Original Article

Check for updates

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

Malacca Pharmaceutics

Vol. 2, No. 2, 2024

QSAR Modeling for Predicting Beta-Secretase 1 Inhibitory Activity in Alzheimer's Disease with Support Vector Regression

Teuku Rizky Noviandy [1], Ghifari Maulana Idroes [2], Trina Ekawati Tallei [3], Dian Handayani [4] and Rinaldi Idroes [5,*]

[1] Interdisciplinary Innovation Research Unit, Graha Primera Saintifika, Aceh Besar 23771, Indonesia; trizkynoviandy@gmail.com (T.R.N.)
[2] Department of Nuclear Engineering and Engineering Physics, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia; ghifarimaulana145@gmail.com (G.M.I.)
[3] Department of Biology, Faculty of Mathematics and Natural Sciences, Sam Ratulangi University, Manado, Indonesia; trina_tallei@unsrat.ac.id (T.E.T.)
[4] Sumatran Biota Laboratory, Faculty of Pharmacy, Universitas Andalas, 25163 Padang, Indonesia; dianhandayani@phar.unand.ac.id (D.H.)
[5] Department of Pharmacy, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala Banda Aceh 23111, Indonesia; rinaldi.idroes@usk.ac.id (R.I.)

* Correspondence: rinaldi.idroes@usk.ac.id

Article History

Received 19 July 2024
Revised 15 September 2024
Accepted 24 September 2024
Available Online 30 September 2024

Keywords:
BACE1
Machine learning
Molecular descriptors
Supervised learning

Abstract

Alzheimer's disease (AD) is a neurodegenerative disorder characterized by cognitive decline, with the accumulation of β-amyloid (Aβ) plaques playing a key role in its progression. Beta-Secretase 1 (BACE1) is a crucial enzyme in Aβ production, making it a prime therapeutic target for AD treatment. However, designing effective BACE1 inhibitors has been challenging due to poor selectivity and limited blood-brain barrier permeability. To address these challenges, we employed a machine learning approach using Support Vector Regression (SVR) in a Quantitative Structure-Activity Relationship (QSAR) model to predict the inhibitory activity of potential BACE1 inhibitors. Our model, trained on a dataset of 7,298 compounds from the ChEMBL database, accurately predicted pIC_{50} values using molecular descriptors, achieving an R^2 of 0.690 on the testing set. The model's performance demonstrates its utility in prioritizing drug candidates, potentially accelerating drug discovery. This study highlights the effectiveness of computational approaches in optimizing drug discovery and suggests that further refinement could enhance the model's predictive power for AD therapeutics.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline and memory impairment, affecting millions of individuals worldwide [1]. A hallmark of AD is the accumulation of β-amyloid (Aβ) plaques in the brain, which are believed to play a crucial role in neuronal dysfunction and death [2]. Beta-Secretase 1 (BACE1) is a key enzyme responsible for the initial step in producing Aβ peptides through the cleavage of amyloid precursor protein (APP) [3]. Consequently, BACE1 has emerged as a promising therapeutic target for developing drugs to reduce Aβ levels and slow the progression of AD [4–6].

Despite extensive research, designing effective BACE1 inhibitors has been challenging due to poor selectivity, limited blood-brain barrier permeability, and adverse

side effects. Traditional drug discovery methods are time-consuming and costly [7], highlighting the need for efficient computational approaches to identify potent BACE1 inhibitors [8, 9]. Quantitative Structure-Activity Relationship (QSAR) modeling is a computational technique that correlates the chemical structures of compounds with their biological activities, enabling the prediction of the activity of new, untested molecules [10–12].

Machine learning algorithms have significantly enhanced QSAR modeling by capturing complex, nonlinear relationships between molecular descriptors and biological activity [13–16]. Support vector regression (SVR) has shown considerable promise among these algorithms due to its ability to handle high-dimensional data and model complex relationships using kernel functions [17, 18]. SVR is a powerful machine learning technique that extends Support Vector Machines (SVM) to regression problems. SVR works by finding a function that approximates the relationship between molecular descriptors and biological activity with a margin of tolerance (epsilon) while maintaining good generalization capabilities [19].

