivity of atoms or molecules. Electron affinity is the energy needed to capture 1 mole of electrons. Electrophilicity illustrates the ability of a molecule to absorb electrons. Total negative natural charges of all non-hydrogen atoms represent the amount of charge carried by all non-hydrogen atoms of the molecule. A fraction of electrons transferred is the number of electrons flowing from the inhibitor molecules to the metal surface atoms. Electron transfer occurs due to differences in electronegativity between inhibitor molecules and metal surface atoms. NICS(1)B and NICS(1)I are aromaticity parameters calculated from the center of the benzene ring and the center of the imidazole ring, respectively. The total energy refers to the ability of inhibitor molecules to be adsorbed on metal surfaces. In general, the mechanism of corrosion inhibition is related to the interaction between inhibitor molecules and metal surfaces. Inhibitor molecules can be adsorbed on the metal surface through chemisorption and physisorption. Therefore, adsorption energy and binding energy are important molecular descriptors[46,47]. Electronegativity is related to the ability of inhibitor molecules to attract electrons so that electron equilibrium is achieved. Global hardness indicates the resistance of a molecule to transfer charges, while global softness shows the capacity of a molecule to receive charges. Electron donor capacity explains the tendency of molecules to donate charges, while electron acceptor capacity explains the tendency of molecules to receive charges. Natural bonding orbital is an analysis of interacting charges, which can be used to show the type and value of the atomic charge. Hydrophobicity is related to the ability of molecules to form an adsorbed layer through a hydrophobic mechanism. Solvent accessible surface area is the surface area of the molecule that is accessible by a solvent. This parameter is also related to the ability of the molecule to prevent the access of corrosive agents to the metal surface. Molecular weight is considered a parameter related to the size of the molecule. The acid dissociation constant is related to the acidity of a solution. The octanol-water partition coefficient is the concentration of a particular substance in the liquid phase of an octanol-water mixture (between the hydrophobic phase and the hydrophilic phase). Water solubility is related to the solubility of a substance in water. The polar surface area is the surface of the molecule associated with the accumulation of charge.

KNN algorithm

The KNN prediction of the testing set is based on the result of the K-nearest neighbor's distance to the training set, which is calculated using the Euclidean distance function as in Eq. (1)[48]

$$d(x,p) = \sqrt{\sum_{i=1}^{m}(x_i - p_i)^2} \tag{1}$$

with x and p are descriptors of training and testing sets, respectively, while m is the number of descriptors. The K target values from the training set are averaged to obtain predictive results as in Eq. (2)[48]

$$y = \frac{1}{K}\sum_{i=1}^{K} y_i \tag{2}$$

where y_i is the i^{th} target and y is the result of testing prediction.

Virtual sample generation method

The VSG method is used to generate virtual samples using bush topology adopted from ref. [49], with a slight modification to get the best virtual samples by selecting the best paths of the smallest errors. Virtual samples are generated on those paths by interpolation techniques using Eqs. (3) and (4)

$$\boldsymbol{x_{vs}} = \boldsymbol{x_i} + t(\boldsymbol{x_{i+1}} - \boldsymbol{x_i}) \tag{3}$$

$$\boldsymbol{y_{vs}} = \boldsymbol{y_i} + t(\boldsymbol{y_{i+1}} - \boldsymbol{y_i}) \tag{4}$$

where, x_i and $x_{(i+1)}$ are the ith and $(i+1)$th feature vectors, while x_{vs} is the virtual sample feature vector, and y_i and y_{i+1} are the ith and the $(i+1)$th targets, while y_{vs} is the virtual sample target. The value of $t = [0,1]$ is used to determine the number of virtual samples inserted between two data points.

