JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Classifying Beta-Secretase 1 Inhibitor Activity for Alzheimer's Drug Discovery with LightGBM

Teuku Rizky Noviandy[1,*], Khairun Nisa[2,3], Ghalieb Mutig Idroes[4], Irsan Hardi[4] and Novi Reandy Sasmita[5]

[1] Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; e-mail : trizkynoviandy@gmail.com
[2] Clinical Pharmacy, Stikes Jabal Ghafur, Sigli 24151, Indonesia; e-mail : nisadara96@gmail.com
[3] RSUD dr. Fauziah Bireuen, Bireun 24261, Indonesia
[4] Interdisciplinary Innovation Research Unit, Graha Primera Saintifika, Aceh Besar 23771, Indonesia; e-mail : ghaliebidroes@outlook.com, irsan.hardi@gmail.com
[5] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Syiah, Banda Aceh 23111, Indonesia; e-mail : novireandys@usk.ac.id
* Corresponding Author: Teuku Rizky Noviandy

Abstract: This study explores the utilization of LightGBM, a gradient-boosting framework, to classify the inhibitory activity of beta-secretase 1 inhibitors, addressing the challenges of Alzheimer's disease drug discovery. The study aims to enhance classification performance by focusing on overcoming the limitations of traditional statistical models and conventional machine-learning techniques in handling complex molecular datasets. By sourcing a dataset of 7298 compounds from the ChEMBL database and calculating molecular descriptors for each compound as features, we employed LightGBM in conjunction with a set of carefully selected molecular descriptors to achieve a nuanced analysis of compound activities. The model's efficiency was benchmarked against traditional machine-learning algorithms, revealing LightGBM's superior accuracy (84.93%), precision (87.14%), sensitivity (89.93%), specificity (77.63%), and F1-score (88.17%) in classifying beta-secretase 1 inhibitor activity. The study underscores the critical role of molecular descriptors in understanding drug efficacy, highlighting LightGBM's potential in streamlining the virtual screening process. Conclusively, the findings advocate for LightGBM's adoption in computational drug discovery, offering a promising avenue for advancing Alzheimer's disease therapeutic development by facilitating the identification of potential drug candidates with enhanced precision and reliability.

Keywords: ChEMBL; Machine-learning; Molecular descriptor; QSAR; Virtual screening.

Received: February, 11th 2024
Revised: February, 24th 2024
Accepted: March, 9th 2024
Published: March, 10th 2024

1. Introduction

Alzheimer's disease is a devastating neurodegenerative disorder characterized by progressive cognitive decline, memory loss, and impaired daily functioning [1]. It affects millions of individuals worldwide, and its prevalence is expected to rise significantly as the global population ages [2]. The pathological hallmarks of Alzheimer's disease include the accumulation of amyloid-beta plaques and neurofibrillary tangles in the brain, which are believed to contribute to the disease's progression [3].

Beta-secretase 1 also known as β-site amyloid precursor protein cleaving enzyme 1, is crucial in producing Aβ peptides [4], [5]. Inhibition of beta-secretase 1 has emerged as a promising therapeutic strategy for Alzheimer's disease, potentially reducing Aβ production and slowing down the disease's progression [6]. Consequently, developing effective beta-secretase 1 inhibitors has become a focal point in Alzheimer's disease research.

Classification of the inhibitory activity of compounds is crucial in drug discovery and development [7]. Accurate classification is essential to prioritize compounds for further investigation, ultimately leading to the identification of potential drug candidates [8]. However, the classification challenge is non-trivial, given the extensive chemical diversity of potential inhibitors and the need to achieve a delicate sensitivity-specificity balance [9].

In recent years, there has been a significant leap forward in leveraging machine-learning techniques to address the challenges in drug discovery [10]–[12]. Machine-learning models can analyze vast chemical datasets, classify compound activities, and streamline the identification of potential drug candidates [13]–[16]. Moreover, the integration of machine-learning in drug discovery extends to Quantitative Structure-Activity Relationship (QSAR) modeling. Machine-learning models harness molecular descriptors as features, enabling a more comprehensive analysis of chemical structures and their activities [17], [18]. This approach enhances the accuracy of classification, expedites the identification of promising drug candidates, and revolutionizes the efficiency of the drug discovery process.

Current methodologies for classifying beta-secretase 1 inhibitors predominantly lean on traditional statistical models and conventional machine-learning techniques [19], [20]. While these approaches have yielded foundational insights, they frequently encounter limitations when confronted with high-dimensional data and the intricate nonlinear relationships inherent in molecular datasets. The complexity of molecular interactions often surpasses the capabilities of these methods, leading to incomplete understanding and potentially missed opportunities for identifying novel inhibitors [21]. As a result, there's a growing recognition within the scientific community of the need to explore more advanced methodologies capable of handling the intricacies of molecular data.

