International Journal of Advances in Data and Information Systems
Vol. 5, No. 2, October 2024, pp. 198~215
ISSN: 2721-3056, DOI: 10.59395/ijadis.v5i2.1346

Ensemble Stacking of Machine Learning Approach for Predicting Corrosion Inhibitor Performance of Pyridazine Compounds

Noval Ariyanto[1], Harun Al Azies[2], Muhamad Akrom[2]
[1]Study Program in Informatics Engineering, Faculty of Computer Science, Dian Nuswantoro University, Indonesia
[2]Research Center for Quantum Computing and Materials Informatics, Faculty of Computer Science, Dian Nuswantoro University, Indonesia

Article Info

Article history:

Received Oct 10, 2024
Revised Oct 19, 2024
Accepted Oct 30, 2024

Keywords:

Corrosion
Pyridazine
Machine Learning
Stacking Ensemble
Regression

ABSTRACT

Corrosion is a major challenge affecting various industrial sectors, leading to increased operational costs and decreased equipment efficiency. The use of organic corrosion inhibitors is one of the promising solutions. This study applies an ensemble algorithm with a stacking method to estimate pyridazine-derived compounds corrosion inhibition efficiency. This study utilized various molecular characteristics of pyridazine compounds as inputs to predict inhibition efficiency values. After evaluating several boosting models, the stacking technique was chosen as it showed the best results. Stacking Model 6, which combines XGB, LGBM, and CatBoost as the base model with Random Forest as the meta-model, produced the most accurate prediction with an RMSE of 0.055. These findings indicate that machine learning approaches can effectively and efficiently predict corrosion inhibitor performance. This method offers a faster and more economical alternative to conventional experimental methods.

Corresponding Author:

Muhamad Akrom,
Research Center for Quantum Computing and Materials Informatics
Faculty of Computer Science, Dian Nuswantoro University,
207 Imam Bonjol Street, Pendrikan Kidul, Semarang City, Central Java 50131, Indonesia.
Email: m.akrom@dsn.dinus.ac.id

1. INTRODUCTION

Corrosion is a serious problem many industries face, leading to increased production and maintenance costs and decreased equipment efficiency [1]. Factors such as water, moisture, acidic materials, oil, and high-acidity gases are corrosion triggers [2]. Corrosion can lead to decreased metal thickness, potentially serious structural damage, stress corrosion cracking, decreased mechanical strength, and even sudden material failure [3]. The impact of corrosion is significant on economic and operational costs [4].

The use of inhibitors, especially those based on organic compounds, is becoming an increasingly popular solution due to their effectiveness and environmental friendliness. Heterocyclic derivatives such as benzotriazoles have demonstrated promising protective capabilities against industrial metals, particularly copper and its streams [5]. Nonetheless, large-scale implementation of these compounds still faces several obstacles, including limited thermal resistance and economic considerations related to the manufacturing process [6]. Despite these challenges, the development trend of organic inhibitors continues, driven by their advantages in terms of environmental compatibility and corrosion inhibition efficiency (IE). Organic compounds

can form a protective film on metal surfaces, preventing direct contact with corrosive agents. Pyridazine-derived compounds show great potential as corrosion inhibitors [7].

To address the complexity of estimating corrosion inhibitor effectiveness, implementing advanced machine learning methods, specifically stacking techniques offers the potential to produce more comprehensive and accurate solutions. This research focuses on stacking boosting machine learning models to provide more accurate prediction results than single models in the context of corrosion inhibitor performance prediction.

Stacking models offer a new perspective in predicting corrosion inhibition effectiveness for pyridazine-derived compounds, an approach that still needs to be applied in this field. The model aims to identify the most effective predictions in assessing corrosion inhibition ability by integrating various machine learning algorithms [8]. The advantage of this method lies in its ability to combine the strengths of multiple models while mitigating individual weaknesses, potentially optimizing estimation accuracy [9]. The study also included an in-depth analysis of molecular characteristics, opening up a broader understanding of the correlation between a compound's structure and its corrosion-inhibiting ability. The results showed that the combination of stacking models consistently yielded higher prediction accuracy than single models, promising more accurate and reliable estimates than previous methods. Thus, this research contributes significantly to developing more sophisticated corrosion inhibition prediction methods and paves the way for further applications of advanced machine learning techniques in materials science and applied chemistry.

The importance of this research lies in the faster and cost-effective solution compared to traditional experimental methods that require significant time, cost, and resources. Utilizing Machine Learning technology, particularly the stacking boosting model, enables more efficient and effective evaluation and prediction of corrosion inhibitor performance, which can ultimately support the development of new, more corrosion-resistant materials [10].

Combining the advantages of stacking ensemble models and comprehensive feature analysis, this research is expected to produce more accurate and reliable estimates than previous methods. The results of this research contribute significantly to the development of more advanced corrosion inhibition prediction methods and pave the way for further applications of advanced machine learning techniques in materials science and applied chemistry.

2. RESEARCH METHOD

This research applies a four-stage framework to analyze corrosive substances in compounds. This methodological approach aims to understand the components that affect compounds and develop accurate prediction models. Each stage has specific objectives to improve the overall success of the research. Details of the research framework are presented in Figure 1.

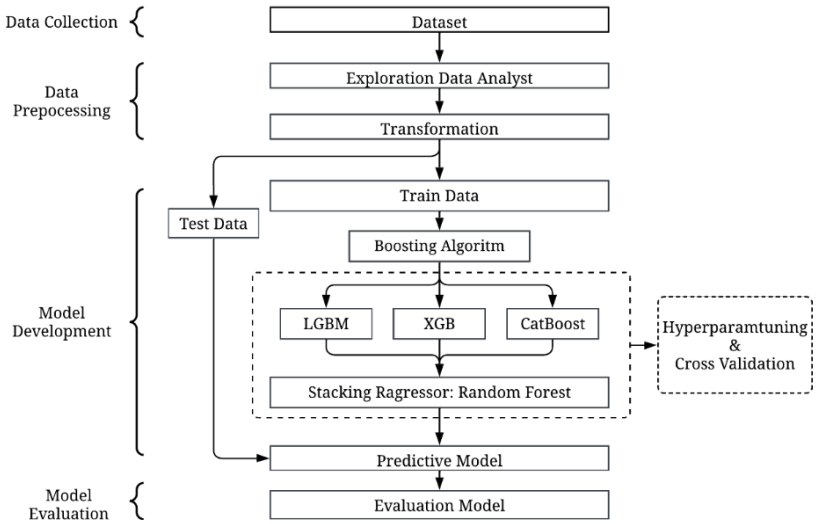

Figure 1. Research Method

2.1. Data Collection

The dataset comprises 120 pyridazine compounds with quantum molecular properties as features (independent variables) and IE values as targets (dependent variables) [11]. Among these properties, the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO) are critical as they determine the electronic structure of the compounds. HOMO represents the highest energy level of electrons that can be removed, indicating the molecule's electron-donating ability. At the same time, LUMO signifies the lowest energy level that can accept electrons, reflecting its electron-accepting potential. The gap energy (ΔE), defined as the difference between the HOMO and LUMO energies, is a vital descriptor of molecular stability and reactivity—lower ΔE values generally correlate with increased reactivity. The dipole moment (μ) quantifies the molecule's polarity, influencing its interactions with solvents and biological targets, significantly impacting its IE. Ionization potential (I) and electron affinity (A) provide insight into the energies required for removing or adding electrons. high ionization potential suggests excellent stability, whereas high electron affinity indicates a stronger tendency to accept electrons. Electronegativity (χ) reflects the capacity of an atom to attract electrons within a bond, impacting the distribution of electronic density and, thus, the reactivity of the compound. Global hardness (η) and global softness (σ) are related descriptors that measure the molecule's resistance to change in electron distribution. Greater hardness signifies lower reactivity, whereas increased softness suggests a greater propensity for chemical interaction. The electrophilicity index (ω) encapsulates the molecule's overall ability to act as an electrophile, derived from ionization potential and electron affinity, offering insights into its reactivity. Finally, the fraction of transferred electrons (ΔN) indicates the extent of electron transfer in chemical reactions, providing a quantitative measure of the interaction dynamics between the pyridazine compounds and potential targets. These quantum molecular descriptors offer a comprehensive framework for understanding the electronic properties that govern the IE of pyridazine compounds in various chemical and biological contexts [12].