The number of inserts are exercised to get an optimum value that increases the number of correlated features, stability of the models, and prediction accuracy. For this purpose, the number of

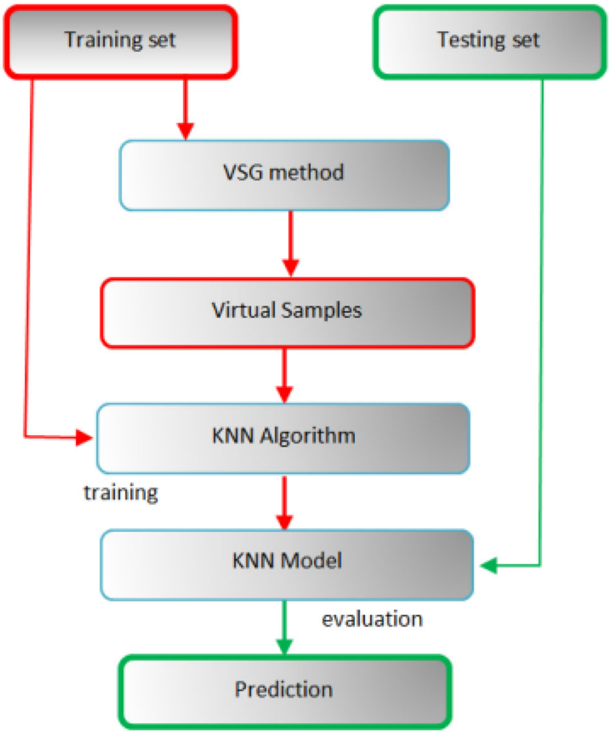

Fig. 4 Flowchart of machine learning model development. Virtual samples are generated for training, while testing sets are used for validation.

virtual samples generated between two data points is determined to be 10. If the number of training samples is n, then the number of virtual samples will be $(n-1)\times 10$. From the overall validation, k models that have the smallest errors have been selected. If there are $(n-1)\times 10$ virtual samples for each models then the total number of virtual samples will be $(n-1)\times 10\times k$. These virtual samples are added into the training set to train the machine learning algorithm to get the final model. In this study, $k=10$ is chosen to represent 10 different experiments.

Model development

In principle, the VSG technique can be implemented on any algorithm. For the preliminary study, three algorithms, namely, K-Nearest Neighbor (KNN), Decision Tree (DT), and Random Forest (RF), are chosen to exercise ML models. For all six datasets used in this research, the KNN produces the most consistent results of error lower than reference values. It also shows the most stable error for varying number of virtual samples (see Supplementary Table 1 and Supplementary Fig. 1). Therefore, the KNN is chosen to build an ML model to predict inhibition efficiencies for the six datasets. This study uses the KNN algorithm of the Python programming language. All other parameters and arrangements are defaults as stipulated in sci-kit learn release 0.23.2[50]. Figure 4 illustrates the model development, where the dataset is split into training and testing sets with a ratio of 70:30, except for the benzimidazole derivatives and commercial drugs, where the training and testing sets were determined following the previous related literature. For Benzimidazole derivatives, sample number 1–16 are used for training and number 17–20 are used for testing. For Commercial Drugs, all samples are used for training except eight samples for testing. At the preprocessing stage, both training and testing sets are normalized to avoid the sensitivity problem of certain features of prediction results. Virtual samples produced by the VSG method are added into the training set to train the KNN algorithm, and this combination model is then evaluated using the testing set to see its accuracy level in making a prediction. The performances of the model are also evaluated to see the effectiveness of adding virtual samples, before they are finally compared to that of other previous models.

Model validation

Both actual and virtual samples are used to train the KNN algorithm by implementing Monte Carlo cross-validation (MCCV), which randomly selects any pair of 70:30 actual sample partition[51], 70 for training and 30 for testing. For each random pair of training-testing partitions, virtual samples are generated along the lines connecting all-two training data for all possible paths depending on the number of training samples. Each pair is associated with $\frac{(n-1)!}{2}$ different path for distributing virtual samples, where n is the number of actual training samples. Using actual and virtual samples on each path, the training produces a temporary model, and its performance of RMSE is measured using the corresponding testing set[52–54]. A threshold θ is determined to select a temporary model that meets the criteria, its RMSE ≤θ. In the beginning, an RMSE value from a related reference model is chosen as θ. For a certain pair, a smaller θ is set up when there are more than k temporary models that meet the criteria. This process is repeated until it generates k models with the smallest RMSE values, and the associated virtual samples are stored. The virtual samples are then added to the training set to train the KNN algorithm to produce a final model.