One promising method to address these limitations is the application of LightGBM (Light Gradient Boosting Machine). LightGBM is an advanced gradient boosting framework that is designed to be efficient, flexible, and powerful [22]. It stands out for its ability to handle large-scale data and efficiently manage high-dimensional spaces typical of molecular datasets [23]. Notably, it can handle imbalanced data effectively [24], which is a common issue in classifying potential drug inhibitors [25]. This capability helps prevent biased classifications toward the majority class. LightGBM addresses this challenge by implementing weighted sampling and gradient-based one-side sampling techniques, ensuring that the model pays more attention to underrepresented classes. This capability is important for accurately identifying potent beta-secretase 1 inhibitors, where the number of effective compounds may be relatively small compared to non-inhibitors.

Moreover, LightGBM's histogram-based algorithm represents a significant improvement over conventional methods by reducing memory usage and accelerating the learning process [22]. This makes it particularly suitable for classifying beta-secretase 1 inhibitors, as it can effectively capture complex nonlinear relationships without compromising on computational efficiency or accuracy, even in the presence of data imbalance. In contrast, traditional models often encounter difficulties with the non-linearity and complexity inherent in these datasets, potentially overlooking subtle yet critical patterns essential for identifying novel inhibitors.

This study aims to utilize LightGBM's robust capabilities to significantly enhance the classification performance of beta-secretase 1 inhibitors, addressing the limitations of traditional statistical models and conventional machine-learning techniques. Molecular datasets pose challenges such as high dimensionality and complex nonlinear relationships, which LightGBM is well-equipped to handle. Renowned for its efficiency, accuracy, and ability to process large-scale data using tree-based learning algorithms, LightGBM offers a promising solution. By leveraging these strengths, this research endeavors to improve the accuracy and reliability of classifying the inhibitory activity of beta-secretase 1 inhibitors, facilitating the identification of potential drug candidates for Alzheimer's disease therapy.

The contributions of this study encompass several key aspects aimed at advancing the field of drug discovery with machine-learning for Alzheimer's disease therapy. These contributions include:

1. Introduction of LightGBM as an innovative tool for classifying beta-secretase 1 inhibitor activity.
2. Implementation of hyperparameter tuning to optimize the performance of the LightGBM model.
3. Analysis of feature importance influencing beta-secretase 1 inhibitor classification.
4. Establishing a benchmark, evaluating the LightGBM model using metrics such as accuracy, precision, sensitivity, specificity, and F1-score.
5. Advancement of drug discovery strategies targeting Alzheimer's disease.

2. Related Works

Early endeavors to classify beta-secretase 1 inhibitor activity predominantly relied on experimental screening and structure-activity relationship studies [26]. These approaches involved synthesizing and testing chemical compounds against beta-secretase 1 to assess their inhibitory effects. While providing valuable insights, traditional methods are resource-intensive, time-consuming, and limited in their ability to analyze large chemical libraries comprehensively.

In recent years, machine-learning techniques have emerged as powerful tools for classifying beta-secretase 1 inhibitor activity, significantly improving efficiency and accuracy. Several studies have explored the application of various machine-learning algorithms. In their study, Ponzoni et al. [27] utilized a combination of neural networks and random forests to identify potential inhibitors of the beta-secretase 1 protein, employing classification methods for model development. Their approach involved utilizing a database containing 215 molecules to train and validate the classification models.

In the study by Nugroho et al. [28], they employed a neural network model to classify beta-secretase 1 activity for a dataset comprising 1531 compounds. They employed three optimization strategies to optimize the neural network: the Bat Algorithm, the Hybrid Bat Algorithm, and the Adaptive Bat Algorithm. The optimized model demonstrated an accuracy of 0.81 and an F1-score of 0.78, indicating its effectiveness in accurately classifying beta-secretase 1 activity.

Our previous study [19] explored the efficacy of four machine-learning models - Random Forest, AdaBoost, Gradient Boosting, and Extra Trees - for classifying beta-secretase 1 inhibitor activity. Among these models, Random Forest emerged as the top performer, demonstrating a testing accuracy of 82.53%. Notably, Random Forest exhibited superior precision, recall, and F1-score compared to the other models evaluated. These findings underscore the effectiveness of Random Forest in accurately classifying beta-secretase 1 inhibitor activity, highlighting its potential for advancing Alzheimer's disease drug discovery efforts.

