Original Article

Check for updates

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

Heca Journal of Applied Sciences

Vol. 1, No. 1, 2023

QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms

Teuku Rizky Noviandy [1], Aga Maulana [1], Talha Bin Emran [2], Ghazi Mauer Idroes [3,*] and Rinaldi Idroes [4]

[1] Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; trizkynoviandy@gmail.com (T.R.N.); agamaulana@usk.ac.id (A.M.)
[2] Department of Pharmacy, BGC Trust University Bangladesh, Chittagong 4381, Bangladesh; talhabmb@bgctub.ac.bd (T.B.E.)
[3] Department of Occupational Health and Safety, Faculty of Health Sciences, Universitas Abulyatama, Aceh Besar 23372, Indonesia; idroesghazi_k3@abulyatama.ac.id (G.M.I.)
[4] Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; rinaldi.idroes@usk.ac.id (R.I.)

***** Correspondence: idroesghazi_k3@abulyatama.ac.id

Article History

Received 22 April 2023
Revised 16 May 2023
Accepted 26 May 2023
Available Online 30 May 2023

Keywords:
Beta-secretase 1
Ensemble machine learning
Molecular descriptors
QSAR

Abstract

This study focuses on the development of a machine learning ensemble approach for the classification of Beta-Secretase 1 (BACE1) inhibitors in Quantitative Structure-Activity Relationship (QSAR) analysis. BACE1 is an enzyme linked to the production of amyloid beta peptide, a significant component of Alzheimer's disease plaques. The discovery of effective BACE1 inhibitors is difficult, but QSAR modeling offers a cost-effective alternative by predicting the activity of compounds based on their chemical structures. This study evaluates the performance of four machine learning models (Random Forest, AdaBoost, Gradient Boosting, and Extra Trees) in predicting BACE1 inhibitor activity. Random Forest achieved the highest performance, with a training accuracy of 98.65% and a testing accuracy of 82.53%. In addition, it exhibited superior precision, recall, and F1-score. Random Forest's superior performance was a result of its ability to capture a wide variety of patterns and its randomized ensemble approach. Overall, this study demonstrates the efficacy of ensemble machine learning models, specifically Random Forest, in predicting the activity of BACE1 inhibitors. The findings contribute to ongoing efforts in Alzheimer's disease drug discovery research by providing a cost-effective and efficient strategy for screening and prioritizing potential BACE1 inhibitors.

1. Introduction

Beta-Secretase 1 (BACE1) is a key enzyme involved in the production of amyloid beta peptide, which is a major component of the plaques found in the brains of individuals with Alzheimer's disease [1]. The development of BACE1 inhibitors has been a major focus in drug discovery research for the treatment of Alzheimer's disease. However, the discovery of effective BACE1 inhibitors is challenging due to the complex nature of the enzyme and the lack of reliable experimental data for many potential inhibitors [2, 3]. To overcome this challenge, researchers use Quantitative Structure-Activity Relationship (QSAR) modeling [4–6].

QSAR is a computational modeling technique that identifies connections between the chemical structure of a compound and its biological activity or other properties [7]. Based on these structural characteristics, it enables scientists to predict and comprehend the behavior and impacts of molecules [8]. In recent years, there have been several studies on the subject of QSAR analysis. These

studies include research on the prediction of small-molecule binding to RNA, predicting the transfer of environmental chemicals across the placenta, and the classification of inhibitor activity [9, 10].

QSAR is a powerful computational approach that can aid in the discovery of new BACE1 inhibitors [11]. QSAR modeling provides a cost-effective and efficient alternative by screening and prioritizing compounds based on their predicted activity. This enables researchers to focus their efforts on a more targeted set of compounds, saving time, resources, and costs in the drug discovery process [12].

One of the methods that can be used in QSAR modeling is the ensemble machine learning algorithm [13]. Ensemble methods combine multiple individual models to improve the overall predictive performance. Ensemble machine learning algorithms offer several advantages in QSAR modeling [14]. They can handle complex relationships between the structural features and the biological activity of BACE1 inhibitors. They are robust and can reduce model variance. Additionally, they can provide insights into feature importance and identify the most influential structural descriptors for BACE1 inhibition. By incorporating ensemble machine learning algorithms into QSAR modeling, researchers can enhance the accuracy and reliability of predictions for potential BACE1 inhibitors [15].

