Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2059~2067
e-ISSN: 2548-6861

Gaussian Mixture-Based Data Augmentation Improves QSAR Prediction of Corrosion Inhibition Efficiency

Darnell Ignasius [1]*, Muhamad Akrom [2], Setyo Budi [3]***
* Study Program in Information Systems, Faculty of Computer Science Dian Nuswantoro University, Semarang 50131, Indonesia
** Research Center for Quantum Computing and Materials Informatics, Faculty of Computer Science Dian Nuswantoro University, Semarang 50131, Indonesia
ignasiusdarnell@gmail.com [1], m.akrom@dsn.dinus.ac.id [2], setyo@dsn.dinus.ac.id [3]

Article Info

Article history:

Received 2025-08-25
Revised 2025-09-08
Accepted 2025-09-19

Keyword:

Corrosion Inhibition,
Data Augmentation,
Gaussian Mixture Model,
Machine Learning,
Small Data.

ABSTRACT

Predicting corrosion inhibition efficiency IE (%) is often hindered by small, heterogeneous datasets. This study proposes a Gaussian mixture–based data augmentation pipeline to strengthen QSAR generalization under data scarcity. A curated set of 70 drug-like compounds with 14 physicochemical and quantum descriptors was cleaned, split 90/10 (train/test), and transformed using a Quantile Transformer followed by a Robust Scaler. A Gaussian Mixture model (GMM) with 2–5 components selected by the variational lower bound was fitted to the transformed training features and used to generate up to 2,500 synthetic samples. Eight regressors (Gaussian Process, Decision Tree, Random Forest, Bagging, Gradient Boosting, Extra Trees, SVR, and KNN) were evaluated on the held-out test set using R2 and RMSE. Augmentation improved performance across several families: for example, Gaussian Process R2 improved from −1.54 to 0.54 (RMSE 11.71 to 5.01) and Decision Tree R2 from −0.33 to 0.63 (RMSE 8.48 to 4.44), Bagging and Random Forest showed R2 increases of 0.67 and 0.40, respectively. The optimal synthetic size varied by model.

I. INTRODUCTION

Corrosion is the progressive physicochemical degradation of materials caused by interactions with the surrounding environment, leading to safety incidents, operational downtime, and substantial environmental and economic burdens across energy, manufacturing, and chemical-processing sectors. Global estimates by NACE place the annual cost of corrosion at USD 2.5 trillion (3.4% of global GDP), underscoring that mitigation is not only a technical necessity but also a strategic lever for sustainability and operational resilience [1].

Organic corrosion inhibitors are widely used due to their cost-effectiveness, but measuring inhibition efficiency (IE%) is labour-intensive, producing small and heterogeneous datasets that increase overfitting risk and weaken external validity QSAR models are especially vulnerable to this small-data bias, where limited samples and high-dimensional descriptors yield unstable estimators and poor generalization across chemical domains. Recent reviews recommend strategies such as transfer learning, multitask frameworks, and generative augmentation to mitigate these challenges in cheminformatics and materials informatics [2], [3], [4].

Data augmentation is a natural response to small-data regression, but widely used oversampling methods such as SMOTE were designed for classification and do not directly respect the continuity of regression targets like IE (%). Even regression-oriented variants (e.g., G-SMOTE-R) can be highly sensitive to data/model characteristics, and without careful calibration they may distort the target distribution and disrupt feature target covariance. Similarly, deep generative approaches such as CTGAN often suffer from instability and mode collapse when applied to small tabular datasets, limiting their reliability for QSAR tasks [5], [6].

Probabilistic generative modeling offers a more principled alternative. Gaussian Mixture Models (GMMs) and their Bayesian variants explicitly capture multimodal joint distributions and latent covariance structures, enabling the synthesis of statistically coherent samples that preserve

correlations among molecular descriptors an essential property for corrosion QSAR modeling. The efficacy of GMM/BGMM-based synthesis has been demonstrated in small-data biomedical applications under stringent quality evaluation, supporting downstream predictive modeling [7], [8].

Evidence specific to corrosion inhibitors supports this direction. Rustad et al. proposed a GMM-based virtual sample generation (GMM-VSG) framework and reported substantial gains on multiple small corrosion datasets, with R^2 values up to 0.99 and marked RMSE reductions [9]. Likewise, Akrom, Rustad, and colleagues showed that augmenting triazole-based inhibitor datasets with GMM-VSG improved Random Forest performance from $R^2 = 0.80$ to 0.99 while reducing RMSE from 9.87 to 0.22, comparable improvements were observed for kernel-based models [10]. These findings suggest that mixture-based augmentation can expand statistical support around the empirical data manifold while preserving distributional fidelity.

This study investigates Gaussian-mixture-based data augmentation to enhance predictive modeling of corrosion inhibition efficiency. We hypothesize that augmenting training data with GMM-generated samples constrained to the domain of the original data and validated statistically improves generalization relative to training on original data alone. We evaluate performance (R^2, RMSE) across multiple ML model.

II. METHOD

Figure 1 presents the workflow of the proposed pipeline for improving corrosion inhibition efficiency IE (%) prediction using Gaussian mixture–based data augmentation.