While previous studies have made significant strides in classifying beta-secretase 1 inhibitor activity using various machine-learning techniques, there remains a gap in adopting more recent and advanced methodologies. Leveraging the latest advancements in machine-learning, particularly LightGBM, offers an opportunity to enhance classification models' efficiency, accuracy, and scalability for Alzheimer's disease drug discovery. By employing LightGBM, we aim to address this gap and further refine the classification of beta-secretase 1 inhibitor activity, ultimately advancing the development of potential therapeutics for Alzheimer's disease.

3. Proposed Method

The workflow of our proposed approach is shown in Figure 1, encompassing three main steps: initial data preparation, followed by the model building phase utilizing LightGBM, and concluding with the evaluation of the model's performance in classifying beta-secretase 1 inhibitors.

3.1. Data Preparation

A total of 7298 compound data was obtained from the ChEMBL database, a comprehensive resource for chemical and biological data [29]. We defined class labels based on the IC_{50} values to create a binary classification task. Compounds with IC_{50} values below 1000 nM were designated "active," totaling 4,574 compounds. Conversely, 2,724 compounds were labeled as "inactive" due to their IC_{50} values equal to or exceeding 1000 nM. These class labels form the basis of our subsequent analysis [30].

Next, we calculated a set of molecular descriptors for each compound in the dataset using the Mordred [31]. Molecular descriptors are numerical representations of a compound's chemical and structural properties, providing valuable information for machine-learning models [32]. These descriptors capture a wide array of chemical information, ranging from simple constitutional properties, such as molecular weight and number of bonds, to more complex 3D molecular geometries and electronic properties. They enable the translation of chemical compounds into a numerical format that machine-learning algorithms can readily

analyze, facilitating the identification of patterns and relationships that are not easily discernible through traditional chemical analysis methods.

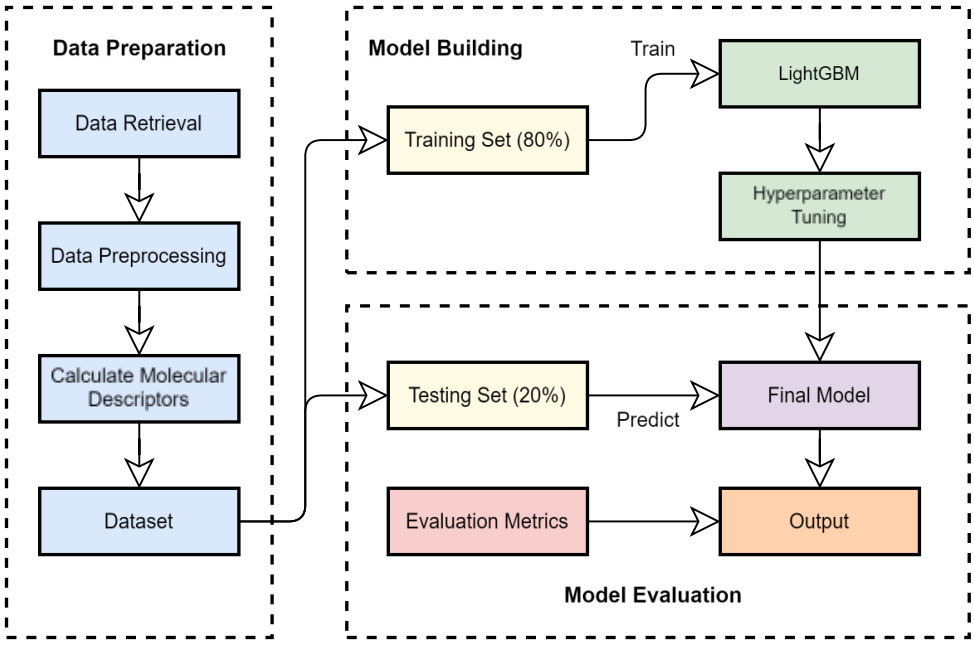

Figure 1. The proposed approach

In the preprocessing phase, we performed a rigorous filtering process to ensure that the selected descriptors were informative and did not introduce multicollinearity issues. Descriptors exhibiting zero variance across the dataset were removed, as they do not contribute to discrimination between active and inactive compounds. Additionally, we applied a threshold of 0.95 for multicollinearity, removing descriptors that were highly correlated with each other [33]. This step aimed to retain only the most relevant and non-redundant descriptors for our classification task. After applying the filtering criteria, we arrived at a final set of 456 molecular descriptors for each compound.

The dataset was then stratified and split into a training set (80%) and a testing set (20%). The distribution of data and class for each subset is shown in Table 1. It can be observed that the number of inactive compounds is higher than that of active compounds. This stratification ensures that both the training and testing sets have a proportionate representation of each class, which is crucial for maintaining the integrity of the model's performance across different data samples. It is particularly important to accurately classify the active compounds to ensure the effectiveness and cost-efficiency of drug screening before proceeding to laboratory validation.