Assessment metrics

Three performance metrics, including Spearman's rank correlation coefficient (r), determination coefficient R-squared (R^2), and root mean square error (RMSE), are used to analyze and assess the performance of the model. The Spearman correlation coefficient, measuring the strength and direction of the monotonous relationship between two variables, is calculated using Eq. (5)

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{5}$$

where,

r = strength of the rank correlation between variables
di = the difference between the x and y variable ranks for each pair of data
n = sample size

The coefficient of determination of R-squared (R^2) measures the fitting degree of a model[44], where a value approaching one indicates a good fit, as formulated in Eq. (6).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}_i\right)^2} \tag{6}$$

The RMSE measures the deviation between the predicted value and the real value[44] and is calculated using Eq. (7)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2\right)} \tag{7}$$

where Y_i, $\overline{Y}$, and $\widehat{Y}_i$are observed, average observed, and predicted values, respectively, and n is the number of samples.

DATA AVAILABILITY

The datasets that support the findings of this study are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

The codes that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 15 November 2022; Accepted: 21 February 2023;
Published online: 17 March 2023

REFERENCES

1. Finšgar, M. & Jackson, J. Application of corrosion inhibitors for steels in acidic media for the oil and gas industry: a review. *Corros. Sci.* **86**, 17–41 (2014).
2. Tiu, B. D. B. & Advincula, R. C. Polymeric corrosion inhibitors for the oil and gas industry: design principles and mechanism. *React. Funct. Polym.* **95**, 25–45 (2015).
3. Kadhim, A. et al. Corrosion inhibitors. A review. *Int. J. Corros. Scale Inhib.* **10**, 54–67 (2021).
4. Miralrio, A. & Vázquez, A. E. Plant extracts as green corrosion inhibitors for different metal surfaces and corrosive media: a review. *Processes* **8**, 8 (2020).
5. Keshavarz, M. H. et al. Simple approach to predict corrosion inhibition efficiency of imidazole and benzimidazole derivatives as well as linear organic compounds containing several polar functional groups. *Zeitschrift fur Anorg. und Allg. Chemie* **642**, 906–913 (2016).
6. Majd, M. T. et al. Probing molecular adsorption/interactions and anti-corrosion performance of poppy extract in acidic environments. *J. Mol. Liq.* **304**, 112750 (2020).
7. Alrefaee, S. H. et al. Challenges and advantages of using plant extract as inhibitors in modern corrosion inhibition systems: Recent advancements. *J. Mol. Liq.* **321**, 114666 (2021).
8. Salleh, S. Z. et al. Plant extracts as green corrosion inhibitor for ferrous metal alloys: a review. *J. Clean. Prod.* **304**, 127030 (2021).
9. El-Haddad, M. N. & Fouda, A. E. A. S. Evaluation of Curam drug as an ecofriendly corrosion inhibitor for protection of stainless steel-304 in hydrochloric acid solution: chemical, electrochemical, and surface morphology studies. *J. Chinese Chem. Soc.* **68**, 826–836 (2021).
10. Farahati, R. et al. Experimental and computational study of penicillamine drug and cysteine as water-soluble green corrosion inhibitors of mild steel. *Prog. Org. Coatings.* **142**, 105567 (2020).
11. Neves, B. J. et al. QSAR-based virtual screening: advances and applications in drug discovery. *Front. Pharmacol.* **9**, 1–7 (2018).
12. Toropov, A. A. & Toropova, A. P. QSPR/QSAR: state-of-art, weirdness, the future. *Molecules* **25**, 1292 (2020).
13. Belghiti, M. E. et al. Computational simulation and statistical analysis on the relationship between corrosion inhibition efficiency and molecular structure of some hydrazine derivatives in phosphoric acid on mild steel surface. *Appl. Surf. Sci.* **491**, 707–722 (2019).