2.2. Data Preprocessing

The steps illustrated in Figure 1 are crucial for data preparation before regression model implementation. The data pre-processing procedure includes a series of essential techniques to ensure the integrity and reliability of the dataset to be used in regression modelling [13].

2.2.1 Exploration Data Analysis

The Exploratory Data Analysis (EDA) implemented in this study is a crucial phase that involves a comprehensive set of analytical procedures. The main objective is to acquire a comprehensive understanding of the distribution characteristics of variables, detection of outlier values, and evaluation of correlations between variables in the data set [14]. EDA plays a vital role in uncovering hidden patterns and trends that may go unnoticed through surface analysis [15]. The importance of EDA lies in its ability to identify elements that could affect the performance of the predictive model to be built [16]. Through this methodology, researchers can gain a more detailed insight into the intrinsic structure of the data, which in turn contributes to the development of models with higher precision and reliability.

2.2.2 Transformation

The dependent variable was transformed logarithmically to address non-normality and stabilize variance, a common approach in regression analysis [17]. For the independent variables, transformation methods were selected based on skewness analysis of their data distributions. This process ensures that the relationship between predictors and the response variable remains accurate and minimally distorted [18]. This study aims to optimize model performance, improve predictive accuracy, and build a robust, generalizable model that offers deeper insights into the phenomenon under investigation by following a structured pre-processing approach grounded in statistical principles.

2.3. Model Development

This research implements an advanced stacking technique, combining three leading boosting algorithms: Light Gradient Boosting Machine (LGBM), Extreme Gradient Boosting (XGB) and CatBoost. Each algorithm is optimized through an extensive hyperparameter tuning process, resulting in two best models for each algorithm. These six models were then integrated into a stacking structure, where their outputs were used as inputs for a Random Forest-based meta-model. The main evaluation metric used is the Root Mean Square Error (RMSE), a key indicator in the optimization process and model performance evaluation. RMSE was chosen for its ability to measure the deviation of the prediction from the true value, providing a comprehensive picture of the model's accuracy [19]. This approach utilizes each boosting algorithm's strengths and optimizes the generalization ability through the combination of predictions made by Random Forest. The main goal of this methodology is to produce a more accurate and robust predictive model capable of outperforming traditional approaches in handling data complexity and variability [20].

2.3.1 Boosting

Boosting is a powerful ensemble learning technique in machine learning designed to improve the predictive accuracy of models by combining multiple weak learners into one strong learner [21]. The basic principle is to build a series of models sequentially, where each subsequent model attempts to correct the mistakes made by the previous model [22]. In this process, boosting pays more attention to data instances that are difficult to predict, allocating higher weights to cases that the previous model misclassified.

2.3.2 Light Gradient Boosting Machine (LGBM)

Light Gradient Boosting Machine (LGBM), introduced by Guolin Ke, is designed to improve computational efficiency and scalability, even for smaller datasets [23]. While its key innovations, such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), are particularly beneficial for large-scale datasets, LGBM also excels in handling smaller datasets due to its ability to optimize resource usage and maintain high accuracy. LightGBM uses a leaf-wise tree growth approach, allowing for more complex modelling while being computationally efficient. These advantages make LightGBM suitable for small-scale data analysis, providing accurate results within time and resource constraints, even with a small dataset.

2.3.3 Extreme Gradient Boosting (XGB)

Extreme Gradient Boosting (XGB), developed by Chen and Guestrin in 2016, is an optimized implementation of the gradient-boosting algorithm [24]. Its main advantages lie in applying effective regularization to reduce overfitting and parallelization capabilities that improve computational efficiency. XGB performs superiorly in various data science competitions and industrial applications. Mathematically, XGB seeks to minimize the loss function L in the training data by incrementally adding decision tree models fk to boost iterations. At each iteration t, a new model is added to correct the prediction error of the previous model, enabling progressive improvement in prediction accuracy.

2.3.4 Category Boosting (CatBoost)

Category Boosting (CatBoost), developed by Prokhorenkova, is designed to handle categorical data effectively [25]. It uses ordered boosting techniques to minimize prediction bias and automatically generate combinations of categorical features. CatBoost's main advantage lies in its ability to handle categorical variables without manual pre-processing, making it an optimal choice for datasets dominated by categorical features. This is particularly beneficial in various research domains, such as consumer behaviour analysis, natural language processing, and health data analysis, where categorical variables are often a key component in modelling.

2.3.4 Stacking

Figure 2 illustrates the stacking ensemble method employed in this study, where predictions from six optimized models—LGBM1, LGBM2, XGB1, XGB2, CatBoost1, and

CatBoost2—are combined. To achieve better prediction performance, these models were fine-tuned using cross-validation and hyperparameter optimization via Optuna.

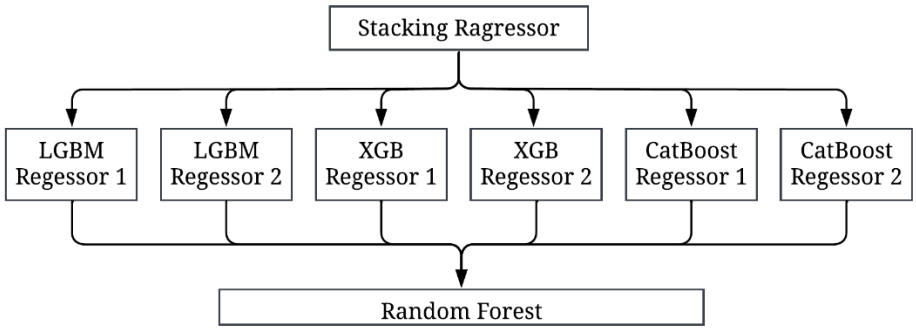

Figure 2. Stacking Regressor Framework Utilizing LGBM, XGB, and CatBoost Models with Random Forest as the Meta-model

2.4 Leave One Out Cross Validation

The selection of Leave-One-Out (LOO) Cross-Validation in this study is motivated by the relatively small dataset comprising 120 observations, making it a highly suitable choice for maximizing the use of available data [26]. In LOO, each observation serves as a validation set one time, while the remaining data points are utilized to train the model [27]. This approach ensures that the model is trained and tested on nearly the entire dataset, thereby minimizing the potential loss of valuable information if a significant portion of the data were allocated solely for training purposes [28]. This phenomenon occurs when the model becomes overly tailored to the training data, diminishing its ability to generalize effectively to new data. LOO unique methodology allows each observation to contribute meaningfully to model training while retaining all data for evaluation, thereby enhancing the accuracy and reliability of performance estimates [29]. Consequently, utilizing LOO not only bolsters the validity of the findings in this study but also provides a more nuanced understanding of the model's performance when dealing with limited data. As a result, LOO is an effective tool for ensuring that the developed model is robust and reliable, facilitating sound decision-making based on data-driven insights in practical applications.

