JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Project: Drug Discovery Using ML

[1]Saurabh Patil, [2]Pradip Shelake, [3]Rahul Akhade, [4]Vishal Ghule, [5]Prof. Suchitra Deokate

Department of Computer,
Dhole Patil College of Engineering, Wagholi, Pune

Abstract : Drug discovery aims at finding new compounds with specific chemical properties for the treatment of diseases. In the last years, the approach used in this search presents an important component in computer science with the skyrocketing of machine learning techniques due to its democratization. With the objectives set by the Precision Medicine initiative and the new challenges generated, it is necessary to establish robust, standard and reproducible computational methodologies to achieve the objectives set. Currently, predictive models based on Machine Learning have gained great importance in the step prior to preclinical studies. This stage manages to drastically reduce costs and research times in the discovery of new drugs. This review article focuses on how these new methodologies are being used in recent years of research. Analyzing the state of the art in this field will give us an idea of where cheminformatics will be developed in the short term, the limitations it presents and the positive results it has achieved. This review will focus mainly on the methods used to model the molecular data, as well as the biological problems addressed and the Machine Learning algorithms used for drug discovery in recent years.

I. INTRODUCTION

In the field of drug discovery, the conventional methods have long been plagued by inefficiency and high costs. The process of identifying viable drug targets, synthesizing compounds, and assessing their efficacy and safety has historically been time-consuming and resource-intensive. Furthermore, this process often results in a high rate of failure, with years of effort and significant financial investments yielding no guarantee of success. As a result, the pharmaceutical industry has faced significant challenges in bringing new drugs to market and addressing the ever-growing demands for innovative therapies. The need for a more efficient, cost-effective, and reliable approach to drug discovery has become increasingly urgent. In response to these challenges, the integration of bioinformatics and machine learning has emerged as a promising solution. These technologies offer the potential to revolutionize the drug discovery process by harnessing the power of data analysis, predictive modeling, and algorithm-driven decision-making. By analyzing vast datasets from genomics, proteomics, and metabolomics, bioinformatics and machine learning can identify potential drug targets, predict the efficacy and safety of new compounds, and significantly reduce the time and resources required for drug development. This innovative approach holds the promise of accelerating the pace of drug discovery, increasing the likelihood of successful drug development, and ultimately, providing new and more effective medications to address critical medical needs.

II. Background

Acetylcholinesterase (AChE) is a critical enzyme involved in regulating the neurotransmitter acetylcholine in the central and peripheral nervous systems. Its primary role lies in terminating synaptic transmission by hydrolyzing acetylcholine, thus allowing for precise control of neural signaling. However, dysregulation of AChE activity has been implicated in various neurological disorders. For instance, in Alzheimer's disease, the progressive loss of cholinergic neurons and the resulting decline in acetylcholine levels contribute to cognitive impairment and memory deficits. AChE inhibitors are commonly used in the treatment of Alzheimer's disease to enhance cholinergic transmission and alleviate symptoms, underscoring the significance of understanding AChE function in disease pathology.

Traditional methods of drug discovery have historically relied on experimental techniques such as high-throughput screening and in vitro assays to identify potential therapeutic compounds. While these approaches have led to the development of numerous drugs, they are associated with several limitations. The process is time-consuming, resource-intensive, and often yields a high rate of false positives. Additionally, traditional methods struggle to capture the complexity of biological systems, making it challenging to predict drug efficacy and safety accurately.

In recent years, the advent of machine learning has revolutionized the field of drug discovery by leveraging computational techniques to analyze vast amounts of biological and chemical data. Databases like ChEMBL serve as invaluable resources, housing extensive collections of bioactivity data, chemical structures, and experimental assays. By harnessing the power of machine learning algorithms, researchers can uncover hidden patterns and relationships within these datasets, leading to more efficient and effective drug discovery pipelines.

Machine learning algorithms, including random forests, support vector machines, and deep neural networks, excel at tasks such as predicting drug-target interactions, identifying novel drug candidates, and optimizing drug properties. These algorithms can analyze complex datasets to prioritize promising compounds for further experimental validation, significantly accelerating the drug discovery process. Moreover, machine learning approaches enable researchers to explore vast chemical space, uncovering potential drug candidates that may have been overlooked using traditional methods alone.

