Original Article

Check for updates

Available online at
www.heca-analitika.com/hjas

Heca Journal of Applied Sciences

Vol. 4, No. 1, 2026

Comparative Analysis of Ensemble Machine Learning Models for QSAR-Based Prediction of Anticoagulant Activity in Thrombotic Disorders

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

[1] Department of Information Systems, Faculty of Engineering, Universitas Abulyatama, Aceh Besar 23372, Indonesia; rizky_si@abulyatama.ac.id (T.R.N.); rahmatsufri_si@abulyatama.ac.id (R.S); ryan_si@abulyatama.ac.id (R.Se.); anisah_si@abulyatama.ac.id (A.A.)

* Correspondence: rizky_si@abulyatama.ac.id

Article History

Received 9 January 2026
Revised 17 March 2026
Accepted 24 March 2026
Available Online 31 March 2026

Keywords:
Thrombin inhibitors
Molecular descriptors
Hyperparameter tuning

Abstract

Thrombotic disorders remain a major cause of global morbidity and mortality, with dysregulation of blood coagulation pathways playing a central role in disease progression. In particular, Thrombin is a key therapeutic target for anticoagulant drug development, making accurate prediction of inhibitory activity highly relevant for accelerating discovery efforts. Despite advances in computational drug discovery, there is still a need for systematic evaluation of machine learning approaches for QSAR-based prediction of anticoagulant activity. Many existing studies focus on single models or lack consistent comparison frameworks, limiting insights into the relative performance of different ensemble techniques. To address this gap, this study explores the application of multiple ensemble machine learning methods, including Random Forest, XGBoost, Gradient Boosting, and Extra Trees, combined with hyperparameter optimization using random search. The main objective of this work is to conduct a comparative analysis of these ensemble models to predict pIC50 values for thrombin inhibitors using molecular descriptors derived from chemical structures. The results show that the Extra Trees model achieved the best overall performance, with an R^2 of 0.697, RMSE of 0.851, and MAE of 0.615 after tuning. Additionally, Gradient Boosting and XGBoost demonstrated significant improvement following hyperparameter optimization, highlighting the importance of model tuning in QSAR tasks. Overall, the study confirms that ensemble learning methods yield reliable, accurate predictions of anticoagulant activity, with Extra Trees emerging as the most effective approach for this dataset.

1. Introduction

Thrombotic disorders remain a major global health challenge, contributing significantly to morbidity and mortality through conditions such as deep vein thrombosis, pulmonary embolism, and ischemic stroke [1, 2]. Central to the pathophysiology of these disorders is the dysregulation of blood coagulation pathways, particularly the excessive activation of clotting factors [3, 4]. Among these, Thrombin plays a pivotal role as a serine protease that converts fibrinogen into fibrin, ultimately leading to clot formation [5]. Given its central role, thrombin has emerged as an important therapeutic target for developing anticoagulant agents to prevent or treat thrombotic events [6].

In addition to its central role in fibrin formation, thrombin also amplifies the coagulation cascade by activating multiple clotting factors and platelets, making it a critical convergence point in hemostasis. Compared with other anticoagulant targets, such as Factor Xa or protease-activated receptor-1 (PAR-1), thrombin inhibition

provides a more direct mechanism to suppress clot formation at its final stage. However, currently available thrombin-targeting anticoagulants present several limitations [7]. For instance, warfarin, although widely used, requires careful monitoring due to its narrow therapeutic window and significant drug–food interactions. Direct thrombin inhibitors such as dabigatran offer improved specificity but are still associated with risks of bleeding complications and variability in patient response [8]. These limitations highlight the need to identify novel thrombin inhibitors with improved safety and efficacy profiles. Therefore, developing accurate predictive models specifically targeting thrombin inhibition remains an important objective in computational drug discovery.

Traditional drug discovery approaches for anticoagulants are often time-consuming, costly, and resource-intensive, involving extensive experimental screening and optimization [9]. In recent years, computational methods have gained increasing attention as efficient alternatives to accelerate the drug discovery pipeline. Among these, Quantitative Structure–Activity Relationship (QSAR) modeling has proven a valuable tool for predicting the biological activity of chemical compounds using molecular descriptors [10]. QSAR enables researchers to uncover relationships between chemical structure and pharmacological effect, thereby reducing the need for exhaustive laboratory testing and facilitating the identification of promising lead compounds [11].

