International Journal of
Science Inventions Today
www.ijsit.com
ISSN 2319-5436
Research Article

DESIGN AND IMPLEMENTATION FOR PREDICTING ACTIVITIES OF ACETYLCHOLINESTERASE INHIBITORS USING MACHINE LEARNING

Bolsunbek Meimanov[1*] and Tajallah Khan[2]

[1]*School of Software Engineering, University of South China, Hengyang, Hunan, 421001, China*

[2]*Hengyang Medical College, University of South China, Hengyang, Hunan, 421001, China*

ABSTRACT

Objective: This study aims to develop effective predictive models using machine learning, specifically Support Vector Machine (SVM), k-nearest neighbor (k-NN), and Random Forest (RF), to target the acetylcholinesterase enzyme (AChE) for Alzheimer's disease treatment. The goal is to distinguish AChE inhibitors from non-inhibitors, enhancing synaptic acetylcholine levels for improved memory and mental acuity.

Methods: Utilizing SVM, k-NN, and RF algorithms, we trained models and evaluated them with an external dataset and 10-fold cross-validation. Descriptor analysis identified key features, including MACCS keys and fingerprint properties. Models built from fingerprint data using the RF technique outperformed others.

Results: The RF-based models demonstrated superior accuracy, with the best model achieving 85.38 percent accuracy on test data. Important identifiers influencing inhibitory activities were identified, emphasizing the efficacy of ensemble learning models, particularly those using random forest techniques.

Conclusion: Ensemble learning models, especially random forest-based ones, showed promise in predicting AChE inhibitors from diverse datasets. The blended model, AChEI-EL, combined top-performing models, serving as a potent tool for identifying potential AChE inhibitors. The approach holds potential for Alzheimer's disease research, supported by the development of a web-based academic prediction tool using the most robust model.

Keywords: Acetylcholinesterase (AChE), Support Vector Machine (SVM), K-nearest neighbor (k-NN), Random Forest (RF), Machine learning.

INTRODUCTION

Alzheimer's disease (AD), marked by amyloid beta cascade protein alterations and tau protein hyperphosphorylation, is the leading cause of clinical dementia affecting nearly 10% of seniors over 65. With a current global impact on 35 million individuals, projections estimate 115 million cases by 2050 [1]. Early signs include memory loss, cognitive decline, and language impairment. While AD has complex origins, risk factors include age, depression, high blood pressure, and genetics. The cholinergic hypothesis links AD to impaired acetylcholine synthesis [2]. Acetylcholinesterase (AChE) inhibition, enhancing acetylcholine levels, is a symptomatic treatment for cognitive improvement and AD prevention. Only four AChE inhibitors are currently approved for AD treatment.

Traditional discovery of AChE inhibitors is costly, time-consuming, and labor-intensive. Computational techniques, such as quantitative structure-activity relationships (QSAR), molecular docking, and drug repurposing, offer efficient alternatives. Machine learning, explored in limited studies, presents a promising approach. Using a larger dataset, this work employs SVM, k-NN, and RF methods, incorporating recursive feature elimination and Boruta for dimensionality reduction. Descriptor analysis identifies crucial features, and the best model achieves 85.38% accuracy on the test set. PaDEL and rank-obtained descriptors and fingerprints contribute to model creation [3]. This study advances predictive categorization models, emphasizing the significance of machine learning in identifying potential AChE inhibitors.

