Optimizing Machine Learning Models for Cardiovascular Drug Candidate Screening with XGBoost and Successive Halving

Teuku Rizky Noviandy[1,*], Ryan Setiawan[1], Rahmat Sufri[1], Anisah Anisah[1]

[1]*Department of Information Systems, Faculty of Engineering, Universitas Abulyatama, Aceh Besar 23372, Indonesia*

rizky_si@abulyatama.ac.id, ryan_si@abulyatama.ac.id, rahmatsufri_si@abulyatama.ac.id, anisah_si@abulyatama.ac.id

Article history

Received:
14 September 2025

Received in revised form:
19 September 2025

Accepted:
10 November 2025

Published online:
26 December 2025 2

*Corresponding author
rizky_si@abulyatama.ac.id

Abstract

Thrombin, a key enzyme in the blood coagulation process, is an important target in the development of anticoagulant therapies. This study proposes an efficient and accurate quantitative structure–activity relationship (QSAR) modeling framework that combines the XGBoost algorithm with the Successive Halving (SH) method for hyperparameter optimization. A dataset of 3,145 compounds was collected from the ChEMBL database, and molecular descriptors were generated using the Mordred calculator. After preprocessing and feature selection, the SH-tuned XGBoost model achieved the highest performance, with an accuracy of 85.69%, F1-score of 80.93%, and ROC-AUC of 0.917. Compared to baseline and Random Search-tuned models, the SH-tuned model demonstrated superior predictive accuracy and significantly reduced training time. Confusion matrix analysis confirmed the model's strong sensitivity and balanced classification of active and inactive compounds. These results illustrate the effectiveness of combining gradient boosting with efficient hyperparameter optimization for virtual screening. The proposed framework reduces computational cost without compromising model quality and can support early-stage drug discovery. Future extensions may include regression modeling, expanded datasets, and integration of biological or pharmacokinetic data to further improve performance and applicability.

Keywords: *Classification; Drug discovery; Hyperparameter optimization; QSAR modeling; Virtual screening*

1. Introduction

The growing prevalence of thrombotic disorders, including stroke and deep vein thrombosis, has intensified the demand for effective anticoagulant therapies [1,2]. Thrombin, a key enzyme in the blood coagulation cascade, remains a prime target for the development of novel therapeutic agents [3]. Traditional drug discovery approaches are often time-consuming and costly, prompting researchers to increasingly turn to computational methods to streamline the identification and optimization of thrombin inhibitors [4,5]. Among these approaches, quantitative structure–activity relationship (QSAR) modeling has emerged as a powerful tool,

*Corresponding author. rizky_si@abulyatama.ac.id

leveraging molecular descriptors to predict the biological activity of chemical compounds [6].

QSAR modeling relies on the fundamental principle that a compound's chemical structure determines its biological activity [7]. Over the years, machine learning techniques have significantly enhanced QSAR models by improving prediction accuracy and enabling the analysis of complex, high-dimensional data. Among the many algorithms available, ensemble learning methods, particularly gradient boosting frameworks, have gained popularity due to their robustness and performance [8]. However, efficient model selection and hyperparameter tuning remain challenging, especially when dealing with large chemical datasets and numerous molecular descriptors.

Despite the advancements in machine learning-based QSAR modeling, the process of identifying optimal model configurations often requires extensive computational resources and time-consuming cross-validation [9,10]. Traditional grid or random search methods for hyperparameter tuning can be inefficient, especially when applied to large datasets with thousands of potential descriptor combinations [11]. This inefficiency poses a significant barrier to the rapid and scalable development of predictive models for drug discovery, particularly in scenarios demanding both high accuracy and speed, such as virtual screening of thrombin inhibitors.

To address these challenges, this study proposes an efficient QSAR modeling framework for thrombin inhibitors that integrates the Extreme Gradient Boosting (XGBoost) algorithm with the Successive Halving (SH) technique for hyperparameter optimization. XGBoost is a high-performance ensemble method known for its predictive accuracy and scalability [12,13], while SH is a resource-efficient tuning strategy that quickly eliminates underperforming configurations [14]. By combining these tools, the proposed approach aims to significantly reduce computational cost without compromising model quality, thereby accelerating the drug discovery process.