In this study, an ensemble machine learning approach for the classification of BACE1 inhibitors in QSAR analysis is proposed. In the approach, four ensemble machine learning models, namely AdaBoost, Extra Trees (ET), Gradient Boosting (GB), and Random Forest (RF) are utilized to assess their performance and effectiveness in the analysis. The performance of each model is compared individually to gain a clear understanding of their unique strengths, limitations, and suitability to classify the activity of BACE1 inhibitors.

2. Materials and Methods

2.1. Data Collection

In this study, a total of 10,551 chemical compounds that have been tested as inhibitors of BACE1 protein, along with their IC50 values, were collected from the ChEMBL database. The collected compounds were filtered to make sure that there were no duplicates. The number of compounds that remained after the filtering process was 7,298. Then, a class labeling process was performed, where chemical compounds with IC50 values < 1000 were grouped into the active class, or else they were grouped into the inactive class [16].

2.2. Molecular Descriptor Calculation

The molecular descriptor is a mathematical representation of the chemical and physical properties of a molecule. It is a quantitative value or set of values that are calculated based on the molecular structure of a compound and is used to characterize and compare molecules [17]. Molecular descriptors are used as features to construct a machine learning model. A total of 1661 2D molecular descriptors were successfully calculated using Mordred for each compound [18]. Molecular descriptors with low variance and those with multicollinearity > 0.95 were removed, resulting in a final number of 456 molecular descriptors for each compound.

2.3. Feature Selection

A large number of molecular descriptors can be irrelevant or redundant for the classification task, so it is necessary to employ a feature selection method. Feature selection is the process of selecting a subset of relevant features from the original set of features in order to improve the performance of machine learning models. Feature selection can help to reduce the dimensionality of the data and prevent overfitting, improve the interpretability of the model, and reduce the computational cost of training models [19, 20].

In this study, a genetic algorithm (GA) was used to select the most optimal molecular descriptor [21, 22]. To initiate the GA-based feature selection process, an initial population containing 50 random molecular descriptors subsets was generated. The GA then used crossover and mutation operations to create new combinations of features. Crossover occurred with a 90% probability between selected subsets, while mutation happened with a 5% probability for each subset. Independent probabilities were assigned to crossover and mutation to encourage exploration. The GA process ran for 200 generations, evaluating the fitness of each subset using accuracy through 10-fold cross-validation. The best subsets became parents for the next generation, ensuring the gradual replacement of weaker subsets. To prevent the GA from getting stuck, a stopping criterion was implemented. If the best subset remained unchanged for ten consecutive generations, the algorithm terminated, indicating convergence gene.

rdk
it

2.4. Ensemble Machine Learning Model

In this study, four ensemble machine learning models were used, namely AdaBoost, ET, GB, and RF. These