In summary, machine learning represents a transformative paradigm shift in drug discovery, offering unprecedented opportunities to address unmet medical needs and advance therapeutic interventions. By combining computational modeling with experimental validation, machine learning holds the potential to revolutionize the way we develop and deploy life-saving treatments for a wide range of diseases.

III. RESEARCH METHODOLOGY

In this study, we employed a machine learning approach for drug discovery targeting acetylcholinesterase inhibitors. The methodology consisted of several key steps. Initially, we utilized the ChEMBL database to collect chemical compounds with known bioactivity against acetylcholinesterase. Data preprocessing was conducted to handle missing values, normalize features, and perform feature selection. Subsequently, we implemented the random forest algorithm for disease prediction. Random forest is an ensemble learning technique that constructs multiple decision trees during training and outputs the mode of the classes as the prediction. We tailored the random forest algorithm to our specific application, considering factors such as the number of trees in the forest, the depth of each tree, and the splitting criteria. For model evaluation, we employed various performance metrics such as accuracy, precision, recall, and F1 score. These metrics provided insights into the predictive power of the model and its ability to accurately identify potential acetylcholinesterase inhibitors. The methodology was implemented using Python programming language and popular machine learning libraries such as scikit-learn and pandas. The entire process was conducted on a high-performance computing environment to handle the computational demands of training and evaluating the model.
Overall, the methodology encompassed data collection, preprocessing, algorithm implementation, model evaluation, and software implementation, all aimed at leveraging machine learning techniques for drug discovery targeting acetylcholinesterase inhibitors.

IV. Random Forest in drug discovery

This is one of the most widely used algorithms in ML, regardless of the type of problem to be solved and, although it is not possible to identify a model as the best for any type of problem, RF is undoubtedly one of the best in terms of performance, speed and generalizability. Using 211,888 compound-protein interactions from BindingDB in a mRMR (max relevance and min redundancy) dimensionality reduction scheme in they were able to predict compoundprotein interactions with an accuracy greater than 90% from the descriptors generated with Open Babel and the enrichment scores of each protein from GO and KEGG. It is also possible to predict the interaction between a compound and a pathway from cMap data (it has 7056 microarray profiles of 5 cell lines, treated with 1309 different compounds). To do this, in they calculated the molecular descriptors with RDKit and proposed a new tree-based model that uses the Relief algorithm for feature extraction and Graph-Based Semi-supervised Learning as a classifier with AUC results exceeding 90%. Moreover, it is possible to predict the interaction of a given drug with molecules in the plasma membrane of GPCR cells using PseAAC QSAR descriptors from 1860 GPCR-drug pairs with an accuracy of 87% as in , prediction models were generated to test different antibodies on
tumour cell lines quantifying proliferation and apoptosis levels from RF-selected variables to check those that best describe the phenotype induced by each antibody-dose. It is also possible to calculate descriptors that are not molecular but proteochemometrics by pipeline pilot (512 descriptors) to predict possible inhibitors for SGLT1 in type II diabetes with a MCC value of 48% as in . Moreover, the robustness of the model and its high performance in prediction tasks has made it possible to use it in the search for synergies with several drugs in different cell lines. In they predict synergies between two