In this study, we aim to develop an SVR-based QSAR model to predict the inhibitory activity of potential BACE1 inhibitors for Alzheimer's disease therapeutics. By utilizing a dataset of known BACE1 inhibitors with experimentally determined activities, our goal is to create a predictive model that can accurately estimate the activity of novel compounds. This computational approach will help prioritize promising candidates for further synthesis and testing, ultimately accelerating the drug discovery process and reducing the associated costs and resources.

2. Materials and Methods

2.1. Dataset

The dataset used in this study was retrieved from the ChEMBL database [20], specifically targeting BACE1 inhibitors (ChEMBL target ID: CHEMBL4822). The dataset consists of 7,298 compounds, each represented by its Simplified Molecular Input Line Entry System (SMILES) notation and their respective IC_{50} values, which indicate the inhibitory concentration required to reduce BACE1 activity by 50%. SMILES is a text-based format that encodes the structure of chemical compounds, serving as the input for molecular descriptor calculation [21]. The IC_{50} values are then converted to pIC_{50}, a more convenient measure of potency. The pIC_{50} scale allows for easier comparison, with higher values indicating stronger inhibition. This transformation standardizes the data, making it more suitable for further analysis and modeling [22].

Molecular descriptors are quantitative representations of the chemical properties and structural attributes of molecules [23]. These descriptors capture various aspects of molecular structure, such as size, shape, functional groups, and electronic distribution [24]. They are often used as features in computational models to correlate chemical structure with biological activity. In this study, we used the Mordred software to calculate 2D molecular descriptors from the SMILES representations of the compounds [25], yielding 1,613 descriptors per compound. The 2D descriptors encompass various categories, including topological, constitutional, and electronic properties. We focused on 2D descriptors because they are computationally less intensive than 3D descriptors and are widely used in QSAR studies, where 2D representations often provide sufficient information for predictive modeling.

We applied feature selection methods to reduce redundancy and enhance the model's performance. First, a variance threshold of 0.1 was used to remove descriptors with low variability across the dataset, ensuring that only informative features were retained. Next, descriptors with high multicollinearity were removed by applying a correlation threshold of 0.80, discarding highly correlated features. After these filtering steps, 156 molecular descriptors remained for use in the QSAR model.

The dataset was split into training (80%) and testing (20%) sets using a stratified approach, ensuring that both sets maintained the same distribution of pIC_{50} values as the original dataset. This stratification also helped preserve chemical diversity, which is crucial for the model's ability to generalize across a wide range of compounds and activity levels. Specifically, 80% of the compounds were allocated for training the model, while the remaining 20% were reserved for testing and validating its predictive performance [26, 27]. The distribution of compounds in the training and testing sets is shown in Figure 1. The training set shows a normal-like distribution centered around pIC_{50} values between 6 and 8. The testing set follows a similar distribution, indicating that the data split preserved the overall distribution of inhibitory activity across both sets.

2.2. Support Vector Regression

SVR is a machine learning algorithm designed to predict continuous target variables, making it well-suited for tasks such as predicting pIC_{50} values. SVR works by finding a function that makes predictions within a specified margin of tolerance, denoted as ϵ, ensuring that