Table 1. Hyperparameter space for LightGBM

Subset	Inactive Class	Active Class
Training Set	3655	2183
Testing Set	919	541

3.2. Model Building

The LightGBM training process involved a fine-tuning step to enhance its classification performance. We employed a random search approach with 10-fold cross-validation for this purpose. This method enables efficient exploration of the hyperparameter space, including parameters such as maximum depth, learning rate, subsample ratio, column subsample ratio, L1 regularization term (reg alpha), and L2 regularization term (reg lambda). The random state parameter was consistently set to 42 throughout the experiments to maintain reproducibility. The hyperparameter space explored is detailed in Table 2.

Table 2. Hyperparameter space for LightGBM

Hyperparameter	Description	Range
max_depth	Maximum depth of the tree	3 - 51
learning_rate	Learning rate for boosting	0.01-0.2
subsample	Subsample ratio of the training instance	0.6-1.0
colsample_bytree	Subsample ratio of columns when constructing each tree	0.6-1.0
reg_alpha	L1 regularization term	0.0-1.0
reg_lambda	L2 regularization term	0.0-1.0

3.3. Model Evaluation

To comprehensively evaluate model performance, we consider a set of metrics, including accuracy, precision, sensitivity, specificity, and F1-score. Accuracy gauges the overall correctness of our classifications, while precision measures the accuracy of positive classifications. Sensitivity quantifies the model's ability to capture all positive instances, and the F1-score balances precision and recall. These metrics ensure a thorough assessment of our model's ability to accurately classify beta-secretase 1 inhibitor activity, enabling us to identify the optimal model configuration for our research objectives. The equations for accuracy, precision, sensitivity, specificity, and F1-score are presented in Equations (1)-(5).

$$Accuracy = \frac{TP + FN}{FP + FN + TP + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$F1 - Score = 2\frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

We further compared the performance of our proposed approach with six other machine-learning algorithms: Random Forest, Support Vector Machine, Logistic Regression, Naïve Bayesian, Neural Network and XGBoost to evaluate its effectiveness in classifying beta-secretase 1 inhibitor activity.

4. Results and Discussion

The LightGBM model has been trained using specific hyperparameters to optimize its performance. These include a maximum depth of 41, a learning rate of 0.2, a subsample ratio of 1.0, a column subsample ratio of 1.0, L1 regularization (reg alpha) of 0.0, and L2 regularization (reg lambda) of 1.0. Based on these hyperparameters, the trained LightGBM model exhibits several characteristics. A large maximum depth value indicates that the model can capture complex relationships within the data, potentially leading to a higher capacity for learning intricate patterns. The relatively high learning rate suggests that the model adapts quickly to new information during training. The subsample ratio value indicates that all training instances are used for each tree, meaning no subsampling exists. The column subsample ratio value means all features are considered for splitting at each node. The L1 regularization term value indicates that there's no additional penalty for large coefficients, and the L2 regularization term value indicates a balanced approach to controlling overfitting.

The performance of the LightGBM model and its comparison with other machine-learning models are illustrated in Table 3. The LightGBM model achieved an accuracy of

84.93%, a precision of 87.14%, a sensitivity of 89.93%, a specificity of 77.63%, and an F1-score of 88.17%. These results underscore the robustness of the LightGBM model, especially in the F1-score, which translates to its ability to classify compounds into active and inactive categories accurately. While the Random Forest model shows a superior recall rate, indicating its strength in identifying active compounds, the LightGBM model demonstrates a more balanced performance. This balance is crucial in applications such as drug discovery, where accurate classification of compounds streamlines the virtual screening process.

Table 3. Performance of machine-learning models

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
LightGBM	**84.93**	**87.14**	89.93	**77.63**	**88.17**
Random Forest	82.19	81.47	**92.82**	64.14	86.78
Support Vector Machine	82.12	84.20	88.14	71.90	86.12
Logistic Regression	81.23	83.28	87.81	70.86	85.49
Naïve Bayesian	75.00	76.18	87.70	53.42	81.54
Neural Network	83.49	85.61	88.68	74.68	87.12
XGBoost	84.93	86.79	88.68	77.08	87.73

Bold values indicate best results

Table 4. Confusion matrix of machine-learning models

Model	Actual	Predicted	
		Active	Inactive
LightGBM	Active	820	99
	Inactive	121	420
Random Forest	Active	853	66
	Inactive	194	347
Support Vector Machine	Active	810	109
	Inactive	152	389
Logistic Regression	Active	807	112
	Inactive	162	379
Naïve Bayesian	Active	806	113
	Inactive	252	289
Neural Network	Active	815	104
	Inactive	137	404
XGBoost	Active	815	104
	Inactive	124	417

