Original Article

Check for updates

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

Malacca Pharmaceutics

Vol. 3, No. 1, 2025

Evaluation of Machine Learning Methods for Identifying Carbonic Anhydrase-II Inhibitors as Drug Candidates for Glaucoma

Teuku Rizky Noviandy [1], Eva Imelda [2,3], Ghazi Mauer Idroes [4], Rivansyah Suhendra [5] and Rinaldi Idroes [6,*]

[1] Department of Information Systems, Faculty of Engineering, Universitas Abulyatama, Aceh Besar 23372, Indonesia; rizky_si@abulyatama.ac.id (T.R.N.)
[2] Department of Ophthalmology, General Hospital Dr. Zainoel Abidin, Banda Aceh 23126, Indonesia; evaimeldaspmpo@gmail.com (E.I.)
[3] Department of Ophthalmology, Faculty of Medicine, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
[4] Department of Occupational Health and Safety, Faculty of Health Sciences, Universitas Abulyatama, Aceh Besar 23372, Indonesia; idroesghazi_k3@abulyatama.ac.id (G.M.I)
[5] Department of Information Technology, Faculty of Engineering, Universitas Teuku Umar, Aceh Barat 23681, Indonesia; rivansyahsuhendra@utu.ac.id (R.S.)
[6] School of Mathematics and Applied Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; rinaldi.idroes@usk.ac.id (R.I.)

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

Article History

Received 28 December 2024
Revised 19 February 2025
Accepted 26 February 2025
Available Online 4 March 2025

Keywords:
CA-II inhibitors
Virtual screening
Glaucoma drug discovery
Artificial intelligence

Abstract

Glaucoma is a leading cause of irreversible blindness, primarily managed by lowering intraocular pressure (IOP). Carbonic Anhydrase-II (CA-II) inhibitors play a crucial role in this treatment by reducing aqueous humor production. However, existing CA-II inhibitors often suffer from poor selectivity, side effects, and limited bioavailability, highlighting the need for more efficient and targeted drug discovery approaches. This study uses machine learning-driven Quantitative Structure-Activity Relationship (QSAR) modeling to predict CA-II inhibition based on molecular descriptors, significantly enhancing screening efficiency over traditional experimental methods. By evaluating multiple machine learning models, including Support Vector Machine, Gradient Boosting, and Random Forest, we identify SVM as the most effective classifier, achieving the highest accuracy (83.70%) and F1-score (89.36%). Class imbalance remains challenging despite high sensitivity, necessitating further improvements through resampling and hyperparameter optimization. Our findings underscore the potential of machine learning-based virtual screening in accelerating CA-II inhibitor identification and advocate for integrating AI-driven approaches with traditional drug discovery techniques. Future directions include deep learning enhancements and hybrid machine learning-docking frameworks to improve prediction accuracy and facilitate the development of more potent and selective glaucoma treatments.

1. Introduction

Glaucoma is a progressive neurodegenerative disease that leads to irreversible vision loss, primarily caused by increased intraocular pressure (IOP) due to impaired aqueous humor drainage [1]. One of the key therapeutic targets for glaucoma treatment is carbonic anhydrase-II (CA-II), an enzyme involved in aqueous humor production [2, 3]. Inhibiting CA-II reduces the production of bicarbonate ions, leading to decreased aqueous humor secretion and, consequently, a reduction in IOP, making CA-II inhibitors a crucial class of drugs for glaucoma management [4].

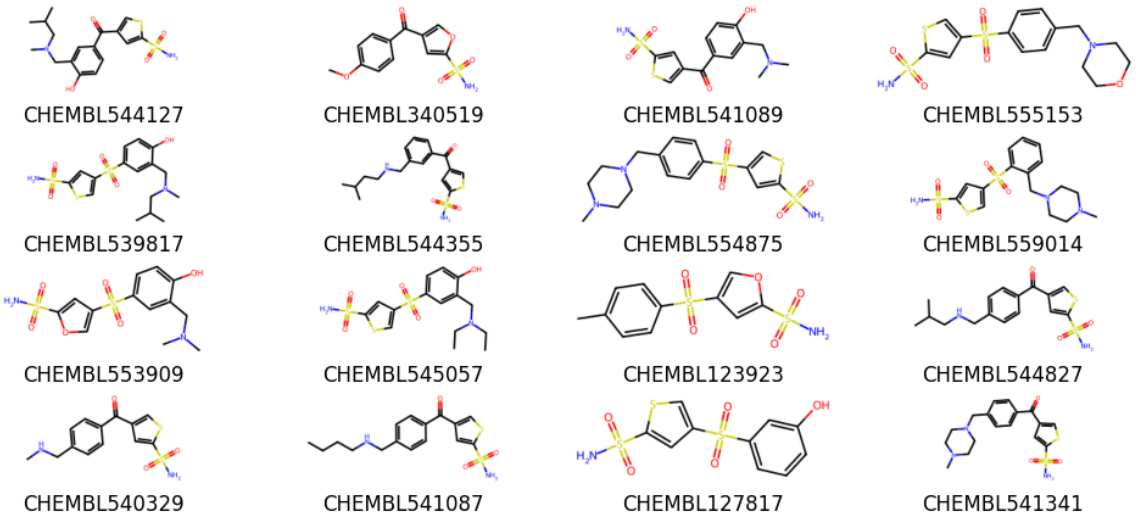

Figure 1. Examples of chemical structures from the dataset utilized in this study.

Despite significant advancements in drug development, identifying novel and effective CA-II inhibitors remains a challenge. Many existing inhibitors exhibit poor selectivity, undesirable side effects, or limited bioavailability, reducing their therapeutic potential [5–7]. Furthermore, traditional drug screening methods require extensive biochemical and structural analysis, which limits the speed of drug discovery [8]. A more efficient, cost-effective, and accurate approach is needed to identify potential CA-II inhibitors with improved pharmacological properties.