The rapid advancement of machine learning techniques has further enhanced the predictive capabilities of QSAR models [12, 13]. Ensemble machine learning methods, in particular, have demonstrated superior performance by combining multiple base learners to improve accuracy, robustness, and generalization [14]. Algorithms such as Random Forest, XGBoost, Gradient Boosting, and Extra Trees have been widely applied in cheminformatics for their ability to capture complex nonlinear relationships in high-dimensional data [14, 15].

Despite the growing application of ensemble learning in drug discovery, limited studies have specifically focused on thrombin as a target while simultaneously benchmarking multiple ensemble algorithms under consistent experimental conditions. Furthermore, existing QSAR studies often do not explicitly address the challenges associated with currently available thrombin inhibitors, such as safety concerns and variability in therapeutic response, which underscores the need for improved predictive modeling approaches tailored to this target. Such comparative analyses are essential for identifying the most suitable modeling approach for accurate prediction and for providing insights into the strengths and limitations of each method when applied to QSAR datasets.

In this study, we present a comparative analysis of multiple ensemble machine learning models for QSAR-based prediction of thrombin inhibitory activity. To the best of our knowledge, this is one of the first studies to directly benchmark several widely used ensemble algorithms, including Random Forest, Gradient Boosting, XGBoost, and Extra Trees, within a unified experimental framework using the same dataset, descriptor set, and validation strategy. The results not only identify the most effective algorithm for predicting thrombin inhibition but also highlight the relative strengths and limitations of different ensemble strategies. Therefore, this work provides a clearer understanding of model selection in QSAR modeling and offers practical guidance for future computational drug discovery efforts targeting thrombin.

The findings of this research aim to contribute to the field of computational drug discovery by identifying the most effective ensemble learning approach for QSAR-based prediction of thrombin inhibitors. By providing a detailed comparison of model performance and highlighting optimal parameter configurations, this study seeks to support the development of more accurate and efficient predictive models. Ultimately, such advancements can facilitate the discovery of novel anticoagulant agents, improve therapeutic strategies for thrombotic disorders, and reduce the overall cost and time associated with drug development.

2. Materials and Methods

The overall workflow of this study is illustrated in Figure 1. The process begins with data collection from the ChEMBL database, followed by the calculation of molecular descriptors to transform chemical structures into numerical features. These descriptors are then used as input for multiple ensemble machine learning models, including Random Forest, XGBoost, Gradient Boosting, and Extra Trees. Each model is trained and optimized through hyperparameter tuning. Finally, model performance is evaluated using standard regression metrics, and the best-performing model is selected based on its predictive accuracy.

2.1. Data Collection

The dataset used in this study was retrieved from the ChEMBL database, focusing on compounds associated with the Thrombin target (CHEMBL204) and their reported IC_{50} values [16]. To maintain data reliability, any compounds lacking IC_{50} information were removed from the dataset. In cases where a compound had multiple IC_{50}

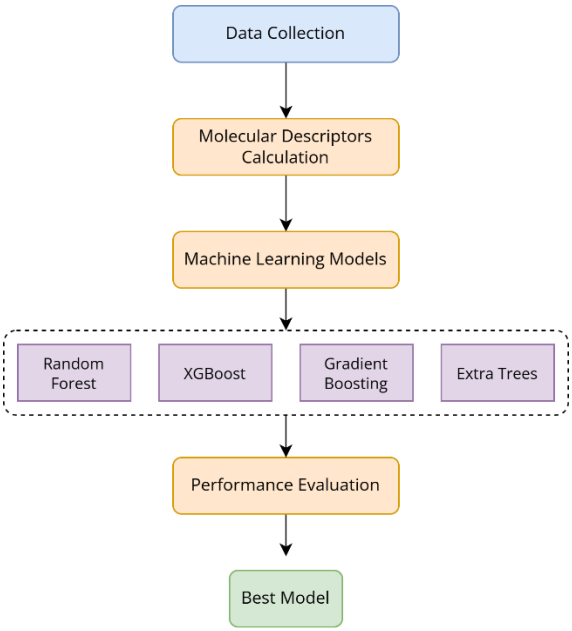

Figure 1. Workflow of the QSAR-based machine learning framework for predicting thrombin inhibitory activity, including data collection, descriptor calculation, model development, and performance evaluation.

measurements, the median value was selected to represent its activity, ensuring a more robust estimate.