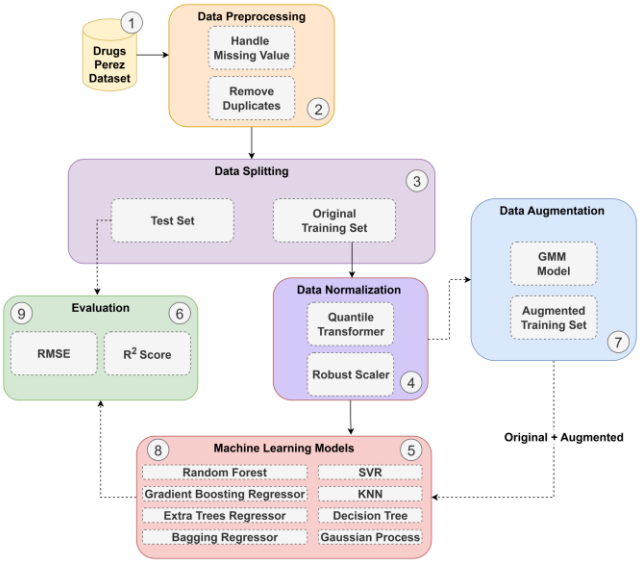

Figure 1 Workflow of the proposed GMM based data augmentation framework for QSAR modelling

A. Dataset

The dataset used in this study was obtained from [11], comprising 70 organic corrosion inhibitors derived from drug-like compounds. These molecules primarily include heterocyclic scaffolds, triazole derivatives, imidazole-based systems, and substituted aromatic frameworks, which are known to exhibit corrosion inhibition activity.

Each record in the dataset consists of a target value corrosion inhibition efficiency IE (%) along with 14 molecular descriptors capturing physicochemical and quantum-mechanical properties. These descriptors include features such as atomic partial charges, polarizability, HOMO–LUMO energy gap, and molecular surface area

B. Data Preprocessing

Prior to modeling and augmentation, the dataset was subjected to a structured preprocessing phase to enhance data quality and consistency. Two primary operations were applied. First, all instances containing missing values in any descriptor or target field were removed. Handling missing values through deletion is considered appropriate when their frequency is low and the dataset remains sufficiently representative, particularly in QSAR studies where imputation can introduce artificial bias [12].

Second, duplicate entries defined as compounds with identical molecular descriptors and inhibition efficiency IE (%) were identified and eliminated. The presence of duplicates can skew learning algorithms by overrepresenting specific chemical patterns and falsely inflating performance metrics [13], [14].

C. Data Splitting

Following preprocessing, the cleaned dataset was partitioned into training and test sets to facilitate supervised learning and assess the generalization ability of machine learning models. Given the relatively small size of the dataset (70 compounds), a 90:10 train–test split was adopted. This allocation ensures that the training set retains sufficient statistical diversity for model fitting, while the test set remains untouched to provide an unbiased estimate of predictive performance [15]. Using a higher proportion of training data is particularly advantageous in low-sample QSAR studies, as smaller test sets may produce unreliable or unstable validation metrics [16].

D. Feature Transformation

To improve model convergence and ensure consistent input scales, all numerical features were transformed using a two-stage process combining distribution normalization and outlier robustness. In the first stage, a Quantile Transformer with output distribution set to *'normal'* was applied. This non-parametric method reshapes the empirical distribution of each feature to approximate a standard normal distribution, effectively reducing skewness and aligning feature properties with the assumptions of many machine learning algorithms,

particularly Gaussian Process Regression and kernel-based models [17], [18].

In the second stage, the transformed features were scaled using a Robust Scaler, which centers data on the median and scales it according to the interquartile range (IQR). This approach is less sensitive to outliers than standard z-score normalization, making it more appropriate for molecular descriptors that may include extreme values due to conformational or electronic variations [19].

To prevent information leakage, both the Quantile Transformer and Robust Scaler were fitted only on the training set and subsequently applied to both training and test data. This ensures that statistical parameters from the test set are not inadvertently introduced into the model during training, thus preserving the integrity of the external evaluation [20].

E. GMM-Based Augmentation

To enhance predictive performance in low-data regimes, we employed a generative data augmentation strategy based on GMM. This approach synthesizes statistically plausible samples by modeling the joint distribution of molecular descriptors in a multivariate, probabilistic framework [21].

The augmentation pipeline followed a four-step process. First, the GMM was trained on the transformed training set features (after Quantile and Robust scaling). The model was initialized with *random_state* = 42 to ensure reproducibility, and the covariance type was set to "*tied*" so that all mixture components shared the same covariance structure. The number of mixture components was tuned in the range of 2 to 5, with the optimal configuration selected using the variational lower bound criterion to balance model complexity and log-likelihood [22].

Second, synthetic samples were drawn from the fitted GMM. For each generated sample, individual features were clipped to the range of the original training data to ensure domain validity and prevent unrealistic molecular descriptor values.

Third, since GMM only models the input features, a surrogate regressor was employed to assign inhibition efficiency IE (%) values to the generated samples. We selected a Gradient Boosting Regressor (GBR) with hyperparameters tuned via GridSearchCV and 3-fold cross-validation. The parameter grid included n_estimators {50, 100}, max_depth {3, 5}, and learning_rate {0.05, 0.1}, with the best configuration selected based on R^2 performance..The use of surrogate labeling allows regression-compatible augmentation without violating the statistical properties of the original label space [23].

Fourth, labeled synthetic samples were combined with the original training data to form an expanded dataset. Multiple augmentation sizes were explored (e.g., 50, 100, 500, 1000, 2500 synthetic samples) to assess sensitivity and saturation effects in downstream model performance.