The confusion matrix table depicting the performance of various machine-learning models, including LightGBM, Random Forest, Support Vector Machine, Logistic Regression, Naïve Bayesian, Neural Network, and XGBoost, in classifying the activity of beta-secretase 1 inhibitors is shown in Table 4. Specifically focusing on the LightGBM model, it correctly classified 820 active and 420 inactive compounds while misclassifying 99 active and 121 inactive compounds. Comparatively, the Random Forest model achieved higher accuracy in classifying active compounds but exhibited lower accuracy in classifying inactive compounds. This disparity resulted in a less balanced performance, as evidenced by the higher false positive and false negative rates. In contrast, the LightGBM model demonstrated a more balanced performance, achieving a higher F1-score, which signifies a better trade-off between precision and recall.

This balanced performance is crucial in drug discovery scenarios, where accurately identifying both inactive and active compounds is essential for streamlining the virtual screening process and identifying potential drug candidates effectively. In this domain, achieving an optimal balance between recall and specificity is important for ensuring efficiency and cost-effectiveness in laboratory experiments. Specifically, false positives can lead to costly

and time-consuming experimental validation processes, making specificity a crucial metric. On the other hand, false negatives may result in overlooking potentially promising drug candidates. Given these considerations, we utilized the F1-score as our primary evaluation metric. The F1-score provides a balanced measure of both precision and recall, which aligns well with our objectives of maximizing the overall accuracy of our drug screening process while minimizing the risk of false positives and false negatives.

The Receiver Operating Characteristic (ROC) curves for the various machine-learning models evaluated in this study are shown in Figure 2. The ROC curve represents the true positive rate (sensitivity) against the false positive rate (1 - specificity) for different classification thresholds. Essentially, it illustrates the trade-off between sensitivity and specificity across different threshold values. In this context, a higher area under the ROC curve (AUC) indicates better overall performance of the model in distinguishing between true positive and false positive cases. It is evident that the ROC curve for the LightGBM model exhibits the highest AUC value, measuring at 0.91. This signifies that the LightGBM model outperforms the other machine-learning models in terms of its ability to accurately classify beta-secretase 1 inhibitors, achieving a superior balance between sensitivity and specificity.

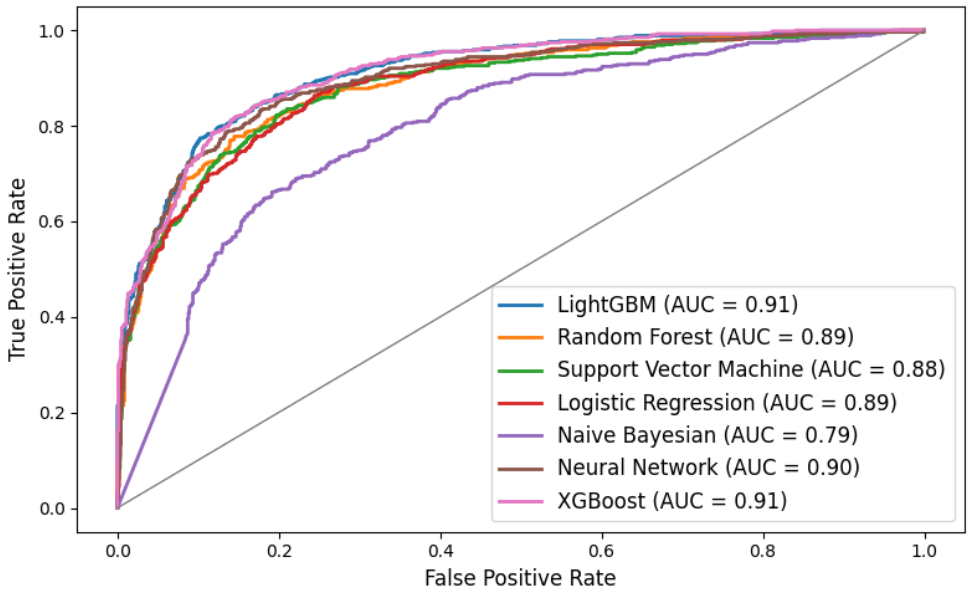

Figure 2. ROC plot of machine-learning models

For further analysis, we visualized the top five most important molecular descriptors in the LightGBM model to understand the LightGBM model's decision-making process for classifying beta-secretase 1 inhibitor activity (Figure 3). This LightGBM model uses a split feature importance method, which measures the importance of a feature by counting the number of times the feature is used in a model's decision trees. The most important descriptor appears to be PEOE_VSA8, which pertains to a specific atom type's steric and electronic environment and its impact on molecular interactions with beta-secretase 1. Following this, ATSC8dv is a topological autocorrelation descriptor that highlights the molecule's shape and connectivity, suggesting its influence on binding affinity. Similarly, ATSC8Z accounts for the atomic number influences of atom type 8, reflecting the role of elemental composition. IC2 measures the informational complexity of the molecular structure, indicating a nuanced aspect of molecular diversity. Lastly, ATSC8se brings attention to specific electronic properties, which could affect how a molecule interacts with the target site. These descriptors collectively contribute to the model's effectiveness in identifying potential inhibitors, with each descriptor providing a unique piece of the puzzle in understanding molecular activity against beta-secretase 1.