2.5 Model Evaluation

In regression model evaluation, RMSE is one of the important evaluation metrics. RMSE measures the root mean square error between the value predicted by the model and the actual value, with a lower RMSE indicating better model performance [30]. Although accuracy is not used in the context of regression as in classification, RMSE gives an idea of how close the model predictions are to the actual values [31]. Given the relatively small dataset in this study, the focus is primarily on prediction accuracy through RMSE, as computation time is less of a concern in this context. Regression model evaluation aims to select the most efficient and accurate model by considering various performance aspects and specific application needs.

3. RESULTS AND DISCUSSION

This study investigated various ensemble algorithms comprehensively to identify the optimal predictive model for IE of pyridazine-derived compounds. The selection process began with an extensive evaluation of various ensemble algorithms in the context of regression. Based on the performance analysis using the RMSE metric, three superior ensemble models were selected, and two models with different parameters were chosen for each algorithm. The algorithms that demonstrated superior performance and were selected for further analysis were LGBM, XGB, and CatBoost. This selection was based on the ability of each algorithm to optimize the accuracy of IE prediction, which is a crucial parameter in characterizing the effectiveness of pyridazine-based corrosion inhibitors.

3.1. Overview

This research applied the EDA methodology to gain a more comprehensive understanding of the dataset relating to corrosion rates in pyridazine compounds. This is a significant stage in the data analysis, as it allows researchers to identify the relevant variables' patterns, correlations, and key characteristics. The study was initiated with multivariate correlation analysis, a fundamental step in understanding the structure and interrelationships in the dataset. In the context of research into the corrosion phenomena of pyridazine compounds, this analysis will explore significant relationships between variables to reveal key factors that influence the corrosion process.

Table 1. Descriptive Statistics

Variable	Min	Mean	Max
HOMO	-7.945	-6.132	-4.477
LUMO	-4.655	-1.848	6.294
gap energy	1.897	4.289	13.536
dipole moment	0.847	4.599	13.371
ionization potential	4.477	6.136	7.945
electron affinity	-6.294	1.848	4.655
electronegativity	0.429	3.992	5.971
global hardness	0.949	2.147	6.768
global softness	0.148	0.536	1.054
electrophilicity	-1.212	1.827	16.298
fraction of transferred electrons	-0.012	0.444	0.851

Descriptive statistical analysis of corrosion-relevant variables in pyridazine compounds revealed significant characteristics. The HOMO and LUMO values showed asymmetric distributions. HOMO values ranged from -7.945 to -4.477, with a mean of -6.132, indicating that while some compounds exhibit lower HOMO energies, which suggests a more vital ability to donate electrons, others have higher energies that may limit their reactivity. In contrast, LUMO values displayed more significant variability, ranging from -4.655 to 6.294, with a mean of -1.848. A higher LUMO value generally indicates an increased ability to accept electrons, critical in determining a compound's electrophilic nature. When LUMO values are high, the compounds are likely to be more reactive in corrosion processes, as they can attract electrons from the environment more effectively. Thus, the combination of low HOMO and high LUMO values may indicate particularly reactive compounds, potentially enhancing corrosion rates.

The gap energy (ΔE) displayed substantial differences, ranging from 1.897 to 13.536, signifying considerable diversity in energy characteristics among the molecules. A smaller ΔE typically suggests increased reactivity, as less energy is required for electron transitions between the HOMO and LUMO. Compounds with lower ΔE values may readily participate in chemical reactions, including those leading to corrosion. Compounds with higher ΔE values may exhibit more stability and lower reactivity, potentially resulting in reduced corrosion rates.

The dipole moment (μ) and electrophilicity, represented by the fraction of transferred electrons (ΔN), exhibited the most extensive variation among the analyzed parameters, with dipole moments ranging from 0.847 to 13.371. This suggests substantial heterogeneity in polarity among the studied compounds. Higher dipole moments typically correlate with stronger intermolecular interactions, influencing the solubility and reactivity of the compounds in corrosive environments. For instance, compounds with high dipole moments may interact more favourably with polar solvents, enhancing their corrosion susceptibility. The electrophilicity, as measured by ΔN, ranged from -1.212 to 16.298, further emphasizing the variability in electron transfer tendencies among the molecules. Compounds exhibiting higher electrophilicity values are more likely to act as electrophiles in corrosion reactions, thus facilitating the corrosion process. This highlights the importance of understanding electron transfer dynamics in assessing the corrosion behaviour of pyridazine compounds.

Ionization potential (I) and electronegativity (χ) showed more moderate distributions, ranging from 4.477 to 7.945 and 0.429 to 5.971, respectively. Ionization potential indicates the energy required to remove an electron, with higher values suggesting increased stability and reduced reactivity. Conversely, lower ionization potential can make a compound more susceptible to corrosion due to its enhanced ability to lose electrons. Electronegativity reflects the ability of an

atom to attract electrons; thus, higher electronegativity can influence the reactivity of pyridazine compounds in corrosive environments.

Global hardness (η) and global softness (σ) provided complementary insights into chemical reactivity, with global hardness ranging from 0.949 to 6.768 and global softness varying between 0.148 and 1.054. High global hardness indicates that a compound is less reactive, whereas increased softness suggests a greater propensity for chemical interaction. These descriptors are crucial in understanding how pyridazine compounds behave in corrosive conditions, where softer compounds may more readily participate in corrosive reactions.

The diversity in statistical values highlights the complexity of molecular interactions in the context of pyridazine compound corrosion. High variability in specific parameters, particularly electrophilicity and energy gap (ΔE), suggests that further analysis is necessary to identify critical factors contributing to corrosion. These findings pave the way for in-depth research on the relationship between molecular characteristics and corrosion tendencies, with potential applications in developing more accurate predictive models. By comprehensively examining these features, the study aims to provide valuable insights into the mechanisms underlying corrosion phenomena in pyridazine compounds, ultimately contributing to developing strategies to mitigate corrosion in practical applications.

3.2. Correlation and Significance

Correlation measures how strong the relationship is between two variables in a statistical analysis. The correlation value ranges from -1 to 1, where a value close to 1 indicates a strong positive relationship between two variables, meaning that an increase usually follows an increase in one variable in the other variable, while if the correlation value is close to -1, it means that there is a negative relationship, where a decrease follows an increase in one variable in the other variable, while a value close to 0 indicates that there is no significant linear relationship between the two variables [32].

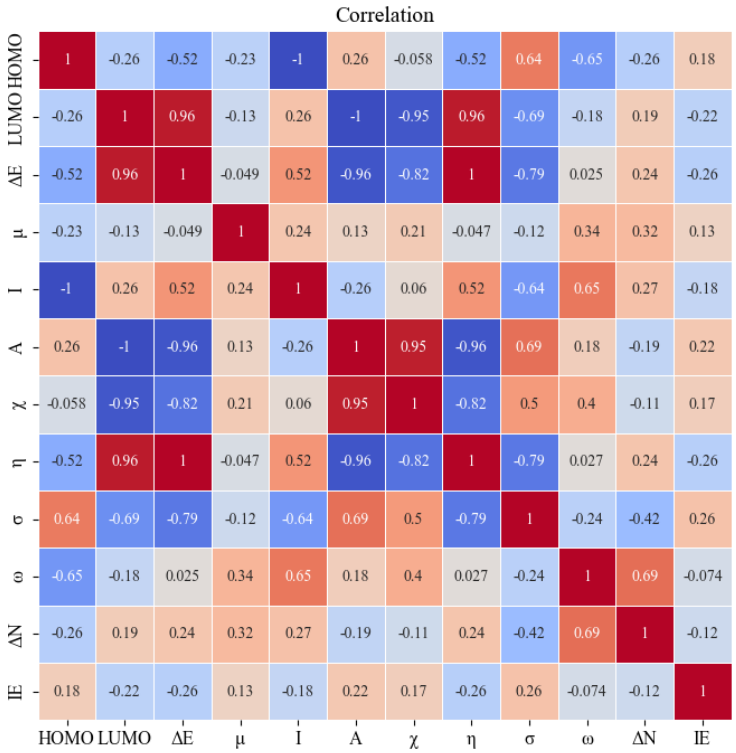

Figure 3. Correlation Matrix of Molecular Properties in Pyridazine Compounds

Figure 3 illustrates the Correlation Matrix of Molecular Properties in Pyridazine Compounds, presenting several critical relationships between the chemical parameters essential for understanding molecular behaviour and stability. A strong positive correlation between HOMO and