Machine learning has emerged as a powerful tool in drug discovery, enabling the rapid prediction of chemical compounds' bioactivity, toxicity, and pharmacokinetic properties [9, 10]. Machine learning algorithms leverage large datasets to recognize patterns in molecular structures and predict their interactions with target proteins [11, 12]. Machine learning-based models offer a more efficient alternative by rapidly evaluating candidate molecules, prioritizing the most promising compounds, and significantly reducing the need for exhaustive experimental trials.

A key computational approach in drug discovery is Quantitative Structure-Activity Relationship (QSAR) modeling, which uses mathematical and statistical techniques to correlate molecular descriptors with biological activity [13, 14]. QSAR modeling is advantageous because it provides interpretable relationships between molecular features and biological responses, allowing researchers to design and optimize drug candidates rationally. By integrating machine learning with QSAR modeling, predictive accuracy can be further enhanced, overcoming the limitations of traditional QSAR methods that rely on linear or predefined models [15].

Several studies have demonstrated the effectiveness of machine learning- driven QSAR models in identifying inhibitors for various drug targets, such as kinase inhibitors for cancer treatment [16–18] and antiviral compounds for infectious diseases [19–22]. However, our search did not yield studies specifically applying machine learning-driven QSAR models for CA-II inhibitors. This highlights a gap in the literature, underscoring the need for further investigation, which this study aims to address.

This study aims to evaluate the effectiveness of different machine learning methods in identifying potential CA-II inhibitors as drug candidates for glaucoma treatment. Specifically, we will develop and compare multiple machine learning models to predict CA-II inhibition based on molecular descriptors and chemical features. By leveraging computational techniques, this study seeks to enhance the efficiency and accuracy of CA-II inhibitor discovery, ultimately contributing to the development of improved therapeutic options for glaucoma.

2. Materials and Methods

2.1. Data Collection and Preparation

The dataset used in this study was obtained from ChEMBL [23], explicitly targeting CA-II (CHEMBL205). A total of 920 compounds were selected, each represented by its SMILES notation along with its IC_{50} value. The IC50 values were used to define compound activity to facilitate the classification task. Compounds with IC_{50} values lower than 1000 nM were categorized as active, while those with IC_{50} values equal to or greater than 1000 nM were classified as inactive [24]. The chemical structures from the dataset used in this study can be seen in Figure 1. The figure serves to provide an overview of the molecular

diversity within the dataset, highlighting the structural variations among the selected compounds.

2.2. Calculate Fingerprint

To represent the molecular structures numerically, PubChem fingerprints were generated using PaDEL version 2.2.1, with preprocessing steps that included the removal of salts and standardizing nitro groups. These fingerprints are vectorized representations of molecules, encoding key structural and chemical features essential for computational analysis [25]. PubChem fingerprints were selected due to their widespread use and ability to capture key structural and chemical features relevant for computational analysis. This fingerprint consists of 881 binary descriptors, where each descriptor indicates the presence or absence of specific molecular substructures, providing a standardized and efficient representation for further analysis.

Preprocessing steps were applied to improve the quality of the features. First, descriptors with zero variance were removed as they provided no discriminatory power [26]. Next, a multicollinearity threshold of 95% was applied, eliminating highly correlated features to reduce redundancy [27]. After these refinements, a final set of 387 descriptors was retained for further analysis, ensuring a more informative and efficient feature space for machine learning models.

2.3. Exploratory Data Analysis

To gain insights into the dataset and assess potential separability between active and inactive compounds, Principal Component Analysis (PCA) was applied to the PubChem fingerprint descriptors. PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving variance [28]. The first two principal components (PC1 and PC2) were extracted and visualized in a 2D scatter plot to explore the distribution of active and inactive compounds.

Additionally, Lipinski's descriptors were computed to analyze the physicochemical properties of the compounds [29]. The selected descriptors included molecular weight (MW), which measures the size of the molecule, LogP, representing lipophilicity and indicating solubility in lipids and water, number of hydrogen bond donors (NumHDonors), which reflects potential hydrogen-bonding interactions, and number of hydrogen bond acceptors (NumHAcceptors), indicating the molecule's capacity to form hydrogen bonds. These descriptors provide essential insights into the drug-likeness of compounds and their potential bioavailability. Descriptive statistics were computed to compare the distribution of these descriptors between active and inactive compounds, including mean, median, standard deviation, and interquartile range. A Mann-Whitney U test, a non-parametric statistical test, was performed to assess whether there were significant differences in these descriptors between active and inactive compounds [30]. Finally, boxplots were generated for each Lipinski descriptor to visualize variations between the two classes, highlighting any distinguishing physicochemical characteristics that could influence compound activity.

2.4. Machine Learning Models

Multiple machine learning models were implemented to classify compounds as active or inactive inhibitors of CA-II. The selected models included Logistic Regression, Random Forest, Support Vector Machine, Gradient Boosting, and k-nearest Neighbors. Each offers distinct advantages in handling molecular data and classification tasks.

Logistic Regression was selected due to its simplicity and effectiveness in binary classification problems [31]. Random Forest, an ensemble-based model, was chosen for its ability to handle non-linearity, mitigate overfitting, and improve accuracy by aggregating multiple decision trees [32]. Support Vector Machine was included because of its strong performance in high-dimensional spaces, a common characteristic of molecular fingerprint data, where it efficiently finds the optimal hyperplane to separate active and inactive compounds [33].