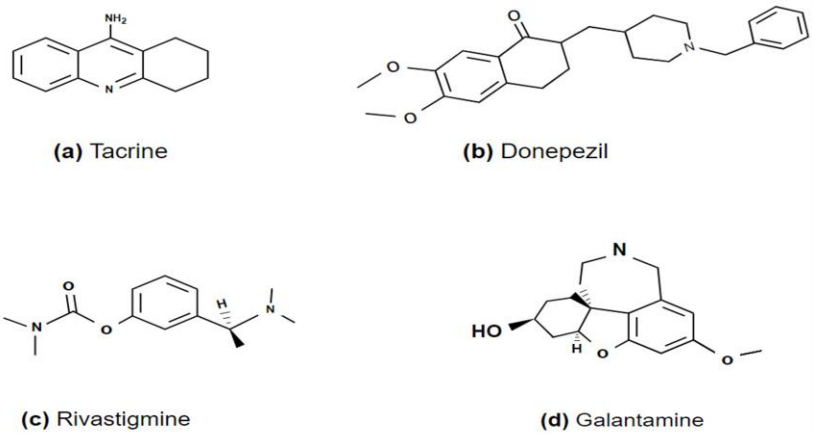

Figure 1: Alzheimer's disease acetyl-cholinesterase inhibitors approved by FDA.

Acetylcholinesterase (AChE, EC 3.1.1.7) is a serine protease responsible for catalyzing the breakdown of acetylcholine in neurotransmitters, playing a vital role in terminating synaptic transmission. Found predominantly in muscle nerve connections and choline synapses in the brain, one AChE molecule can catalyze the breakdown of acetylcholine at a remarkable rate of 25,000 molecules per second. The enzyme's active pocket contains two binding sites, with the residues Ser203, His447, and Glu334 forming a deep "canyon" leading to the primary catalytic active site, and peripheral anionic areas (PAS) located near specific residues on the enzyme's surface.

In the context of Alzheimer's disease (AD), which affects one in every 85 individuals by 2050, the

"cholinergic hypothesis" links mental impairment to decreased acetylcholine production. A promising therapeutic target for AD treatment is AChE. FDA-approved acetylcholinesterase inhibitors (AChEIs) for AD include Tacrine, the first approved AChEI, which exhibits hepatotoxicity; Donepezil, a reversible central-acting AChEI effective in mild and moderate AD; Rivastigmine, a physostigmine derivative treating mild and moderate AD and Parkinson's disease-related dementia; and Galantamine, a reversible AChEI synthesized from snowdrop bulbs, effectively treating mild to moderate Alzheimer's disease and other forms of memory loss. Each AChEI has specific indications, benefits, and potential side effects, contributing to the therapeutic landscape for Alzheimer's disease.

MATERIALS AND METHODS

Dataset Acquisition:

A comprehensive dataset comprising 5,692 molecules was utilized for the investigation, focusing on their potential to inhibit human acetylcholinesterase (AChE). The dataset was curated from the BindingDB database, a repository housing binding affinities for various protein-ligand complexes.

Refinement Procedures:

Several refinement steps were implemented to ensure dataset quality. Initial measures involved eliminating redundant molecules, ensuring each was considered only once. Prohibitions against salt forms, heavy metals, and fragments further refined the dataset. Following these procedures, a refined set of 4,140 distinct molecules was subjected to in-depth analysis. A specific criterion, the half-inhibitory concentration (IC50) cutoff value, was set at 1,000 nM. Molecules were categorized as inhibitors if their IC50 was less than or equal to 1,000 nM, while those with higher values were classified as non-inhibitors.

Inhibitory Activity Classification:

Inhibitory activity against AChE was assessed for the final dataset, identifying 1,862 molecules as inhibitors and 2,278 as non-inhibitors based on the established IC50 cutoff. This categorization is essential for distinguishing compounds affecting the key enzyme in acetylcholine degradation. The resulting dataset, comprising 1,862 inhibitor compounds and 2,278 non-inhibitors, serves as a valuable resource for researchers in drug discovery and development.

Molecular Standardization:

To standardize the dataset for further analysis, MolVS, an open-source toolkit based on the RDKit chemistry framework, was employed to create simplified molecular-input line-entry system (SMILES) descriptors. Duplicate compounds' IC50 values were averaged, and those with a large standard deviation (10 mM) were excluded. The compounds were classified as blockers (hERG-positive) if their IC50 was less than or equal to 10 M and as non-blockers (hERG-negative) if it was greater than or equal to 10 M.