This paper demonstrates that the integration of XGBoost with SH enables the development of highly accurate and computationally efficient QSAR models for predicting thrombin inhibitory activity. Through rigorous evaluation on benchmark datasets, the proposed method outperforms traditional approaches in both speed and predictive performance. These findings highlight the potential of advanced machine learning and optimization techniques to streamline early-stage drug discovery and offer a scalable solution for modeling complex biological activities.

2. Methods

2.1. Data Collection

The dataset for this study was obtained from the ChEMBL database [15], specifically targeting the Thrombin protein (CHEMBL204) and associated IC_{50} values. Compounds with missing IC_{50} data were excluded to ensure data quality. For compounds with multiple IC_{50} entries, the median value was used to obtain a consistent measurement [16]. Compounds with IC_{50} values greater than or equal to 1000 nM were labeled as active, while those with lower values were labeled as

inactive [17]. After preprocessing, the final dataset comprised 3,145 unique compounds, each assigned a binary activity label suitable for machine learning classification.

The dataset was then split into training and testing sets using an 80:20 stratified split to preserve the original class distribution. A summary of the class distribution across the full dataset, training set, and test set is presented in Table 1.

Table 1. Class Distribution Summary

Dataset	Active (Count)	Active (%)	Inactive (Count)	Inactive (%)	Total Samples
Full	1101	35.01%	2044	64.99%	3145
Train	887	35.25%	1629	64.75%	2516
Test	214	34.02%	415	65.98%	629

2.2. Molecular Descriptors Calculation

To represent the chemical structure of each compound numerically for machine learning, molecular descriptors were used as features. Molecular descriptors are quantitative representations of molecular properties derived from a compound's structure, capturing information such as topology, geometry, atom types, and electronic properties [18]. These descriptors were calculated for each compound using the Mordred descriptor calculator, which generates a wide range of two-dimensional and three-dimensional descriptors [19]. To enhance model performance and reduce redundancy, descriptors with a variance lower than 0.1 were removed, as they provide little discriminative power [20]. Additionally, descriptors with a pairwise Pearson correlation greater than 0.80 were excluded to avoid multicollinearity, which can distort model interpretation and learning [21]. After filtering, a final set of 309 molecular descriptors was retained for model training.

2.3. XGBoost Model

To classify potential drug candidates for cardiovascular applications, the XGBoost algorithm was employed due to its proven efficiency, scalability, and exceptional performance in structured data scenarios. XGBoost is a powerful ensemble learning technique based on gradient-boosted decision trees. It operates by building an additive model in a forward stage-wise fashion, where each new tree corrects the residual errors of the previous ensemble. This sequential learning approach allows the model to focus increasingly on difficult-to-predict instances, improving overall predictive accuracy [12,22].

One of XGBoost's key advantages lies in its incorporation of both L1 (Lasso) and L2 (Ridge) regularization, which helps reduce model complexity and prevent overfitting. Additionally, XGBoost supports parallel and distributed computing, enabling faster training times even on large-scale datasets, which is essential in the context of drug discovery, where high-throughput screening data can be substantial [23]. These features make XGBoost an ideal choice for robust and scalable drug candidate screening pipelines in cardiovascular drug development.

2.4. Successive Halving

To optimize model performance, hyperparameter tuning was conducted using the Successive Halving (SH) algorithm, a resource-efficient and scalable method for hyperparameter optimization. SH is particularly effective when evaluating computationally expensive configurations, as it balances exploration of a wide search space with exploitation of the most promising candidates [24].

The SH algorithm operates by evaluating a large number of hyperparameter configurations using a small amount of computational budget. After each evaluation round, a fixed proportion of the worst-performing configurations is discarded, and the remaining configurations receive an increased budget for further evaluation. This process is repeated iteratively until only the top-performing configurations remain [25].