The empirical evidence presented highlights the effectiveness of the LightGBM model in identifying beta-secretase 1 inhibitors, demonstrating well-tuned hyperparameters and emphasizing the significant role of molecular descriptors in classifying drug efficacy. With its remarkable accuracy, precision, and balanced performance metrics compared to alternative models, LightGBM emerges as a valuable asset in the drug discovery toolkit. Examining the

importance of features has yielded additional insights into the essential molecular characteristics crucial for inhibitor binding, offering valuable guidance for enhancing virtual screening and drug design strategies.

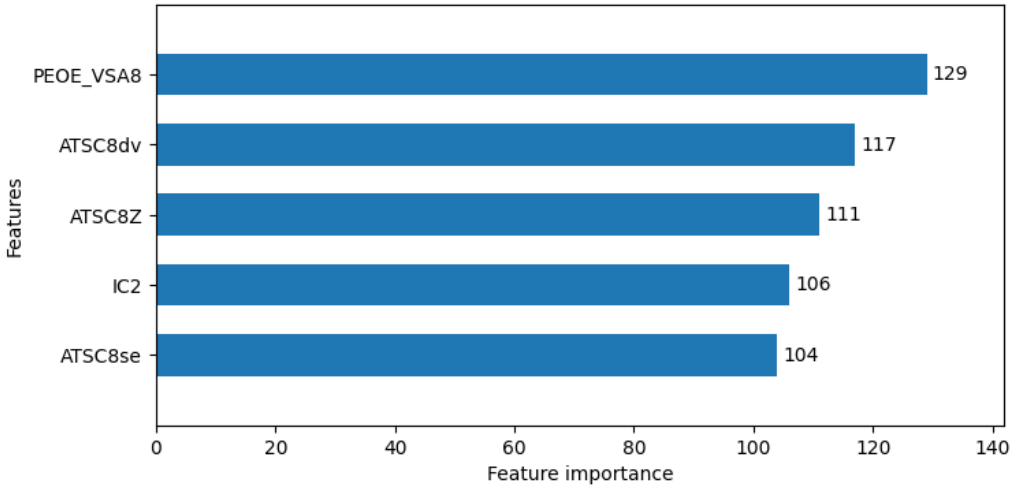

Figure 3. Feature importance of the LightGBM model

5. Conclusions

The study conducted with the LightGBM model has demonstrated its efficacy in classifying the activity of beta-secretase 1 inhibitors, marking a significant advancement in the field of computational drug discovery. The main findings reveal that the model, through its optimal hyperparameter settings, yields high accuracy, precision, specificity, and a balanced F1-score, which aligns with the research objectives to improve virtual screening processes. The relationship between the top molecular descriptors and inhibitor activity provides a deeper understanding of the drug-binding mechanisms. While the results are promising, the limitations in molecular descriptors' scope call for further research to explore additional properties and external validation with larger datasets to generalize the findings. This study paves the way for future investigations that could expand the utility of machine-learning in drug development, potentially leading to more targeted and effective treatments for diseases like Alzheimer's.

Author Contributions: Conceptualization: T.R.N., K.N. and N.R.S.; methodology, T.R.N.; software: T.R.N.; validation: T.R.N., K.N., and I.H..; formal analysis: T.R.N. and G.M.I.; investigation: T.R.N. and G.M.I.; resources: K.N. and I.H.; data curation: K.N., I.H. and N.R.S.; writing—original draft preparation: T.R.N. and I.H.; writing—review and editing: T.R.N., K.N. and G.M.I.; visualization: G.M.I.; supervision: T.R.N.; project administration: T.R.N. and N.R.S.; funding acquisition: T.R.N.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study is available upon reasonable request to the corresponding author.

Acknowledgments: The authors express their gratitude to their respective institutions for their invaluable support throughout this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] A. Gholami, "Alzheimer's disease: The role of proteins in formation, mechanisms, and new therapeutic approaches," *Neurosci. Lett.*, vol. 817, p. 137532, Nov. 2023, doi: 10.1016/j.neulet.2023.137532.