The rationale for adopting GMM over deterministic or rule-based augmentation lies in its superior ability to preserve local density, inter-feature covariance, and multimodal structures, which are frequently observed in chemical descriptor datasets.

F. Machine Learning Models

This study applied a diverse set of regression models to assess whether the benefits of GMM-based augmentation generalized across different machine learning paradigms. The models covered ensemble methods, kernel-based learners, instance-based approaches, and probabilistic techniques each offering distinct inductive biases relevant to cheminformatics.

Ensemble models, including Random Forest, Gradient Boosting, Extra Trees, and Bagging Regressor, were selected for their proven ability to capture non-linear patterns and manage high-dimensional descriptors. While Random Forest and Extra Trees reduce variance through averaging, Gradient Boosting incrementally improves predictions by correcting residuals, and Bagging enhances stability via bootstrap aggregation [24].

Support Vector Regression (SVR) and Gaussian Process Regression (GPR) were chosen to capture non-linear dependencies and uncertainty, respectively. SVR leverages kernel functions for flexible fitting, while GPR models outputs probabilistically, benefiting from Gaussian-distributed features [25].

To complete the comparative framework, k-Nearest Neighbors (KNN) was included for its local instance-based reasoning, and a Decision Tree Regressor served as a baseline due to its simplicity and interpretability under both original and augmented settings

G. Evaluate

To ensure an accurate and comprehensive assessment of the model performance, this study employed two widely accepted regression metrics: the coefficient of determination (R^2) and the root mean squared error (RMSE). These metrics jointly quantify the proportion of variance explained and the average prediction error in the same unit as the target variable [26].

The R^2 score evaluates the proportion of the total variance in the observed data that is captured by the model's predictions. It is mathematically defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

Where:

y_i : Actual observed values
$\hat{y}_i$: Model predicted values
$\bar{y}$: Mean of actual observed values
n : Number of observations

RMSE is used to quantify the average prediction error magnitude. It penalizes large deviations more heavily due to the squaring of differences, making it sensitive to outliers. RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

Where the variables are as previously defined. RMSE retains the unit of the target variable (IE%), thus making interpretation straightforward. A lower RMSE indicates a model whose predictions are closer to actual values.

III. Result and Discussion

A. Visualization Validation of Augmentation

Validation of the quality of the augmented data is done by applying the Principal Component Analysis (PCA) dimension reduction technique. The main purpose of this process is to evaluate the extent to which the synthetic data generated through the GMM is able to represent the multivariate distribution structure of the original data. By reducing the dimensionality of the features to two principal components, it is possible to visualize high dimensional data in a two-dimensional form without losing too much important information contained in the variance of the data.

Figure 2 presents a scatter plot of the results of the PCA projection of the two principal components. In this visualization, the original data is represented by blue dots, while the synthetic data is shown with red dots. As seen in the graph, the distribution between the two types of data appears homogeneous and overlaps without forming visually separate clusters. This similar distribution pattern indicates that the virtual data not only mimics the univariate values of the original data, but also manages to capture more complex multivariate structures, including the correlation between features and the spatial distribution within the reduced feature space.

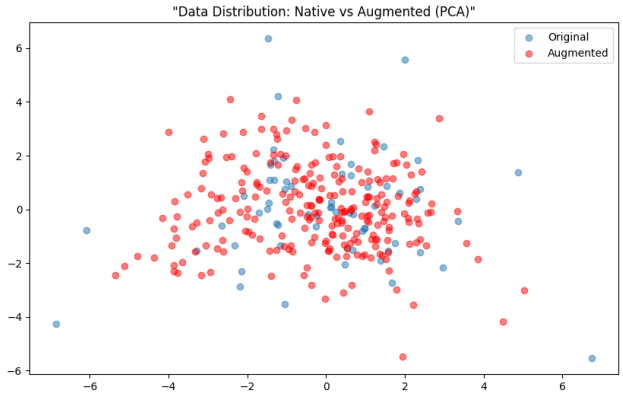

Figure 2 PCA projection showing the overlap between original and GMM augmented datasets

The lack of cluster separation between the original and virtual data is a strong indicator that the augmentation process performed by GMM has been effective and has successfully retained the important statistical properties of the original data. This is very important in the context of predictive modelling, as the existence of a uniform data distribution between the original and virtual data can help avoid model bias and improve generalizability. In other words, the augmented data can be functionally treated as a valid representation of the original data, both in the process of model training and predictive performance evaluation.

To evaluate the effectiveness of the data augmentation process using the GMM, a univariate distribution analysis of several important features was conducted, both before and after the augmentation process. Visualization was done using an overlay histogram depicting the distribution of the original data (purple) and the augmented data (yellow), as shown in Figure 3. The features analyzed include molecular weight (g/mol), pKa, Log P, and Log S, which have an important role in influencing the efficiency of compounds as corrosion inhibitors.

