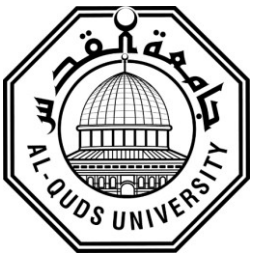
Deanship of Graduate Studies

Al-Quds University



Machine Learning Model for COVID-19 Drug Discovery

Claudia Elias Rafat Alawi

MSc Thesis

Jerusalem-Palestine

1444-2023

Machine Learning Model for COVID-19 Drug Discovery

Prepared By:

Claudia Elias Rafat Alawi

B.Sc.: Computer Systems Engineering,

Birzeit University, Palestine.

Supervisor: Dr. Rashid Jayousi

Co-Supervisor: Dr. Yousef Najajreh

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Computer Science/ Department of Computer Science & Information Technology/Faculty of Science & Technology/Graduate Studies.

1444-2023

Al-Quds University

Deanship of Graduate Studies

Master of Computer Science

Thesis Approval

Machine Learning Model for COVID-19 Drug Discovery

Prepared By: Claudia Elias Rafat Alawi

Registration No: 21911799

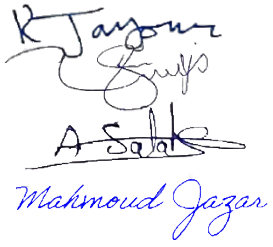
Supervisor: Dr. Rashid Jayousi

Co-Supervisor: Dr. Yousef Najajreh

Master thesis submitted and accepted, Date: 27/03/2023.

The name and signatures of examining committee members are as follows:

1. Head of committee: Dr. Rashid Jayousi Signature:
2. Co-Supervisor: Dr. Yousef Najajreh Signature:
3. Internal Examiner: Dr. Saeed Salah Signature:
4. External Examiner: Dr. Mahmoud Jazzar Signature:



Jerusalem-Palestine

1444-2023

Dedication

I dedicate this thesis to God Almighty, my creator, he has been the source of my strength throughout my studies. This thesis is also dedicated to my loving mother, who has always encouraged and supported me through the difficulties of graduate school and life, and who taught me to work hard for the things I want to achieve. I'm grateful beyond words to have you in my life.

Claudia Alawi

Signed….Claudia Alawi

Date: 27 March 2023

Declaration

I certify that this thesis submitted for the degree of Master, is the result of my own research, except where otherwise acknowledged, and that this study (or any part of the same) has not been submitted for a higher degree to any other university or institution.

Signed……

Claudia Alawi

Date: 27 March 2023

Acknowledgments

I would like to express my gratitude to my primary supervisor, Dr. Rashid Jayousi for his support and guidance throughout this project, for the thoughtful comments and recommendations on this dissertation. I would like to also acknowledge the help provided by my co-supervisor Dr. Yousef Najajjreh for his feedback, recommendations and all the biochemical knowledge he provided to me to complete this project. I would also like to thank my family who supported me and offered deep insight into the study.