Mathematically, given a total computational budget $\boldsymbol{B}$, number of configurations n, and a minimum resource r, the SH algorithm allocates resources as shown in Equation 1:

$$For\ i = 0,1, \dots, \lfloor log_{\eta}(n) \rfloor: \quad n_i = \left\lfloor \frac{n}{\eta^i} \right\rfloor, \quad r_i = \eta^i r \tag{1}$$

where η is the reduction factor, n is the initial number of configurations, r is the minimum allocated resource, n_i is the number of configurations in round i, and r_i is the resource per configuration in round i. This approach ensures that computational resources are concentrated on high-potential configurations while maintaining efficiency by quickly eliminating underperformers.

The specific hyperparameter search ranges used during tuning are detailed in Table 2. These ranges were selected based on prior literature and empirical experimentation to balance model complexity, generalization, and convergence speed.

Table 2. XGBoost hyperparameter grid

No.	Hyperparameter	Range	Description
1	learning_rate	0.001 – 0.3	Step size shrinkage used to prevent overfitting
2	max_depth	3 – 19	Maximum depth of each tree
3	subsample	0.5 – 1.0	Fraction of training data used for each boosting round
4	colsample_bytree	0.5 – 1.0	Fraction of features used when constructing each tree
5	gamma	0 – 5	Minimum loss reduction required to make a split
6	reg_alpha	0 – 5	L1 regularization term on weights
7	reg_lambda	0 – 5	L2 regularization term on weights

2.5. Performance Evaluation

The performance of the trained XGBoost model with hyperparameters optimized via SH was compared against two baseline models: a default XGBoost model without tuning and an XGBoost model tuned using Random Search, using the same hyperparameter ranges defined in Table 2. All models were evaluated on a consistent test set. The following metrics were used to assess model performance: accuracy, precision, sensitivity (recall), specificity, and F1-score [26]. These metrics are defined in Equations 2–5:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{6}$$

Where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

In addition to these metrics, this study also calculated and visualized the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC), which provides an aggregate measure of performance across all classification thresholds.

3. Results and Discussion

The XGBoost model optimized using SH was successfully trained and demonstrated the strongest overall performance in this study. The detailed evaluation results are presented in Table 3. The SH-tuned model achieved the highest accuracy at 85.69%, with excellent scores across all key metrics: precision (76.71%), sensitivity (85.65%), specificity (85.71%), and F1-score (80.93%). This indicates that the model was highly effective at distinguishing between active and inactive compounds. Furthermore, it completed training in just 13.3 seconds, highlighting its computational efficiency.

In comparison, the Randomized Search (RF) tuned model also improved upon the baseline, achieving 83.94% accuracy and an F1-score of 77.40%, but required a significantly longer training time (73.3 seconds). While effective, it offered a less favorable balance between performance and efficiency than SH.

The baseline XGBoost model, trained with default hyperparameters and no tuning, achieved an accuracy of 82.03% and an F1-score of 74.49%, with very low training time (0.14 seconds). Although it performed reasonably well, its predictive power was lower than the tuned models.

Table 3. Performance comparison of the baseline XGBoost model and tuned variants using Randomized Search (RF) and Successive Halving (SH)

Model	Time (s)	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
XGboost (Base)	0.14	82.03	75.00	73.99	86.45	74.49
XGBoost + RF	73.3	83.94	77.23	77.58	87.44	77.40
XGBoost + SH	13.3	85.69	76.71	85.65	85.71	80.93

Figure 1 illustrates the confusion matrices for the three XGBoost models evaluated in this study. These matrices provide a visual summary of each model's ability to correctly classify active and inactive compounds.

The baseline model correctly predicted 165 active and 351 inactive compounds. However, it misclassified 58 active compounds as inactive (false negatives) and 55 inactive compounds as active (false positives). This indicates a tendency to underpredict active compounds, which can be problematic in a drug screening where identifying actives is crucial.

The Randomized Search-tuned model showed improved performance over the baseline. It correctly classified 173 active and 355 inactive compounds, reducing false negatives to 51 and false positives to 50. This reflects a better balance between sensitivity and specificity, suggesting that hyperparameter tuning helps the model generalize more effectively.