Gradient Boosting was incorporated due to its sequential learning approach, which iteratively refines weak classifiers to enhance predictive performance. This makes it particularly effective for capturing subtle patterns in molecular structures [34]. Lastly, k-Nearest Neighbors was chosen as a distance-based method that classifies compounds based on molecular similarity, which aligns well with structure-activity relationships in drug discovery [35].

2.5. Model Training and Validation

The dataset was divided into training and testing sets using an 80:20 split to ensure a robust evaluation of model performance. The training set, comprising 80% of the data, was used to train the machine learning models, while the remaining 20% was reserved for testing to assess generalization capabilities [36]. All models were implemented using the Scikit-Learn library, with default hyperparameters applied during training. This approach provided a baseline performance measure before considering further optimization [37].

Figure 2. The PCA scatter plot of the first two principal components illustrates the distribution of active and inactive CA-II inhibitors.

2.6. Performance Evaluation

The performance of the trained machine learning models was evaluated using four key classification metrics: accuracy, precision, recall, and F1-score. Accuracy measured the overall correctness of the model by calculating the proportion of correctly classified compounds. Precision indicated the reliability of positive predictions by assessing the proportion of correctly identified active compounds among all predicted active compounds. Recall, also known as sensitivity, measured the model's ability to identify all actual active compounds within the dataset. The F1-score, a harmonic mean of precision and recall, provided a balanced assessment, particularly useful in cases of class imbalance. These metrics were computed for each model on the test dataset, allowing for a comparative analysis of their effectiveness in distinguishing active and inactive CA-II inhibitors. The results helped determine the most suitable approach for predicting potential drug candidates for glaucoma treatment [38].

3. Results and Discussion

3.1. Exploratory Data Analysis

PCA was performed to explore the distribution of active and inactive compounds within a lower-dimensional space. Figure 2 illustrates the scatter plot of the first two principal components, highlighting the spatial distribution of the compounds categorized as active and inactive. The plot demonstrates a high degree of overlap between the two classes, suggesting that the molecular descriptors alone may not provide a clear boundary for distinguishing active CA-II inhibitors from inactive ones. This indicates that simple linear separability is challenging, reinforcing the need for more advanced classification techniques such as non-linear machine learning models.

While some regions exhibit a noticeable clustering of active or inactive compounds, the overall spread suggests that multiple molecular features contribute to activity, and their interactions are not easily captured in two-dimensional space. The observed dispersion also underscores the chemical space's complexity and PCA's limitations in fully resolving class separability. This motivates the application of machine learning models that can learn non-linear relationships and leverage high-dimensional patterns within the dataset.

Table 1 summarizes the descriptive statistics and Mann-Whitney U test results for Lipinski's Rule of Five descriptors. The analysis reveals significant differences between active and inactive compounds regarding MW, LogP, nHA, and nHD. The Mann-Whitney U test yielded p-values lower than 0.05 for all descriptors, indicating statistically significant differences between the two groups.

Active compounds generally exhibited higher molecular weights, with a median value of 358.93 compared to 295.73 for inactive compounds. This suggests that larger molecular structures may be more favorable for CA-II inhibition. Similarly, the mean LogP value for active compounds was lower (1.6) than for inactive compounds (2.08), implying that more hydrophilic compounds tend to

Table 1. Descriptive statistics and Mann-Whitney U test results for Lipinski's Rule of Five descriptors.

	MW		LogP		nHA		nHD	
	Inactive	**Active**	**Inactive**	**Active**	**Inactive**	**Active**	**Inactive**	**Active**
p-value	1.79×10^{-14}		5.30×10^{-4}		4.95×10^{-3}		9.11×10^{-36}	
Min	65.39	149.09	-1.72	-3.14	0	0	0	2
Max	654.72	780.85	6.43	7.05	6	7	10	16
Median	295.73	358.93	2.04	1.52	2	2	4	6
Mean	311.03	372.79	2.08	1.6	1.97	2.13	4.23	6.05
Skew	0.72	0.97	0.11	0.2	0.8	1.56	0.63	1.37
Kurtosis	0.61	0.79	-0.92	0.55	-0.09	3.07	0.31	3.22

Note: The p-value represents the result of the Mann-Whitney U test.