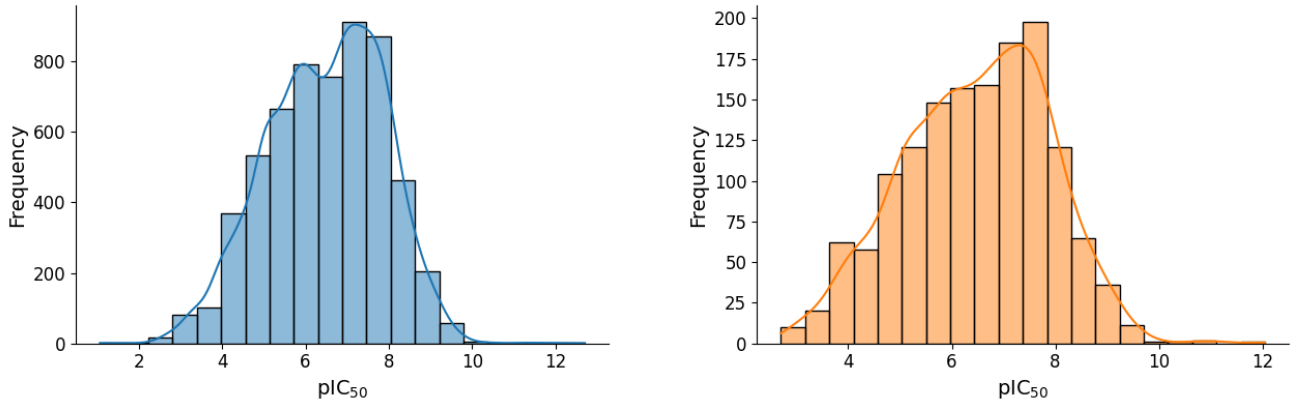


Figure 1. Distribution of compounds in (a) the training set and (b) the testing set.

most predictions are close to the actual values. The algorithm also strives to balance prediction accuracy with model simplicity, reducing the risk of overfitting, where the model performs well on training data but poorly on unseen data [28].

We chose SVR for this study because of its effectiveness in handling high-dimensional data and its ability to model complex relationships between molecular descriptors and biological activity [29]. This makes SVR particularly suitable for tasks like QSAR modeling, where the relationship between chemical structure and activity may not always be straightforward or linear. To account for potential non-linearities in the structure-activity relationship, we used a non-linear kernel function in SVR, which allows the model to capture more complex patterns in the data. Compared to other machine learning algorithms, SVR was selected for its ability to deliver accurate predictions while maintaining robustness in the face of high-dimensional data and small training datasets, which are common in cheminformatics [30, 31].

The SVR algorithm aims to fit a function. $f(x)$ that has at most ϵ deviation from the actual target values for all training data points. Mathematically, the function $f(x)$ is expressed as a linear combination of kernel functions, as shown in Equation 1:

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b \qquad (1)$$

where α_i and α_i^* are the Lagrange multipliers, $K(x_i, x)$ represents the kernel function, x_i are the input data points, and b is the bias term. The kernel function $K(x_i, x)$ allows SVR to model nonlinear relationships by mapping the input data into a higher-dimensional feature space.

In this study, we use the Radial Basis Function (RBF) kernel, widely used in SVR models for its ability to capture complex nonlinear patterns in the data effectively. The RBF kernel is defined as shown in Equation 2:

$$K(x_i, x) = \exp\left(-\gamma \left|| x_i - x \right|\right|^2\right) \qquad (2)$$

where γ is a hyperparameter that determines the width of the kernel. The RBF kernel transforms the input data into a higher-dimensional space, making it suitable for modeling nonlinear relationships between molecular descriptors and biological activity.

The SVR model minimizes the following objective function, as shown in Equation 3:

$$\frac{1}{2} || w ||^2 + C \sum_{i=1}^{n} max(0, |y_i - f(x_i)| - \epsilon) \qquad (3)$$

where $|| w ||^2$ represents the regularization term that penalizes model complexity, C is a hyperparameter that controls the trade-off between model complexity and prediction accuracy, y_i is the actual target value, and ϵ is the margin of tolerance.

In this study, we employ SVR with the RBF kernel to model the relationship between molecular descriptors and the biological activity (pIC_{50} values) of BACE1 inhibitors. The hyperparameters of the SVR model, including C and γ, were optimized through grid search with 5-fold cross-validation to ensure the best predictive performance [32, 33]. The grid search explored the values of C as [0.1, 1, 10, 100] and gamma as ['scale', 'auto']. The 'scale' option for gamma sets the value based on the data's variance, while 'auto' sets gamma based on the number of features. We set ϵ to 0.1 to control the margin of tolerance in the predictions.

2.3. Performance Evaluation

The performance of the SVR model was evaluated using three key metrics: the coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics assess the model's ability to predict the biological activity of BACE1 inhibitors.

The R^2 measures how well the predictions from the model capture the variance in the actual data. It indicates the proportion of the variance in the target variable explained by the model, with higher values (closer to 1) suggesting a better fit [34]. The RMSE quantifies the model's prediction error by calculating the square root of the average squared differences between the predicted and actual values. RMSE is sensitive to large errors, making it useful for capturing the overall magnitude of the prediction error [35]. The MAE measures the average of the absolute differences between predicted and actual values. Unlike RMSE, MAE does not disproportionately penalize larger errors, providing a more straightforward interpretation of the prediction error [36]. The corresponding equations for these metrics are shown in Equations 4, 5, and 6, respectively:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

where y_i represents the actual values, $\hat{y}_i$ are the predicted values, $\bar{y}$ is the mean of the actual values, and n is the total number of compounds used in the dataset.