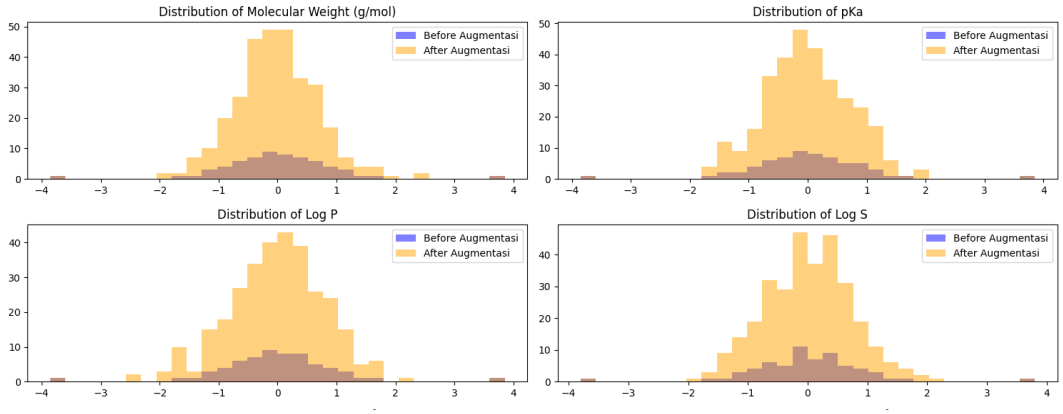

Figure 3 Overlay histograms of original and synthetic data for key molecular descriptors.

The Molecular Weight distribution in the top left figure shows that the augmented data is able to follow the distribution pattern of the original data quite well. The peaks of the distribution are around similar mean values, and the shape of the distribution tends to be symmetrical and resembles a normal distribution. This indicates that GMM

successfully preserves the basic statistical characteristics of this feature, including the center of distribution (mean) and dispersion (standard deviation).

Furthermore, the pKa feature demonstrated a relatively similar distribution pattern between the original and virtual datasets, with minimal differences observed in the distribution tails. The peaks of the distribution remained approximately at the same value, indicating that the augmentation did not result in a significant shift in the distribution. This observation is crucial, given that the pKa value is associated with the ionization ability of the compound, which is a critical parameter in the chemical activity of inhibitor molecules.

For the Log P feature, which describes the lipophilicity of the compound, the augmented distribution also follows the shape of the original data distribution well. The histogram shows that the GMM is able to produce a balanced variation of values between the right and left of the center point of the distribution, showing similarity in shape and distribution of values. The ability to maintain the Log P distribution is important in the context of molecular transport and permeability to the metal surface.

Finally, the Log S or solubility distribution of the compound also shows a good fit between the real and virtual data. Although there is a slight difference in the density of values around the peak of the distribution, the overall shape still resembles the original distribution, showing that the augmentation does not cause a large deviation from the original univariate structure.

B. Sensitivity Analysis to Augmentation Size

The model performance was evaluated by comparing two primary metrics: the coefficient of determination (R^2) and the Root Mean Square Error (RMSE). This comparison was performed across eight machine learning algorithms: Gaussian Process (GP), Decision Tree (DT), Bagging Regressor (bagging), Random Forest (RF), Extra Trees (ET), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and Gradient Boosting (GB). The analysis was performed under two conditions: prior to and following data augmentation using the GMM based Virtual Sample Generation method. The objective of this evaluation was to ascertain the degree to which the augmentation method enhanced the generalization capacity and predictive accuracy of each model

Figure 4 illustrates the dynamics of the R^2 and RMSE values as the number of virtual samples increases from 0 to 2500. The R^2 curve (left) indicates that most models exhibit a pronounced increase within the range of 0 1000 samples, followed by a tendency to stabilize or decline beyond 1500 or 2000 samples. For instance, the Gaussian Process and KNN models demonstrated a threshold beyond which the addition of synthetic samples ceased to enhance performance and may even have resulted in a decline, potentially due to overfitting to synthetic data that did not adequately capture the natural variability of the original dataset. In contrast, models such as Decision Tree and Gradient Boosting continue to show an upward trend in performance up to 2500 samples, reflecting their ability to leverage data diversity in a more adaptable and sustainable manner.

In the RMSE graph (right), it can be seen that the sharpest pattern of error reduction occurs in the range of 0 1000 samples, indicating that most of the benefits of augmentation are obtained in the early stage of data addition. Thereafter, the RMSE tended to stagnate or slightly increase, indicating that the marginal benefit of data augmentation diminished as the maximum model capacity was approached

C. Model Performance Before and After GMM Augmentation

Table 1 presents the coefficient of determination (R^2) for eight machine learning models before and after applying GMM based data augmentation Notably, all models exhibited an increase in R^2, indicating a substantial improvement in their capacity to explain variance in the corrosion inhibition efficiency IE (%) after augmentation.