14. Winkler, D. A. Predicting the performance of organic corrosion inhibitors. *Metals* **7**, 1–8 (2017).
15. Mendoza, R. L. C. et al. Density functional theory and electrochemical studies: structure-efficiency relationship on corrosion inhibition. *J. Chem. Inf. Model.* **55**, 2391–2402 (2015).
16. Li, L. et al. The discussion of descriptors for the QSAR model and molecular dynamics simulation of benzimidazole derivatives as corrosion inhibitors. *Corros. Sci.* **99**, 76–88 (2015).
17. Khaled, K. F. & Al-Mobarak, N. A. A predictive model for corrosion inhibition of mild steel by thiophene and its derivatives using artificial neural network. *Int. J. Electrochem. Sci.* **7**, 1045–1059 (2012).
18. Khaled, K. F. & Sherik, A. Using neural networks for corrosion inhibition efficiency prediction during corrosion of steel in chloride solutions. *Int. J. Electrochem. Sci.* **8**, 9918–9935 (2013).
19. Quadri, T. W. et al. Development of QSAR-based (MLR/ANN) predictive models for effective design of pyridazine corrosion inhibitors. *Mater. Today Commun.* **30**, 103163 (2022).
20. Ser, C. T. et al. Prediction of corrosion inhibition efficiency of pyridines and quinolines on an iron surface using machine learning-powered quantitative structure-property relationships. *Appl. Surf. Sci.* **512**, 145612 (2020).
21. Liu, Y. et al. 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).
22. Zhi, Y. et al. Long-term prediction on atmospheric corrosion data series of carbon steel in China based on NGBM(1,1) model and genetic algorithm. *Anti-Corrosion Methods Mater* **66**, 403–411 (2019).
23. De Masi, G. et al. Machine learning approach to corrosion assessment in subsea pipelines. *MTS/IEEE Ocean. 2015 - Genova Discovering Sustainable Ocean Energy for a New World* 8–13 (2015).
24. Beltran-Perez, C. et al. A general use QSAR-ARX model to predict the corrosion inhibition efficiency of drugs in terms of quantum mechanical descriptors and experimental comparison for lidocaine. *Int. J. Mol. Sci.* **23**, 5086 (2022).
25. Chen, Z. S. et al. A PSO based virtual sample generation method for small sample sets: applications to regression datasets. *Eng. Appl. Artif. Intell.* **59**, 236–243 (2017).
26. Roy, M. A. et al. Machine-learning-guided descriptor selection for predicting corrosion resistance in multi-principal element alloys. *npj Mater. Degrad.* **6**, 9 (2022).
27. Coelho, L. B. et al. Reviewing machine learning of corrosion prediction in a data-oriented perspective. *npj Mater. Degrad.* **6**, 8 (2022).
28. Chen, Z. S. et al. Integrating virtual sample generation with input-training neural network for solving small sample size problems: application to purified terephthalic acid solvent system. *Soft Comput.* **25**, 6489–6504 (2021).
29. Li, D. C. et al. A new approach for manufacturing forecast problems with insufficient data: the case of TFT-LCDs. *J. Intell. Manuf.* **24**, 225–233 (2013).
30. Luo, H. & Paal, S. G. Reducing the effect of sample bias for small data sets with double-weighted support vector transfer regression. *Comput. Civ. Infrastruct. Eng.* **36**, 248–263 (2021).
31. Asanya, K. C. et al. Robust Bayesian approach to logistic regression modeling in small sample size utilizing a weakly informative student's t prior distribution. *Commun. Stat. Theory Methods.* **52**, 1–11 (2021).
32. Wang, X. & Yao, J. Linear regression estimation methods for inferring standard values of snow load in small sample situations. *Math. Probl. Eng.* **2020**, 1–10 (2020).
33. Liu, Q. et al. A new support vector regression model for equipment health diagnosis with small sample data missing and its application. *Shock Vib.* **2021** (2021). https://doi.org/10.1155/2021/6675078.
34. Liu, B. et al. Small dataset modeling and application of plant medicine extraction. *Commun. Comput. Inform. Sci.* **1006**, 381–392 (2019).
35. Li, D. C. et al. Using virtual samples to improve learning performance for small datasets with multimodal distributions. *Soft Comput.* **23**, 11883–11900 (2019).
