2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS) | 979-8-3503-8436-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICACCS60874.2024.10717206

Computational Web Drug Discovery Application ForSARS- Cov-2 Disease Using Predictive and Analytical Approach

Hemkiran S
Dept. of Computer Science and Engineering
PSG Institute of Technology and Applied Research
Coimbatore, India
hemkiran@psgitech.ac.in

R Kartik
Dept. of Computer Science and Engineering
PSG Institute of Technology and Applied Research
Coimbatore, India
kartikravichandran75@gmail.com

S Rathan Aswath
Dept. of Computer Science and Engineering
PSG Institute of Technology and Applied Research
Coimbatore, India
rathanaswaths@gmail.com

K Kabilan
Dept. of Computer Science and Engineering
PSG Institute of Technology and Applied Research
Coimbatore, India
kabilan22kk@gmail.com

Sudha Sadasivam G
Dept. of Computer Science and Engineering
PSG College of Technology
Coimbatore, India
gss.cse@psgtech.ac.in

***Abstract*—In response to the urgent need for coronavirus treatments, this research focuses on leveraging bioactivity data collection and processing for efficient drug discovery, employing computational methods to predict potential antiviral compounds. Exploratory data analysis was performed to identify patterns and trends, and various molecular descriptors were computed for the Mpro inhibitors in the descriptor dataset preparation step. A regression model was trained using the developed random forest algorithm to predict the bioactivity of new compounds against the standard value and compared with other regression models. Additionally, a web-based application was created using the random forest model to allow users to obtain predicted bioactivity values against Mpro by providing information about molecular structures. Machine learning-based bioactivity pre- diction offers an intriguing plan for drug discovery, and the proposed work provides a comprehensive workflow for COVID- 19 drug discovery. The web application provides a user-friendly interface for drug discovery researchers to evaluate the potential of compounds against the Mpro target quickly.**

Keywords—Random Forest Model, Exploratory Data Analysis, Drug Discovery, Descriptor calculation, Docking, Data Preprocessing, Lipinski Rule of Five.

I. INTRODUCTION

The COVID-19 pandemic, which is caused by the virus known as SARS- CoV-2, has become a global health emergency, necessitating the urgent development of efficient med- ications for its treatment. The conventional methods of drug discovery are time-consuming, expensive, and frequently yielding inadequate outcomes. As a result, there has been an increasing inclination towards employing computational approaches to expedite the process of drug discovery. This study proposes an innovative computational methodology that combines the Lipinski rule of five with a regression model uti- lizing the random forest algorithm, to find possible medication candidates to the SARS-CoV-2.

A. Bioinformatics

Bioinformatics is an interdisciplinary field that integrates computer science, mathematics, and biology to create techniques for analyzing and interpreting biological data include data on gene regulation, protein structures, and genetic sequences. An important application of bioinformatics lies in drug discovery, where it facilitates the discovery of possible therapeutic targets, prediction of drug efficacy, and optimization of drug design.

B. Regression Model

The effective machine learning method known as the random forest algorithm is utilised for regression in this research work. Numerous decision trees are combined together using an ensemble learning technique to produce a more precise and reliable prediction model. The random forest approach is able to manage huge, complicated datasets that have substantial dimensionality, making it particularly ideal for drug discoveryapplications.

1) Data preparation: The initial step in constructing a regression model for drug discovery involves collecting data on the proteins of the virus and known drugs. The protein data encompasses sequence and structural information, as well as binding site data. The known drug data includes information on the chemical structure, physicochemical properties, andefficacy against SARS-CoV-2.

2) Feature engineering: The technique of choosing andretrieving pertinent features from the data to improve the model's precision is referred to as feature engineering. In drug discovery, features such as physicochemical properties, molecular descriptors, and protein-ligand interaction energies are often utilized.

979-8-3503-8436-9/24/$31.00©2024 IEEE

3) Model training: Following data preparation and feature engineering, the regression model is trained using the random forest algorithm. To create the final forecast, the model builds a number of decision trees and aggregates the projections.

4) Model evaluation: After training the regression model, it is essential to evaluate its performance by testing across dif- ferent data. This involves evaluation of the model's accuracy, precision, recall, along with some relevant measures usingappropriate evaluation techniques.