Abstract

COVID-19 was a big issue facing the world, and the development of an effective drug for the virus is still under research. However, developing a new drug is a lengthy and costly process that might take up many years. Artificial intelligence can have a vital role for faster and more cost-effective drug discovery. The primary protease that is essential to SARS-CoV-2 replication is 3CLpro. In this thesis a machine learning model that can be used to predict the inhibitory activity of 3CLpro was developed based on decision tree regressor. The descriptors that represent the chemical molecules were obtained using PADEL descriptor software, and these descriptors were fed into the decision tree model to train it and predict the bioactivity of unknown compounds with the target protein. The model was optimized using pruning and ensemble methods, where the decision tree was combined with SVM to improve the model performance. The research focused on both external and internal approaches for validating model performance. The model successfully discovered 26 unknown compounds from Zinc natural product data source that showed bioactivity with the target protein. Moreover, Lipinski rule of five (RO5) was applied to prioritize drug-like compounds resulting in 25 of the discovered compounds having drug like properties and can be used in clinical trials. The model was validated using 10-folds cross validation and was also validated using external dataset from different data source than the data source used in training the model, on both external and internal datasets, the proposed model has proven to be effective, however, the model showed higher performance on the external validation with accuracy of 0.89, precision of 0.75, recall of 0.6 and f1 score of 0.67 for the internal validation, while for external validation 0.98 accuracy, 0.99 precision, recall of 0.93 and f1-score of 0.96. Compared to similar studies using deep learning, our machine learning model showed better performance. In conclusion the proposed model can be useful in the drug discovery of new compounds for the COVID-19 virus.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full Word
AI	Artificial Intelligence
HTS	High-Throughput Screening
FDA	Food and Drug Administration
SBDD	Structure-Based Drug Discovery
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity
3CLpro	3C-Like Protease
PLPro	Papain-Like Protease
DT	Decision Tree
RO5	Rule of Five
ML	Machine Learning
SVM	Support Vector Machine
RL	Reinforcement Learning
QSAR	Quantitative Structure–Activity Relationship
NN	Neural Networks
CNN	Convolutional Neural Networks
RF	Random Forest
SMILES	Simplified Molecular Input Line Entry System
TP	True Negative
FP	False Positive
TP	True Positive
FN	False Negative

Chapter 1

Introduction and Background

1.1 Introduction

COVID-19 was a big issue facing the world, and the development of an effective drug for the virus is still under research. However, a typical drug discovery cycle might take up to 14 years [1] and cost up to 800 million dollars [2] to complete from target identification to Food and Drug Administration (FDA) approval. The median cost of the effectiveness trials for the 59 new drugs that the FDA had approved during the 2015–2016 time period was $19 million [3]. As a result, it is critical to replace the inefficient methodologies used in traditional drug development with more efficient, low-cost, and broad-spectrum computational alternatives. This is where Artificial Intelligence (AI) can take critical role in accelerating the drug discovery process, where intelligent and computational drug design offers a fresh perspective on the system-centric approach, beginning with the definition of a new drug's scope [4]. The ability to manage new scales and levels of data complexity is what the AI methods deliver. Various approaches for statistical calculations normally work within the constraints of pre-built or fixed assumptions, but AI has the potential to be useful in a broad situation where it can assist in determining whether or not a molecule meets some criteria. AI has the ability to help in the fight against COVID-19 by assisting in the discovery of novel drugs and vaccinations. Indeed, even before the COVID-19 outburst, AI was known for its enormous ability to aid in the development of new drugs [5].

The process of identifying novel drugs for new diseases is known as drug discovery. The process includes the following steps: target identification, target validation, lead identification, and lead optimization. The process of locating a protein with a particular role in a disease is known as target identification. Target validation is the process of verifying a target according to the inventor's thought process. Lead identification is the process of determining the best compounds for a given target protein. The process of assuring drug-related features of compounds is known as lead optimization. Bioavailability, specificity, and toxicity of discovered chemicals must all be guaranteed by the inventor. Experimental High-Throughput Screening (HTS) was used to identify promising leads in the past, but it is time consuming and

costly [6]. In contrast to earlier techniques for drug discovery, rational drug design is effective and affordable. The approach is often referred to as reverse pharmacology [7] since the initial phase in the rational drug design process is to identify interesting target proteins that are then used to screen small-molecule libraries. The ability to identify binding cavities due to the availability of 3D structures of therapeutically important proteins has cleared the way for Structure-Based Drug Discovery (SBDD). SBDD is a more focused, effective, and rapid approach for lead discovery and optimization because it uses information of the disease at the molecular level and the 3D structure of a target protein [8]. Computational resources are a useful tool for speeding up the drug development process, which involves screening processes, combinatorial chemistry, and calculations of parameters including Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) [9]. SBDD is a multi-cycle procedure that leads to the development of an optimized drug candidate for clinical trials.