The SH-tuned model demonstrated the best performance. It correctly identified 191 active and 348 inactive compounds. False negatives dropped substantially to 32, highlighting the model's strong sensitivity in detecting active compounds. Although the number of false positives slightly increased to 58, the overall classification results were more balanced, aligning with this model's highest F1-score and accuracy.

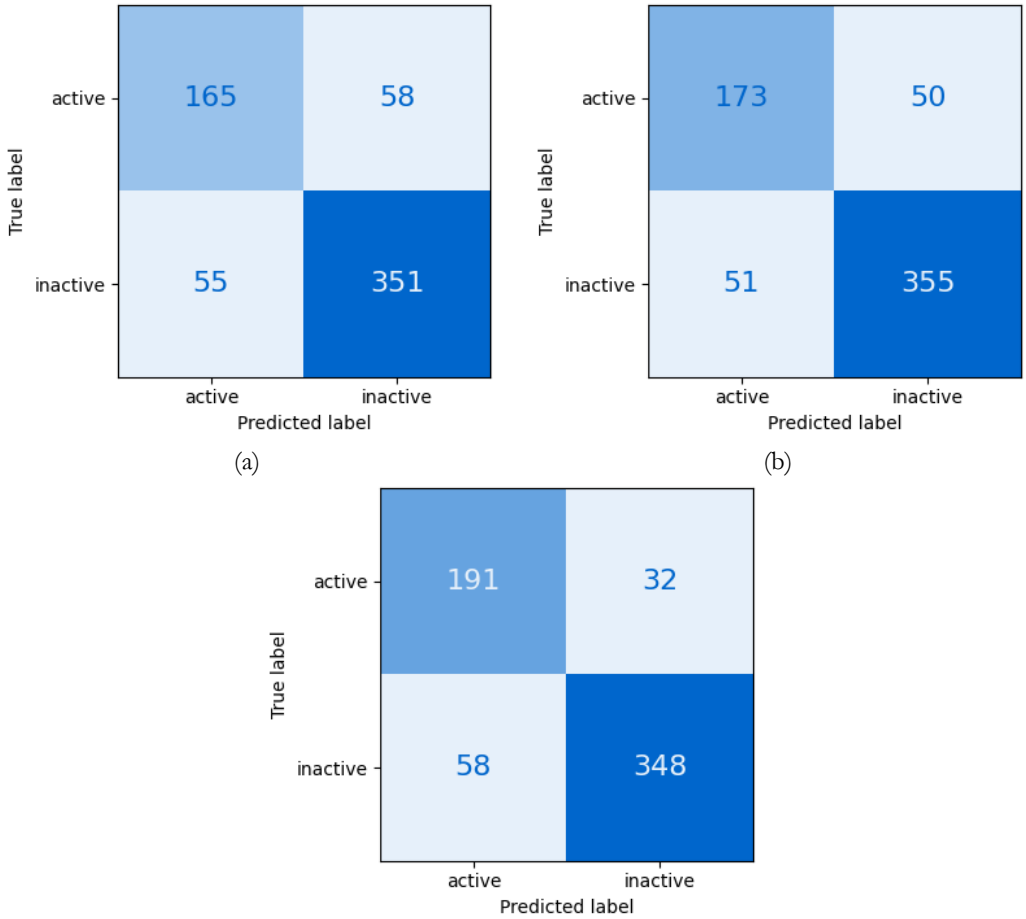

Figure 1. Confusion matrices of XGBoost models: (a) baseline without tuning, (b) tuned with Randomized Search (RF), and (c) tuned with Successive Halving (SH)

These confusion matrices support the conclusion that the SH-tuned model provides the most effective classification, particularly in terms of correctly identifying active compounds with minimal compromise in specificity.

Figure 2 presents the ROC (Receiver Operating Characteristic) curves for the three XGBoost models. The baseline model achieved an AUC (Area Under the Curve) of 0.885, indicating strong classification ability. However, both tuning approaches improved upon this result. The RS-tuned model increased the AUC to 0.905, showing enhanced capability in distinguishing active from inactive compounds. The best performance was again observed with the SH-tuned model, which achieved the highest AUC of 0.917, reflecting superior overall classification performance.