5) Model optimization: Finally, the regression model can be optimized by tuning the algorithm's parametric variables for the random forest. This process entails adjusting parameters such as the quantity of trees and the greatest depth of the trees to enhance the model's accuracy and generalization capabilities.

II. Literature Survey

There is limited existing literature within this research domain that substantiates the findings derived from the re- search. Ali Abdelkrim., [2] discussed use of machine learning techniques in the search for new drugs. These sources provide review by covering different methods and strategies applied in this field.The studies contribute to the understanding and advancement of bioinformatics and biomedicine, fostering innovation in the field.

Bhardwaj et al. [3] proposed the analysis of COVID-19 data available in India including forecasting effect rates, daily new cases, and total completed cases using Regressor Random Forest and Random Forest Classification. These techniques showed better performance than others, including Support Vector Machine, KNN, Multilinear Regression, Gaussian Clas- sifier, Decision Tree Classification, Logistic Classifiers, and Extreme Gradient Boosting.

Cui et al [4] mentioned the use of machine learning to assess a compound's anti-SARS-CoV-2 activity. The research aimed to contribute to the understanding and development of potential treatments for COVID-19.

Dadhwal et al [5] proposed the use of artificial intelligence methods to evaluate and foresee drugs.

Huang et al [6] proposed the concept of COVID-19 knowl- edge graph centred on the production of medications and vaccines. The authors highlight the importance of leveraging knowledge graphs to enhance research efforts and identify potential targets for therapeutic interventions and medication. Long, et al [7] developed a novel approach using a hetero- geneous graph attention network to identify potential drugs for COVID-19 treatment and constructed a graph using drug- target and drug-disease association data and applied a graph attention mechanism to understand the connections among drugs, destinations and diseases.

Monteiro et al. [8] suggested a unique deep learning archi- tecture model which employs Simplified Molecular Input Line Entry System (SMILES) strings of chemicals and Convolu- tional Neural Networks (CNNs) to extract 1D representations from protein sequences. These representations, which capture local relationships, are used as a binary classifier in a Fully Connected Neural Network (FCNN). The exhibiting enhanced efficacy in accurately categorizing Drug-Target Interactions (DTIs) that are positive or negative.

Motohashi et al. [9] discussed about the various machine learning techniques to develop regression models and ranking methods for identifying potential p53 inhibitor candidates. The work aimed to improve the selection process of molecules with inhibitory effects on p53, a protein involved in cancer development.

Zamitalo et al. [11] considered a machine learning re- gression model for forecasting drug targets in relation to COVID-19. The study aimed to explore the potential of using computational methods to identify effective targets for drug development.

III. Proposed Methodology

The proposed method involves in the process of collecting and curating data on the disease and potential drug targets from various sources, such as scientific literature, clinical trials, and public databases. Preprocessing the data is used to ensure a clean, consistent, and structure suitable for machine learning. Developing machine learning models to identify potential drug targets and predict the activity of compounds against these targets. The process of creating a user-friendly web page that allows researchers to access the data, machine learning models, and analytics tools to facilitate drug discovery for the SARS- CoV 2 effectively and identify promising drug candidates for further investigation.

Fig 1 demonstrates the overview of the proposed method- ology The system flow begins with the target protein search, where the desired protein for analysis is identified. The ChEMBL database is then utilized to retrieve bioactivity data related to the target protein. SARS coronavirus 3C like proteinase, a single protein is considered as a target virus (CHEMBL3927) [1]. Around 1310 bioactivity data for CHEMBL3927 are gathered as IC 50 values in nM unit. The notation, which represents the molecular structure of compounds, is cleaned to ensure accurate analysis.

After data preprocessing the data, Lipinski's descriptors are calculated, which are important molecular properties used in drug design. Exploratory data analysis (EDA) is performed to understand the dataset and spot any trends or patterns.

The Padel descriptor is used to calculate the molecular fingerprint of each compound.

The resulting data is read as a framework for subsequent steps. Next, low variance features are removed to eliminate redundant or uninformative variables. The dataset is subse- quently separated into training (80% of data) and testing (20% of data) subsets for model building and evaluation. Regression models are constructed using the training data to