After preprocessing, the dataset contained 3,145 unique compounds. The IC_{50} values were then transformed to pIC_{50}, defined as the negative logarithm (base 10) of the IC_{50}, to normalize the scale and improve model performance [17]. Finally, the dataset was divided into training and testing subsets using an 80:20 split, with stratification applied to maintain the original distribution of activity values across both sets.

2.2. Molecular Descriptors Calculation

To convert chemical structures into a format suitable for machine learning, molecular descriptors were used as input features. These descriptors provide numerical representations of different molecular characteristics, including structural patterns, atom composition, and physicochemical properties [18]. By encoding such information, they allow models to learn relationships between molecular structure and biological activity.

In this study, descriptors were generated using the Mordred tool, which computes a comprehensive set of both two-dimensional and three-dimensional descriptors for each compound [19]. This step resulted in a large number of features describing various aspects of the molecules.

To improve model efficiency and avoid redundant information, feature selection was applied [20]. Descriptors with very low variance (below 0.1) were removed, as they contribute minimal useful information for prediction [21]. In addition, highly correlated descriptors (Pearson correlation coefficient> 0.80) were excluded to reduce multicollinearity, which can negatively affect model stability and interpretation [22]. These thresholds were selected based on common practices in QSAR studies to balance dimensionality reduction and information retention. After applying these filtering steps, a total of 309 molecular descriptors were retained and used as input features for subsequent model development.

2.3. Machine Learning Models

In this study, several ensemble-based machine learning algorithms were employed to model the relationship between molecular descriptors and anticoagulant activity. Ensemble methods combine multiple learning algorithms to improve predictive performance and robustness compared to single models. The models evaluated include Random Forest, XGBoost, Gradient Boosting, and Extra Trees.

Initially, each model was trained with its default hyperparameters to establish a baseline performance. This step provides a reference point for evaluating the extent of improvement achievable through optimization. Following this, hyperparameter tuning was performed using random search with 3-fold cross-validation and 20 iterations, which explores a range of parameter combinations by randomly sampling from predefined distributions. Compared with an exhaustive grid search, this approach is more efficient while still delivering strong performance [23]. The ranges of hyperparameters explored for each model are summarized in Table 1.

Table 1. Hyperparameter search space used for randomized search in ensemble machine learning models.

Model	Hyperparameter	Values Tested
XGBoost	n_estimators	100, 200, 300
	max_depth	3, 5, 7, 10
	learning_rate	0.01, 0.05, 0.1
	subsample	0.7, 0.8, 1.0
	colsample_bytree	0.7, 0.8, 1.0
Random Forest	n_estimators	100, 200, 300
	max_depth	None, 10, 20, 30
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 2, 4
Gradient Boosting	n_estimators	100, 200, 300
	learning_rate	0.01, 0.05, 0.1
	max_depth	3, 5, 7
	subsample	0.7, 0.8, 1.0
Extra Trees	n_estimators	100, 200, 300
	max_depth	None, 10, 20, 30
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 2, 4

The tuning process was carried out using cross-validation to ensure reliable model evaluation and to reduce the risk of overfitting. Key hyperparameters such as the number of estimators, tree depth, and learning rate were optimized for each model. The best-performing configuration identified during the search was then used to train the final model, which was subsequently evaluated on the test dataset.

2.4. Performance Evaluation

The performance of the developed models was assessed using several standard regression metrics to evaluate prediction accuracy [24]. In this study, three commonly used metrics were applied: the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). The mathematical definitions of R^2, RMSE, and MAE are presented in Equations (1)–(3), respectively:

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

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

where y_i represents the observed (experimental) value, $\hat{y}_i$ denotes the predicted value, $\bar{y}$ is the mean of the observed values, and n is the total number of samples.

The R^2 score measures how well the model explains the variance in the observed data, with values closer to 1 indicating better predictive performance. In QSAR modeling, an R^2 value greater than 0.6 is generally considered acceptable, indicating moderate predictive capability, while values above 0.7 reflect good model performance and strong agreement between predicted and experimental values.

RMSE quantifies the average magnitude of prediction errors, with greater sensitivity to larger errors due to the squaring of residuals. Lower RMSE values indicate better model accuracy; in QSAR studies, RMSE values below 1.0 are typically considered acceptable for continuous bioactivity prediction, depending on the dataset's variability.