3. Results and Discussion

The SVR-based QSAR model for predicting BACE1 inhibitory activity demonstrated strong performance across the training, cross-validation, and testing datasets. Table 1 summarizes the key performance metrics: R^2, RMSE, and MAE, which were used to evaluate the model's ability to generalize and predict the biological activity of potential inhibitors.

On the training dataset, the model achieved an R^2 value of 0.911, indicating that 91.1% of the variance in BACE1 inhibitory activity could be explained by the molecular descriptors selected for the QSAR model. The low RMSE (0.417) and MAE (0.209) values further reflect the model's ability to accurately fit the training data, suggesting that the SVR algorithm effectively captured the key relationships between chemical structure and biological activity.

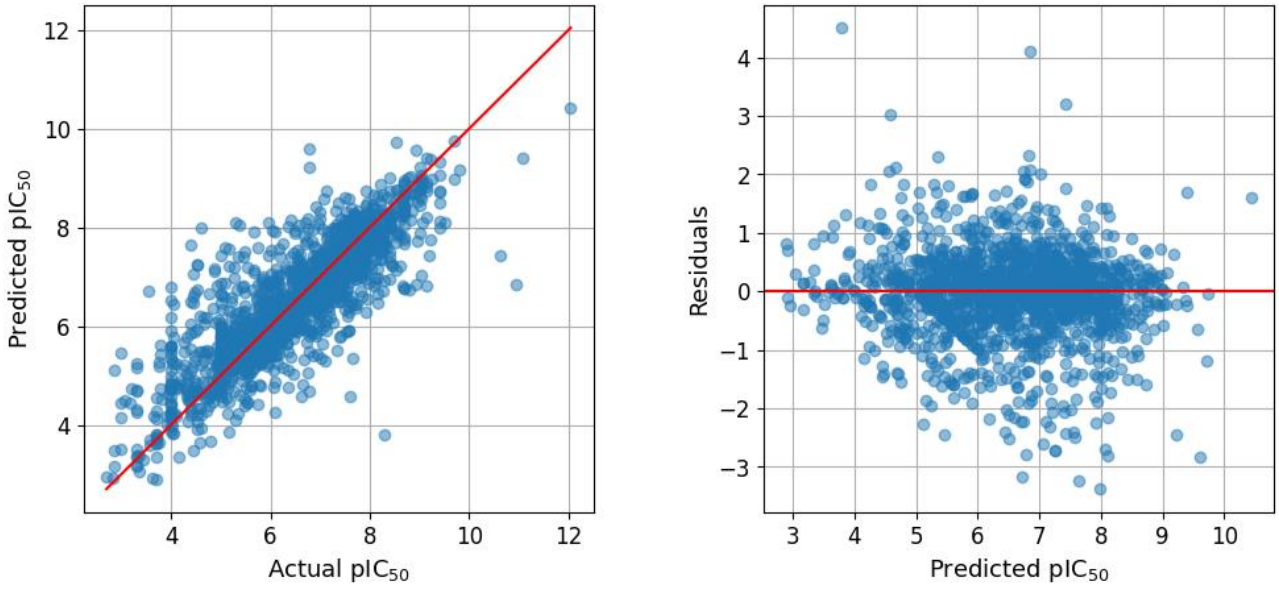
However, the decrease in model performance during cross-validation (R^2 = 0.689, RMSE = 0.779, MAE = 0.553) suggests that the model's predictive power diminished when applied to unseen data. This decrease is typical in machine learning applications, where models tend to perform better on data they have been trained on than new, unobserved data. The cross-validation results highlight the trade-off between model complexity and generalization capability, indicating that while the model fits the training data well, it still faces challenges when extrapolating to new compounds.

The testing set results were consistent with the cross-validation findings, with an R^2 value of 0.690, RMSE of 0.781, and MAE of 0.553. This stability between cross-validation and testing suggests that the model is not significantly overfitted and can generalize reasonably well to new data. The similarity between these two results also reinforces the model's robustness in predicting the inhibitory activity of potential BACE1 inhibitors.