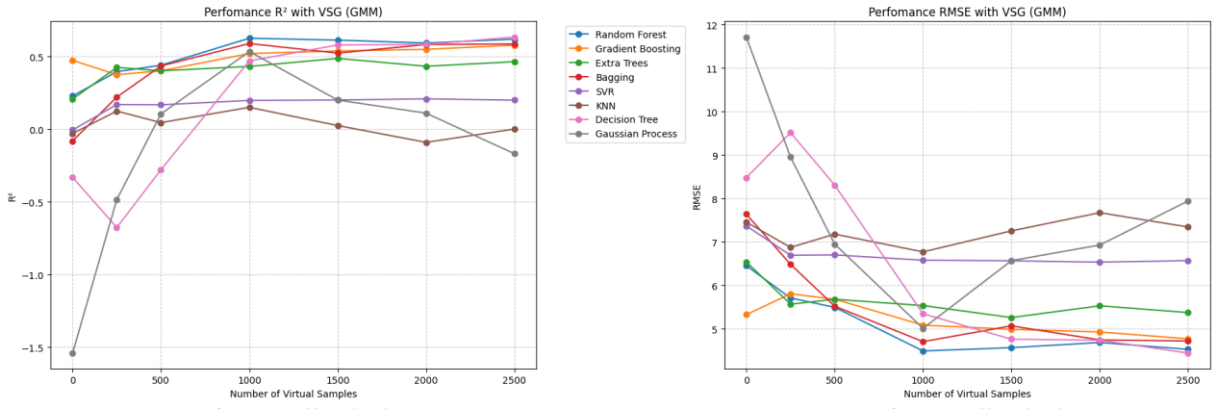

Figure 4 Model performance trends (R^2 and RMSE) as a function of augmented sample size.

The Gaussian Process Regressor (GP) showed the most significant gain, with R^2 increasing from –1.54 to 0.54, yielding an absolute improvement of +2.08, suggesting that GMM augmentation is particularly beneficial for probabilistic models with strong distributional assumptions. Decision Tree (DT) and Bagging Regressor (BG) also experienced meaningful improvements of +0.97 and +0.67, respectively.

Among ensemble models, Random Forest (RF) and Extra Trees (ET) improved by +0.40 and +0.28, respectively, highlighting their responsiveness to richer training distributions. Although Gradient Boosting (GB) and k-Nearest Neighbors (KNN) saw smaller improvements (+0.10 and +0.18), the gains still confirm the general efficacy of synthetic sample augmentation in low-sample QSAR tasks.

TABLE 1 R^2 PERFORMANCE BEFORE AND AFTER AUGMENTATION

Model	R^2 before	R^2 after	R^2 improvement	Optimal Sample Size
GP	**-1.54**	**0.54**	**2.08**	**1000**
DT	**-0.33**	**0.63**	**0.97**	**2500**
Bagging	**-0.08**	**0.59**	**0.67**	**1000**
RF	0.23	0.63	0.40	1000
ET	0.21	0.49	0.28	1500
SVR	-0.01	0.21	0.22	2000
KNN	-0.03	0.15	0.18	1000
GBR	0.47	0.58	0.10	2500

Table 2 shows the Root Mean Square Error (RMSE) for the same models before and after data augmentation. Consistent with the R^2 analysis, all models demonstrated a decrease in RMSE, confirming better predictive accuracy with the inclusion of augmented data.

Once again, Gaussian Process (GP) led the improvements with an RMSE reduction from 11.71 to 5.01, representing a –6.70 decrease, the highest among all models. This reinforces the finding that GP models when aligned with normalized Gaussian-like features benefit significantly from GMM-based data synthesis.

Other notable reductions were observed in Decision Tree (–4.04) and Bagging (–2.93), suggesting that non-parametric and ensemble learners can also harness benefits from synthetic population expansion. While RMSE reductions were modest for SVR (–0.83), KNN (–0.68), and Gradient Boosting (–0.56), they still reflect enhanced precision, particularly in high-variance prediction scenarios.

Interestingly, models such as Random Forest and Extra Trees, despite already having relatively low RMSE prior to augmentation, still improved further with –1.96 and –1.28 reductions, demonstrating that even strong learners can benefit from well-structured data augmentation.

TABLE 2 RMSE PERFORMANCE BEFORE AND AFTER AUGMENTATION

Model	RMSE before	RMSE after	RMSE improvement	Optimal Sample Size
GP	**11.71**	**5.01**	**6.70**	**1000**
DT	**8.48**	**4.44**	**4.04**	**2500**
Bagging	**7.64**	**4.70**	**2.93**	**1000**
RF	6.45	4.49	1.96	1000
ET	6.54	5.26	1.28	1500
SVR	7.37	6.53	0.83	2000
KNN	7.45	6.77	0.68	1000
GBR	5.33	4.77	0.56	2500

The results validate that GMM-based augmentation successfully enhances the generalization performance of diverse learning algorithms, particularly those sensitive to sample diversity and data distribution.

D. Model-Agnostic Effects of GMM-Augmented Data

This study aimed to evaluate the effectiveness of GMM based data augmentation in enhancing predictive performance across various machine learning models in a model-agnostic fashion. To illustrate this, we focused on three representative algorithms Gaussian Process Regression (GPR), Decision Tree Regressor (DT), and Bagging Regressor each belonging to a distinct learning paradigm: probabilistic, interpretable rule-based, and ensemble-based, respectively.

These models were intentionally selected for visualization and deeper analysis due to their contrasting behavior under data-scarce conditions and their varied sensitivity to data distribution, sample diversity, and overfitting. GPR is known for its reliance on smooth Gaussian-like feature distributions and performs poorly with skewed or sparse data. In contrast, DTs are highly interpretable but prone to overfitting in small datasets, while Bagging Regressors benefit from sample diversity and provide insight into ensemble behavior under data augmentation