[2] A. Gustavsson *et al.*, "Global estimates on the number of persons across the Alzheimer's disease continuum," *Alzheimer's Dement.*, vol. 19, no. 2, pp. 658–670, Feb. 2023, doi: 10.1002/alz.12694.

[3] Y. Wang, Y. Zhang, and E. Yu, "Targeted examination of amyloid beta and tau protein accumulation via positron emission tomography for the differential diagnosis of Alzheimer's disease based on the A/T(N) research framework," *Clin. Neurol. Neurosurg.*, vol. 236, p. 108071, Jan. 2024, doi: 10.1016/j.clineuro.2023.108071.
[4] R. Singh *et al.*, "Classification of beta-site amyloid precursor protein cleaving enzyme 1 inhibitors by using machine learning methods," *Chem. Biol. Drug Des.*, vol. 98, no. 6, pp. 1079–1097, Dec. 2021, doi: 10.1111/cbdd.13965.
[5] K. S. Orobets and A. L. Karamyshev, "Amyloid Precursor Protein and Alzheimer's Disease," *Int. J. Mol. Sci.*, vol. 24, no. 19, p. 14794, Sep. 2023, doi: 10.3390/ijms241914794.
[6] I. Hajdú, B. M. Végh, A. Szilágyi, and P. Závodszky, "Beta-Secretase 1 Recruits Amyloid-Beta Precursor Protein to ROCK2 Kinase, Resulting in Erroneous Phosphorylation and Beta-Amyloid Plaque Formation," *Int. J. Mol. Sci.*, vol. 24, no. 13, p. 10416, Jun. 2023, doi: 10.3390/ijms241310416.
[7] T. R. Noviandy *et al.*, "Ensemble Machine learning Approach for Quantitative Structure Activity Relationship Based Drug Discovery: A Review," *Infolitika J. Data Sci.*, vol. 1, no. 1, pp. 32–41, Sep. 2023, doi: 10.60084/ijds.v1i1.91.
[8] T. R. Noviandy *et al.*, "Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery," *Malacca Pharm.*, vol. 1, no. 2, pp. 48–54, 2023, doi: 10.60084/mp.v1i2.60.
[9] K. Roy and S. Kar, "How to Judge Predictive Quality of Classification and Regression Based QSAR Models?," Z. Ul-Haq and J. D. B. T.-F. in C. C. Madura, Eds. Bentham Science Publishers, 2015, pp. 71–120. doi: https://doi.org/10.1016/B978-1-60805-979-9.50003-2.
[10] T. R. Noviandy, A. Maulana, G. M. Idroes, I. Irvanizam, M. Subianto, and R. Idroes, "QSAR-Based Stacked Ensemble Classifier for Hepatitis C NS5B Inhibitor Prediction," in *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, Aug. 2023, pp. 220–225. doi: 10.1109/COSITE60233.2023.10250039.
[11] M. Azizah, A. Yanuar, and F. Firdayani, "Dimensional Reduction of QSAR Features Using a Machine learning Approach on the SARS-Cov-2 Inhibitor Database," *J. Penelit. Pendidik. IPA*, vol. 8, no. 6, pp. 3095–3101, Dec. 2022, doi: 10.29303/jppipa.v8i6.2432.
[12] K. Mondal and S. K. S, "QSAR Classification Models for Predicting 3CLPro-protease Inhibitor Activity," in *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, Sep. 2021, pp. 1–6. doi: 10.1109/GUCON50781.2021.9573896.
[13] N. B. Maulydia, K. Khairan, and T. R. Noviandy, "Prediction of Pharmacokinetic Parameters from Ethanolic Extract Mane Leaves (Vitex pinnata L.) in Geothermal Manifestation of Seulawah Agam Ie-Seu'um, Aceh," *Malacca Pharm.*, vol. 1, no. 1, pp. 16–21, Jun. 2023, doi: 10.60084/mp.v1i1.33.
[14] M. Agustia *et al.*, "Application of Fuzzy Support Vector Regression to Predict the Kovats Retention Indices of Flavors and Fragrances," in *2022 International Conference on Electrical Engineering and Informatics (ICELTICs)*, Sep. 2022, pp. 13–18. doi: 10.1109/ICELTICs56128.2022.9932124.
[15] R. Idroes *et al.*, "Application of Genetic Algorithm-Multiple Linear Regression and Artificial Neural Network Determinations for Prediction of Kovats Retention Index," *Int. Rev. Model. Simulations*, vol. 14, no. 2, p. 137, 2021.
[16] T. R. Noviandy *et al.*, "The Prediction of Kovats Retention Indices of Essential Oils at Gas Chromatography Using Genetic Algorithm-Multiple Linear Regression and Support Vector Regression," *J. Eng. Sci. Technol.*, vol. 17, no. 1, pp. 306–326, 2022.