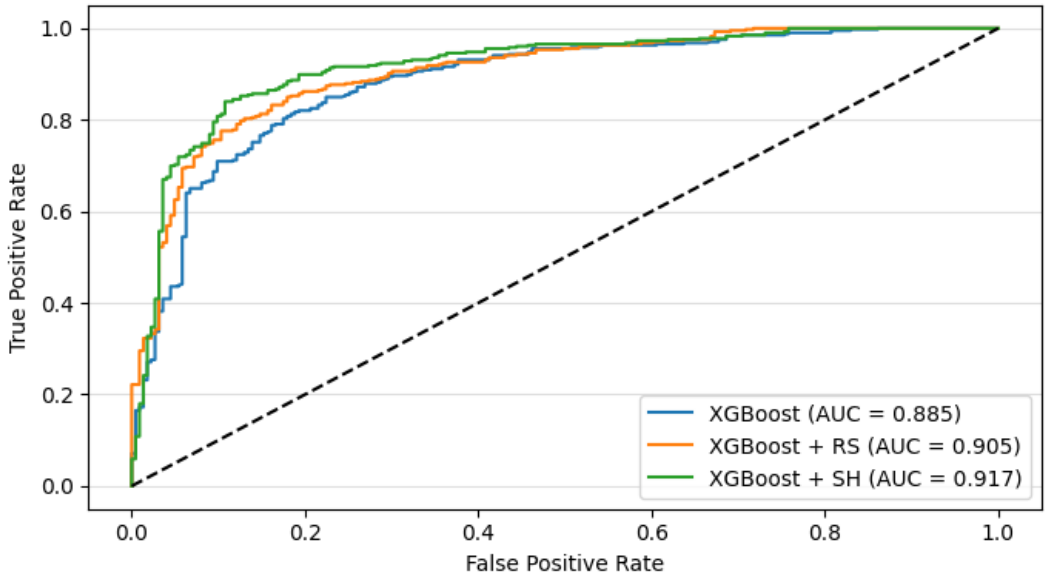

Figure 2. ROC curves for the XGBoost models: baseline (blue), tuned with Randomized Search (orange), and tuned with Successive Halving (green)

This study demonstrates the effectiveness of combining the XGBoost algorithm with SH for optimizing QSAR models aimed at predicting thrombin inhibitory activity. The SH-tuned XGBoost model not only achieved the highest performance across all evaluated metrics, including accuracy, F1-score, and ROC-AUC, but also did so with a significantly reduced computational cost compared to traditional hyperparameter tuning methods like Randomized Search.

The superior performance of the SH-tuned model can be attributed to the efficiency of SH in exploring a wide hyperparameter space while eliminating suboptimal configurations early in the process. This approach allowed for a faster convergence to optimal model parameters without exhaustive searches, which is particularly advantageous in drug discovery tasks where time and computational resources are limiting factors. Importantly, the enhanced sensitivity (recall) of the SH-tuned model highlights its utility in minimizing false negatives, critical in virtual screening applications where missing a potential active compound could hinder downstream drug development efforts.

Despite these promising results, several limitations should be acknowledged. First, while the dataset used from ChEMBL was curated and filtered for quality, it still represents a relatively small portion of the chemical space relevant to thrombin inhibition. The final dataset, though balanced for a binary classification task, may not fully capture the structural diversity and complexity of all potentially bioactive molecules. Moreover, IC_{50} values were converted to binary classes based on a fixed threshold (1000 nM), which could oversimplify the gradient of biological activity and obscure nuanced structure–activity relationships.

Additionally, while Mordred descriptors offer comprehensive molecular characterization, the exclusion of low-variance and highly correlated descriptors might inadvertently eliminate features with subtle but meaningful contributions to

model performance. The study also focused exclusively on XGBoost; alternative ensemble methods or deep learning architectures could be explored for further improvements.

In terms of future work, expanding the dataset to include more diverse and newly identified thrombin inhibitors would enhance the model's generalizability. Incorporating continuous regression models alongside classification approaches may offer a more nuanced understanding of activity trends. Furthermore, combining SH with multi-objective optimization strategies could simultaneously balance predictive performance with interpretability or domain-specific constraints.