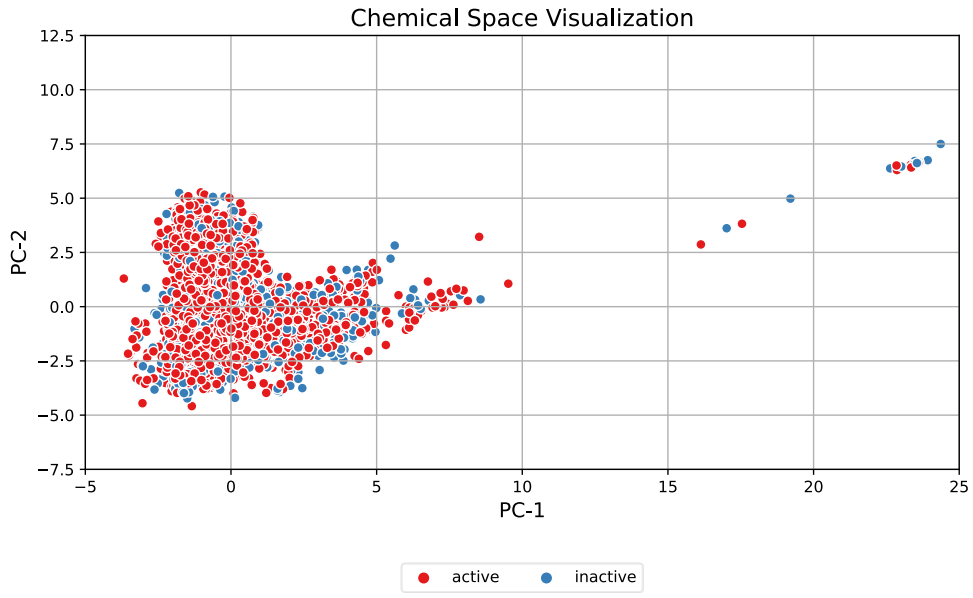

Figure 1. Chemical space visualization using 16 molecular descriptors.

models were selected based on their high accuracy and robustness in handling complex data sets. The utilization of these four ensemble machine learning models allows for a comprehensively analysis their performance and determination of the most suitable approach to classify BACE1 inhibitors.

AdaBoost is an ensemble model that iteratively trains weak learners and adjusts the weights of misclassified samples to focus on difficult cases. Weak learners are combined to create a strong learner that makes accurate predictions [23]. GB is an ensemble model that builds a sequence of weak learners, such as decision trees, in a step-wise manner. Each subsequent learner is trained to correct the mistakes made by the previous ones, leading to a strong predictive model [24]. ET is an ensemble model similar to RF but with a higher level of randomness. It creates multiple decision trees using random splits and averages their predictions to make the final prediction [25]. RF is an ensemble model that combines multiple decision trees to make predictions. Each tree is trained on a random subset of the data, and the final prediction is determined by averaging the predictions of all trees [19].

2.5. Model Evaluation

A comprehensive performance evaluation was conducted using four metrics: accuracy, precision, recall, and F1-score, in order to assess the effectiveness of the proposed model. Accuracy measures the proportion of correctly classified instances out of the total number of instances, and is useful when the classes in the dataset are balanced, i.e., have an equal number of instances. Precision, on the other hand, measures the proportion of true positives out of all positive predictions. Recall measures the proportion of true positives out of all actual positives in the dataset. F1 score is a harmonic mean of precision and recall and is a more robust metric for imbalanced datasets. The equation used to calculate the metric can be seen in Equations 1, 2, 3, and 4, respectively [26].

$$Accuracy = \frac{TP + FN}{FP + FN + TP + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{FN + TP} \quad (3)$$

$$F1 - Score = 2\frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

where TP is the number of true positive classifications, FN is the number of false negative classifications, FP is the number of false positive classifications, and TN is the number of true negative classifications.

3. Results and Discussions

The 16 most optimal molecular descriptors were selected from a total of 456 molecular descriptors using GA. These selected descriptors are listed in Table 1. It can be observed that the selected descriptors are organized into 12 groups and 4 classes based on their characteristics and properties. Among the selected molecular descriptors, the largest number of descriptors belongs to

Table 1. Most optimal molecular descriptors selected by GA.

Descriptor Type	Descriptor Class	Definition	Descriptor Names
Atom Counts	Molecular Composition	Number of nitrogen (N) and oxygen (O) atoms in the molecule	nN, nO
Averaged Modified Burden Eigenvalues	Topological	Average modified burden eigenvalues for specific atoms	AATS6dv, AATS0v
Atom-type E-state indices	Topological	Atom-type E-state indices for specific atoms	ATSC5d
Geary Autocorrelation	Connectivity	Geary autocorrelation for specific atoms	GATS4Z, GATS5v
CATS	Connectivity	CATS descriptor	C1SP2
X Contribution	Connectivity	X contribution descriptor	Xp-3dv
Atom-type Counts	Molecular Composition	Atom-type counts descriptors	NtsC, SssssC
Information Indices	Topological	Information indices descriptor	IC1
Partial Charges	Physicochemical Properties	Partial charges descriptor	PEOE_VSA3
SlogP/VSA	Physicochemical Properties	SlogP/VSA descriptor	SMR_VSA4
Ring Counts	Topological	Ring counts descriptor	n6ARing
Graph-based Indices	Topological	Graph-based indices descriptor	GGI7

Table 2. Performance of ensemble machine learning model.