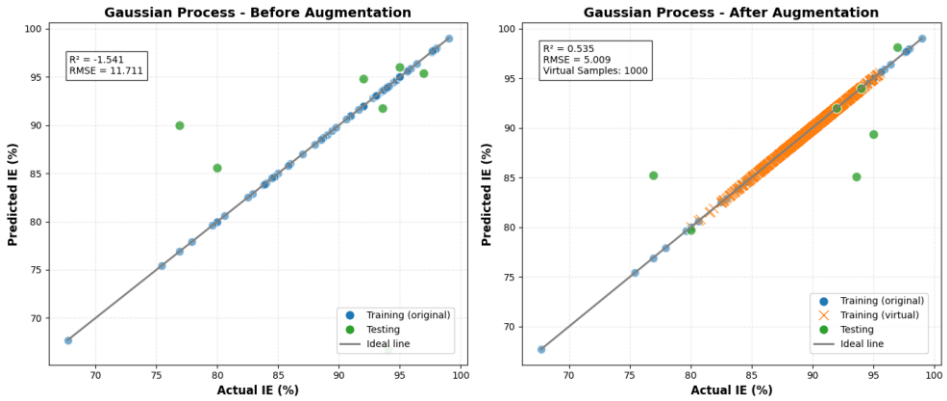

Figure 5 Performance of Gaussian Process model before and after augmentation

As shown in Figure 5, GPR performed poorly prior to augmentation, with an R^2 value of –1.541 and an RMSE of 11.711. These metrics indicate that the model not only failed to capture the underlying structure of the inhibition efficiency (IE) data but also produced predictions that deviated significantly from actual values, especially on the held-out test set. This can be attributed to the model's strong dependence on the smoothness and Gaussianity of input feature distributions conditions that are often violated in small, noisy QSAR datasets.

After augmenting the training set with 1000 synthetic samples generated from the Bayesian GMM, the model's performance improved substantially. The R^2 increased to 0.535 and RMSE dropped to 5.009. As seen in the right panel of Figure 6, the predicted IE values aligned more closely with the ideal line, both for training and testing points. The improvement suggests that the virtual samples helped reinforce the covariance structure of the feature space and allowed the Gaussian kernel to better approximate the functional relationship between molecular descriptors and IE(%)

Figure 6 depicts the performance of the Decision Tree Regressor before and after augmentation. Initially, DT achieved an R^2 of –0.33 and an RMSE of 8.448. The visualization shows severe overfitting: predictions on the training data were nearly perfect, while predictions on the test set were erratic and scattered far from the ideal line. This behavior is expected in small-sample regression tasks where trees create overly specific partitions that do not generalize well to unseen data.

With the addition of 2500 virtual samples, the model's generalization improved significantly. R^2 reached 0.63 and RMSE dropped to 4.44. Notably, test predictions became more concentrated along the ideal line, indicating better generalization. The larger augmented dataset enabled the tree to form deeper, more balanced partitions, capturing more nuanced structure in the data without overfitting.

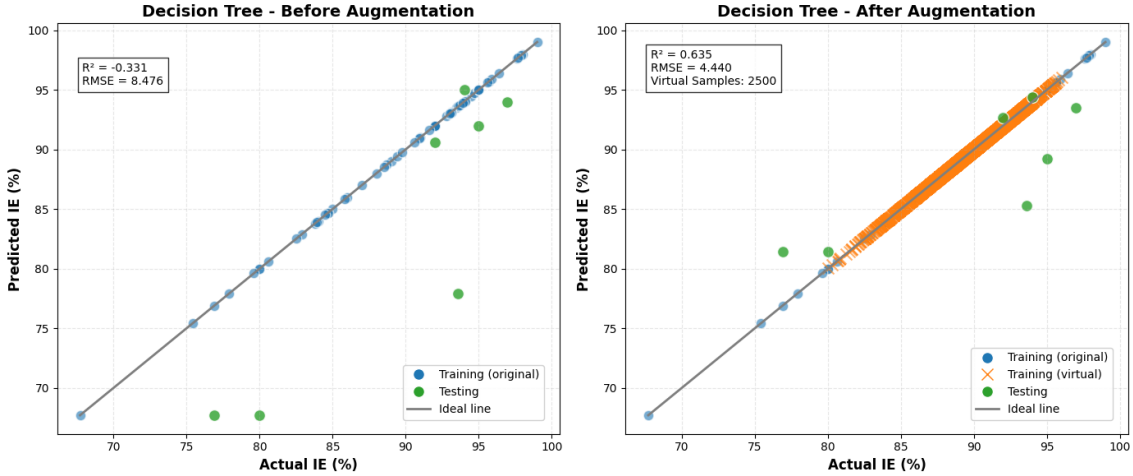

Figure 6 Performance of Decision Tree model before and after augmentation

.

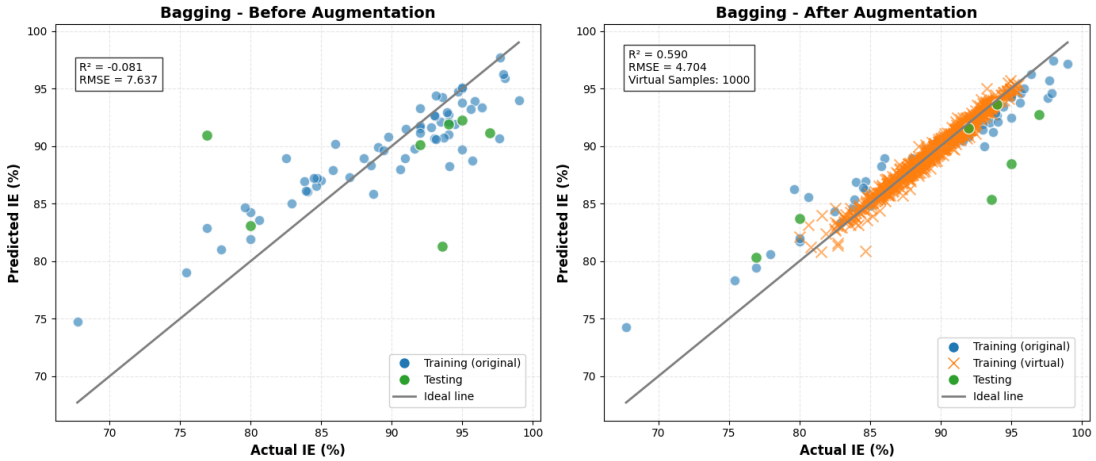

Figure 7 Performance of Bagging Regressor before and after augmentation.