Biological Implications:

Understanding the inhibitory impact of compounds against AChE holds significance, particularly for

conditions like Alzheimer's associated with cholinergic system dysfunction. This dataset provides insights into the structure-activity relationship, facilitating the development of treatments targeting AChE and contributing to advancements in therapeutic research.

The dataset employed in this study originated from the Binding DB database, encompassing 5,692 compounds. Following refinement stages that eliminated duplicates, salt forms, heavy metals, and fragments, a curated set of 4,140 distinct molecules was obtained. Utilizing an IC50 cutoff value of 1,000 nM, 1,862 compounds were identified as inhibitors, while 2,278 were classified as inactive. This dataset stands as a valuable repository for researchers exploring acetylcholinesterase inhibition and the development of novel pharmaceutical interventions.

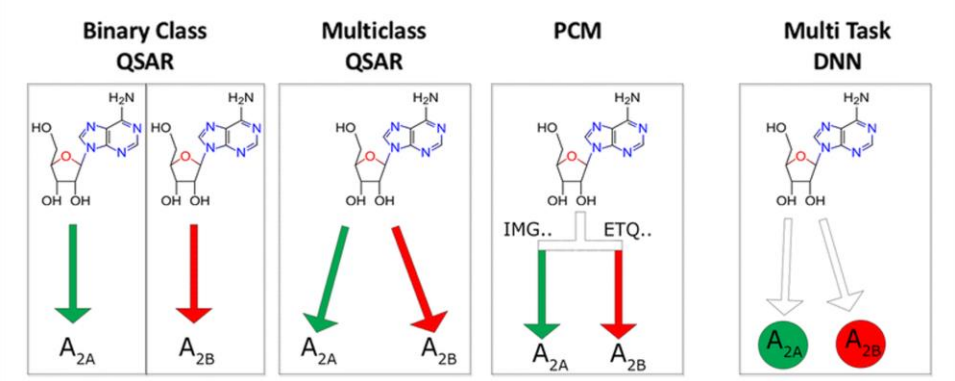

Figure 2: The adenosine ligand is more active (labeled as 'active') on the adenosine A2A receptor than the A2B receptor (marked as 'inactive,' using PChEMBL > 6.5 as a cutoff), demonstrating the differences in the approaches for modeling bioactivity data.

	Organism	Pref_name	Score	Species_group_flag	Target_chembl_id	Target_type	Tax_id
1	Homosapiens	Acetylcholinesterase	27	FALSE	CHEMBL220	SINGLE PROTEIN	9606
2	Homosapiens	Cholinesterases; ACHE & BCHE	27	FALSE	CHEMBL2095233	SELECTIVITY GROUP	9606
3	Drosophila melanogaster	Acetylcholinesterase	18	FALSE	CHEMBL2242744	SINGLE PROTEIN	7227
4	Bemisia tabaci	AChE2	16	FALSE	CHEMBL2366409	SINGLE PROTEIN	7038
5	Leptinotarsa decemlineata	Acetylcholinesterase	16	FALSE	CHEMBL2366490	SINGLE PROTEIN	7539
6	Torpedo californica	Acetylcholinesterase	15	FALSE	CHEMBL4780	SINGLE PROTEIN	7787
7	Mus musculus	Acetylcholinesterase	15	FALSE	CHEMBL3198	SINGLE PROTEIN	10090
8	Rattus norvegicus	Acetyicholinesterase	15	FALSE	CHEMBL3199	SINGLE PROTEIN	10116
9	Electrophorus electricus	Acetylcholinesterase	15	FALSE	CHEMBL4087	SINGLE PROTEIN	8005

Table 1: Dataset for acetylcholinesterase inhibition

Descriptor Selection:

In the realm of structure-activity relationship (SAR) and quantitative structure-activity relationship (QSAR) studies, molecular descriptors play a pivotal role in representing the physical attributes of molecules. For this analysis, a total of 1,444 descriptors (combining 1D and 2D descriptors) were computed using PaDEL (version 2.21). To avoid complexities associated with developing 3D descriptors, the focus remained on 1D and 2D features. Additionally, two sets of fingerprints, totaling 245 bits, were generated using the rcdk package. The dataset was then divided proportionally into training, test, and external sets. Normalization was achieved using the 'range' approach to bring values within the 0 to 1 range. Zero-variance descriptors were eliminated from the dataset, as they did not contribute to the identification of desired compound features.