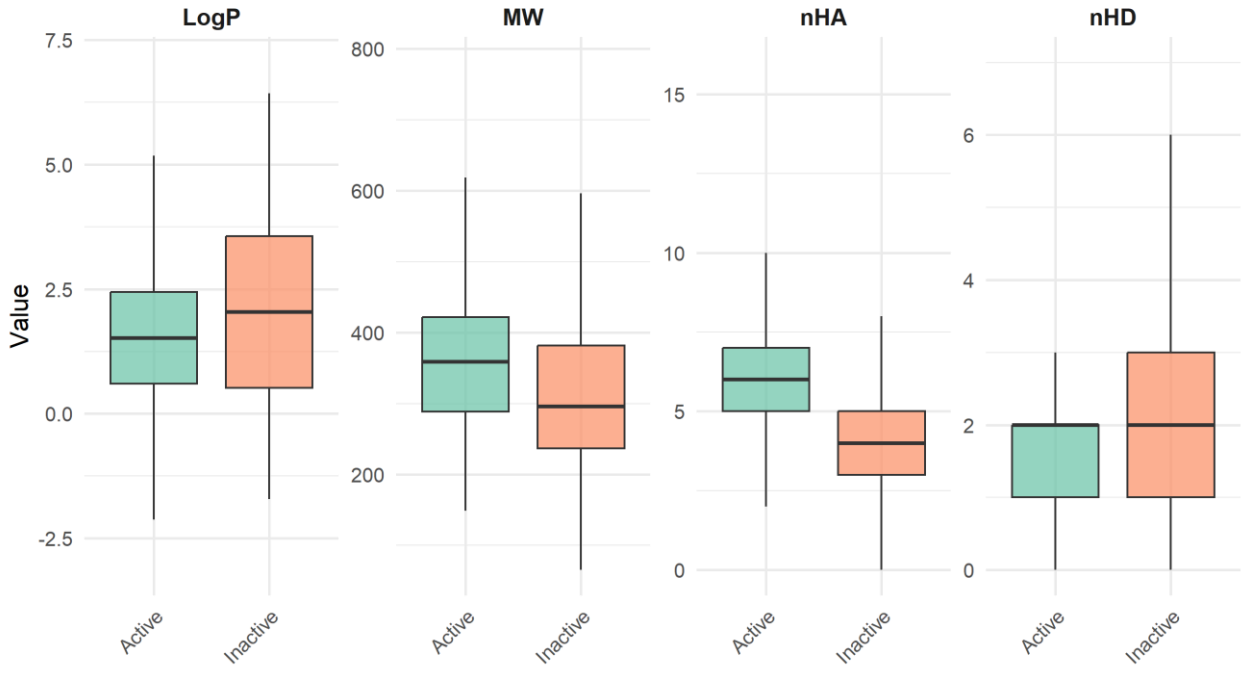

Figure 3. Box plots comparing Lipinski's Rule of Five descriptors (MW, LogP, nHD, nHA) between active and inactive CA-II inhibitors.

be better inhibitors. The differences in nHA and nHD were also notable, with active compounds showing a higher median and mean for both descriptors. Specifically, active compounds had more hydrogen bond donors (median = 6) and acceptors (median = 2), which may contribute to their ability to form key interactions with the CA-II enzyme.

Additionally, the skewness and kurtosis values indicate that the distribution of these descriptors is not perfectly normal, with active compounds displaying greater asymmetry and higher peaks in their distributions. The results highlight that physicochemical properties play an essential role in determining CA-II inhibition potential and reinforce the importance of incorporating these descriptors in predictive modeling. The statistically significant differences suggest that molecular weight, lipophilicity, and hydrogen bonding characteristics should be key features in developing machine learning models for CA-I inhibitor classification.

Figure 3 presents box plots comparing the distribution of Lipinski's Rule of Five descriptors between active and inactive compounds. These visualizations further support the statistical findings from Table 1, reinforcing the significant differences in physicochemical properties between the two groups.

The MW distribution shows that active compounds tend to have a higher median and interquartile range than inactive compounds, suggesting that larger molecules may be more effective as CA-II inhibitors. The LogP box plot reveals that active compounds generally have lower lipophilicity, which aligns with the idea that more hydrophilic molecules may enhance CA-II inhibition potential. The distribution of nHD and nHA also indicates a notable distinction, with active compounds exhibiting higher median values for both descriptors. This trend suggests that the ability to form hydrogen bonds plays a crucial role in the bioactivity of potential inhibitors.

Table 2. Performance of ensemble machine learning models on the classification of CA-II inhibitors.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F-1 Score (%)
Logistic Regression	82.61	88.06	88.06	68.00	88.06
Random Forest	81.52	85.71	89.55	60.00	87.59
Support Vector Machine	83.70	85.14	94.03	56.00	89.36
Gradient Boosting	83.15	86.01	91.79	60.00	88.81
K-Nearest Neighbors	81.52	86.23	88.81	62.00	87.50

3.2. Machine Learning Models

Table 2 summarizes the performance of different machine learning models for classifying CA-II inhibitors based on accuracy, precision, sensitivity, specificity, and F1-score. Among the models evaluated, the Support Vector Machine achieved the highest accuracy (83.70%) and F1-score (89.36%), indicating its superior capability in distinguishing active from inactive compounds. The Gradient Boosting model followed closely, with an accuracy of 83.15% and an F1-score of 88.81%, demonstrating its effectiveness in capturing complex patterns in the data. While slightly behind, Logistic Regression performed well with an accuracy of 82.61% and an F1-score of 88.06%, serving as a strong baseline method.

Random Forest and k-Nearest Neighbors exhibited similar performance, achieving an accuracy of 81.52% but differing slightly in F1-score, with Random Forest at 87.59% and k-Nearest Neighbors at 87.50%. Sensitivity values were relatively high across all models, suggesting they effectively identified active compounds. However, specificity was lower, particularly for the Support Vector Machine (56.00%) and Random Forest (60.00%) models, indicating that some inactive compounds were misclassified as active.

Based on the F1-score, the best-performing model was the SVM, demonstrating its effectiveness in handling the classification task. Despite lower specificity, the Support Vector Machine's high accuracy suggests that it effectively captures patterns distinguishing active compounds but may be biased toward classifying compounds as active. This trade-off between sensitivity and specificity could stem from the dataset composition or the nature of molecular descriptors.

Figure 4 presents the confusion matrices for five machine learning models that classify CA-II inhibitors, highlighting their predictive performance and class imbalance challenges. Class 0 represents active compounds in this study, while class 1 represents inactive compounds. The Support Vector Machine model achieved the highest true active count (126) but misclassified 22 inactive compounds as active. Despite this, Support Vector Machine demonstrated strong specificity and an overall high F1-score. Gradient Boosting and Logistic Regression showed more balanced classification, while Random Forest and K-Nearest Neighbors had slightly lower performance distinguishing inactive compounds, reducing their F1-scores. These results align with Table 2, where the Support Vector Machine exhibited the best balance in classification, though the trade-off between specificity and sensitivity remains a challenge. In drug discovery, false positives (misclassified inactive compounds) are particularly concerning as they may lead to the rejection of potentially effective inhibitors, delaying new treatments. Conversely, false negatives (misclassified active compounds) could waste resources on ineffective candidates, increasing development costs.