Model	Accuracy (%)		Precision (%)		Recall (%)		F1-score (%)	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
AdaBoost	77.34	76.64	77.09	76.53	77.34	76.64	76.58	75.86
Extra Trees	98.65	82.47	98.67	82.34	98.65	82.47	98.64	82.25
Gradient Boosting	81.55	80.07	81.44	79.92	81.55	80.07	81.13	79.70
Random Forest	98.65	82.53	98.65	82.39	98.65	82.53	98.64	82.37

the topological descriptor class. This class includes descriptors such as Averaged Modified Burden Eigenvalues, Atom-type E-state indices, Geary Autocorrelation, Ring Counts, and Graph-based Indices. These descriptors provide insights into the connectivity, structure, autocorrelation, and graph-based properties of the molecules. In addition to the topological descriptor class, the GA selected descriptors from various other descriptor classes, such as physicochemical properties, molecular composition, and connectivity. The physicochemical properties descriptor class consists of two descriptors, namely Partial Charges and SlogP/VSA. These descriptors capture information about the distribution of partial charges within the molecule and provide insights into lipophilicity and molecular surface properties. The molecular composition descriptor class includes the Atom Counts descriptor, representing the number of nitrogen (N) and oxygen (O) atoms in the molecule. The connectivity descriptor class consists of the CATS and X Contribution descriptors, which provide information about specific connectivity patterns and the contribution of the X atom to overall molecular properties, respectively. The selected molecular descriptors are then used as features in ensemble machine learning model to capture the relationships between molecular characteristics and the BACE1 inhibitor activity.

To visualize the chemical space of the dataset, principal component analysis (PCA) was applied to 16 selected molecular descriptors. Figure 1 provides an overview of the relationships and similarities among the compounds in terms of their molecular descriptors. Compounds that are closer in the plot are considered more similar in their characteristics, while molecules that are farther apart exhibit greater dissimilarity. It can be seen that the compounds create a cluster that shares similar molecular descriptors. However, amidst these clusters, there are also compounds that stand out as outliers, positioned farther away from the main cluster. These outliers signify compounds with distinct chemical features or unique combinations of molecular descriptors that differentiate them from the majority of the dataset.

The dataset was divided into a training set and a testing set, with proportions of 80% and 20% respectively. The training set was used to train the ensemble machine learning models, while the testing set was used to evaluate their performance on unseen data. Table 2. presents an overview of the performance obtained from the ensemble machine learning model. The models'

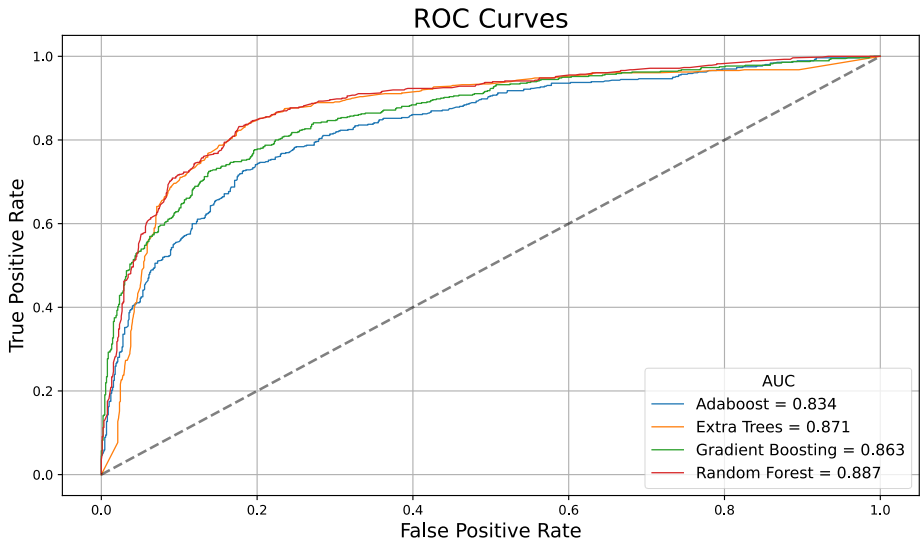

Figure 2. ROC curves of the models.