Descriptor Calculation:

To address multicollinearity and facilitate feature selection, Recursive Feature Elimination (RFE) and Boruta were employed. RFE assigns weights to input features based on their significance, building models iteratively with a decreasing set of elements. Boruta, wrapping random forest, prioritizes features by

determining their relative value. The dataset preprocessing, including constant feature removal, elimination of highly correlated features, and normalization for feature selection, was conducted using the training set to prevent information leakage.

Descriptor Dataset Refinement:

The process involved calculating descriptors for the dataset using the PaDEL descriptor toolkit, resulting in datasets labeled D1a and D2a, each with specific features. Variance and correlation filters were applied to generate refined datasets D1b and D2b. Feature scaling was performed to produce datasets D1c, D2c, D1d, and D2d. A meticulous approach was taken to ensure data cleaning and feature selection were conducted with precision.

Molecular Descriptor Utility:

In the broader context, the computed molecular descriptors contribute to understanding structure-activity relationships and quantitative aspects in drug discovery. Leveraging these descriptors aids researchers in elucidating ligand-protein interactions across the proteome. The dataset, sourced from the Binding database, provides a valuable resource for further investigations into acetylcholinesterase inhibition and the development of novel therapeutic interventions.

Machine Learning Models:

The study further involved the development of two sets of machine learning models based on Support Vector Classification (SVC). The first set aimed to predict active substances with IC50 values below 1000 nM, utilizing datasets D1a-D1d. The second set identified compounds with estimated IC50 values below 100 nM using datasets D2a-D2d. Model fitting and hyperparameter optimization were performed on the training set, with subsequent evaluation on the test set to ensure robustness and prevent overfitting. The regularization parameter and gamma were tuned for optimal performance. The comprehensive approach undertaken in this study aligns with contemporary practices in computational drug discovery and contributes to advancing research in the field.

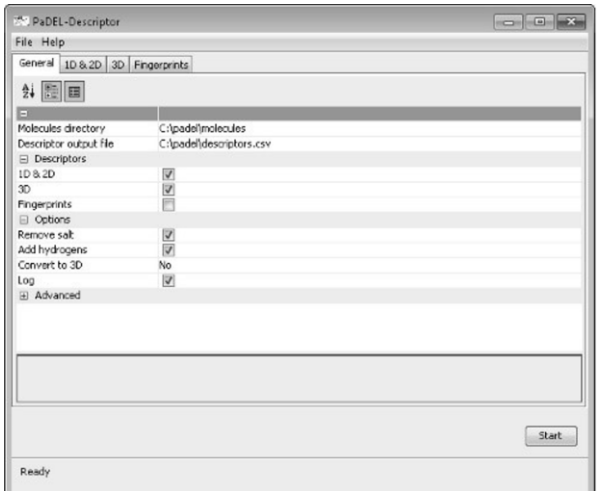

Figure 3: GUI for PaDEL-Descriptor.

Upon restarting the program, users have the flexibility to configure it either automatically or manually based on predefined settings stored in the configuration file. The preferences for selected fingerprints and descriptors can be conveniently saved to an XML file, streamlining future selections. The implementation leverages the Apache Commons CLI library to create a command line interface, as depicted in Figure 3. This interface facilitates program execution in computer clusters using a software task scheduler. The graphical user interface (GUI) generates both a configuration file and an XML file (as illustrated in Figure 4 and Figure 5, respectively), enabling the command line interface to efficiently perform tasks such as adjusting settings and specifying the types of descriptors and fingerprints for calculation.