COVID-19 drug design methods are progressing at the same time as computational artificial intelligence and molecular chemistry. This approach is proving to be a useful tool in medicinal chemistry for identifying the beginning points for COVID-19 hit compounds. This method cuts down on the time and money spent on drug research and development. The applications that use an AI-based method for drug creation are specifically concerned with the molecular structure of the medications. AI-based apps are critical for finding new drug candidates and optimizing drug repurposing by retrieving data and information from engines. The COVID-19 is becoming the benchmark for "Artificial Intelligence and Computational Drug Designing" approaches, bringing up new options for drug development [10].

The Papain-Like Protease (PLPro) and the 3C-like protease (3CL or "Main") are two appealing targets for small-molecule therapeutic intervention that are found in coronaviruses. The viral life cycle depends on both of these cysteine proteases, which are non-structural enzymes. It is feasible that protease inhibitors created to inhibit these viral proteins will have low toxicity because mammals lack proteases with similar substrate preferences. As the first step in the rational drug design process is to determine the target protein, the 3C-like protease (3CLpro) of the SAR-CoV [11] is one of the most potential protein targets. Protease inhibitors are the most effective at preventing replication [12-14]. As a result, the 3CLpro enzyme appears to be a suitable target for drug development, and hence, it is a prospective target for generating efficient COVID-19 inhibitors.

Validation of Machine Learning (ML) models is a common approach for verifying the efficacy and generalizability of models. According Ramspek, Jager, Dekker, Zoccali, and Ddiepen in their article about the use external validation [67], internal validation techniques like cross-validation and bootstrap, cannot ensure the quality of a ML model due to possibly biased training data and the complexity of the validation procedure itself. They suggested using external data sources from elsewhere as validation datasets to better evaluate a learnt model's generalization capacity. So, in this thesis, external database was used to perform the model validation.

This thesis proposes ML model for identifying new COVID-19 drugs against 3CLpro enzymes. A decision tree-based (DT) model was developed to predict the bioactivity of unknown compounds with the identified target protein 3CLpro. The descriptors that represent the chemical molecules were obtained using PADEL descriptor software, and these descriptors were then fed into the DT model to train it and predict active compounds. PubChem Bioassays were used to collect experimental datasets with bioactivity on 3CLpro.The data collected was not large though, and in order to ovoid model overfitting, 10-K cross validation was used, as for model hyperparameter tuning, grid search method was used, and for model optimization pruning and ensemble methods were applied. ChEMBL database was used to obtain external dataset for validation. Furthermore, the model was deployed to predict unknown compounds retrieved from ZINC database, for predicted compounds that have active bioactivity with the target protein, Lipinski RO5 was employed to prioritize drug-like compounds.

The aim of this study is to inspect the use ML model like DT in order to predict the bioactivity of unknown molecules that has drug-like properties which could be used as inhibitors for the 3CLpro enzyme and hence be used as drugs for COVID-19 virus.

1.2 Background

1.2.1 AI-Driven Drug Discovery

Artificial intelligence has been used to create molecules that are chemically correct and effective against new ailments. AI systems can be taught to learn the essential characteristics of a well-known drug. A well-trained AI algorithm can learn to assemble new molecules, leading to the creation of valuable compounds. In their study [15], the

authors applied AI approaches to improve and speed up the drug selection process. In their study [16, 17], the authors talked about applying artificial intelligence to target and design drugs.

Artificial Intelligence in computational drug designing [18] seeks high-quality research on drug and clinical research on artificial intelligence techniques to leveraging the power of computational drug designing by combining AI and core chemistry. Computational drug designing is a growing field of study that focuses on the design and testing of molecular characteristics, interactions, and behavior in order to create better materials, processes, and systems for specific activities. ML is a subfield of AI that uses statistical learning methods. The use of AI in drug discovery can be seen as the automated integration of ML algorithms to find new compounds by analyzing, learning, and interpreting pharmaceutical large data [19].