Figure 5 presents the Receiver Operating Characteristic (ROC) curves for the machine learning models used to classify CA-II inhibitors, evaluating their ability to distinguish between active and inactive compounds. The area under the curve (AUC) serves as a key performance metric, with higher values indicating better classification. Among the models, Random Forest (AUC = 0.865) and Gradient Boosting (AUC = 0.862) demonstrate the highest AUC values, confirming their strong predictive performance. Support Vector Machine follows with an AUC of 0.845, while K-Nearest Neighbors (AUC = 0.836) and Logistic Regression (AUC = 0.828) exhibit slightly lower classification accuracy. These results align with previous performance metrics, validating Random Forest and Gradient Boosting as the most effective models for identifying potential CA-II inhibitors.

3.3. Discussion

The results from this study demonstrate the potential of machine learning methods in identifying CA-II inhibitors for glaucoma treatment. Through extensive model evaluation, we observed that advanced machine learning techniques, particularly Support Vector Machine, achieved the highest classification accuracy, F1-score, and AUC, making them promising tools for virtual screening of novel drug candidates. However, several critical observations emerge from our findings, highlighting the approaches' strengths and limitations.

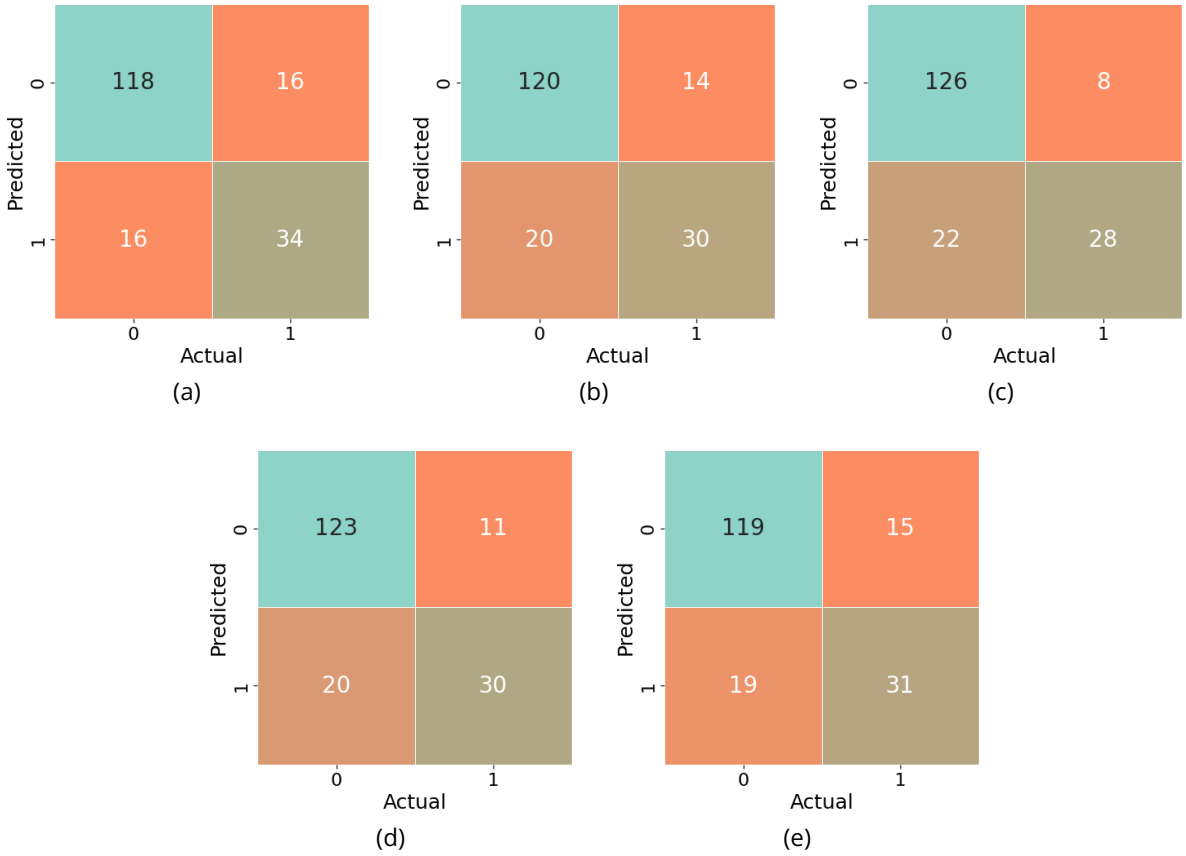

Figure 4. Confusion matrices of the machine learning models for CA-II inhibitor classification: (a) Logistic Regression, (b) Random Forest, (c) SVM, (d) Gradient Boosting, (e) KNN.