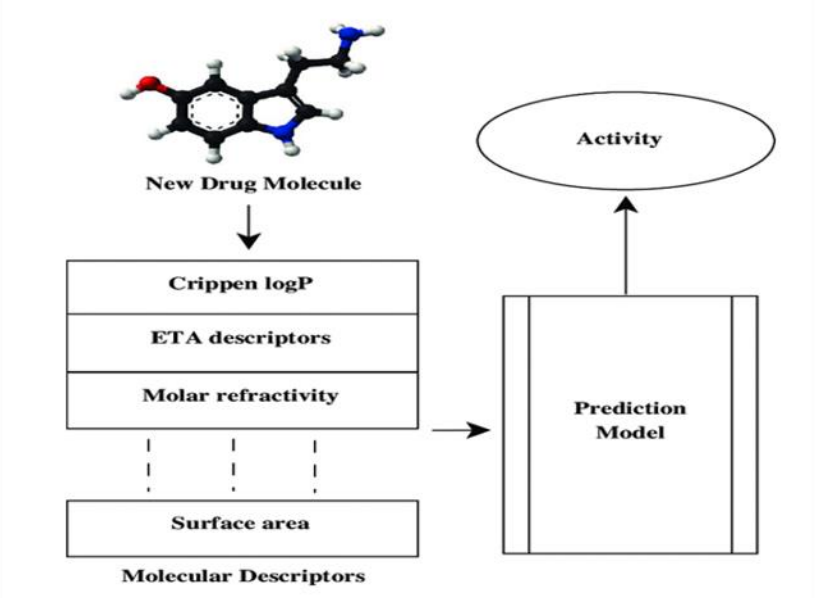

Figure 4: PaDEL-Description

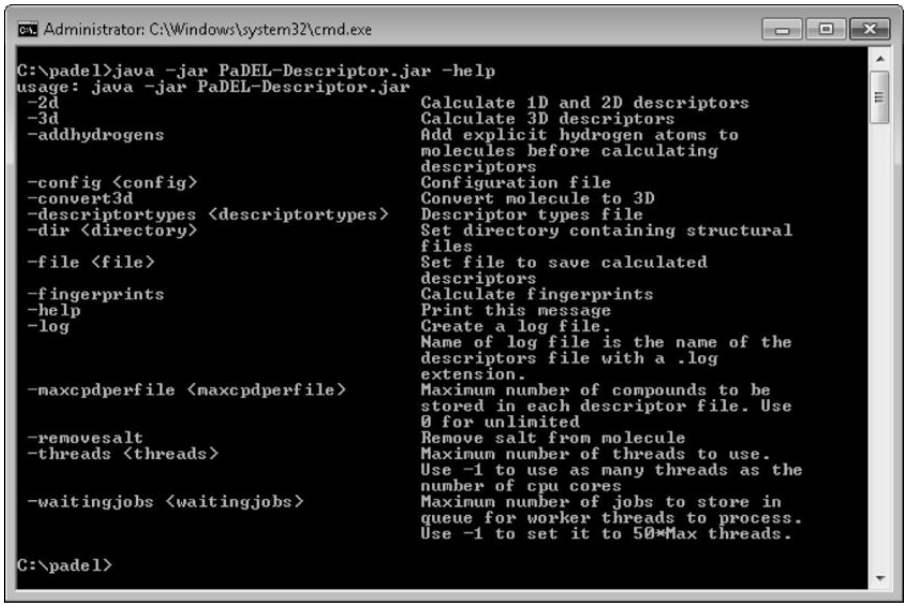

Figure 5: Command line interface for PaDEL-Descriptor

Lipinski Descriptor and Rule of Five:

Lipinski's rule of five, a foundational concept in the pharmaceutical industry introduced by Pfizer in 1997, plays a crucial role in defining "drug-likeness." This rule guides the assessment of a molecule's bioavailability when administered orally, emphasizing five key physiochemical parameters: molecular weight, lipophilicity, polar surface area, hydrogen bonding, and charge [4]. The Rule of Five stipulates that an orally active drug should not simultaneously meet more than one of the following conditions: having over 5 hydrogen bond donors (encompassing bonds between nitrogen and oxygen) and a hydrogen-bond acceptor count exceeding 10 (involving all nitrogen or oxygen atoms).