The scatter plot (Figure 2a) displays the relationship between actual and predicted pIC_{50} values. The red diagonal line represents the ideal case where predictions match the actual values perfectly. The points scattered around this line indicate the model's prediction accuracy. Most data points are closely aligned with the line, suggesting good predictive performance, though there is some scatter, particularly for higher pIC_{50} values, indicating areas where the model struggled slightly.

The residual plot (Figure 2b) shows the difference between predicted and actual values (residuals) plotted against predicted pIC_{50} values. The red horizontal line represents zero residual error. A symmetric spread of residuals around this line is visible, indicating no significant systematic bias in the model's predictions. However, some clusters of residuals suggest that predictions for compounds with certain activity levels might have larger errors, especially for lower or higher pIC_{50} values. Overall, these plots suggest that while the model performs well, there is room for improvement, particularly in predicting extreme values.

The prediction error distribution of the SVR model on the testing set is illustrated in Figure 3. The histogram shows that the majority of prediction errors are centered around zero, with most errors falling between -1 and 1, indicating that the model's predictions are generally accurate and close to the actual values. There is a slight



(a) (b)

Figure 2. Scatter plots showing: a) Actual vs. predicted values for the SVR model on the testing set, and b) residuals on the testing set.

Table 1. Performance metrics for SVR model.

Subset	R^2	RMSE	MAE
Training	0.911	0.417	0.209
Cross Validation	0.689	0.779	0.553
Testing	0.690	0.781	0.553

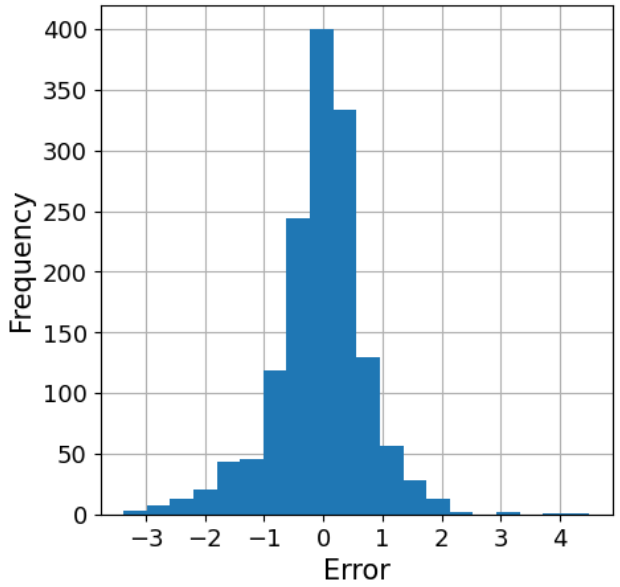


Figure 3. Prediction error distribution for the SVR model on the testing set.

skew towards negative errors, suggesting that the model may slightly underestimate the inhibitory activity for some compounds. The presence of a few outliers, both positive and negative, indicates that for a small number of cases, the model predictions deviate more significantly from the actual values. Overall, the distribution shows that the model's errors are concentrated within a narrow range, demonstrating good predictive performance.

The QSAR model developed in this study holds potential for accelerating the discovery of BACE1 inhibitors. The model's ability to predict the inhibitory activity of compounds with reasonable accuracy means that researchers can prioritize the most promising candidates for synthesis and experimental validation.

One of the key advantages of the SVR-based QSAR model is its efficiency in screening large chemical libraries. This model can be applied to thousands of compounds, allowing for the rapid prediction of BACE1 inhibitory potential. Traditional experimental methods, such as high-throughput screening (HTS), are resource-intensive and time-consuming. In contrast, the QSAR model can quickly identify promising compounds, enabling researchers to focus on the most likely candidates for further experimental testing. This accelerates the discovery process while also reducing associated costs.

Additionally, the stratified sampling approach used in this study ensures that the model captures a wide range of chemical diversity. This allows the model to generalize across different molecular structures, an essential factor in drug discovery. Inhibitors with diverse chemical scaffolds are often needed to address challenges like drug resistance or poor pharmacokinetics. Ensuring that the model encompasses a broad chemical space can help researchers identify potent inhibitors that may have otherwise been overlooked due to structural novelty or complexity.