The effectiveness of ML has frequently been demonstrated in classification, generative modeling, and Reinforcement Learning (RL). ML is broken down into three groups: reinforcement learning, unsupervised learning, and supervised learning. The model is predicted using input and output data sources by the subcategory of supervised learning, classification, and regression methods. Binary activity prediction is used by Support Vector Machine (SVM) with supervised ML algorithms to distinguish between a drug and a nondrug [20,21] or between particular and nonspecific substances [22,23]. As a result of a clustering strategy for an unsupervised learning category, a disease subtype can be found, whereas a target in a disease can be found using a feature-finding method [24,25]. Decision-making Modeling and quantum chemistry help RL improve its performance in de novo drug design. Dataset learning is less important in RL. RL can be used to influence the intended physical and biological features of freshly produced chemical compounds [26].

ML is used to create drugs that take advantage of the link between biological action and chemical structure. Quantitative Structure-Activity Relationship (QSAR) models, pharmacophore models, molecular docking analyses, and ranking/scoring functions in similarity searches can all be implemented using machine learning techniques and statistically validated [27]. The diversity of the training dataset, the capability to handle imbalanced datasets of active and inactive compounds in the library, and the definition

of precise parameters to cover the entire chemical space, including active and inactive molecules, are just some factors that affect the output of machine learning methods [28]. Effective machine learning models can be created to screen large libraries with few false positives and a large number of active chemicals in the output. This can be accomplished by employing a variety of training datasets that include anticipated inactive substances [29,30].

1.2.2 Deep Learning for Drug Discovery

Molecular features can be examined in a rational and systematic manner using contemporary computational tools. The information gathered from each chemical can be analyzed from a variety of angles [31]. There has most likely been a rise in the size of data generation in the current era of technology. The resulting data may contain errors, duplications, missing or incorrect data, and other inconsistencies that might affect how accurately simulation and analytical processes work. Advanced stages such as preliminary analysis and curation are necessary to assure fairness, accuracy, and experimental efficacy [32]. In deep learning, artificial neurons are employed to process data, and this method is used to accelerate the drug development process in drug discovery. Deep learning is also used in the drug development process, and it is most commonly used in virtual screening [33].

1.2.3 Lipinski Rule of five (RO5)

Lipinski's rule of five is a general guideline for determining a molecule's drugability. This criterion aids in determining whether a biologically active molecule has the chemical and physical qualities required for oral bioavailability. Pharmacokinetic drug features such as absorption, distribution, metabolism, and excretion are based on specific molecular qualities such as:

- There should be no more than 5 hydrogen bond donors.
- There should be no more than ten hydrogen bond acceptors.
- Less than 500 Da molecular mass
- Not more than 5 partition coefficients

A compound is predicted to be a non-orally accessible medication if two or more of

these conditions are violated. The term "rule of five" stems from the fact that all of the determinant criteria are multiples of five [65,66].

1.2.4 *De Novo* Drug Design

De novo drug design is a technique for creating new chemical compounds from the ground up. The basic idea behind this method is to create chemical structures for tiny molecules that attach to the target binding cavity with high affinity [34]. For *de novo* design, a stochastic technique is typically utilized, and it is critical to consider the search space information in the design algorithm. The positive and negative designs are being used. A search is narrowed to certain regions of chemical space in the former design, with a higher possibility of obtaining results with the needed properties. In the negative mode, on the other hand, the search criteria are established to prevent the selection of false positives [35]. Computational approaches can be used to develop chemical compounds that mimic synthetic chemistry, while scoring functions can be used to perform binding experiments [36].

One of the assessment tools for a candidate's critical evaluation, which is essential for the design process, is the scoring function. For multi-objective drug design [37], which analyzes numerous aspects at once, multiple scoring algorithms can be used in parallel.

1.2.5 Molecular Docking

Docking is a virtual simulation tool for molecular interactions [38]. Because molecular docking accurately predicts the conformation and binding of ligands inside a target active site, it is the most widely used approach in SBDD [39,40]. This approach may be used to investigate essential molecular processes like ligand-binding posture and intermolecular interactions, which are important for a complex's stability [41]. Furthermore, docking algorithms use several scoring methods to predict binding energies and rank ligands [41,42]. The proper ligand-binding conformation is determined by two factors: a wide conformational space defining various binding poses, and an explicit prediction of binding energy associated with each conformation. Multiple repetitions are carried out until the minimal energy state is reached, at which

point ligand-binding is evaluated using a variety of scoring systems [43].