Considerations for Ligands Binding Quadruplexes:

In the context of ligands binding to quadruplexes, specific considerations become paramount. First, reducing the ligand's molecular mass is preferred, with compounds approaching the Lipinski limit of 500 Daltons being advantageous for cellular uptake. Notably, instances exist where much larger molecules have successfully entered cells. Second, enhancing absorption and permeability is crucial, especially when targeting telomeric or promoter quadruplexes. Experimental determinations of log P and pKa are preferred, and various programs, such as those available on www.chemaxon.com and www.molinspiration.com, can provide valuable estimations during motif-selection and lead-optimization phases. The log P values of well-characterized quadruplex ligands, including those with reported in vivo anti-cancer efficacy, are detailed in Table 2. It's noteworthy that none of these compounds strictly adheres to the Rule of Five. For instance, RHSP4, a polycyclic acridine molecule, exhibits efficient uptake into the nuclei of diverse cancer cells despite a relatively weak log P score attributed to its low molecular weight [5]. Similarly, the large polycyclic ligand BRACO-19 (3+) and the porphyrin TMPyP4 (4+) demonstrate facile diffusion into small experimental tumors, displaying anti-cancer effects and uptake into cancer cells in culture. While their high polarity suggests a potential active transport mechanism aiding uptake, their limited penetration into large tumor masses is hindered by their substantial charge. Therefore, optimizing the lipophilic/lipophobic balance within a compound and monitoring log P values becomes imperative for effective drug design [5].

Compound	Log P	Mol. wt (daltons)	Polar surface area (Å^2)	Volume (Å^3)	Reference
BRACO-19	1.86	597	96	575	Read *et al.*, 2001
AS1410	1.31	604	93	556	Martins *et al.*, 2007
Telomestatin	1.76	582	194	440	Kim *et al.*, 2002
Diaryl polyamide derivative	0.94	554	122	500	Rahman *et al.*, 2009
RHPS4	−0.04	333	9	287	Cookson *et al.*, 2005
6-Substituted indeno isoquinoline derivative	−0.85	361	42	340	Bejugam *et al.*, 2010
Bis-phenyl ethynyl-amide derivative	−0.99	596	44	585	Dash *et al.*, 2008
Quarfloxin	−0.28	605	97	536	Drygin *et al.*, 2009

Table 2: Selected physio-chemical properties for some representative G-quadruplex ligands, calculated using the facilities on the www.molinspiration.com website

Support Vector Machine (SVM):

In the realm of toxicological prediction and computational approaches, the Support Vector Machine (SVM) stands out as a cutting-edge innovation in machine learning. SVM is particularly valuable in forecasting chemical properties and understanding structure-activity relationships. Unlike traditional linear methods, SVM excels in capturing subtle nonlinear correlations within experimental data, crucial in the complex and dynamic context of living systems. The SVM's unique ability to create models with numerous molecular property descriptors and a small set of training data enhances its applicability, preventing overfitting while maintaining generalization performance [6-8].

k-Nearest Neighbor (k-NN):

Addressing the challenge of predicting the danger of industrial byproducts, the k-Nearest Neighbor (k-NN) algorithm emerges as an innovative machine learning approach. Leveraging proximity-based

classification, k-NN proves valuable in assessing the harmful potential of compounds. With its adaptability to varying datasets and high generalization performance, k-NN finds applications in diverse domains, including handwritten digit recognition, drug discovery, and disease diagnostics. Notably, its ability to create models with minimal experimental data, while avoiding overfitting, enhances its effectiveness in predictive tasks [9].