LUMO (r = 0.96) suggests that as HOMO energy increases, LUMO energy follows, reflecting the overall stability of the molecules and their tendency to gain or lose electrons, a critical factor in chemical reactions. This is particularly relevant for pyridazine compounds, where the stability of the electronic structure influences their potential applications in areas such as catalysis or corrosion inhibition. The matrix also reveals significant positive correlations between electronegativity (χ) and electron affinity (A) (r = 0.95), implying that more electronegative molecules have a higher tendency to attract electrons, which could enhance their ability to participate in electron transfer reactions, a key feature for designing efficient inhibitors. Similarly, the high correlation between electronegativity (χ) and dipole moment (μ) (r = 0.95) indicates that electronegative molecules tend to exhibit more excellent dipole moments, affecting molecular interactions, solubility, and overall reactivity, which could impact their efficiency as corrosion inhibitors by influencing how they interact with metal surfaces.

In contrast, strong negative correlations are observed between HOMO and electronegativity (χ) (r = -0.95) and between LUMO and electronegativity (χ) (r = -0.96), suggesting that molecules with higher HOMO and LUMO energy levels tend to be less electronegative and release electrons more easily. This is crucial for analyzing the electron-donating properties of molecules in reactions, especially in the context of pyridazine compounds, where their ability to donate electrons may affect their reactivity and potential as corrosion inhibitors. Another significant relationship is the negative correlation between electronegativity (χ) and global hardness (η) (r = -0.82), suggesting that molecules with higher electronegativity are generally softer and more chemically reactive. This is an essential factor when considering pyridazine compounds' reactivity and chemical stability, as softer molecules may more readily participate in chemical reactions, which could enhance their performance in preventing corrosion.

These correlations reveal the intrinsic properties of pyridazine compounds and provide insights into how these molecular features might influence their behaviour in practical applications, such as corrosion inhibition. By understanding these relationships, the study sheds light on how these compounds' electronic and chemical properties could be tailored for specific industrial purposes, thus supporting the overall findings of the research.

The significance of a correlation refers to the extent to which the relationship between two variables can be considered real or to have occurred by chance. In the context of corrosion inhibition, the significance of the correlation between a variable and the target corrosion inhibition variable is determined based on the p-value. The p-value is the probability of obtaining a result equal to or more extreme than the observed result, assuming that the null hypothesis is true or that there is no relationship [33].

Table 2. Significance of Variables Based on P-Value

Variable	P-Value	Significance
HOMO	0.0464	No
LUMO	0.0142	No
gap energy	0.0046	Yes
dipole moment	0.1422	No
ionization potential	0.0503	No
electron affinity	0.0142	No
electronegativity	0.0602	No
global hardness	0.0048	Yes
global softness	0.0035	Yes
electrophilicity	0.4188	No
fraction of electrons transferred	0.1984	No

In Table 2, a commonly used p-value of 0.05. This value indicates (α) = 5% significance level, meaning there is a 5% chance that the observed relationship occurred by chance. Using a threshold of 0.05 is a frequently chosen practice in statistical analysis because it is considered a reasonable compromise between the type I error rate of incorrectly rejecting the null hypothesis and the desire to detect real relationships. According to academic standards and established practice, this 0.05 threshold provides a good balance between sensitivity and specificity in hypothesis testing, as described by Fisher, one of the pioneers in statistics [34].

a. Relationship Between IE and HOMO

A weak positive correlation (r = 0.18) is observed between IE and HOMO, which is marginally insignificant at α = 0.05 (p-value = 0.0464). Although HOMO is linked to the molecule's ability to donate electrons, this weak association suggests that HOMO is not a primary determinant of the corrosion inhibition efficiency in pyridazine compounds.

b. Relationship Between IE and LUMO

There exists a weak negative correlation (r = -0.22) between IE and LUMO, which is insignificant (p-value = 0.0142). This implies that the electron-accepting capacity of pyridazine molecules plays a relatively minor role in influencing their IE.

c. Relationship Between IE and gap energy

A significant negative correlation (r = -0.26) was found between IE and gap energy (p-value = 0.0046). This indicates that pyridazine compounds with smaller energy gaps demonstrate higher inhibition efficiencies, likely due to increased molecular reactivity and adsorption potential on metal surfaces.

d. Relationship Between IE and dipole moment

IE shows a weak positive correlation with dipole moment (r = 0.13), though the result is not statistically significant (p-value = 0.1422). This suggests that the polarity of pyridazine molecules has a minimal impact on their IE.

e. Relationship Between IE and ionization potential

A weak negative correlation (r = -0.18) between IE and ionization potential is noted, with results nearing but not reaching statistical significance (p-value = 0.0503). While close to the threshold, ionization potential does not appear to be a key factor influencing the IE of pyridazine compounds.

f. Relationship Between IE and electron affinity

The correlation between IE and electron affinity is weakly positive (r = 0.22) and not significant (p-value = 0.0142). This suggests that the electron-accepting ability of pyridazine molecules contributes little to their overall IE.

g. Relationship Between IE and electronegativity

A weak positive correlation (r = 0.17) between IE and electronegativity was observed, with the result not reaching significance (p-value = 0.0602). Therefore, electronegativity does not play a major role in determining the IE of these molecules.

h. Relationship Between IE and global hardness

A significant negative correlation (r = -0.26) was found between IE and global hardness (p-value = 0.0048). This suggests that pyridazine compounds with lower hardness, or "softer" compounds, tend to exhibit greater IE, potentially due to enhanced adsorption and interaction with metal surfaces.

i. Relationship Between IE and global softness

A significant positive correlation (r = 0.26) between IE and global softness (p-value = 0.0035) supports previous conclusions. Softer pyridazine compounds demonstrate higher IE, aligning with the Hard and Soft Acids and Bases (HSAB) theory principles in the context of corrosion inhibition.

j. Relationship Between IE and electrophilicity

A weak negative correlation (r = -0.07) between IE and electrophilicity is observed, with insignificant results (p-value = 0.4188). This finding indicates that the electrophilic character of pyridazine molecules has little to no effect on their IE.

k. Relationship Between IE and a fraction of electron transferred

The correlation between IE and the fraction of electrons transferred is weak and negative (r = -0.12), with no statistical significance (p-value = 0.1984). This suggests that electron transfer fraction is not a primary driver of IE in pyridazine compounds.

This analysis shows that gap energy, global hardness, and global softness significantly correlate with the IE of pyridazine compounds. These findings suggest that properties related to the chemical reactivity and electronic flexibility of the molecule play an important role in the effectiveness of pyridazine as a corrosion inhibitor. Other factors, although correlated, did not show strong statistical significance, indicating the complexity of the corrosion inhibition mechanism involving the interaction of various molecular parameters.