Integration of more biologically relevant features, such as docking scores, physicochemical properties, or ADMET (absorption, distribution, metabolism, excretion, and toxicity) profiles, could also be explored to create a more holistic prediction framework. Lastly, deploying the optimized models in a real-time virtual screening pipeline could validate their practical utility and guide experimental validation efforts in early-stage drug discovery.

4. Conclusion

This study presents an efficient and accurate QSAR modeling framework for thrombin inhibitor screening by integrating the XGBoost algorithm with the SH method for hyperparameter optimization. The SH-tuned model outperformed both the default and Randomized Search-tuned variants in terms of predictive accuracy, sensitivity, and computational efficiency, demonstrating its potential for scalable virtual screening applications. By significantly reducing the time and resources required for model tuning without sacrificing performance, this approach offers a valuable tool for accelerating early-stage drug discovery and supports the broader adoption of machine learning in pharmaceutical research.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

References

[1] Roopkumar, J., Swaidani, S., Kim, A. S., Thapa, B., Gervaso, L., Hobbs, B. P., Wei, W., Alban, T. J., Funchain, P., Kundu, S., Sangwan, N., Rayman, P., Pavicic, P. G., Diaz-Montero, C. M., Barnard, J., McCrae, K. R., & Khorana, A. A. (2021). Increased incidence of venous thromboembolism with cancer immunotherapy. Med, 2(4), 423-434.e3. https://doi.org/10.1016/j.medj.2021.02.002

[2] Lutsey, P. L., & Zakai, N. A. (2023). Epidemiology and prevention of venous thromboembolism. Nature Reviews Cardiology, 20(4), 248–262. https://doi.org/10.1038/s41569-022-00787-6

[3] Mackman, N., Bergmeier, W., Stouffer, G. A., & Weitz, J. I. (2020). Therapeutic strategies for thrombosis: new targets and approaches. Nature Reviews Drug Discovery, 19(5), 333–352. https://doi.org/10.1038/s41573-020-0061-0

[4] Tiwari, P. C., Pal, R., Chaudhary, M. J., & Nath, R. (2023). Artificial Intelligence Revolutionizing Drug Development: Exploring Opportunities and Challenges. Drug Development Research, 84(8), 1652–1663. https://doi.org/10.1002/ddr.22115

[5] Choudhuri, S., Yendluri, M., Poddar, S., Li, A., Mallick, K., Mallik, S., & Ghosh, B. (2023). Recent Advancements in Computational Drug Design Algorithms through Machine Learning and Optimization. Kinases and Phosphatases, 1(2), 117–140. https://doi.org/10.3390/kinasesphosphatases1020008

[6] Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Afidh, R. P. F., & Idroes, R. (2024). An Explainable Multi-Model Stacked Classifier Approach for Predicting Hepatitis C Drug Candidates. Sci, 6(4), 81. https://doi.org/10.3390/sci6040081

[7] Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V, Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., Isayev, O., Curtalolo, S., Fourches, D., Cohen, Y., Aspuru-Guzik, A., Winkler, D. A., Agrafiotis, D., Cherkasov, A., & Tropsha, A. (2020). QSAR without borders. Chemical Society Reviews, 49(11), 3525–3564. https://doi.org/10.1039/D0CS00098A

[8] Noviandy, T. R., Maulana, A., Idroes, G. M., Maulydia, N. B., Patwekar, M., Suhendra, R., & Idroes, R. (2023). Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery. Malacca Pharmaceutics, 1(2), 48–54. https://doi.org/10.60084/mp.v1i2.60

[9] Niazi, S. K., & Mariam, Z. (2023). Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review. International Journal of Molecular Sciences, 24(14), 11488. https://doi.org/10.3390/ijms241411488

[10] Tropsha, A., Isayev, O., Varnek, A., Schneider, G., & Cherkasov, A. (2024). Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. Nature Reviews Drug Discovery, 23(2), 141–155. https://doi.org/10.1038/s41573-023-00832-0

[11] Ali, Y., Awwad, E., Al-Razgan, M., & Maarouf, A. (2023). Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity. Processes, 11(2), 349. https://doi.org/10.3390/pr11020349

[12] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 785–794.