Random Forest (RF):

In the landscape of ensemble learning, Random Forest (RF) emerges as a potent method for classification and nonlinear regression models. Combining the strength of decision trees, RF mitigates the risk of overfitting inherent in individual trees. Through random feature selection and data instance sampling, RF introduces diversity, ensuring robustness against overfitting. This approach is particularly beneficial in handling large datasets and capturing complex interactions among attributes, making it well-suited for predicting molecular properties. RF's adaptability and reliability contribute to its prominence in pharmaceutical applications, where predictive models play a vital role [10].

Model Validation:

Model validation in this study revolves around a meticulous 10-fold cross-validation procedure to evaluate machine learning models predicting inhibitory activity against human acetylcholinesterase (AChE). Cross-validation is a widely embraced practice for its effectiveness in enhancing model performance by testing on diverse data samples. The assessment metrics include sensitivity, specificity, accuracy, F1 score, and Matthews' Correlation Coefficient (MCC). These metrics provide a comprehensive evaluation of the models' accuracy and effectiveness in distinguishing between inhibitors and non-inhibitors. The emphasis on such validation practices ensures the reliability and applicability of the developed machine learning models [11].

RESULTS

In the study, three sets of features, denoted as Sets A, B, and C, were utilized for model development and evaluation. Set A comprised 30 features identified through Recursive Feature Elimination (RFE), while Set B included 30 features determined by the Boruta software. Set C consisted of 116-bit fingerprints. Models were trained using Random Forest (RF), k-Nearest Neighbors (k-NN), and Support Vector Machine (SVM) algorithms, optimizing their respective parameters. Comparative analyses revealed that the RF model trained with Set C features demonstrated superior performance, exhibiting balanced sensitivity and specificity values across 10-fold cross-validation.

Furthermore, external dataset predictions were conducted to assess the models' generalizability. The RF model trained with Set C characteristics consistently outperformed SVM (Set A) and RF (Set B) models, showcasing its robustness and efficacy in predicting novel instances. These results were reinforced by metrics such as accuracy, sensitivity, specificity, and Matthews Correlation Coefficient (MCC). The model's superior performance, especially in distinguishing between positive and negative instances, was supported by Receiver Operating Characteristic (ROC) curve analysis with a high Area Under the Curve (AUC) value of

0.917.

The comprehensive evaluation of different feature sets and machine learning algorithms emphasized the efficacy of the RF model trained with Set C fingerprints. Its ability to generalize well to external datasets makes it a promising and reliable choice for predicting inhibitory activity against acetylcholinesterase, showcasing the potential of the developed model for practical applications in drug discovery and toxicity prediction.

The study conducted an extensive analysis of Acetylcholinesterase inhibitors (AChEI) using a dataset sourced from the ChEMBL Database, encompassing more than two million chemicals and over 76,000 records. The curated dataset, which involved the construction of a machine learning model employing scikit-learn, was subsequently integrated into a Streamlit web application for user interaction. The web application allows users to input molecular properties through widgets, and the trained machine learning model predicts the inhibitory activity of AChE inhibitors based on the provided information.

The execution process involved preparing the ChEMBL bioactivity data, implementing necessary data preprocessing steps, and using machine learning classification models and molecular docking calculations to identify potential AChE inhibitors from the SistematX database. Notably, the study addressed the challenges associated with ADMET prediction and emphasized the importance of structural optimization to enhance ADMET profiles.

Furthermore, the research delved into the intricate process of web application development, emphasizing the significance of flexible and adaptive methodologies in the Ecuadorian market. The systematic literature review (SLR) compared various development approaches, shedding light on the prevailing trends and challenges faced by developers. The study incorporated insights from the SistematX database, contributing valuable information on secondary metabolites for chemoinformatics and computer-aided drug design.

In the context of web application development approaches in the Ecuadorian market, the results revealed a predominant preference for hybrid methods or procedures informed by experience, as opposed to adherence to specific methodologies such as OOHDM, RMM, WebML, SOHDM, OOHD, UWE, or Hera.