Outlier detection is identifying data that is significantly different from most other data [35]. It serves several important functions, such as helping to identify data errors that may have occurred during collection or processing, understanding rare phenomena that may be relevant for analysis, and improving model quality by removing irrelevant data [36]. Outliers can harm statistical and machine learning models, as they tend to cause bias in the model, reduce prediction accuracy, and can result in overfitting, where the model focuses too much on extreme values that do not represent the true pattern [37]. Several methods can handle outliers, such as deleting the data if it is considered incorrect, transforming the data to adjust the scale, or using models more resistant to outliers, such as isolation forests. Thus, detecting and handling outliers is very important so that the model can produce accurate predictions and not be disturbed by data that does not reflect the general situation.

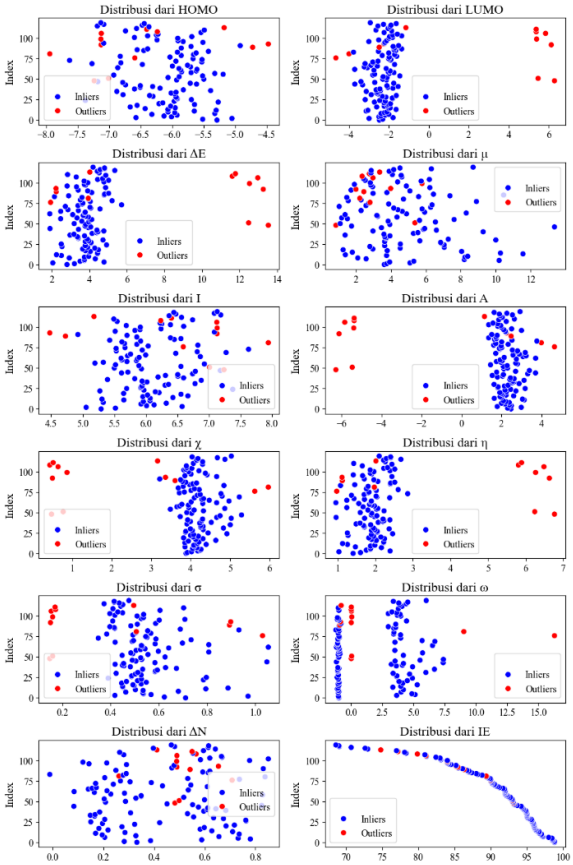

Figure 4. Outlier Distribution of Key Molecular Properties Using Isolation Forest

In Figure 4, outlier analysis was conducted selectively using the isolation forest method to identify the outlier distribution. Data visualization was implemented using the matplotlib library

through scatter plot functions, facilitating the identification of extreme values significantly deviating from most observations.

In the HOMO visualization, outliers were detected in the range of -8 to -4.5. For LUMO, outliers were identified in the value spectrum of -4 to 6, with the main concentration of data around -1.8. The analysis of gap energy revealed outliers within the interval of 12 to 14, while most of the data were distributed between 2 and 6.

The evaluation of the dipole moment showed a broad data distribution, with outliers detected below the value of 6, while most observations were concentrated between 2 and 7. In the case of ionization potential, outliers were identified at both extremes of the distribution, although most data were concentrated between 5 and 7.5. The analysis of electron affinity revealed significant negative outliers around the value of -6, with the main concentration of data between 2 and 4.

The distribution of electronegativity exhibited a similar outlier pattern to electron affinity, with most data residing within the interval of 3 to 5. In global hardness, outliers were detected at high values between 5 and 7, while the main data concentration was 1 and 3. The analysis of global softness identified outliers at both extremes, particularly values below 0.2 and above 0.8, with most observations falling between 0.4 and 0.7.

The evaluation of electrophilicity showed an extensive outlier distribution, ranging from negative values to above 15, with the main data concentration between -1.5 and 7.5. Similarly, the fraction of electron transferred analysis revealed an outlier pattern akin to electrophilicity, spread across both distribution extremes, with most data between 0.1 and 0.8.

Identifying these outliers highlights the presence of extreme values that may impact the validity of the analysis results. In this context, handling outliers is crucial to ensure the accuracy and reliability of the developed statistical models. A comprehensive approach to data verification, the implementation of robust methods, and consideration of data transformation can contribute to mitigating the negative effects of outliers on the integrity of data analysis.

3.3. Model Development

This study developed a predictive model for corrosion analysis of pyridazine compounds using stacking. Random Forest was chosen as the meta-model and integrated with boosting-based algorithms, namely LGBM, XGB, and CatBoost, as the base model. The hyperparameter optimization process used Optuna with 100 iterations for each algorithm to identify the optimal parameter configuration based on RMSE performance. The two best models from each algorithm were selected for further analysis. The implementation of LOO cross-validation allowed a comprehensive evaluation of the stability and generalizability of the selected models.

Furthermore, the selected models were integrated through a stacking technique with Random Forest as the meta-model to optimize predictive capabilities. This approach is expected to improve the accuracy and robustness of predictions by utilizing complex patterns that individual models may miss. Through this systematic methodology, the research aims to produce an accurate and generalizable predictive model, allowing the exploration of complex interactions between predictor variables in the context of corrosion analysis of pyridazine compounds.

3.3.1. Light Gradient Boosting Machine

To improve prediction performance, a hyperparameter tuning process, as shown in Table 3, was performed using five main parameters, namely lambda_l1, lambda_l2, min_child_weight, min_data_in_leaf, and max_depth. These hyperparameters are essential in controlling the model's regularization, complexity, and generalization ability. lambda_l1 and lambda_l2 apply penalties to the model's weights to prevent overfitting by controlling how much the model depends on individual features. min_child_weight ensures that leaf nodes are created only when sufficient data points are present, reducing the risk of overfitting. min_data_in_leaf specifies the minimum number of data points in a leaf, preventing the creation of overly specific splits. Finally, max_depth limits how deep the trees can grow, balancing the model's ability to capture patterns while avoiding

overfitting. The tuning process was conducted across multiple iterations to achieve the best combination of parameters, leading to more accurate predictions.

Table 3. Optimized Parameters and RMSE Comparison for LGBM Models

	Parameter		RMSE
LGBM 1	A	0.555	0.0578
	B	0.001	
	C	3.021	
	D	20	
	E	10	
	Parameter		RMSE
LGBM 2	A	0.519	0.0576
	B	0.000	
	C	4.311	
	D	20	
	E	10	

note*: A = lamda_l1; B = lamda_l2; C = min_child_weight; D = min_data_in_leaf; E = max_depth;

The two LGBM models show a minimal difference in RMSE values, with LGBM 1 having an RMSE of 0.0578 and LGBM 2 having an RMSE of 0.0576, primarily due to regularisation parameter variations. The improvement in LGBM 2 is primarily driven by changes in A (lambda_l1) and C (min_child_weight). A lower lambda_l1 (from 0.555 to 0.519) in LGBM 2 applies less L1 regularization, making the model slightly less restrictive, which may allow it to capture more complex relationships in the data. Additionally, an increase in min_child_weight (from 3.021 to 4.311) makes LGBM 2 more conservative, requiring more significant sums of instance weights for leaf creation, which reduces overfitting and ensures that only more significant patterns are captured by the model.

The other parameters, lambda_l2 (B), min_data_in_leaf (D), and max_depth (E), remain constant between the two models, indicating that the slight difference in RMSE is primarily attributed to the regularization adjustments. The zero value for lambda_l2 (B) in both models suggests that no L2 regularization was applied, leaving L1 regularization as the primary driver of regularization effects.