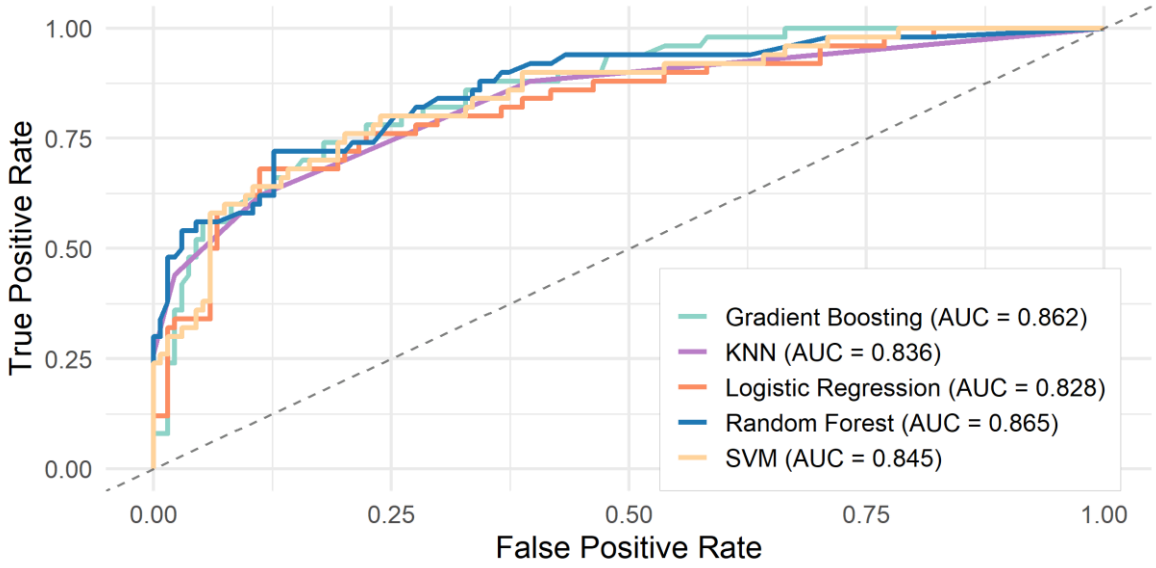

Figure 5. ROC curves of the machine learning models, comparing their ability to distinguish active from inactive CA-II inhibitors based on classification thresholds.

Despite the robust performance of machine learning models, the EDA suggests that the molecular space of active and inactive compounds exhibits significant overlap, as indicated by the PCA plot. This suggests that linear separation of classes is not straightforward, emphasizing the need for non-linear classification models, such as Support Vector Machine and Gradient Boosting, which successfully captured these complex relationships. Additionally, the results from Lipinski's Rule of Five analysis showed significant physicochemical differences between active and inactive compounds, reinforcing the importance of molecular weight,

hydrogen bonding, and lipophilicity in CA-II inhibition. These descriptors played a key role in feature selection and likely contributed to model performance.

While Support Vector Machine exhibited the highest accuracy (83.70%) and F1-score (89.36%), it also demonstrated lower specificity (56.00%), meaning that it tended to classify some inactive compounds as active. This trade-off suggests that while the model effectively identifies potential inhibitors, it may introduce false positives, increasing experimental validation costs. Similarly, Random Forest and K-Nearest Neighbors exhibited slightly lower performance due to their reliance on decision trees and distance metrics, which may not generalize well across diverse chemical spaces.

One interesting observation is that Gradient Boosting, despite achieving high accuracy (83.15%) and F1-score (88.81%), had relatively balanced sensitivity and specificity compared to Support Vector Machine. This suggests that ensemble learning techniques may offer a more generalizable approach when screening for potential drug candidates. However, the computational cost of training ensemble methods like Gradient Boosting is typically higher than simpler models like Logistic Regression.

The models exhibited high sensitivity but relatively lower specificity, suggesting a bias towards identifying active compounds. This imbalance may arise due to the inherent dataset composition or the high-dimensional nature of molecular fingerprints. Although we applied feature selection techniques, overfitting remains a concern, particularly for models trained on relatively small datasets. Future work could explore data augmentation strategies, such as generative models or molecular similarity-based augmentation, to improve generalizability. Furthermore, employing resampling techniques could help balance class distributions and improve specificity. Another approach to mitigate overfitting is hyperparameter optimization, where fine-tuning model parameters, particularly for tree-based methods like Random Forest and Gradient Boosting, could enhance classification performance.

Despite their promise, machine learning models should not be considered standalone replacements for traditional structure-based drug discovery methods. Instead, machine learning-based predictions should be integrated with computational docking and experimental validation to streamline the discovery pipeline. Given that machine learning models can rapidly screen thousands of molecules, they can serve as a pre-filtering step to prioritize candidates for more computationally expensive docking simulations.

4. Conclusions

This study demonstrates the potential of machine learning models in identifying CA-II inhibitors for glaucoma treatment, with Support Vector Machine and Gradient Boosting emerging as the most effective classifiers. The exploratory data analysis highlighted significant physicochemical differences between active and inactive compounds, reinforcing the importance of molecular descriptors such as molecular weight, hydrogen bonding, and lipophilicity in CA-II inhibition. While machine learning models exhibited strong predictive performance, class imbalance and lower specificity remain challenges, necessitating future improvements through resampling techniques, hyperparameter optimization, and explainability methods. Additionally, machine learning-based virtual screening should be integrated with traditional drug discovery techniques, such as molecular docking and pharmacophore modeling, to enhance predictive accuracy and reduce false positives. Future directions include deep learning approaches, hybrid machine learning-docking models, and expanding training datasets to improve model robustness and generalizability. Overall, this research highlights the feasibility of AI-driven virtual screening for accelerating the discovery of novel, potent, and selective CA-II inhibitors, paving the way for more effective glaucoma treatments.

Author Contributions: Conceptualization, T.R.N. and R.I.; methodology, T.R.N. and R.I.; software, T.R.N. and G.M.I.; validation, E.I. and R.S.; formal analysis, T.R.N. and G.M.I.; investigation, T.R.N.; resources, E.I. and G.M.I.; data curation, E.I. and R.I.; writing—original draft preparation, T.R.N. and G.M.I.; writing—review and editing, E.I., R.S. and R.I.; visualization, R.S.; 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 that support the findings of this study are available from the corresponding author upon reasonable request.