1.2.6 Scoring Functions

A scoring method can be used by a docking program to delve deeper into the ligand-binding area. Once a relevant binding conformation is determined, the scoring function determines binding affinity. As a result, docking is likely to be significantly impacted by scoring functions. A training dataset of compounds from a related class for which experimental binding affinity data is available is used to develop scoring functions.

There are four types of scoring functions: force field, empirical, knowledge-based, and ML [44–46]. Dynamic methodologies for developing and optimizing models that predict binding posture and affinity are provided by machine learning techniques, which are essentially model-based approaches. ML is increasingly being used to develop novel scoring functions [47]. These methods account for interactions between a ligand and its target but disregard interactions that are prone to mistake. In order to cope with nonlinear dependence among binding interactions, various ML approaches, including Random Forests (RF), Support Vector Machines (SVM), and Neural Networks (NN), are used. As a result, in estimations of binding energy, ML-based scoring functions perform better than competing scoring methods. Consensus scoring is a scoring function that makes use of group scores in order to lower the probability of individual score inaccuracy and raise the possibility of true positive selection [48].

Chapter 2

Literature Review

2.1 Related work

This section has an overview of some of the related papers that have been published which discuss AI and drug discovery for COVID-19. Not many studies implement machine learning models specifically in the field but rather do an overview for the computational methods and techniques for COVID-19 drug design. S. Lalmuanawma, J. Hussain, and L. Chhakchhuak [49] in their article, they provide an insight of recent studies that use ML and AI. It also discusses a few common pitfalls and difficulties encountered when applying such algorithms to real-world applications. The article also includes model designers, medical experts, and policymakers' recommendations for dealing with the COVID-19 pandemic now and in the future. They came with a conclusion that for the COVID-19 epidemic, continuous advances in AI and machine learning have dramatically improved treatment, medication, screening, prediction, forecasting, contact tracing, and drug/vaccine research, while reducing human participation in medical practice. However, most of the models haven't been tested enough to show how well they work in the real world, but they're still capable of combating the COVID-19 outbreak.

E. N. Muratov [50] in their study offered a critical overview of the most important computational methods and their applications for the discovery of COVID-19 small-molecule therapies that have been described in the scientific literature. It stated that, following the first year of the COVID-19 pandemic, drug repurposing appears to have failed to deliver speedy and worldwide remedies. It assumed that truly effective computational tools must provide actionable, experimentally testable hypotheses that enable the discovery of novel drugs and drug combinations, and that open science and rapid sharing of research results are critical for accelerating the development of novel drugs and drug combinations.
M. Batool, B. Ahmad, and S. Choi in their article [51] also focused on the currently available methodologies and algorithms for structure-based drug design, such as virtual screening and de novo drug design, with a special emphasis on AI- and deep-learning-based drug discovery methods.

In their opinion, A. Chandra Kaushik and U. Raj [4], an artificial intelligence (AI) based

method that can predict drugs/peptides directly from infected patients' sequences and hence have better affinity with the target and contribute to COVID-19 vaccine formulation would be beneficial. However, they stated that testing of these proposed vaccines/drugs will be required to ensure their safety and feasibility in combating COVID-19.

Table (2.1) below discusses the related work that implements machine learning COVID-19 drug discovery in their approach, publication year and main contribution in the field.

Table 2.1: Summary of related work papers that implement machine learning COVID-19 drug discovery in their approach.