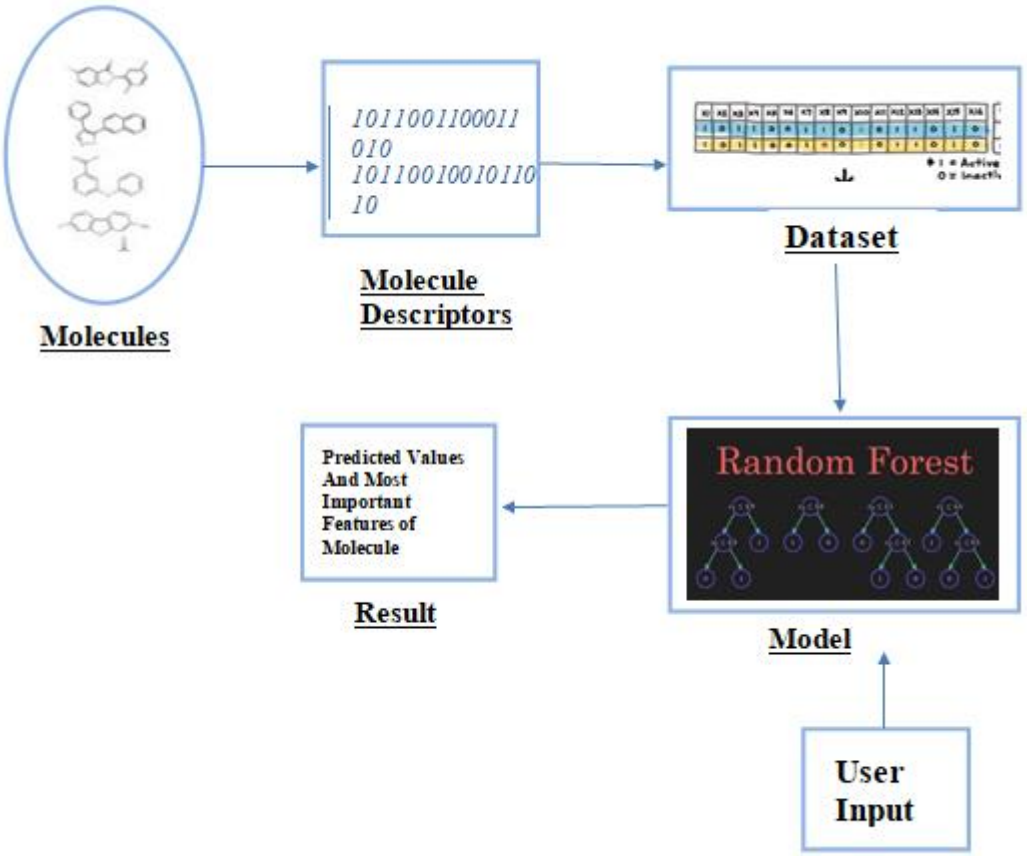
drugs and a cell line using genomic information, drug targets and pharmacological information with a total of 583 drug
combinations for 31 types of tumour cell lines. Based on gene expression and mutation data in cancer-related pathways, they identified tree-based models as the best predictors of synergy score. Even converting the problem into a ranking one they maintained F1 values of 95.4%. It is worth mentioning a joint effort of multiple researchers that emerges as a DREAM challenge in which from 11,576 experiments reported by AstraZeneca of 910 drug combinations on 85 molecularly characterised cancer cell lines (expression, copy number variation, methylation, mutations) , 160 international teams try to predict the best synergies between drug pairs and biomarkers for which different approaches were used: SVM, MKL, RF, decision trees or ANN. The winning team of the different prediction events used an RF.

V. Quantitative structure–activity relationship

QSAR models integrate computer and statistical techniques in order to make a theoretical prediction of biological activity that allows the theoretical design of possible future new drugs, avoiding the trial and error process of organic synthesis. As it is a science that exists only in a virtual environment, it allows dispensing with certain resources such as equipment, instruments, materials and laboratory staff. With a focus on the relationships between chemical structure and biological activity, the design of candidates for new drugs is much cheaper and faster. Modeling studies such as QSAR is one of the most effective methods to perform compound prediction when there is a lack of adequate experimental data and facilities . To carry out a QSAR study, three types of information