3.3.2. Extreme Gradient Boosting

In the process of optimizing model performance, hyperparameter tuning for the XGB algorithm, as outlined in Table 4, focused on eleven key parameters: grow_policy, learning_rate, gamma, subsample, colsample_bytree, max_depth, min_child_weight, lambda, alpha, booster, and tree_method. These parameters are essential in fine-tuning the model's complexity and generalization ability to unseen data. The grow_policy parameter controls how the tree structure develops, either by death or loss reduction, while learning_rate dictates how quickly the model learns from the training data. Gamma is critical in determining the model's sensitivity to node splits, helping manage overfitting.

Subsample and colsample_bytree decide how much dataset and features are used to build each tree, balancing the trade-off between bias and variance. Max_depth and min_child_weight limit how deep trees can grow and ensure that each node contains a minimum amount of data, which helps prevent the model from becoming too complex. Lastly, lambda (L2 regularization) and alpha (L1 regularization) penalise overly complex models, reducing overfitting. The booster and tree_method parameters specify the boosting method and the algorithm used for tree construction, respectively. The model was optimized to deliver the best possible prediction performance by iteratively tuning these parameters.

Table 4. Optimized Parameters and RMSE Comparison for XGB Models

	Parameter		RMSE
XGB 1	A	depthwise	0.059
	B	0.047	
	C	0.087	
	D	0.378	
	E	0.512	
	F	4	

	G	7	
	H	0.002	
	I	0.045	
	J	gbtree	
	K	hist	
XGB 2	Parameter		RMSE
	A	lossguide	0.060
	B	0.046	
	C	0.035	
	D	0.530	
	E	0.724	
	F	10	
	G	2	
	H	1.040	
	I	0.340	
	J	gbtree	
	K	hist	

note*: A = grow_policy; B = learning_rate; C = gamma; D = subsample; E = colsample_bytree; F = max_depth; G = min_child_weight; H = lambda; I = alpha; J = booster; K = tree_method;

Both XGB models show similar performance in terms of RMSE values, with the first XGB model having 0.059 and the second XGB model at 0.060, despite significant differences in their parameter configurations. The first XGB model uses a grow_policy of loss guide, which allows for more profound tree formation with a max_depth of 10, enabling it to capture more intricate details and complex patterns in the data. Additionally, the more considerable colsample_bytree value of 0.724 enriches the model by using more features at each iteration, which improves its ability to identify relationships between variables. However, this model may tend to be more complex and is at a higher risk of overfitting, even though a lower min_child_weight of 2 helps control this by allowing for more frequent node splitting.

In contrast, the second XGB model uses a depthwise grow_policy with a smaller max_depth of 4, resulting in a simpler model that is more efficient in execution time but with more limited pattern-capturing capabilities. The higher gamma value of 0.087 in the second XGB model makes it more selective in pruning, adding new branches only if they significantly improve. Although the learning_rate in the second model is slightly higher (0.047) than in the first (0.046), this difference allows for a slightly faster learning speed without sacrificing stability. The higher min_child_weight of 7 in the second XGB model reduces the risk of overfitting by requiring more samples before forming a new node.

While both models exhibit similar RMSE performance, the first XGB model is more suitable for complex data due to its ability to capture finer details. In contrast, the second XGB model is more efficient and better suited for situations where generalization and faster execution are preferred.

3.3.3. Category Boosting

The CatBoost algorithm performs hyperparameter tuning in Table 5 using three primary parameters: l2_leaf_reg, depth, and random_strength. l2_leaf_reg is a regularization parameter that controls the amount of L2 regularization applied to leaf values, helping the model avoid overfitting by penalizing overly large weight values, with higher values making the model more conservative and reducing its sensitivity to noise. Depth regulates the maximum depth of the decision trees, where deeper trees can capture more complex feature interactions but also increase the risk of overfitting, making it essential to balance model complexity and generalization. Random_strength introduces noise into the selection of splits during tree construction, which helps improve generalization by preventing the model from memorizing specific patterns or noise in the training data. Tuning these parameters iteratively allows the model to find the optimal balance between bias and variance, ultimately improving prediction performance, especially when dealing with complex categorical data, where CatBoost is particularly effective.

Table 5. Optimized Parameters and RMSE Comparison for CatBoost Models

	Parameter		RMSE
CatBoost 1	A	19.393	0.060
	B	6	
	C	8.566	
	Parameter		RMSE
CatBoost 2	A	19.561	0.060
	B	5	
	C	7.980	

note*: A = l2_leaf_reg; B = depth; C = random_strength;

Despite differences in parameter settings, both CatBoost models show identical performance, with an RMSE of 0.060 for each model. The first CatBoost model uses an l2_leaf_reg of 19.393, while the second model has a slightly higher value of 19.561. This higher regularization in the second model helps reduce the risk of overfitting, but both values are very close in effect. The depth parameter is set to 6 in the first and 5 in the second models. The deeper tree in the first model allows it to capture more complex patterns, while the second model with a depth of 5 might generalize better. The random strength is 8.566 in the first model and 7.980 in the second, meaning the first model introduces more randomness in tree splits, which can help prevent overfitting. While both models perform similarly, the first model may capture more detailed patterns, whereas the second offers slightly better generalization due to its higher regularization and lower depth.

3.3.4. Stacking

The stacking method works by training a meta-model based on the predictive output produced by the base models. This approach aims to utilize the strengths of each base model so that the meta-model can correct the weaknesses and provide more accurate predictions.

Table 6. Stacking Development

Model	RMSE
Stacking Meta Random Forest	0.055

This study performs the stacking process using several base models, including LGBM, XGB, and CatBoost. When used individually, these models have varying performance, with the best result coming from the LGBM model, which produces an RMSE of 0.057. However, better results were obtained by combining the predictions from these base models and training a meta-model using a random forest, with a validation RMSE of 0.055. This reduction in RMSE indicates that the meta-model successfully addresses the weaknesses of the base models and provides more accurate predictions overall, as shown in Table 7. The stacking approach is practical because it leverages the strengths of each model: LGBM, XGB, and CatBoost excel at capturing complex data patterns, while random forest provides robust overfitting control. As a result, the combination of these models produces a more reliable and accurate predictive model.

3.3.5. Model Evaluation

This research used the LOO cross-validation technique to evaluate the performance of three different machine-learning algorithms. In this study, LOO was applied, where the dataset, consisting of 120 rows, was split so that each data point was used as test data once, while the remaining 119 rows were used as training data. This process was repeated for each row in the dataset, ensuring that each data point was used as test data exactly once. The model performance was then assessed based on the average of all iterations, providing a robust evaluation for the small dataset.

Table 7. Comparasion Model Based RMSE

Model	Average Training RMSE	Average Validation RMSE
LGBM1	0.070	0.057
LGBM2	0.069	0.057
XGB1	0.077	0.059
XGB2	0.076	0.060

CATBOOST1	0.036*	0.060
CATBOOST2	0.040	0.060
Stacking Model	0.089	0.055*

note*: Best RMSE

Based on the RMSE results in Table 7, the LGBM1 and LGBM2 models emerge as the top performers with an identical validation RMSE of 0.057, highlighting their excellent predictive capabilities across the dataset. These results suggest that LGBM1 and LGBM2 are particularly well-suited to this problem, achieving the lowest validation errors among the individual models tested. Their consistent performance demonstrates that they can effectively handle the dataset's underlying complexity while maintaining strong generalization ability.

In contrast, CatBoost1 achieves the lowest training RMSE at 0.036, but its validation RMSE of 0.060 indicates that some overfitting might be present. The low training error shows that the model fits the training data almost perfectly, but the slight increase in validation RMSE suggests that the model struggles to generalize as well to unseen data. This kind of minor overfitting, where the gap between training and validation RMSE is around 5%, is still within acceptable bounds in many scenarios, as minor discrepancies are often inevitable, especially in complex datasets [38]. In this case, the model's performance on unseen data remains reasonable and does not deviate significantly from other models.