Paper	**Year**	**Contribution**
Gianchandani et al.	*2020*	*Ensemble deep transfer learning models were proposed to diagnosis coronavirus infections from radiography [52].*
Singh et al.	*2021*	*Proposed an automated COVID-19 screening model based on densely linked convolutional networks [53].*
Kumari et al.	*2020*	*Used machine learning methods such as random forest (RF), support vector machine (SVM), and DT for the classification of anti-tubercular compounds [54].*
Chen et al.	*2020*	*Developed a deep learning-based approach in order to detect new coronavirus pneumonia from a picture [55].*
J. Peng, J. Li, X. Shang	2020	*Convolutional neural network (CNN) models were used to predict drug-target interactions [56].*
S. Hu, C.P. Chen, J. Zhang, B. Wang	2019	*CNN models were used for predicting drug-target interactions from drug structure [57].*
Meyer et al.	*2019*	*Used CNN and RF models to deduce pharmacological functionalities from chemical structures [58].*
Kumari M., Subbarao N.	*2021*	*Developed a deep learning CNN model to predict drugs for 3CLpro enzymes to cure COVID-19 infections [59].*

2.1.1 Related Work Results Discussion

This subsection provides an insight on the closely related work to our thesis starting from the least relevant to the most relevant, it also mentioned if any of these studies have contributed to our work.

S. Hu, C.P. Chen, and J. Zhang [57] used three benchmark datasets in their investigation to examine potential interactions between drugs and target proteins. Two datasets were created using the DrugBank database, and one dataset was constructed using the KEGG DRUG database. They suggested a CNN-based deep learning approach in their research to predict drug-target interactions only based on knowledge of drug structures and protein sequences. The final results demonstrated that, for the target families of enzymes, ion channels, GPCRs, and nuclear receptors in their dataset, respectively, their technique can perform with accuracies up to 0.92, 0.90, 0.92, and 0.91. To further evaluate the generality of the model, a different dataset collected from DrugBank was employed, which produced an accuracy of 0.9015.

To anticipate the drug-target interactions, J. Peng, J. Li, and X. Shang [56] introduced the DTI-CNN, a learning-based approach based on feature representation learning and deep neural networks. They begin by employing the Jaccard similarity coefficient and a restart random walk model to extract the pertinent characteristics of medicines and proteins from heterogeneous networks. Then, in order to shrink the dimensions and isolate the crucial characteristics, they used a denoising autoencoder model. subsequently created a CNN model to forecast how medications would interact with proteins. The evaluation's findings revealed that the DTI-CNN's average AUROC and AUPR scores were 0.9416 and 0.9499, respectively.

In their work M. Kumari and N. Subbarao [59], they developed a deep learning-based CNN model that was to do virtual screening for the 3CLpro target protein to predict anti-SARS-CoV drug candidates and compare. They compared their model with other classification methods, including RF, NB, DT, and SVM modelling. The model was trained on 282 compounds and predicted an external validation test set of 141 compounds with an accuracy of 0.86, a sensitivity of 0.45, a specificity of 0.96, a precision of 0.73, a recall of 0.45, and an F-measure of 0.55. The CNN model screened 17 out of 918 phytochemical compounds; 60 out of 423 natural products from the NCI divest IV; 17,831 out of 1,12,267 natural compounds from the ZINC natural product database; and 315 out of 1556 FDA-approved drugs as anti-SARS-CoV agents. This study was the most related study to our thesis, first because of targeting the same protein and also because of having small data set which is similar to our case, however, the

method for calculating the descriptors were different as we used PADEL, they have also used the 3D structure of the chemical compounds while we used the Simplified Molecular Input Line Entry System (SMILES) of the compounds, however it was based the on deep learning and the use of CNN.

D. Singh, V. Kumar and M. Kaur [53] research work used machine learning-based predictive modelling for virtual screening on a big dataset of compounds which retrieved from ChEMBL dataset. In their research, three classifiers; RF, DT and SVM were used to build predictive models. The comparative analysis of predictive models revealed that RF showed the best performance compared with the J48 DT, and SVM, however, they concluded that this performance depends on the type of data to be used for modelling. The RF model exhibited an accuracy of 0.938. The second-ranked predictive model was J48 DT showing accuracy of 0.928. The last Lib SVM model with accuracy 0.906 Their results showed that a systematically designed computational model for bioactivity based on IC50 value works very well to prioritize specific compounds. Their findings were used along with domain expert's recommendations when our external validation data was retrieved from ChEMBL as it was also filtered only based on the IC50 values, so we used their study as a reference, also in the use of molecule's SMILES, however, we made distinction by using PADEL descriptors directly from SMILES unlike their study that converted SMILES to 3D structure and the use of other software to generate the descriptors .