Despite its promise, several challenges remain in developing BACE1 inhibitors using computational models. First, while the QSAR model provides strong predictions, its ability to generalize to entirely novel chemotypes, particularly those that fall outside the chemical space of the training data, remains a concern. This issue highlights the need for continuous expansion of the dataset with diverse chemotypes, as well as further optimization of feature selection and descriptor calculation methods. Additionally, BACE1 inhibitor development must contend with issues such as blood-brain barrier permeability and off-target effects. While the current QSAR model is focused on predicting BACE1 inhibitory activity, future work could involve incorporating additional pharmacokinetic and pharmacodynamic properties into the model to enhance its utility in identifying drug-like candidates.

4. Conclusions

The SVR-based QSAR model developed in this study demonstrated strong predictive capability in estimating the inhibitory activity of BACE1 inhibitors, as evidenced by its performance on training, cross-validation, and testing datasets. While the model captured a substantial portion of the variance in pIC_{50} values, as shown by an R^2 of 0.690 on the testing set, some errors were observed, particularly in predicting higher and lower activity compounds. The residual analysis showed no significant bias, confirming the model's generalizability. Despite these promising results, there is room for improvement, particularly in refining the feature selection process and incorporating additional molecular descriptors to enhance predictive accuracy. This approach provides an efficient computational method for prioritizing potential drug candidates, potentially accelerating the drug discovery process for Alzheimer's disease therapeutics.

Author Contributions: Conceptualization, T.R.N. and R.I.; methodology, T.R.N., T.E.T. and R.I.; software, T.R.N. and G.M.I.; validation, T.E.T., D.H. and R.I.; formal analysis, T.R.N. and G.M.I.; investigation, T.R.N. and G.M.I.; resources, T.E.T. and R.I.; data curation, D.H. and R.I.; writing—original draft preparation, T.R.N. and G.M.I.; writing—review and editing, T.E.T., D.H. and R.I.; visualization, G.M.I.; supervision, R.I.; project administration, R.I.; funding acquisition, R.I. 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 analyzed during the current study are available from the corresponding author upon reasonable request.