36. Raikwal, J. S. & Saxena, K. Performance evaluation of SVM and K-nearest neighbor algorithm over medical data set. *Int. J. Comput. Appl.* **50**, 35–39 (2012).
37. Kumar, S. & Chong, I. Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. *Int. J. Environ. Res. Public Health.* **15**, 2907 (2018).
38. Vettoretti, M. & Di Camillo, B. A variable ranking method for machine learning models with correlated features: In-silico validation and application for diabetes prediction. *Appl. Sci.* **11**, 7740 (2021).
39. Moedjahedy, J. et al. CCrFS: combine correlation features selection for detecting phishing websites using machine learning. *Futur. Internet.* **14**, 229 (2022).
40. Ying, X. An overview of overfitting and its solutions. *J. Phys. Conf. Ser.* **1168**, 022022 (2019).
41. Yan, L. et al. 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).
42. Giola, C. et al. Learning curves: a novel approach for robustness improvement of load forecasting. *Eng. Proc.* **5**, 38 (2021).
43. Brumen, B. et al. Overview of machine learning process modelling. *Entropy* **23**, 1123 (2021).
44. Hassan, A. H. M. et al. Visualization & prediction of COVID-19 future outbreak by using machine learning. *Int. J. Inf. Technol. Comput. Sci.* **13**, 16–32 (2021).
45. Song-Qing, H. et al. 3D-QSAR study and molecular design of benzimidazole derivatives as corrosion inhibitors. *Chem. J. Chinese Univ.* **32**, 2402 (2011).
46. Kozlica, D. K. et al. Synergistic effect of 2-mercaptobenzimidazole and octylphosphonic acid as corrosion inhibitors for copper and aluminium – An electrochemical, XPS, FTIR and DFT study. *Corros. Sci.* **182**, 109082 (2021).
47. Kokalj, A. Corrosion inhibitors: physisorbed or chemisorbed? *Corros. Sci.* **196**, 109939 (2022).
48. Imandoust, S. B. & Bolandraftar, M. Application of K-nearest neighbor (KNN) approach for predicting economic events: theoretical background. *Int. J. Eng. Res. Appl.* **3**, 605–610 (2013).
49. Sutojo, T. et al. Investigating the impact of synthetic data distribution on the performance of regression models to overcome small dataset problems. *Proc. 2020 International Seminar on Application for Technology of Information and Communication* (*iSemantic*). 125–130 (IEEE, 2020).
50. Xu, Q. S. & Liang, Y. Z. Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.* **56**, 1–11 (2001).
51. Jun Lv, Y. et al. Steel corrosion prediction based on support vector machines. *Chaos Solitons Fractals.* **136**, 109807 (2020).
52. Zhou, C. et al. A novel stacking heterogeneous ensemble model with hybrid wrapper-based feature selection for reservoir productivity predictions. *Complexity* **2021**, 1–12 (2021).
53. Zhang, Y. et al. Data augmentation strategy for small sample short-term load forecasting of distribution transformer. *Int. Trans. Electr. Energy Syst.* **30**, e12209 (2019).
54. Scikit-learn. Scikit-learn user guide - Release 0.23.2. (2020).

ACKNOWLEDGEMENTS

The author sincerely thanks LPPM Dian Nuswantoro University Semarang Indonesia, for the support. T.S. acknowledges scholarship support from Dian Nuswantoro University Semarang Indonesia. M.A. acknowledges scholarship support from LPDP (Lembaga Pengelola Dana Pendidikan) Indonesia.

AUTHOR CONTRIBUTIONS

T.S. designed the algorithms, provided technical expertize on machine learning, and wrote the manuscript. M.A. extracted data from the literature, performed DFT analysis, and drafted the manuscript. S.R. and H.K.D. contributed to the main idea and methodology, interpreted the results and discussion, and reviewed the manuscript. A.S. and G.F.S. performed the data analysis.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41529-023-00336-7.

Correspondence and requests for materials should be addressed to Supriadi Rustad.

Reprints and permission information is available at http://www.nature.com/reprints

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