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

References

1. Artero-Castro, A., Rodriguez-Jimenez, F. J., Jendelova, P., VanderWall, K. B., Meyer, J. S., and Erceg, S. (2020). Glaucoma as a Neurodegenerative Disease Caused by Intrinsic Vulnerability Factors, *Progress in Neurobiology*, Vol. 193, 101817. doi:10.1016/j.pneurobio.2020.101817.

2. García-Llorca, A., Carta, F., Supuran, C. T., and Eysteinsson, T. (2024). Carbonic Anhydrase, Its Inhibitors and Vascular Function, *Frontiers in Molecular Biosciences*, Vol. 11. doi:10.3389/fmolb.2024.1338528.

3. Supuran, C. T., Altamimi, A. S. A., and Carta, F. (2019). Carbonic Anhydrase Inhibition and the Management of Glaucoma: A Literature and Patent Review 2013–2019, *Expert Opinion on Therapeutic Patents*, Vol. 29, No. 10, 781–792. doi:10.1080/13543776.2019.1679117.

4. Mincione, F., Nocentini, A., and Supuran, C. T. (2021). Advances in the Discovery of Novel Agents for the Treatment of Glaucoma, *Expert Opinion on Drug Discovery*, Vol. 16, No. 10, 1209–1225. doi:10.1080/17460441.2021.1922384.

5. Supuran, C. T. (2021). Emerging Role of Carbonic Anhydrase Inhibitors, *Clinical Science*, Vol. 135, No. 10, 1233–1249. doi:10.1042/CS20210040.

6. Kumar, S., Rulhania, S., Jaswal, S., and Monga, V. (2021). Recent Advances in the Medicinal Chemistry of Carbonic Anhydrase Inhibitors, *European Journal of Medicinal Chemistry*, Vol. 209, 112923. doi:10.1016/j.ejmech.2020.112923.

7. Tiwari, P. C., Pal, R., Chaudhary, M. J., and Nath, R. (2023). Artificial Intelligence Revolutionizing Drug Development: Exploring Opportunities and Challenges, *Drug Development Research*, Vol. 84, No. 8, 1652–1663. doi:10.1002/ddr.22115.

8. Sadybekov, A. V., and Katritch, V. (2023). Computational Approaches Streamlining Drug Discovery, *Nature*, Vol. 616, No. 7958, 673–685. doi:10.1038/s41586-023-05905-z.

9. Staszak, M., Staszak, K., Wieszczycka, K., Bajek, A., Roszkowski, K., and Tylkowski, B. (2022). Machine Learning in Drug Design: Use of Artificial Intelligence to Explore the Chemical Structure–Biological Activity Relationship, *WIREs Computational Molecular Science*, Vol. 12, No. 2. doi:10.1002/wcms.1568.

10. Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M., and Ahsan, M. J. (2022). Machine Learning in Drug Discovery: A Review, *Artificial Intelligence Review*, Vol. 55, No. 3, 1947–1999. doi:10.1007/s10462-021-10058-4.

11. 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.

12. Dhudum, R., Ganeshpurkar, A., and Pawar, A. (2024). Revolutionizing Drug Discovery: A Comprehensive Review of AI Applications, *Drugs and Drug Candidates*, Vol. 3, No. 1, 148–171. doi:10.3390/ddc3010009.

13. 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.

14. Sinsulpsiri, S., Nishii, Y., Xu-Xu, Q.-F., Miura, M., Wilasluck, P., Salamteh, K., Deetanya, P., Wangkanont, K., Suroengrit, A., Boonyasuppayakorn, S., Duan, L., Harada, R., Hengphasatporn, K., Shigeta, Y., Shi, L., Maitarad, P., and Rungrotmongkol, T. (2025). Unveiling the Antiviral Inhibitory Activity of Ebselen and Ebsulfur Derivatives on SARS-CoV-2 Using Machine Learning-Based QSAR, LB-PaCS-MD, and Experimental Assay, *Scientific Reports*, Vol. 15, No. 1, 6956. doi:10.1038/s41598-025-91235-1.

15. Priya, S., Tripathi, G., Singh, D. B., Jain, P., and Kumar, A. (2022). Machine Learning Approaches and Their Applications in Drug Discovery and Design, *Chemical Biology & Drug Design*, Vol. 100, No. 1, 136–153. doi:10.1111/cbdd.14057.

16. 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.

17. El Rhabori, S., Alaqarbeh, M., El Allouche, Y., Naanaai, L., El Aissouq, A., Bouachrine, M., Chtita, S., and Khalil, F. (2025). Exploring Innovative Strategies for Identifying Anti-Breast Cancer Compounds by Integrating 2D/3D-QSAR, Molecular Docking Analyses, ADMET Predictions, Molecular Dynamics Simulations, and MM-PBSA Approaches, *Journal of Molecular Structure*, Vol. 1320, 139500. doi:10.1016/j.molstruc.2024.139500.

18. Khan, S., Sarfraz, A., Prakash, O., and Khan, F. (2024). Machine Learning-Based QSAR Modeling, Molecular Docking, Dynamics Simulation Studies for Cytotoxicity Prediction in MDA-MB231 Triple-Negative Breast Cancer Cell Line, *Journal of Molecular Structure*, Vol. 1315, 138807. doi:10.1016/j.molstruc.2024.138807.

19. Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Afidh, R. P. F., and Idroes, R. (2024). An Explainable Multi-Model Stacked Classifier Approach for Predicting Hepatitis C Drug Candidates, *Sci*, Vol. 6, No. 4, 81. doi:10.3390/sci6040081.