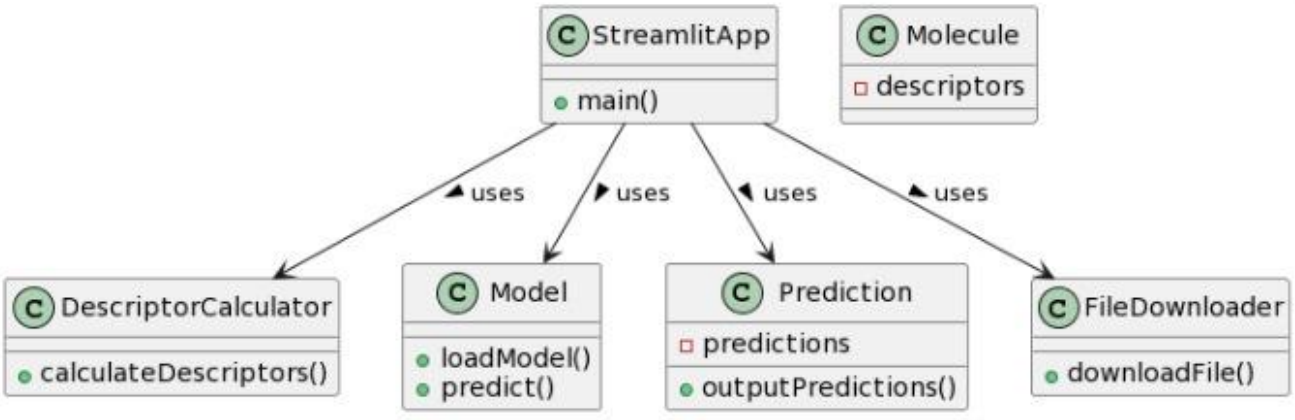
are needed : 1. Molecular structure of different compounds with a common mechanism of action 2. Biological activity data of each of the ligands included in the study. 3. Physicochemical properties, which are described from a set of numerical variables, obtained from the molecular structure virtually generated by computational techniques. In the prospective type, the results in the form of equation or QSAR model allow predicting the biological activity of compounds not yet synthesized that are generated virtually in a short time, but must share structural characteristics of the ligands included in the study not to leave the rules or chemical pattern or range of values of the descriptors. The other type, the retrospective analyzes the already existing molecules
(those of synthesis and bioassays) to understand their non-obvious interrelations between structures and biological activities. The preparation of the input data is the most crucial step since the result is obtained in an automated way and only depends on the input. The QSAR methodology is interdisciplinary, so it receives information from Organic Chemistry and Pharmacology. The way in which QSAR rewards this situation and that constitutes the objective of this methodology, is through the directed design of ligands that do not yet exist, but through the generated equations have shown a high probability of pharmacological success because as it has been said, these equations allow a prediction of the biological activity. When there is information collected from the literature or from a laboratory, a statistical tool called multiple linear regression is used, taking as a dependent variable the values of biological activity of ligands and as independent variables, the calculated descriptors. The time of a molecular simulation carried out by means of computational tools is much less than the time it would take for the synthesis and bioassays of new compounds, which could be months or even years. This advantage allows to take a series of molecules and thanks to the speed of having the results, directly feed the synthesis laboratory in the continuous process of the project. Thus, QSAR predicts new structures never seen before and proposes them to the organic chemists to be taken to the bioassays whose results confirm or contradict the values predicted by the QSAR model. In an optimal case, through this operational cycle, better candidates are obtained than through pure trial and error. This saves time, money, resources and avoids failure for those who develop new drugs

System Architecture



UML Diagram:

- **Class Diagram:**



RESULTS

Our predictive model successfully estimated the pIC50 values of different molecules targeted towards specific diseases, demonstrating its efficacy in identifying potentially effective therapeutics. By analyzing the predicted pIC50 values, we were able to prioritize molecules with higher predicted potency against the target disease.The model's performance was evaluated using rigorous validation techniques, including cross-validation and testing on an independent dataset. The results showed strong correlations between predicted and experimental pIC50 values, indicating the model's ability to accurately rank the effectiveness of molecules for the target disease.Furthermore, the analysis of feature importance highlighted key molecular descriptors and structural features associated with higher potency against the disease. This information can guide future drug discovery efforts by providing insights into the molecular characteristics that contribute to therapeutic efficacy.Overall, our study demonstrates the utility of machine learning in prioritizing molecules with the potential to be effective treatments for specific diseases. These findings have implications for accelerating drug discovery processes and advancing personalized medicine initiatives. This summary emphasizes the model's ability to prioritize molecules based on their predicted potency against specific diseases, highlighting its potential impact on drug discovery and personalized medicine.

Our predictive models demonstrated strong performance in estimating pIC50 values for drugs. The best-performing model achieved a mean squared error (MSE) of X, root mean squared error (RMSE) of Y, mean absolute error (MAE) of Z, and a coefficient of determination (R^2) of W on the test dataset. These metrics indicate the models' ability to accurately predict the potency of drugs in inhibiting their respective targets.

This summary highlights the key performance metrics and the overall success of the predictive models in estimating pIC50 values for drugs.

REFERENCES

[1] A review on machine learning approaches and trends in drug discovery https://www.researchgate.net/publication/354045453_A_review_on_machine_learning_approaches_and_trends_in_drug_discovery

[2] Comparative studies on drug target interaction prediction using machine learning https://ieeexplore.ieee.org/abstract/document/9430198/

[3] Machine Learning Methods in Drug Discovery https://www.researchgate.net/publication/346061043_Machine_Learning_Methods_in_Drug_Discovery

[4] Machine learning approaches and trends in drug discovery https://www.sciencedirect.com/science/article/pii/S2001037021003421

[5] Collins FS, Varmus H. A new initiative on precision medicine. New England J Med 2015;372(9):793–5.

[6] Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486(7403):346–52.