Bagging, an ensemble of regressors trained on bootstrapped subsets of the data, also demonstrated a significant uplift in performance following GMM-based augmentation. As shown in Figure 7, the model initially produced weak results (R^2 = –0.08, RMSE = 7.617), indicating that the limited size and variability of the training set constrained the ensemble's effectiveness. After augmenting with 1000 synthetic samples, performance improved markedly to R^2 = 0.59 and RMSE = 4.70. This suggests that the bagging framework capitalized on the increased variability and density introduced by the GMM-generated data. With richer training support, the base regressors were able to learn from more representative and diverse samples, yielding more accurate and robust ensemble predictions.

The pronounced improvements observed for Gaussian Process (GP) and Decision Tree (DT) models can be attributed to their sensitivity to distributional properties and sample density. Gaussian processes rely on smooth, Gaussian-like marginals, which were reinforced by the quantile–normal transformation and further densified by GMM augmentation. This alignment reduced posterior variance and stabilized kernel hyperparameter estimation, explaining the large gains observed. Decision Trees, conversely, are prone to severe overfitting under small-sample regimes, producing fragmentary partitions that fail to generalize. With synthetic density added by GMM, trees were able to form deeper yet less idiosyncratic splits, reducing variance and markedly improving R^2 and RMSE.

IV. Conclusion

This study demonstrates that Gaussian Mixture Model (GMM)-based data augmentation can substantially enhance predictive modeling of corrosion inhibition efficiency (IE%) in small, chemically diverse datasets. Through a reproducible pipeline involving feature transformation, statistical validation, and cross-model evaluation, we showed that Bayesian GMM-generated samples consistently improved generalization across diverse machine learning families. Gaussian Process Regression achieved the largest gain R^2 −1.54 to 0.54, RMSE 11.71 to 5.01, with comparable improvements for Decision Tree and Bagging Regressor. These results highlight the ability of GMM to enrich training density while preserving descriptor covariance without introducing distributional artifacts. Sensitivity analysis further revealed that performance typically plateaued or declined beyond 1500–2500 synthetic samples, underscoring the need for careful tuning. Compared with SMOTE and GAN-based approaches, GMM demonstrated superior stability and statistical fidelity for tabular regression under data scarcity.

From a chemical standpoint, the generated data should be regarded as statistically valid but not yet chemically verified. Although PCA projections, univariate distributions, and statistical tests (KS, MMD, TSTR) confirmed strong overlap with the original dataset, inhibitor design ultimately requires domain-specific constraints such as HOMO–LUMO gaps, polarizability, and lipophilicity, alongside experimental validation using electrochemical impedance spectroscopy or weight-loss assays. Accordingly, GMM-based augmentation should be considered a complementary tool to strengthen QSAR modeling rather than a replacement for laboratory studies. Future work should integrate chemical feasibility constraints, benchmark against alternative augmentation techniques, and pursue experimental validation to bridge the gap between statistical synthesis and practical molecular design.

Acknowledgement

The author gratefully acknowledges the support provided by the Research Center for Quantum Computing and Materials Informatics, whose technical environment and

academic resources were instrumental in the successful completion of this work.

Special personal gratitude is also extended to my future partner, whose inspiration, though yet unknown, continues to motivate the pursuit of meaningful research and discovery.

REFERENCES

[1] "The Global Cost and Impact of Corrosion." [Online]. Available: https://inspectioneering.com/news/2016-03-08/5202/nace-study-estimates-global-cost-of-corrosion-at-25-trillion-ann

[2] A. A. A. Serrano, A. Miralrio, and C. Beltran-Perez, "Models for predicting corrosion inhibition efficiency of common drugs on steel surfaces: A rationalized comparison among methodologies," Appl. Surf. Sci. Adv., vol. 22, p. 100621, Aug. 2024, doi: 10.1016/j.apsadv.2024.100621.

[3] I. Baskin and Y. Ein-Eli, "Chemoinformatics for corrosion science: Data-driven modeling of corrosion inhibition by organic molecules," Mol. Inform., vol. 43, no. 11, p. e202400082, 2024, doi: 10.1002/minf.202400082.

[4] C. Özkan et al., "Laying the experimental foundation for corrosion inhibitor discovery through machine learning," Npj Mater. Degrad., vol. 8, no. 1, p. 21, Feb. 2024, doi: 10.1038/s41529-024-00435-z.

[5] L. Camacho, G. Douzas, and F. Bacao, "Geometric SMOTE for regression," Expert Syst. Appl., vol. 193, p. 116387, May 2022, doi: 10.1016/j.eswa.2021.116387.