accuracy, precision, recall, and F1 score are provided for both the training and testing set. The highest performance is achieved by the RF model. It attains an accuracy of 98.65% on the training set and 82.53% on the testing set. Moreover, it demonstrates a precision of 98.65% on the training set and 82.39% on the testing set. The recall scores for this model are 98.65% on the training set and 82.53% on the testing set. The F1 score is 98.64% on the training set and 82.37% on the testing set. The exceptional performance of the RF model can be attributed to its unique characteristics. RF constructs multiple decision trees using random subsets of features and random thresholds for splitting to introduce diversity among the individual trees and helps to reduce the correlation between them, making the model less prone to overfitting. It allows RF to capture a wide range of patterns and make robust predictions.

On the other hand, the model with the lowest performance is the AdaBoost. It achieves an accuracy of 85.30% on the training set and 82.33% on the testing set. The precision scores are 85.30% on the training set and 82.27% on the testing set. In terms of recall, it achieves 85.30% on the training set and 82.33% on the testing set. The F1 score is 85.03% on the training set and 82.02% on the testing set. The lower performance of the AdaBoost model could be attributed to various factors, such as the complexity of the weak learners or the difficulty of the dataset. It is possible that the model complexity was not properly controlled, leading to overfitting. Overfitting occurs when the model becomes too specialized to the training data and performs poorly on unseen data.

Table 3. shows the confusion matrix of each model. The confusion matrix provides a detailed breakdown of the performance of a classification model by showing the number of correct and incorrect predictions for each class. The ET model demonstrated remarkable success in accurately predicting the active class, with 802 compounds correctly classified and only 98 compounds misclassified. On the other hand, the RF model excelled in classifying the inactive class, achieving 409 correct predictions while making 151 incorrect predictions. This indicates that the RF model might have been particularly adept at identifying patterns associated with inactive compounds, leading to its success in classifying the inactive class.

The receiver operating characteristics (ROC) curves of each model can be seen in Figure 2. ROC curves are graphical representations that depict the performance of a binary classification model across various classification thresholds. The x-axis represents the false positive rate (FPR), while the y-axis represents the true positive rate (TPR), also known as sensitivity or recall. In these ROC curves, we can observe the area under the curve (AUC), which quantifies the overall performance of a model. A higher AUC score indicates better discrimination power and more accurate classification. The AUC scores of the AdaBoost, ET, GB, and RF model is 0.834, 0.871, 0.863,

Table 3. Confusion matrix of the models.

Model	Actual	Predicted	
		Active	Inactive
AdaBoost	Active	797	103
	Inactive	238	322
Extra Trees	Active	802	98
	Inactive	158	402
Gradient Boosting	Active	797	103
	Inactive	188	372
Random Forest	Active	796	104
	Inactive	151	409

and 0.887, respectively. This result shows the good performance of the RF model to capture complex patterns and achieve accurate classifications.

These results showed that the RF model outperforms the other models with the highest performance to classify BACE1 inhibitor activity, benefiting from its randomized ensemble approach and robust generalization. Conversely, the AdaBoost model, while still achieving reasonably high metrics, may suffer from overfitting due to its tendency to increase complexity during training.