This thesis was performed using DT model in comparison to other studies that focused on deep learning and CNN, as related work and previous studies have most of their studies based on using deep learning, which in their studies proved to outperform the use machine learning model like DT, however, in this study we explore the possibility to enhance the DT model in order improve its performance and explore its ability to be used in drug development, the model optimization that was used include the use of cross validation , pruning and ensemble method with SVM. This thesis also has added the use of PaDEL descriptors and the use PubChemPy API, as before building the model, critical descriptor vectors were extracted for bioactivity prediction using PaDEL software. This work has also focused on implementing internal and external validation data sets. The validated model's results suggested acceptable and good values for various internal and external validations.

Chapter 3

Methodology

The purpose of this chapter is to outline the process for applying ML to the problem of finding new drugs. The data collection and preprocessing, model selection and training, model optimization, and evaluation and validation of the models' performance are all covered in detail in this chapter. This chapter also provides insights into best practices and potential traps to avoid for other researchers working in the area of ML for drug discovery.

The aim of this thesis is to use ML model and examine its ability to predict new drugs for our target protein. There are many crucial steps in the methodology for using ML in drug discovery. First, a collection of chemical compounds and their corresponding properties are collected. The data is then cleaned, normalized, and the chemical structures transformed in a way that is appropriate to be used in the ML model. Then, using this pre-processed data, the ML model is trained. The model is then optimized using hyperparameter tuning and validation procedures to ensure that it performs as well as possible when predicting a compound's activity against the target protein. The model is then deployed to screen libraries of compounds and identify promising drug candidates for additional clinical experiments.

To identify promising molecules that could be used as inhibitors for the target protein, The molecule should have certain features to be good candidate, first it should have some general features which qualify it to be a drug, and this where Lipinski's RO5 was used to identify the drug like molecules. The molecule should also have local features which are unique building blocks that describe the molecule, as each molecule is comprised of several building blocks and the way these blocks are connected will create a unique property for the molecule, which is described in the chemical structure of the compound, so we need to find molecules that has certain chemical structure that suits the binding cavity of the target protein, these building blocks are characterized in the canonical SMILES of each compound, represents the 2D molecular structure of a chemical compound as a unique string of characters.

3.1 Dataset

The data used in this thesis is composed of two main sets, the first set is used for model training and testing and was collected from PubChem bioassays [60] and the second data set is used for external validation and was collected from ChEMBL [61] database. Data collection was performed separately for each data set as each set was obtained from a different data source.

3.1.1 Data Collection

PubChem Bioassays were used to gather publicly available experimental datasets with bioactivity on 3CLpro of SARS-CoV target protein. For this study, five assays were selected based on the total number of tested compounds, where assays with the greatest number of compounds were chosen, one conformational high throughput screening bioassay, two dose-response bioassays and two late-stage bioassays, the used bioassays are shown in Table (3.1).

Table 3.1: List of PubChem bioassays of 3CLpro of COVID-19 that were used to collect data.

BioAssay AID	Total No. of Compounds	Active Compounds	Inactive Compounds	BioAssay Type
1879	380	136	244	Confirmational HTS assay
1890	101	44	57	Dose response assay
488958	14	9	5	Dose response assay
488967	32	15	17	Late Response assay
488984	103	10	93	Late Response assay

The assays have a total number of 428 compounds, including 69 active and 359 inactive compounds. Experimental bioassays had already been used to classify the chemicals' activity of the compounds, so each of these bioassays has several compounds and their corresponding bioactivity with the 3CLpro. Compounds are labeled as either *Active* or *Inactive*, a sample of the labeled raw data is shown in Table (3.2).

Table 3.2: Sample of one of the bioassays data that was collected from PubChem.