[13] Noviandy, T. R., Idroes, G. M., Maulana, A., Afidh, R. P. F., & Idroes, R. (2024). Optimizing Hepatitis C Virus Inhibitor Identification with LightGBM and Tree-structured Parzen Estimator Sampling. Engineering, Technology & Applied Science Research, 14(6), 18810–18817. https://doi.org/10.48084/etasr.8947

[14] Soper, D. S. (2022). Hyperparameter Optimization Using Successive Halving with Greedy Cross Validation. Algorithms, 16(1), 17. https://doi.org/10.3390/a16010017

[15] Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012). ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Research, 40(D1), D1100–D1107. https://doi.org/10.1093/nar/gkr777

[16] Noviandy, T. R., Idroes, G. M., & Hardi, I. (2025). An interpretable Bayesian-optimized XGBoost framework for neuropsychiatric drug candidate classification. Iran Journal of Computer Science. https://doi.org/10.1007/s42044-025-00297-x

[17] Noviandy, T. R., Idroes, G. M., Mohd Fauzi, F., & Idroes, R. (2024). Application of Ensemble Machine Learning Methods for QSAR Classification of Leukotriene A4 Hydrolase Inhibitors in Drug Discovery. Malacca Pharmaceutics, 2(2), 68–78. https://doi.org/10.60084/mp.v2i2.217

[18] Grisoni, F., Consonni, V., & Todeschini, R. (2018). Impact of Molecular Descriptors on Computational Models (pp. 171–209). https://doi.org/10.1007/978-1-4939-8639-2_5

[19] Moriwaki, H., Tian, Y. S., Kawashita, N., & Takagi, T. (2018). Mordred: A Molecular Descriptor Calculator. Journal of Cheminformatics, 10(1), 1–14. https://doi.org/10.1186/s13321-018-0258-y

[20] Noviandy, T. R., Imelda, E., Idroes, G. M., Suhendra, R., & Idroes, R. (2025). Evaluation of Machine Learning Methods for Identifying Carbonic Anhydrase-II Inhibitors as Drug Candidates for Glaucoma. Malacca Pharmaceutics, 3(1), 32–41. https://doi.org/10.60084/mp.v3i1.271

[21] Olayemi, Olanrewaju, S. (2020). Effects of Multicollinearity and Correlation between the Error Terms on Some Estimators in a System of Regression Equations. Global Journal of Science Frontier Research, 1(1), 77–94. https://doi.org/10.34257/GJSFRFVOL20IS4PG77

[22] Supriatna, D. J. I., Saputra, H., & Hasan, K. (2023). Enhancing the Red Wine Quality Classification Using Ensemble Voting Classifiers. Infolitika Journal of Data Science, 1(2), 42–47. https://doi.org/10.60084/ijds.v1i2.95

[23] Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Afidh, R. P. F., & Idroes, R. (2024). Machine Learning for Antiviral Drug Discovery: Application of XGBoost-TPE for Predicting Bioactivity of Hepatitis C Inhibitors. 2024 International Conference on Electrical Engineering and Informatics (ICELTICs), 62–67. https://doi.org/10.1109/ICELTICs62730.2024.10776122

[24] Brazdil, P., van Rijn, J. N., Soares, C., & Vanschoren, J. (2022). Metalearning for Hyperparameter Optimization (pp. 103–122). https://doi.org/10.1007/978-3-030-67024-5_6

[25] Rojas-Delgado, J., Jiménez, J. A., Bello, R., & Lozano, J. A. (2023). Hyper-parameter Optimization Using Continuation Algorithms (pp. 365–377). https://doi.org/10.1007/978-3-031-26504-4_26

[26] Noviandy, T. R., Maulana, A., Emran, T. B., Idroes, G. M., & Idroes, R. (2023). QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms. Heca Journal of Applied Sciences, 1(1), 1–7. https://doi.org/10.60084/hjas.v1i1.12