MAE measures the average absolute difference between predicted and actual values, providing a more interpretable estimate of prediction error that is less influenced by outliers compared to RMSE. Lower MAE values indicate more accurate and consistent predictions, and values below 0.7 generally indicate good predictive performance in similar QSAR applications.

Together, these metrics offer a comprehensive evaluation of model performance. While R^2 reflects the overall goodness of fit, RMSE and MAE provide insight into the scale of prediction errors. Using all three metrics allows for a more balanced comparison between models and helps identify the most reliable approach for predicting anticoagulant activity.

3. Results and Discussion

The performance of the ensemble models before and after hyperparameter tuning is summarized in Table 2. Overall, the models demonstrated moderate to strong predictive ability in estimating anticoagulant activity against Thrombin, with noticeable improvements after applying hyperparameter optimization.

Table 2. Performance comparison of ensemble machine learning models before and after hyperparameter tuning for predicting pIC_{50} values of compounds targeting Thrombin.

Model	R^2	RMSE	MAE
Random Forest	0.655	0.909	0.685
Random Forest (Tuned)	0.659	0.903	0.679
XGBoost	0.643	0.924	0.671
XGBoost (Tuned)	0.666	0.894	0.649
Gradient Boosting	0.570	1.014	0.778
Gradient Boosting (Tuned)	0.675	0.882	0.654
Extra Trees	0.690	0.861	0.617
Extra Trees (Tuned)	0.697	0.851	0.615

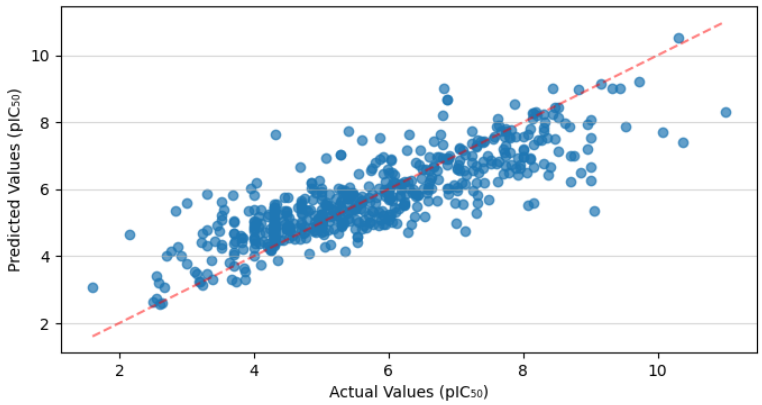

Figure 2. Actual vs predicted pIC_{50} values for the Extra Trees model

Among the baseline models, Extra Trees achieved the best performance, with an R^2 of 0.690, an RMSE of 0.861, and an MAE of 0.617. This suggests that the model captured the underlying relationships between molecular descriptors and biological activity more effectively than the other untuned models. Random Forest and XGBoost showed comparable performance, though slightly lower than that of Extra Trees. At the same time, Gradient Boosting produced the weakest results in its default configuration, with the lowest R^2 (0.570) and highest error values.

After applying hyperparameter tuning using randomized search, all models exhibited improvements in predictive performance, although the magnitude of improvement varied. The most notable enhancement was observed for Gradient Boosting, where the R^2 increased significantly from 0.570 to 0.675, accompanied by substantial reductions in RMSE and MAE. This indicates that Gradient Boosting is highly sensitive to parameter configuration and can perform competitively when properly optimized. Similarly, XGBoost showed meaningful gains, improving its R^2 from 0.643 to 0.666 and reducing prediction errors, highlighting the importance of tuning for boosting-based algorithms.

For tree-based bagging methods, the improvements were more modest. Random Forest showed only slight enhancement after tuning, with R^2 increasing marginally from 0.655 to 0.659. Extra Trees, which already performed strongly in its default state, achieved the highest overall performance after tuning, with an R2 of 0.697, an RMSE of 0.851, and an MAE of 0.615. This suggests that Extra Trees is inherently more robust and less dependent on extensive hyperparameter optimization than boosting methods.

Across all models, the tuned Extra Trees model emerged as the best-performing approach for this QSAR task, followed closely by the tuned Gradient Boosting and XGBoost models. The results indicate that ensemble methods based on randomized tree construction and boosting strategies are effective for modeling complex, nonlinear relationships in molecular descriptor data. Additionally, the consistent reduction in RMSE and MAE across the tuned models confirms that hyperparameter optimization yields more accurate and reliable predictions.