4. Conclusions

This study proposes an ensemble machine learning approach for the classification of BACE1 inhibitors in QSAR analysis. The performance of the four ensemble models AdaBoost, ET, GB, and RF to predict the activity of BACE1 inhibitors were evaluated. The RF model outperformed the other three models, achieving an accuracy of 82.53% on the testing set. It demonstrated strong predictive abilities by effectively capturing a wide variety of patterns and minimizing the risk of overfitting. In contrast, the performance of the AdaBoost model was relatively inferior, possibly due to the complexity of the weak learners or the difficulty of the dataset. Additional parameter optimization and fine-tuning may be required to enhance the model's performance. The results of this study demonstrate the predictive accuracy of ensemble machine learning models, particularly the RF model to classify BACE1 inhibitors. The proposed method can aid in the identification of potential BACE1 inhibitors, potentially contributing to the establishment of Alzheimer's disease therapeutic interventions. Future work for this study involves parameter optimization and fine-tuning of the ensemble models to enhance their performance. By focusing on these aspects, it is possible to make adjustments that will lead to improved accuracy in classifying BACE1 inhibitors.

Author Contributions: Conceptualization: T.R.N., A.M. and G.M.I.; methodology: T.R.N., R.I. and G.M.I.; software: T.R.N. and A.M.; validation: R.I., T.B.E. and G.M.I.; formal analysis: T.R.N.; investigation: T.R.N. and T.B.E.; resources: A.M. and T.B.E.; data curation: T.B.E. and R.I.; writing—original draft preparation: T.R.N. and A.M.; writing—review and editing, T.B.E., G.M.I. and R.I.; visualization: A.M.; supervision: T.B.E., G.M.I. and R.I.; project administration: G.M.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 is available by request.

Acknowledgments: The authors express their gratitude to their individual institutions and universities.

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

References

1. Cervellati, C., Trentini, A., Rosta, V., Passaro, A., Bosi, C., Sanz, J. M., Bonazzi, S., Pacifico, S., Seripa, D., Valacchi, G. (2020). Serum beta-secretase 1 (BACE1) activity as candidate biomarker for late-onset Alzheimer's disease, *GeroScience*, Vol. 42, 159–167
2. Moussa-Pacha, N. M., Abdin, S. M., Omar, H. A., Alniss, H., Al-Tel, T. H. (2020). BACE1 inhibitors: Current status and future directions in treating Alzheimer's disease, *Medicinal Research Reviews*, Vol. 40, No. 1, 339–384
3. Vassar, R. (2014). BACE1 inhibitor drugs in clinical trials for Alzheimer's disease, *Alzheimer's Research & Therapy*, Vol. 6, No. 9, 1–14
4. Das, S., Majumder, T., Sarkar, A., Mukherjee, P., Basu, S. (2020). Flavonoids as BACE1 inhibitors: QSAR modelling, screening and in vitro evaluation, *International Journal of Biological Macromolecules*, Vol. 165, 1323–1330
5. Kumar, A., Roy, S., Tripathi, S., Sharma, A. (2016). Molecular docking based virtual screening of natural compounds as potential BACE1 inhibitors: 3D QSAR pharmacophore mapping and molecular dynamics analysis, *Journal of Biomolecular Structure and Dynamics*, Vol. 34, No. 2, 239–249
6. 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., 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, 1–13
7. Wu, Y., Huo, D., Chen, G., Yan, A. (2021). SAR and QSAR research on tyrosinase inhibitors using machine learning methods, *SAR and QSAR in Environmental Research*, Vol. 32, No. 2, 85–110. doi:10.1080/1062936X.2020.1862297
8. Abdullahi, M., Shallangwa, G. A., Uzairu, A. (2020). In silico QSAR and molecular docking simulation of some novel aryl sulfonamide derivatives as inhibitors of H5N1 influenza A virus subtype, *Beni-Suef University Journal of Basic and Applied Sciences*, Vol. 9, No. 1, 2. doi:10.1186/s43088-019-0023-y
9. Huang, T., Sun, G., Zhao, L., Zhang, N., Zhong, R., Peng, Y. (2021). Quantitative Structure-Activity Relationship (QSAR) Studies on the Toxic Effects of Nitroaromatic Compounds (NACs): A Systematic Review, *International Journal of Molecular Sciences*, Vol. 22, No. 16, 8557. doi:10.3390/ijms22168557
10. Hesping, E., Chua, M. J., Pflieger, M., Qian, Y., Dong, L., Bachu, P., Liu, L., Kurz, T., Fisher, G. M., Skinner-Adams, T. S., Reid, R. C., Fairlie, D. P., Andrews, K. T., Gorse, A.-D. J. P. (2022). QSAR Classification Models for Prediction of Hydroxamate Histone Deacetylase Inhibitor Activity against Malaria Parasites, *ACS Infectious Diseases*, Vol. 8, No. 1, 106–117. doi:10.1021/acsinfecdis.1c00355
11. 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. (2020). QSAR without borders, *Chemical Society Reviews*, Vol. 49, No. 11, 3525–3564
12. Puzyn, T., Leszczynski, J., Cronin, M. T. (Eds.). (2010). *Recent Advances in QSAR Studies* (Vol. 8), Springer Netherlands, Dordrecht. doi:10.1007/978-1-4020-9783-6
13. Kwon, S., Bae, H., Jo, J., Yoon, S. (2019). Comprehensive ensemble in QSAR prediction for drug discovery, *BMC Bioinformatics*, Vol. 20, No. 1, 521. doi:10.1186/s12859-019-3135-4