20. Noviandy, T. R., Idroes, G. M., Maulana, A., Afidh, R. P. F., and Idroes, R. (2024). Optimizing Hepatitis C Virus Inhibitor Identification with LightGBM and Tree-structured Parzen Estimator Sampling, *Engineering, Technology & Applied Science Research*, Vol. 14, No. 6, 18810–18817. doi:10.48084/etasr.8947.

21. Winkler, D. A. (2022). The Impact of Machine Learning on Future Tuberculosis Drug Discovery, *Expert Opinion on Drug Discovery*, Vol. 17, No. 9, 925–927. doi:10.1080/17460441.2022.2108785.

22. Noviandy, T. R., Maulana, A., Irvanizam, I., Idroes, G. M., Maulydia, N. B., Tallei, T. E., Subianto, M., and Idroes, R. (2025). Interpretable Machine Learning Approach to Predict Hepatitis C Virus NS5B Inhibitor Activity Using Voting-Based LightGBM and SHAP, *Intelligent Systems with Applications*, Vol. 25, 200481. doi:10.1016/j.iswa.2025.200481.

23. 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.

24. 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.

25. Rudrapal, M., Kirboga, K. K., Abdalla, M., and Maji, S. (2024). Explainable Artificial Intelligence-Assisted Virtual Screening and Bioinformatics Approaches for Effective Bioactivity Prediction of Phenolic Cyclooxygenase-2 (COX-2) Inhibitors Using PubChem Molecular Fingerprints, *Molecular Diversity*, Vol. 28, No. 4, 2099–2118. doi:10.1007/s11030-023-10782-9.

26. Ojha, P. K., and Roy, K. (2011). Comparative QSARs for Antimalarial Endochins: Importance of Descriptor-Thinning and Noise Reduction Prior to Feature Selection, *Chemometrics and Intelligent Laboratory Systems*, Vol. 109, No. 2, 146–161. doi:10.1016/j.chemolab.2011.08.007.

27. Westad, F., and Marini, F. (2022). Variable Selection and Redundancy in Multivariate Regression Models, *Frontiers in Analytical Science*, Vol. 2. doi:10.3389/frans.2022.897605.

28. Robotti, E., and Marengo, E. (2016). *Chemometric Multivariate Tools for Candidate Biomarker Identification : LDA, PLS-DA, SIMCA, Ranking-PCA, Methods in Molecular Biology* (Vol. 1384), Humana Press. doi:10.1007/978-1-4939-3255-9.

29. Chen, X., Li, H., Tian, L., Li, Q., Luo, J., and Zhang, Y. (2020). Analysis of the Physicochemical Properties of Acaricides Based on Lipinski's Rule of Five, *Journal of Computational Biology*, Vol. 27, No. 9, 1397–1406. doi:10.1089/cmb.2019.0323.

30. Yu, T., Huang, T., Yu, L., Nantasenamat, C., Anuwongcharoen, N., Piacham, T., Ren, R., and Chiang, Y.-C. (2023). Exploring the Chemical Space of CYP17A1 Inhibitors Using Cheminformatics and Machine Learning, *Molecules*, Vol. 28, No. 4, 1679. doi:10.3390/molecules28041679.

31. Bai, Q., Su, C., Tang, W., and Li, Y. (2022). Machine Learning to Predict End Stage Kidney Disease in Chronic Kidney Disease, *Scientific Reports*, Vol. 12, No. 1, 8377. doi:10.1038/s41598-022-12316-z.

32. Noviandy, T. R., Idroes, G. M., and Hardi, I. (2024). Enhancing Loan Approval Decision-Making: An Interpretable Machine Learning Approach Using LightGBM for Digital Economy Development, *Malaysian Journal of Computing (MJOC)*, Vol. 9, No. 1, 1734–1745. doi:10.24191/mjoc.v9i1.25691.33. El Orche, A., Mamad, A., Elhamdaoui, O., Cheikh, A., El Karbane, M., and Bouatia, M. (2021). Comparison of Machine Learning Classification Methods for Determining the Geographical Origin of Raw Milk Using Vibrational Spectroscopy, *Journal of Spectroscopy*, Vol. 2021. doi:10.1155/2021/5845422.

34. Suhendra, R., Suryadi, S., Husdayanti, N., Maulana, A., Noviandy, T. R., Sasmita, N. R., Subianto, M., Earlia, N., Niode, N. J., and Idroes, R. (2023). Evaluation of Gradient Boosted Classifier in Atopic Dermatitis Severity Score Classification, *Heca Journal of Applied Sciences*, Vol. 1, No. 2, 54–61. doi:10.60084/hjas.v1i2.85.

35. Sasmita, N. R., Ramadeska, S., Kesuma, Z. M., Noviandy, T. R., Maulana, A., Khairul, M., and Suhendra, R. (2024). Decision Tree versus k-NN: A Performance Comparison for Air Quality Classification in Indonesia, *Infolitika Journal of Data Science*, Vol. 2, No. 1, 9–16. doi:10.60084/ijds.v2i1.179.

36. Rafiei, H., Khanzadeh, M., Mozaffari, S., Bostanifar, M. H., Avval, Z. M., Aalizadeh, R., and Pourbasheer, E. (2016). QSAR Study of HCV NS5B Polymerase Inhibitors Using the Genetic Algorithm-Multiple Linear Regression (GA-MLR), *EXCLI Journal*, Vol. 15, 38–53. doi:10.17179/excli2015-731.

37. Noviandy, T. R., Idroes, G. M., Mohd Fauzi, F., and Idroes, R. (2024). Application of Ensemble Machine Learning Methods for QSAR Classification of Leukotriene A4 Hydrolase Inhibitors in Drug Discovery, *Malacca Pharmaceutics*, Vol. 2, No. 2, 68–78. doi:10.60084/mp.v2i2.217.

38. 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.