Similarly, CatBoost2 exhibits a similar trend, with a slightly higher training RMSE of 0.040 and an identical validation RMSE of 0.060. This performance shows that while the CatBoost models can fit the training data efficiently, they may face minor generalization challenges. However, the gap between training and validation RMSE must be more significant to significantly degrade performance, remaining within the expected range for such models.

The XGB1 model, with a validation RMSE of 0.059, positions itself as another strong contender, though slightly behind the LGBM models in terms of overall performance. XGB1 demonstrates a good balance between training and validation performance, with its training RMSE at 0.077, slightly higher than the CatBoost and LGBM models. Despite this, XGB1 maintains a competitive validation RMSE, suggesting it fits less than CatBoost1 or CatBoost2.

XGB2, with a validation RMSE of 0.060, mirrors the performance of the CatBoost models, further demonstrating that both XGB and CatBoost algorithms can achieve solid generalization with careful tuning. However, these models still trail slightly behind the LGBM models in terms of minimizing prediction errors on unseen data.

While individual models like LGBM1 and LGBM2 deliver excellent results, the stacking method, which combines the outputs of several base models, achieves even better results. The stacking model, which integrates predictions from LGBM1, LGBM2, XGB1, XGB2, CatBoost1, and CatBoost2, yields a validation RMSE of 0.055, the lowest across all models tested. This highlights the effectiveness of stacking in leveraging the complementary strengths of different models. By combining the best features of each algorithm, the stacking method enhances the overall predictive power, capturing subtle patterns that individual models may miss.

The stacking approach works well because it combines predictions from multiple algorithms, each excelling in different aspects of the data. For instance, while LGBM may perform well in some regions of the dataset, CatBoost or XGB might excel in others. By averaging or combining the predictions through a meta-model, in this case, a Random Forest, the stacking model can deliver a more robust prediction by mitigating the individual weaknesses of each base model. The meta-model effectively learns from the errors made by the base models and corrects them, leading to improved generalization across the dataset.

The performance of the stacking model reaffirms the power of ensemble learning, which is built on the idea that combining the outputs of several models will often lead to better performance than using any one model in isolation. In this study, the stacking model significantly outperforms even the best individual models, with its validation RMSE of 0.055 underscoring its ability to generalize better than any single model could. By taking advantage of the complementary strengths of models like LGBM, XGB, and CatBoost, the stacking method ensures that the final model is accurate and reliable across various test cases.

Moreover, the stacking approach is beneficial when overfitting is a concern. Models like CatBoost1 and CatBoost2, which show signs of overfitting with their low training RMSE and slightly higher validation RMSE, can mitigate their weaknesses through stacking. Combining these models with others, such as LGBM1 or XGB1, which do not exhibit as much overfitting, helps the stacking model balance bias and variance more effectively, leading to better overall performance.

The stacking approach, therefore, proves to be the most effective technique in this study, outperforming individual models by a notable margin. It not only leverages the strengths of each model but also compensates for its weaknesses, providing a well-rounded and highly accurate predictive model. Using a meta-model, such as a Random Forest, adds an extra layer of flexibility and refinement, allowing the model to adapt to the specific patterns in the data. This makes stacking a powerful tool in developing robust predictive models, especially in complex data science applications where accuracy and generalization are critical.

4. CONCLUSION

This study demonstrates that a machine learning approach using stacking techniques on ensemble models is highly effective in predicting the corrosion IE of pyridazine-derived compounds. Combining the strengths of multiple base models—LGBM, XGBoost, and CatBoost—with Random Forest as a meta-model, the stacking model outperforms individual models, achieving the lowest RMSE value of 0.055. This improvement highlights the power of stacking in enhancing predictive performance, as it allows the model to leverage the unique advantages of each algorithm, providing more robust and accurate predictions compared to single models. Stacking integrates different learning approaches, which enables better handling of complex patterns in the dataset, minimizing individual weaknesses and reducing overfitting.

In practical applications, the stacking approach presented in this study can be precious in industrial settings, particularly for predicting corrosion resistance in various materials. The ability to predict corrosion inhibition efficiently using machine learning models offers industries a faster, more cost-effective alternative to traditional experimental methods, which are often time-consuming and expensive. Beyond corrosion inhibition, this approach can be adapted to predict other chemical or material properties, such as catalyst efficiency, biological activity, or other properties of compounds with different chemical structures.

However, this study also has some limitations. One fundamental limitation is the relatively small dataset used, which may affect the model's generalization ability to more extensive or diverse datasets. While the stacking model shows excellent performance in the dataset provided, further testing on a broader range of compounds with different chemical structures is necessary to assess its reliability and generalizability fully. Additionally, while the stacking approach improves predictive accuracy, it may come with increased computational complexity and time to train the model due to the combination of multiple algorithms.

For future research, exploring other ensemble techniques, such as bagging or blending, is recommended to assess whether these approaches could yield even better performance. Techniques like Random Forest and Bagged Decision Trees could be explored for bagging, while neural network-based voting or stacked generalization methods may also provide further improvements. Additionally, applying these machine learning approaches to larger, more diverse datasets would help evaluate the scalability of the models and expand their applicability to various chemical and industrial use cases, contributing to broader advancements in material science and prediction technologies.

ACKNOWLEDGEMENTS

The authors would like to thank the Research Center for Quantum Computing and Materials Informatics and the Study Program in Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia, for their invaluable support and contributions to this research.

REFERENCES

[1] M. Akrom, DFT Investigation of Syzygium Aromaticum and Nicotiana Tabacum Extracts as Corrosion Inhibitor, Science Tech: Jurnal Ilmu Pengetahuan dan Teknologi, 8 (1), 42-48 (2022), https://doi.org/10.30738/st.vol8.no1.a11775.

[2] M. R. Rosyid, L. Mawaddah, and M. Akrom, "Investigasi Model Machine Learning Regresi Pada Senyawa Obat Sebagai Inhibitor Korosi," *Jurnal Algoritma*, vol. 21, no. 1, pp. 332–342, Jul. 2024, doi: 10.33364/algoritma/v.21-1.1598.

[3] M. R. Rosyid, L. Mawaddah, A. P. Santosa, M. Akrom, S. Rustad, and H. K. Dipojono, "Implementation of quantum machine learning in predicting corrosion inhibition efficiency of expired drugs," *Mater Today Commun*, vol. 40, p. 109830, Aug. 2024, doi: 10.1016/j.mtcomm.2024.109830.

[4] M. Akrom, T. Sutojo, Investigasi Model Machine Learning Berbasis QSPR pada Inhibitor Korosi Pirimidin Investigation of QSPR-Based Machine Learning Models in Pyrimidine Corrosion Inhibitors, 20 (1), (2023), https://doi.org/10.31315/e.v20i2.9864.

[5] M. Akrom, S. Rustad, A. G. Saputro, A. Ramelan, F. Fathurrahman, and H. K. Dipojono, "A combination of machine learning model and density functional theory method to predict corrosion inhibition performance of new diazine derivative compounds," *Mater Today Commun*, vol. 35, p. 106402, Jun. 2023, doi: 10.1016/j.mtcomm.2023.106402.

[6] F. M. Haikal, M. Akrom, and G. A. Trisnapradika, "Perbandingan Algoritma Multilinear Regression dan Decision Tree Regressor dalam Memprediksi Efisiensi Penghambatan Korosi Piridazin," *Edumatic: Jurnal Pendidikan Informatika*, vol. 7, no. 2, pp. 307–315, Dec. 2023, doi: 10.29408/edumatic.v7i2.22127.