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

References

1. Tiwari, S., Atluri, V., Kaushik, A., Yndart, A., and Nair, M. (2019). Alzheimer's Disease: Pathogenesis, Diagnostics, and Therapeutics, *International Journal of Nanomedicine*, Vol. Volume 14, 5541–5554. doi:10.2147/IJN.S200490.
2. Penke, B., Bogár, F., and Fülöp, L. (2017). β-Amyloid and the Pathomechanisms of Alzheimer's Disease: A Comprehensive View, *Molecules*, Vol. 22, No. 10, 1692. doi:10.3390/molecules22101692.
3. Gholami, A. (2023). Alzheimer's Disease: The Role of Proteins in Formation, Mechanisms, and New Therapeutic Approaches, *Neuroscience Letters*, Vol. 817, 137532. doi:10.1016/j.neulet.2023.137532.
4. Maia, M. A., and Sousa, E. (2019). BACE-1 and γ-Secretase as Therapeutic Targets for Alzheimer's Disease, *Pharmaceuticals*, Vol. 12, No. 1, 41. doi:10.3390/ph12010041.
5. Vassar, R. (2014). BACE1 Inhibitor Drugs in Clinical Trials for Alzheimer's Disease, *Alzheimer's Research & Therapy*, Vol. 6, No. 9, 89. doi:10.1186/s13195-014-0089-7.
6. Coimbra, J. R. M., Resende, R., Custódio, J. B. A., Salvador, J. A. R., and Santos, A. E. (2024). BACE1 Inhibitors for Alzheimer's Disease: Current Challenges and Future Perspectives, *Journal of Alzheimer's Disease*, 1–26. doi:10.3233/JAD-240146.
7. Leelananda, S. P., and Lindert, S. (2016). Computational Methods in Drug Discovery, *Beilstein Journal of Organic Chemistry*, Vol. 12, 2694–2718. doi:10.3762/bjoc.12.267.
8. Noviandy, T. R., Maulana, A., Idroes, G. M., Emran, T. B., Tallei, T. E., Helwani, Z., and Idroes, R. (2023). Ensemble Machine Learning Approach for Quantitative Structure Activity Relationship Based Drug Discovery: A Review, *Infolitika Journal of Data Science*, Vol. 1, No. 1, 32–41. doi:10.60084/ijds.v1i1.91.
9. Chen, W., Liu, X., Zhang, S., and Chen, S. (2023). Artificial Intelligence for Drug Discovery: Resources, Methods, and Applications, *Molecular Therapy - Nucleic Acids*, Vol. 31, 691–702. doi:10.1016/j.omtn.2023.02.019.
10. Noviandy, T. R., Maulana, A., Emran, T. B., Idroes, G. M., and 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*, Vol. 1, No. 1, 1–7. doi:10.60084/hjas.v1i1.12.
11. Ponzoni, I., Sebastián-Pérez, V., Martínez, M. J., Roca, C., De la Cruz Pérez, C., Cravero, F., Vazquez, G. E., Páez, J. A., Díaz, M. F., and Campillo, N. E. (2019). QSAR Classification Models for Predicting the Activity of Inhibitors of Beta-Secretase (BACE1) Associated with Alzheimer's Disease, *Scientific Reports*, Vol. 9, No. 1, 9102. doi:10.1038/s41598-019-45522-3.
12. Aqeel, I., Bilal, M., Majid, A., and Majid, T. (2022). Hybrid Approach to Identifying Druglikeness Leading Compounds against COVID-19 3CL Protease, *Pharmaceuticals*, Vol. 15, No. 11, 1333. doi:10.3390/ph15111333.
13. Noviandy, T. R., Maulana, A., Idroes, G. M., Maulydia, N. B., Patwekar, M., Suhendra, R., and Idroes, R. (2023). Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery, *Malacca Pharmaceutics*, Vol. 1, No. 2, 48–54. doi:10.60084/mp.v1i2.60.
14. Mswahili, M. E., Martin, G. L., Woo, J., Choi, G. J., and Jeong, Y.-S. (2021). Antimalarial Drug Predictions Using Molecular Descriptors and Machine Learning against Plasmodium Falciparum, *Biomolecules*, Vol. 11, No. 12, 1750. doi:10.3390/biom11121750.
15. Noviandy, T. R., Idroes, G. M., and Hardi, I. (2024). Machine Learning Approach to Predict AXL Kinase Inhibitor Activity for Cancer Drug Discovery Using XGBoost and Bayesian

Optimization, *Journal of Soft Computing and Data Mining*, Vol. 5, No. 1, 46–56.

16. Toopradab, B., Xie, W., Duan, L., Hengphasatporn, K., Harada, R., Sinsulpsiri, S., Shigeta, Y., Shi, L., Maitarad, P., and Rungrotmongkol, T. (2024). Machine Learning-Based QSAR and Lb-PaCS-MD Guided Design of SARS-CoV-2 Main Protease Inhibitors, *Bioorganic & Medicinal Chemistry Letters*, Vol. 110, 129852. doi:10.1016/j.bmcl.2024.129852.

17. Shi, Y. (2021). Support Vector Regression-Based QSAR Models for Prediction of Antioxidant Activity of Phenolic Compounds, *Scientific Reports*, Vol. 11, No. 1, 8806. doi:10.1038/s41598-021-88341-1.

18. Cheng, S., and Ding, Y. (2020). Construction of QSAR Model between the Ligand and γ-Aminobutyric Acid Type A Receptor Using Support Vector Regression Algorithm, *2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)*, IEEE, 202–205. doi:10.1109/DCABES50732.2020.00060.

19. Xu, Y., Zomer, S., and Brereton, R. G. (2006). Support Vector Machines: A Recent Method for Classification in Chemometrics, *Critical Reviews in Analytical Chemistry*, Vol. 36, Nos. 3–4, 177–188. doi:10.1080/10408340600969486.

20. 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.

21. Noviandy, T. R., Idroes, G. M., and Hardi, I. (2024). An Interpretable Machine Learning Strategy for Antimalarial Drug Discovery with LightGBM and SHAP, *Journal of Future Artificial Intelligence and Technologies*, Vol. 1, No. 2, 84–95. doi:10.62411/faith.2024-16.