[6] J. G. Avelino, G. D. C. Cavalcanti, and R. M. O. Cruz, "Resampling strategies for imbalanced regression: a survey and empirical analysis," Artif. Intell. Rev., vol. 57, no. 4, p. 82, Mar. 2024, doi: 10.1007/s10462-024-10724-3.

[7] "Bayesian Inference-Based Gaussian Mixture Models With Optimal Components Estimation Towards Large-Scale Synthetic Data Generation for In Silico Clinical Trials," IEEE Open J. Eng. Med. Biol., vol. 3, pp. 108–114, June 2022, doi: 10.1109/OJEMB.2022.3181796.

[8] L. Kühnel et al., "Synthetic data generation for a longitudinal cohort study - evaluation, method extension and reproduction of published data analysis results," Sci. Rep., vol. 14, no. 1, p. 14412, June 2024, doi: 10.1038/s41598-024-62102-2.

[9] S. Rustad, M. Akrom, T. Sutojo, and H. K. Dipojono, "A Feature Restoration for Machine Learning on Anti-Corrosion Materials," July 12, 2024, Social Science Research Network, Rochester, NY: 4892891. doi: 10.2139/ssrn.4892891.

[10] M. Akrom, S. Rustad, and H. Kresno Dipojono, "Prediction of Anti-Corrosion performance of new triazole derivatives via Machine learning," Comput. Theor. Chem., vol. 1236, p. 114599, June 2024, doi: 10.1016/j.comptc.2024.114599.

[11] C. Beltran-Perez 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., vol. 23, no. 9, p. 5086, Jan. 2022, doi: 10.3390/ijms23095086.

[12] D. Gadaleta et al., "SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data," J. Cheminformatics, vol. 11, no. 1, p. 58, Dec. 2019, doi: 10.1186/s13321-019-0383-2.

[13] T. Yu, C. Nantasenamat, S. Kachenton, N. Anuwongcharoen, and T. Piacham, "Cheminformatic Analysis and Machine Learning Modeling to Investigate Androgen Receptor Antagonists to Combat Prostate Cancer," ACS Omega, vol. 8, no. 7, pp. 6729–6742, Feb. 2023, doi: 10.1021/acsomega.2c07346.

[14] P. Ambure and M. N. D. S. Cordeiro, "Importance of Data Curation in QSAR Studies Especially While Modeling Large-Size Datasets," in Ecotoxicological QSARs, K. Roy, Ed., in Methods in Pharmacology and Toxicology. , New York, NY: Springer US, 2020, pp. 97–109. doi: 10.1007/978-1-0716-0150-1_5.

[15] A. Rácz, D. Bajusz, and K. Héberger, "Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification," Molecules, vol. 26, no. 4, p. 1111, Jan. 2021, doi: 10.3390/molecules26041111.

[16] P. De, S. Kar, P. Ambure, and K. Roy, "Prediction reliability of QSAR models: an overview of various validation tools," Arch. Toxicol., vol. 96, no. 5, pp. 1279–1295, May 2022, doi: 10.1007/s00204-022-03252-y.

[17] W. Wang and B.-Y. Jing, "Gaussian process regression: Optimality, robustness, and relationship with kernel ridge regression," J. Mach. Learn. Res., vol. 23, no. 193, pp. 1–67, 2022.

[18] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian Process Regression for Materials and Molecules," Chem. Rev., vol. 121, no. 16, pp. 10073–10141, Aug. 2021, doi: 10.1021/acs.chemrev.1c00022.

[19] "On the Development of Descriptor-Based Machine Learning Models for Thermodynamic Properties: Part 2 Applicability Domain and Outliers." [Online]. Available: https://www.mdpi.com/1999-4893/16/12/573

[20] J. M. H. Pinheiro et al., "The Impact of Feature Scaling In Machine Learning: Effects on Regression and Classification Tasks," 2025, arXiv. doi: 10.48550/ARXIV.2506.08274.

[21] M. Akrom, "Green Corrosion Inhibitors for Iron Alloys: A Comprehensive Review of Integrating Data-Driven Forecasting, Density Functional Theory Simulations, and Experimental Investigation," J. Multiscale Mater. Inform., vol. 1, no. 1, pp. 22–37, Apr. 2024, doi: 10.62411/jimat.v1i1.10495.

[22] T. Huix, A. Korba, A. Durmus, and E. Moulines, "Variational inference, Mixture of Gaussians, Bayesian Machine Learning," June 06, 2024, arXiv: arXiv:2406.04012. doi: 10.48550/arXiv.2406.04012.

[23] A. Abio et al., "A transfer learning method in press hardening surrogate modeling: From simulations to real-world," J. Manuf. Syst., vol. 77, pp. 320–340, Dec. 2024, doi: 10.1016/j.jmsy.2024.09.012.

[24] R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," Eng. Rep., vol. 5, no. 4, p. e12599, 2023, doi: 10.1002/eng2.12599.

[25] P. F. Sadr, M. Ebrahimi, M. Nekoei, and B. Chahkandi, "QSAR study of novel indole derivatives in hepatitis treatment by stepwise- multiple linear regression and support vector machine," Arch. Pharm. Pract., vol. 11, no. 1–2020, pp. 27–37, 2020.

[26] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Comput. Sci., vol. 7, p. e623, July 2021, doi: 10.7717/peerj-cs.623.