The research provided a comprehensive overview of AChEI analysis, web application development, and development methodologies in the Ecuadorian market. The utilization of machine learning techniques, the creation of a user-friendly web application, and insights from the SistematX database collectively contribute to advancing the understanding of AChE inhibitors and their potential applications in drug discovery.

CONCLUSION

Model Construction and Performance:

- Utilized PaDEL-calculated descriptors and rank-calculated fingerprints for model construction.
- Employed machine learning algorithms (SVM, RF, k-NN) and feature selection techniques (Boruta, RFE) for optimal results.

- Random forest models based on fingerprint data demonstrated high performance.

Dataset Size and Comparison:

- Significantly larger dataset used for model training compared to previously published models for AChE inhibitor categorization.
- Highlighted the impact of dataset size on model robustness and accuracy.

Hybrid Approach for AChE Inhibitor Prediction:

- Introduced a hybrid approach combining atomistic simulations with machine learning for predicting effective AChE inhibitors.
- Atomistic simulations (molecular docking, steered-molecular dynamics) provided confirmation of binding affinities and identified critical residues.
- Anticipated the reliability of predicted inhibitors for further experimental studies.

Challenges in Drug Design and Evaluation:

- Addressed challenges in accurately predicting free energy perturbation caused by experimental inhibitors' binding.
- Emphasized the significance of cost-effective and accurate methods like the rapid ligand pulling (FPL) approach with GROMOS force field (FF) for gauging compound affinity.

Classification Threshold Determination:

- Established a classification threshold of IC50 values less than 100 nM for active AChE and BACE1 inhibitors.
- Employed k-mean cluster analysis for creating diverse training and test sets.

Rebalancing Consideration:

- Observed that applying a rebalancing approach to the final data had a statistically insignificant impact on classifier efficacy.
- Demonstrated the importance of dataset characteristics in determining effective modeling strategies.

In conclusion, the study not only presented robust models for AChE inhibitor prediction but also introduced innovative hybrid methodologies and addressed challenges in drug design and evaluation. The findings contribute valuable insights to the field, offering potential advancements in Alzheimer's disease treatment and drug discovery processes.

REFERENCES

1. Hu, Y. and J. Bajorath, Growth of ligand-target interaction data in ChEMBL is associated with increasing and activity measurement-dependent compound promiscuity. J Chem Inf Model, 2012. 52(10): p. 2550-8.
2. Jasial, S., Y. Hu, and J. Bajorath, Assessing the Growth of Bioactive Compounds and Scaffolds over Time: Implications for Lead Discovery and Scaffold Hopping. Journal of Chemical Information and

Modeling, 2016. 56.

3. Yuan, H., et al., Multitask learning improves prediction of cancer drug sensitivity. Scientific Reports, 2016. 6.
4. Ertl, P., B. Rohde, and P. Selzer, Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem, 2000. 43(20): p. 3714-7.
5. H. Yu, J.Y., W. Wang, J. Han, Computer Society Bioinformatics conference. IEEE, 2003. 7(2): p. 502-508.
6. Jacob, L., et al., Virtual screening of GPCRs: An in silico chemogenomics approach. BMC bioinformatics, 2008. 9: p. 363.
7. Broccatelli, F., et al., A novel approach for predicting P-glycoprotein (ABCB1) inhibition using molecular interaction fields. J Med Chem, 2011. 54(6): p. 1740-51.
8. Roy, K.K., A. Dixit, and A.K. Saxena, An investigation of structurally diverse carbamates for acetylcholinesterase (AChE) inhibition using 3D-QSAR analysis. J Mol Graph Model, 2008. 27(2): p. 197-208.
9. al., M.R.e.a.M.Y.M.e.a.E.H.R.e., Medicinal Chemistry. Journal of Medicinal Chemistry, 2004. 7(1): p. 656-661.
10. al., K.H.M.R.e., Chemical Information and Computer Science. Journal of Medicinal Chemistry 1999. 6(10): p. 961-969.
11. al., M.D.G.e., Expert Opinion on Investigational Drugs. Vol. 7. 1999.