The scatter plot in Figure 2 illustrates the relationship between experimentally observed pIC_{50} values and those predicted by the Extra Trees model. Each point represents a compound, with the x-axis corresponding to actual values and the y-axis representing predicted values. The dashed diagonal line indicates the ideal scenario where predicted values perfectly match the experimental data.

As shown in the figure, the data points are generally distributed near the diagonal, indicating good agreement

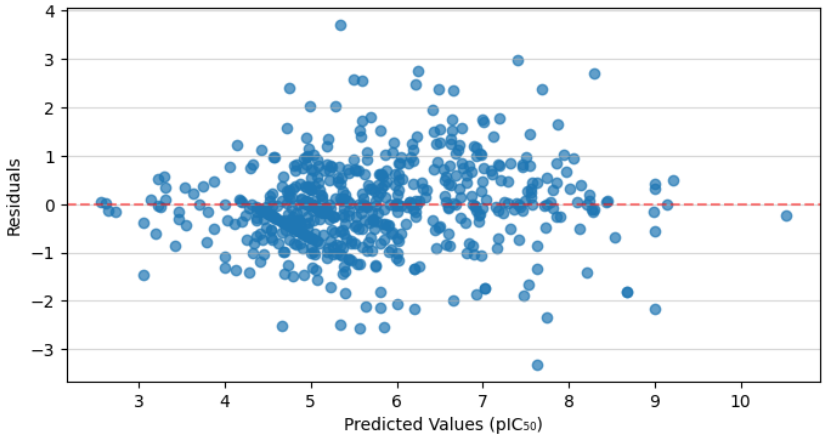

Figure 3. Residual plot for the Extra Trees model.

between predicted and actual values. This pattern reflects the strong predictive performance of the Extra Trees model, which achieved the highest R^2 and lowest error metrics among all evaluated models. The clustering of points around the diagonal suggests that the model captures the underlying relationship between molecular descriptors and anticoagulant activity targeting Thrombin.

However, some deviations from the diagonal line are observed, particularly at higher pIC_{50} values, where a few predictions either underestimate or overestimate the actual activity. This indicates that while the model performs well overall, there is still some variability in predicting compounds with extreme activity values. Such discrepancies are common in QSAR modeling due to structural diversity and limitations in descriptor representation.

Figure 3 presents the residual distribution of the Extra Trees model, where residuals (differences between actual and predicted pIC_{50} values) are plotted against the predicted values. The horizontal dashed line at zero represents perfect predictions, with no error.

The residuals appear to be randomly scattered around the zero line, indicating that the model does not exhibit strong systematic bias across the prediction range. This random distribution suggests that the model is well-fitted and capable of capturing the general relationship between molecular descriptors and anticoagulant activity against Thrombin.

However, a slight increase in the spread is observed at higher predicted values, indicating mild heteroscedasticity. This means prediction errors tend to be larger for more active compounds, a common challenge in QSAR modeling due to increased structural complexity or limited representation of highly active compounds in the dataset. Additionally, a few outliers with relatively large residuals are present, suggesting that certain compounds are more difficult for the model to predict accurately.

The results of this study demonstrate that ensemble learning methods are effective for QSAR-based prediction of anticoagulant activity targeting Thrombin. Among the evaluated models, the Extra Trees model consistently achieved the best performance, both before and after hyperparameter tuning. Its ability to handle high-dimensional descriptor space and capture nonlinear relationships likely contributed to its superior predictive accuracy. The strong alignment between actual and predicted values, along with a stable residual distribution, further supports its robustness.

Hyperparameter tuning played a significant role in improving model performance, particularly for boosting-based methods such as Gradient Boosting and XGBoost. These models showed substantial gains after optimization, highlighting their sensitivity to parameter settings. In contrast, bagging-based methods like Extra Trees and Random Forest demonstrated relatively stable performance even with default configurations, suggesting they are less dependent on extensive tuning. This difference emphasizes the importance of selecting appropriate models based on both performance and computational efficiency.

Overall, the findings indicate that ensemble approaches are well-suited for modeling complex structure–activity relationships in cheminformatics datasets. The combination of diverse molecular descriptors and robust machine learning algorithms enables reliable prediction of biological activity. These results reinforce the value of integrating computational modeling into early-stage drug discovery, where accurate prediction of thrombin inhibition can help prioritize promising compounds and reduce experimental costs.