[17] S. Simeon *et al.*, "Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking," *PeerJ*, vol. 4, p. e2322, Aug. 2016, doi: 10.7717/peerj.2322.
[18] M. Abdullahi, G. A. Shallangwa, and A. Uzairu, "In silico QSAR and molecular docking simulation of some novel aryl sulfonamide derivatives as inhibitors of H5N1 influenza A virus subtype," *Beni-Suef Univ. J. Basic Appl. Sci.*, vol. 9, no. 1, p. 2, Dec. 2020, doi: 10.1186/s43088-019-0023-y.
[19] T. R. Noviandy, A. Maulana, T. B. Emran, G. M. Idroes, and R. Idroes, "QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine learning Algorithms," *Heca J. Appl. Sci.*, vol. 1, no. 1, pp. 1–7, 2023, doi: 10.60084/hjas.v1i1.12.
[20] I. Ponzoni *et al.*, "QSAR Classification Models for Predicting the Activity of Inhibitors of Beta-Secretase (BACE1) Associated with Alzheimer's Disease," *Sci. Rep.*, vol. 9, no. 1, p. 9102, Jun. 2019, doi: 10.1038/s41598-019-45522-3.
[21] A. Fernández-Torras, A. Comajuncosa-Creus, M. Duran-Frigola, and P. Aloy, "Connecting chemistry and biology through molecular descriptors," *Curr. Opin. Chem. Biol.*, vol. 66, p. 102090, Feb. 2022, doi: 10.1016/j.cbpa.2021.09.001.
[22] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[23] T. R. Noviandy, S. I. Nainggolan, R. Raihan, I. Firmansyah, and R. Idroes, "Maternal Health Risk Detection Using Light Gradient Boosting Machine Approach," *Infolitika J. Data Sci.*, vol. 1, no. 2, pp. 48–55, Dec. 2023, doi: 10.60084/ijds.v1i2.123.
[24] L. Sari, A. Romadloni, R. Lityaningrum, and H. D. Hastuti, "Implementation of LightGBM and Random Forest in Potential Customer Classification," *TIERS Inf. Technol. J.*, vol. 4, no. 1, pp. 43–55, Jun. 2023, doi: 10.38043/tiers.v4i1.4355.
[25] P. Kumari, A. Nath, and R. Chaube, "Identification of human drug targets using machine-learning algorithms," *Comput. Biol. Med.*, vol. 56, pp. 175–181, Jan. 2015, doi: 10.1016/j.compbiomed.2014.11.008.
[26] G. B. McGaughey and M. K. Holloway, "Structure-guided design of β-secretase (BACE-1) inhibitors," *Expert Opin. Drug Discov.*, vol. 2, no. 8, pp. 1129–1138, Aug. 2007, doi: 10.1517/17460441.2.8.1129.
[27] I. Ponzoni *et al.*, "QSAR classification models for predicting the activity of inhibitors of beta-secretase (BACE1) associated with Alzheimer's disease," *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, 2019.
[28] A. F. Nugroho, R. Rendian Septiawan, and I. Kurniawan, "Prediction of Human β-secretase 1 (BACE-1) Inhibitors for Alzheimer Therapeutic Agent by Using Fingerprint-based Neural Network Optimized by Bat Algorithm," in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, Feb. 2023, pp. 257–261. doi: 10.1109/ICCoSITE57641.2023.10127718.
[29] A. Gaulton *et al.*, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D1100–D1107, Jan. 2012, doi: 10.1093/nar/gkr777.

[30] S. Simeon *et al.*, "Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking," *PeerJ*, vol. 4, p. e2322, 2016.
[31] H. Moriwaki, Y.-S. Tian, N. Kawashita, and T. Takagi, "Mordred: a molecular descriptor calculator," *J. Cheminform.*, vol. 10, no. 1, p. 4, Dec. 2018, doi: 10.1186/s13321-018-0258-y.
[32] A. Mauri, V. Consonni, and R. Todeschini, "Molecular Descriptors," in *Handbook of Computational Chemistry*, Cham: Springer International Publishing, 2017, pp. 2065–2093. doi: 10.1007/978-3-319-27282-5_51.
[33] T. Yu, C. Nantasenamat, S. Kachenton, N. Anuwongcharoen, and T. Piacham, "Cheminformatic Analysis and Machine learning Modeling to Investigate Androgen Receptor Antagonists to Combat Prostate Cancer," *ACS Omega*, vol. 8, no. 7, pp. 6729–6742, Feb. 2023, doi: 10.1021/acsomega.2c07346.