22. Noviandy, T. R., Nisa, K., Idroes, G. M., Hardi, I., and Sasmita, N. R. (2024). Classifying Beta-Secretase 1 Inhibitor Activity for Alzheimer's Drug Discovery with LightGBM, *Journal of Computing Theories and Applications*, Vol. 2, No. 2, 138–147. doi:10.62411/jcta.10129.

23. Xue, L., and Bajorath, J. (2000). Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening, *Combinatorial Chemistry & High Throughput Screening*, Vol. 3, No. 5, 363–372. doi:10.2174/1386207003331454.

24. Grisoni, F., Consonni, V., and Todeschini, R. (2018). Impact of Molecular Descriptors on Computational Models, 171–209. doi:10.1007/978-1-4939-8639-2_5.

25. 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.

26. Noviandy, T. R., Maulana, A., Idroes, G. M., Irvanizam, I., Subianto, M., and Idroes, R. (2023). QSAR-Based Stacked Ensemble Classifier for Hepatitis C NS5B Inhibitor Prediction, *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, IEEE, 220–225. doi:10.1109/COSITE60233.2023.10250039.

27. Suhendra, R., Husdayanti, N., Suryadi, S., Juliwardi, I., Sanusi, S., Ridho, A., Ardiansyah, M., Murhaban, M., and Ikhsan, I. (2023). Cardiovascular Disease Prediction Using Gradient Boosting Classifier, *Infolitika Journal of Data Science*, Vol. 1, No. 2, 56–62. doi:10.60084/ijds.v1i2.131.

28. Safriandono, A. N., Setiadi, D. R. I. M., Dahlan, A., Rahmanti, F. Z., Wibisono, I. S., and Ojugo, A. A. (2024). Analyzing Quantum Feature Engineering and Balancing Strategies Effect on Liver Disease Classification, *Journal of Future Artificial Intelligence and Technologies*, Vol. 1, No. 1, 51–63. doi:10.62411/faith.2024-12.

29. Zhang, F., and O'Donnell, L. J. (2020). Support Vector Regression, *Machine Learning*, Elsevier, 123–140. doi:10.1016/B978-0-12-815739-8.00007-9.

30. Huang, J., and Fan, X. (2013). Reliably Assessing Prediction Reliability for High Dimensional QSAR Data, *Molecular Diversity*, Vol. 17, No. 1, 63–73. doi:10.1007/s11030-012-9415-9.

31. Algamal, Z. Y., Qasim, M. K., Lee, M. H., and Mohammad Ali, H. T. (2020). High-Dimensional QSAR/QSPR Classification Modeling Based on Improving Pigeon Optimization Algorithm, *Chemometrics and Intelligent Laboratory Systems*, Vol. 206, 104170. doi:10.1016/j.chemolab.2020.104170.

32. Asif, D., Bibi, M., Arif, M. S., and Mukheimer, A. (2023). Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization, *Algorithms*, Vol. 16, No. 6, 308. doi:10.3390/a16060308.

33. Ding, H., Xing, F., Zou, L., and Zhao, L. (2024). QSAR analysis of VEGFR-2 inhibitors based on machine learning, Topomer CoMFA and molecule docking, *BMC Chemistry*, Vol. 18, No. 1, 59. doi:10.1186/s13065-024-01165-8.

34. Idroes, R., Noviandy, T. R., Maulana, A., Suhendra, R., and Sasmita, N. R. (2023). ANFIS-Based QSRR Modelling for Kovats Retention Index Prediction in Gas Chromatography, *Infolitika Journal of Data Science*, Vol. 1, No. 1, 1–14. doi:10.60084/ijds.v1i1.73.

35. Idroes, R., Noviandy, T., Maulana, A., Suhendra, R., Sasmita, N., Muslem, M., Idroes, G. M., Kemala, P., and Irvanizam, I. (2021). Application of Genetic Algorithm-Multiple Linear Regression and Artificial Neural Network Determinations for Prediction of Kovats Retention Index, *International Review on Modelling and Simulations (IREMOS)*, Vol. 14, No. 2, 137.

36. Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Adam, M., Rusyana, A., and Sofyan, H. (2023). Deep Learning-Based Bitcoin Price Forecasting Using Neural Prophet, *Ekonomikalia Journal of Economics*, Vol. 1, No. 1, 19–25. doi:10.60084/eje.v1i1.51.