Despite the promising results, this study has some limitations. The model's performance depends on the quality and diversity of the dataset, and using only molecular descriptors may not fully capture all relevant chemical and biological interactions. Additionally, although sufficient, the dataset size may still limit the model's ability to generalize to entirely new chemical spaces, particularly for compounds with extreme activity values.

Future research could focus on improving model performance by incorporating additional data representations, such as molecular fingerprints or graph-based features, alongside descriptors. The use of advanced optimization techniques, such as Bayesian optimization or deep learning approaches, may further enhance predictive accuracy. Moreover, external validation on independent datasets and experimental verification of the predicted compounds would strengthen the applicability of the proposed models to real-world drug discovery.

4. Conclusions

This study demonstrates the effectiveness of ensemble machine learning approaches for QSAR-based prediction of anticoagulant activity targeting Thrombin. Among the evaluated models, the Extra Trees model achieved the best overall performance, showing strong predictive accuracy and robustness across evaluation metrics. The results also highlight the importance of hyperparameter tuning, particularly for boosting methods, in enhancing model performance. Overall, the findings support the use of ensemble learning as a reliable and efficient tool for identifying potential thrombin inhibitors, thereby accelerating drug discovery for thrombotic disorders.

Author Contributions: Conceptualization, T.R.N. and R.S.; methodology, T.R.N. and R.Se.; software, T.R.N.; validation, T.R.N., R.Se., and A.A.; formal analysis, T.R.N.; investigation, R.S. and R.Se.; resources, A.A.; data curation, A.A.; writing—original draft preparation, T.R.N., R.Se., and A.A.; writing—review and editing, T.R.N. and R.S.; visualization, R.S.; supervision, T.R.N.; project administration, T.R.N.; funding acquisition, T.R.N. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

1. Oleksiuk-Bójko, M., and Lisowska, A. (2023). Venous Thromboembolism: Why Is It Still a Significant Health Problem?, *Advances in Medical Sciences*, Vol. 68, No. 1, 10–20. doi:10.1016/j.advms.2022.10.002.
2. Lutsey, P. L., and Zakai, N. A. (2023). Epidemiology and Prevention of Venous Thromboembolism, *Nature Reviews Cardiology*, Vol. 20, No. 4, 248–262. doi:10.1038/s41569-022-00787-6.
3. Wilhelm, G., Mertowska, P., Mertowski, S., Przysucha, A., Strużyna, J., Grywalska, E., and Torres, K. (2023). The Crossroads of the Coagulation System and the Immune System: Interactions and Connections, *International Journal of Molecular Sciences*, Vol. 24, No. 16, 12563. doi:10.3390/ijms241612563.
4. Al-Koussa, H., AlZaim, I., and El-Sabban, M. E. (2022). Pathophysiology of Coagulation and Emerging Roles for Extracellular Vesicles in Coagulation Cascades and Disorders, *Journal of Clinical Medicine*, Vol. 11, No. 16, 4932. doi:10.3390/jcm11164932.
5. Al-Amer, O. M. (2022). The Role of Thrombin in Haemostasis, *Blood Coagulation & Fibrinolysis*, Vol. 33, No. 3, 145–148. doi:10.1097/MBC.0000000000001130.
6. Mackman, N., Bergmeier, W., Stouffer, G. A., and Weitz, J. I. (2020). Therapeutic Strategies for Thrombosis: New Targets and Approaches, *Nature Reviews Drug Discovery*, Vol. 19, No. 5, 333–352. doi:10.1038/s41573-020-0061-0.
7. Jannati, S., Patnaik, R., and Banerjee, Y. (2024). Beyond Anticoagulation: A Comprehensive Review of Non-Vitamin K Oral Anticoagulants (NOACs) in Inflammation and Protease-Activated Receptor Signaling, *International Journal of Molecular Sciences*, Vol. 25, No. 16, 8727. doi:10.3390/ijms25168727.
8. Sangam, S., and Gudi, S. K. (2025). The Role of Digital Health in the Management of Warfarin Therapy, *Drugs & Therapy Perspectives*, Vol. 41, No. 2, 63–74. doi:10.1007/s40267-024-01134-0.
9. An, Q., Huang, L., Wang, C., Wang, D., and Tu, Y. (2025). New Strategies to Enhance the Efficiency and Precision of Drug Discovery, *Frontiers in Pharmacology*, Vol. 16. doi:10.3389/fphar.2025.1550158.
10. Niazi, S. K., and Mariam, Z. (2023). Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review, *International Journal of Molecular Sciences*, Vol. 24, No. 14, 11488. doi:10.3390/ijms241411488.
11. Altememi, M. A., Favaloro, E. J., Islam, M. Z., and Santhakumar, A. B. (2026). Artificial Intelligence and Machine Learning in Thrombosis and Hemostasis: A Scoping Review of Clinical and Laboratory Applications, Challenges, and Future Directions, *Clinical Chemistry and Laboratory Medicine (CCLM)*, Vol. 64, No. 4, 767–780. doi:10.1515/cclm-2025-1450.
12. De Borja, J. R., and Cabrera, H. S. (2024). In Silico Drug Screening for Hepatitis C Virus Using QSAR-ML and Molecular Docking with Rho-Associated Protein Kinase 1 (ROCK1) Inhibitors, *Computation*, Vol. 12, No. 9, 175. doi:10.3390/computation12090175.
13. Hammoudi, N.-E.-H., Sobhi, W., Attoui, A., Lemaoui, T., Erto, A., and Benguerba, Y. (2021). In Silico Drug Discovery of Acetylcholinesterase and Butyrylcholinesterase Enzymes Inhibitors Based on Quantitative Structure-Activity Relationship (QSAR) and Drug-Likeness Evaluation, *Journal of Molecular Structure*, Vol. 1229, 129845. doi:10.1016/j.molstruc.2020.129845.
14. Noviandy, T. R., Idroes, G. M., Mohd Fauzi, F., and Idroes, R. (2024). Application of Ensemble Machine Learning Methods for QSAR Classification of Leukotriene A4 Hydrolase Inhibitors in Drug Discovery, *Malacca Pharmaceutics*, Vol. 2, No. 2, 68–78. doi:10.60084/mp.v2i2.217.
15. Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Afidh, R. P. F., and Idroes, R. (2024). An Explainable Multi-Model