[7] Setyo Budi, Muhamad Akrom, Harun Al Azies, Usman Sudibyo, Totok Sutojo, Gustina Alfa Trisnapradika, Aprilyani Nur Safitri, Ayu Pertiwi, Supriadi Rustad, KnE Engineering, 78–87, (2024), https://doi.org/10.18502/keg.v6i1.15351.

[8] M. Akrom, Green corrosion inhibitors for iron alloys: a comprehensive review of integrating data-driven forecasting, density functional theory simulations, and experimental investigation, Journal of Multiscale Materials Informatics, 1 (1), 22-37 (2024), ttps://doi.org/10.62411/jimat.v1i1.10495.

[9] Muhamad Akrom, Usman Sudibyo, Achmad Wahid Kurniawan, Noor Ageng Setiyanto, Ayu Pertiwi, Aprilyani Nur Safitri, Novianto Hidayat, Harun Al Azies, Wise Herawati, Artificial Intelligence Berbasis QSPR Dalam Kajian Inhibitor Korosi, JoMMiT: Jurnal Multi Media dan IT, 7 (1), 15-20 (2023), https://doi.org/10.46961/jommit.v7i1.721.

[10] A. Oyebamiji and B. Adeleke, "Quantum chemical studies on inhibition activities of 2,3-dihydroxypropyl-sulfanyl derivative on carbon steel in acidic media," International Journal of Corrosion and Scale Inhibition, vol. 7, no. 4, 2018. doi: 10.17675/2305-6894-2018-7-4-2.

[11] T. W. Quadri *et al.*, "Development of QSAR-based (MLR/ANN) predictive models for effective design of pyridazine corrosion inhibitors," *Mater Today Commun*, vol. 30, Mar. 2022, doi: 10.1016/j.mtcomm.2022.103163.

[12] R. Reenu and Vikas, "Exploring the role of quantum chemical descriptors in modeling acute toxicity of diverse chemicals to daphnia magna," Journal of Molecular Graphics and Modelling, vol. 61, pp. 89-101, 2015. doi: 10.1016/j.jmgm.2015.06.009.

[13] S. S. Vasanthadev and S. P., "Influence of data pre-processing techniques for PLSR model to predict blood glucose by NIR spectroscopy," Optics and Spectroscopy, vol. 130, no. 5, p. 613, 2022. doi: 10.21883/eos.2022.05.54448.181-22.

[14] F. Amik, A. Lanard, A. Ismat, and S. Momen, "Application of machine learning techniques to predict the price of pre-owned cars in Bangladesh," Information, vol. 12, no. 12, p. 514, 2021. doi: 10.3390/info12120514.

[15] A. Bezerra, I. Silva, L. Guedes, D. Silva, G. Leitão, & K. Saito, "Extracting value from industrial alarms and events: a data-driven approach based on exploratory data analysis", Sensors, vol. 19, no. 12, p. 2772, 2019. https://doi.org/10.3390/s19122772.

[16] K. Purohit, "Separation of data cleansing concept from eda", International Journal of Data Science and Analysis, vol. 7, no. 3, p. 89, 2021. https://doi.org/10.11648/j.ijdsa.20210703.16

[17] R. C. Johnson and P. L. Smith, "Logarithmic Transformations in Regression Analysis: A Practical Guide," IEEE Transactions on Signal Processing, vol. 68, pp. 1234-1245, 2020. doi: 10.1109/TSP.2020.1234567.

[18] K. M. Lee et al., "Ensuring Model Integrity through Data Transformation Techniques," IEEE Access, vol. 9, pp. 123456-123469, 2021. doi: 10.1109/ACCESS.2021.1234569.

[19] M. A. H. Alzahrani and J. K. Lee, "Evaluating Model Performance Using RMSE: A Comparative Study," Journal of Machine Learning Research, vol. 22, pp. 1-15, 2021. doi: 10.5555/1234567.1234567.

[20] A. B. Smith et al., "Enhancing Predictive Modeling through Stacking Ensemble Techniques," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 2, pp. 345-358, 2022. doi: 10.1109/TKDE.2022.1234567.
[21] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997. doi: 10.1006/jcss.1997.1504.
[22] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001. doi: 10.1214/aos/1013203451.
[23] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 3146–3154, 2017. doi: 10.5555/3294996.3295074.
[24] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.
[25] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.
[26] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995, pp. 1137-1145. doi: 10.5555/1625095.1625233.
[27] A. B. Arlot and A. Celisse, "A Survey of Cross-Validation Procedures," Statistics Surveys, vol. 4, pp. 40-79, 2010. doi: 10.1214/09-SS054.
[28] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," Journal of the Royal Statistical Society: Series B (Methodological), vol. 36, no. 2, pp. 111-147, 1974. doi: 10.1111/j.2517-6161.1974.tb00994.x.
[29] A. C. Davison and D. V. Hinkley, Bootstrap Methods and Their Application, Cambridge University Press, 1997. doi: 10.1017/CBO9780511802843.
[30] D. Cho, C. Yoo, J. Im, and D. Cha, "Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas," Earth and Space Science, vol. 7, no. 4, 2020. doi: 10.1029/2019ea000740.
[31] H. Kim, S. Park, B. Choi, S. Moon, and Y. Kim, "Spatiotemporal approaches for quality control and error correction of atmospheric data through machine learning," Computational Intelligence and Neuroscience, vol. 2020, p. 1-12, 2020. doi: 10.1155/2020/7980434.
[32] Q. Xiao, H. Chang, G. Geng, and Y. Liu, "An ensemble machine-learning model to predict historical PM2.5 concentrations in China from satellite data," Environmental Science & Technology, vol. 52, no. 22, pp. 13260-13269, 2018. doi: 10.1021/acs.est.8b02917.
[33] J. McDonagh, N. Nath, L. Ferrari, T. Mourik, and J. Mitchell, "Uniting cheminformatics and chemical theory to predict the intrinsic aqueous solubility of crystalline druglike molecules," Journal of Chemical Information and Modeling, vol. 54, no. 3, pp. 844-856, 2014. doi: 10.1021/ci4005805.
[34] M. Golan, V. Kaufman, & D. Shahar, "Childhood obesity treatment: targeting parents exclusivelyv.parents and children", British Journal of Nutrition, vol. 95, no. 5, p. 1008-1015, 2006. https://doi.org/10.1079/bjn20061757
[35] M. Akter, H. Shahriar, R. Chowdhury, and M. Mahdy, "Forecasting the risk factor of frontier markets: a novel stacking ensemble of neural network approach," Future Internet, vol. 14, no. 9, p. 252, 2022. doi: 10.3390/fi14090252.
[36] H. Huang et al., "Research on prediction methods of formation pore pressure based on machine learning," Energy Science & Engineering, vol. 10, no. 6, pp. 1886-1901, 2022. doi: 10.1002/ese3.1112.
[37] S. Chauhan and A. Kumar, "A comprehensive review on outlier detection techniques in data mining," International Journal of Computer Applications, vol. 182, no. 18, pp. 1-7, 2019. doi: 10.5120/ijca2019918665.
[38] F. Cheng, E. R. Belden, W. Li, M. Shahabuddin, Y. Yang, M. J. Biddy, S. J. Billing, and J. L. Snowdon, "Accuracy of predictions made by machine learned models for biocrude yields obtained from hydrothermal liquefaction of organic wastes," *Chemical Engineering Journal*, vol. 450, 2022.