14. Kurniawan, I., Rosalinda, M., Ikhsan, N. (2020). Implementation of ensemble methods on QSAR Study of NS3 inhibitor activity as anti-dengue agent, *SAR and QSAR in Environmental Research*, Vol. 31, No. 6, 477–492

15. Wu, Z., Zhu, M., Kang, Y., Leung, E. L.-H., Lei, T., Shen, C., Jiang, D., Wang, Z., Cao, D., Hou, T. (2021). Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets, *Briefings in Bioinformatics*, Vol. 22, No. 4, bbaa321

16. Simeon, S., Anuwongcharoen, N., Shoombuatong, W., Malik, A. A., Prachayasittikul, V., Wikberg, J. E. S., Nantasenamat, C. (2016). Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking, *PeerJ*, Vol. 4, e2322

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

18. Moriwaki, H., Tian, Y.-S., Kawashita, N., Takagi, T. (2018). Mordred: a molecular descriptor calculator, *Journal of Cheminformatics*, Vol. 10, No. 1, 4. doi:10.1186/s13321-018-0258-y

19. Goodarzi, M., Dejaegher, B., Heyden, Y. Vander. (2012). Feature selection methods in QSAR studies, *Journal of AOAC International*, Vol. 95, No. 3, 636–651

20. Khaire, U. M., Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review, *Journal of King Saud University-Computer and Information Sciences*, Vol. 34, No. 4, 1060–1073

21. Noviandy, T. R., Maulana, A., Sasmita, N. R., Suhendra, R., Irvanizam, I., Muslem, M., Idroes, G. M., Yusuf, M., Sofyan, H., Abidin, T. F., Idroes, R. (2022). The Prediction of Kovats Retention Indices of Essential Oils at Gas Chromatography Using Genetic Algorithm-Multiple Linear Regression and Support Vector Regression, *Journal of Engineering Science and Technology*

22. Idroes, R., Noviandy, T. R., Maulana, A., Suhendra, R., Sasmita, N. R., Muslem, M., Idroes, G. M., Kemala, P., 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. doi:10.15866/iremos.v14i2.20460

23. Ying, C., Qi-Guang, M., Jia-Chen, L., Lin, G. (2013). Advance and prospects of AdaBoost algorithm, *Acta Automatica Sinica*, Vol. 39, No. 6, 745–758

24. Natekin, A., Knoll, A. (2013). Gradient boosting machines, a tutorial, *Frontiers in Neurorobotics*, Vol. 7, 21

25. Geurts, P., Ernst, D., Wehenkel, L. (2006). Extremely randomized trees, *Machine Learning*, Vol. 63, 3–42

26. Tasci, E., Zhuge, Y., Kaur, H., Camphausen, K., Krauze, A. V. (2022). Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with Clinical and Molecular Characteristics, *International Journal of Molecular Sciences*, Vol. 23, No. 22, 14155. doi:10.3390/ijms232214155