Stacked Classifier Approach for Predicting Hepatitis C Drug Candidates, *Sci*, Vol. 6, No. 4, 81. doi:10.3390/sci6040081.

16. Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. (2012). ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery, *Nucleic Acids Research*, Vol. 40, No. D1, D1100–D1107. doi:10.1093/nar/gkr777.

17. Thakur, A., Kumar, A., Sharma, V. kumar, and Mehta, V. (2022). PIC50: An Open Source Tool for Interconversion of PIC50 Values and IC50 for Efficient Data Representation and Analysis, *BioRxiv*, 2010–2022.

18. Mauri, A., Consonni, V., and Todeschini, R. (2017). Molecular Descriptors, *Handbook of Computational Chemistry*, Springer International Publishing, Cham, 2065–2093. doi:10.1007/978-3-319-27282-5_51.

19. Moriwaki, H., Tian, Y. S., Kawashita, N., and Takagi, T. (2018). Mordred: A Molecular Descriptor Calculator, *Journal of Cheminformatics*, Vol. 10, No. 1, 1–14. doi:10.1186/s13321-018-0258-y.

20. Goodarzi, M., Dejaegher, B., and Heyden, Y. Vander. (2012). Feature Selection Methods in QSAR Studies, *Journal of AOAC International*, Vol. 95, No. 3, 636–651. doi:10.5740/jaoacint.SGE_Goodarzi.

21. Noviandy, T. R., Idroes, G. M., Tallei, T. E., Handayani, D., and Idroes, R. (2024). QSAR Modeling for Predicting Beta-Secretase 1 Inhibitory Activity in Alzheimer's Disease with Support Vector Regression, *Malacca Pharmaceutics*, Vol. 2, No. 2, 79–85. doi:10.60084/mp.v2i2.226.

22. 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*, Vol. 1, No. 1, 77–94. doi:10.34257/GJSFRFVOL20IS4PG77.

23. Bergstra, J., and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization., *Journal of Machine Learning Research*, Vol. 13, No. 2.

24. Kalyankar, D. S., Bhagat, C. G., Kadu, A. D., Tambade, A. P., and Dhoran, K. S. (2024). AI-Driven Insights: Paving the Path to Next-Generation Therapeutics, *International Journal of Advanced Research in Science, Communication and Technology*, 372–378. doi:10.48175/IJARSCT-22854.