PUBCHEM_RESULT_TAG	PUBCHEM_SID	PUBCHEM_CID	PUBCHEM_ACTIVITY_OUTCOME	PubChem Standard Value	Standard Type	Standard Relation	Standard Value	Standard	Activity Comment
RESULT_TYPE				FLOAT	STRING	STRING	FLOAT	STRING	STRING
RESULT_DESCR				PubChem standardized va	Standardized activity typ	Qualifier (e.g. >, <, =)	Standardized activity	Selected u	Additional comments
RESULT_UNIT				MICROMOLAR					
RESULT_IS_ACTIVE_CONCENTRATION				TRUE					
RESULT_IS_ACTIVE_CONCENTRATION_QUALIFIER						TRUE			
1	164127741	44634769	Inactive		Inhibition				Not Active
2	164127742	44634812	Inactive		Inhibition				Not Active
3	164127748	51003641	Active	45	IC50	=	45000	nM	
4	164127749	51003639	Active	0.3	IC50	<	300	nM	
5	164130598	44634768	Inactive		Inhibition				Not Active
6	164133419	51003636	Active	5.5	IC50	=	5500	nM	
7	164133420	51003654	Active	0.3	IC50	<	300	nM	
8	164133421	46897843	Inactive		Inhibition				Not Active
9	164133422	46897844	Active	1.6	Ki	=	1600	nM	
10	164133422	46897844	Active	1.5	IC50	=	1500	nM	
11	164136185	71717907	Active	0.5	IC50	=	500	nM	
12	164136186	71717908	Active	13	IC50	=	13000	nM	
13	164136187	44634770	Active	28.1	IC50	=	28100	nM	
14	164136189	71717909	Active	6	IC50	=	6000	nM	
15	164136192	53297485	Active	0.3	IC50	<	300	nM	

3.1.2 Data Preprocessing

The raw data has unnecessary columns that will not be used or useful in the model training, so only PUBCHEM_CID and PUBCHEM_ACTIVITY_OUTCOME columns were taken into consideration and the other columns were dropped. After that duplicates compound were removed, and this resulted in 400 distinct compounds.

The chemical structure of each compound is the most important property, as all feature for model training will be extracted from the chemical structure, however, the raw data doesn't have it, so in order to get the chemical structure information, PubChemPy API [62] was used to fetch the canonical SMILES as these SMILES were used to calculate the important features that are required for the molecule to inhibit the target protein and hence, the canonical SMILES are crucial for model training and testing. Table (3.3) shows sample of the data after removing unnecessary columns and adding canonical SMILES.

3.1.3 External Validation Dataset

For external validation of the model, ChEMBL database was used. ChEMBL contains more than 2 million compounds, and it is compiled from 84,092 documents, the used version is ChEMBL 30. ChEMBL is a database of bioactive drug-like small molecules,

it contains 2-D structures, calculated properties and abstracted bioactivities. The data is abstracted and curated from the primary scientific literature and covers a significant fraction of the discovery of modern drugs. The used data set which contains the biological activity data with the target protein 3CLpro was downloaded directly from the ChEMBL database. The concentration of the molecule needed to achieve a specific level of inhibition of the biological target is usually expressed as an IC50 or EC50, which represents the observed activity. The data was filtered based on the standard type (IC_{50}), so only data that has IC_{50} value was considered. The standard value is the potency of the drug and the lower the value the better the potency of the drug becomes, and likewise the higher the number the worse the potency becomes.

Compounds were cleaned first, so compounds which have missing value for the standard value were dropped. Following that compounds were labeled as active or inactive based on the value of the IC_{50} unit, where compounds with values less than 1000 nM are considered active, whereas those with values larger than 7500 nM are considered inactive, the threshold values were based on domain experts' recommendation. That resulted in a total of 119 compounds, in which 104 were inactive and 15 compounds were active.

After that The IC50 value was converted to the negative logarithmic scale, which is effectively -log10(IC50), to enable more equal distribution of the IC50 data.

3.1.4 Exploratory Data Analysis

In order to have better understanding of the structure of the data, visualization of the bioactivity of the compounds with target protein are shown in Figure (3.1), the figure shows big difference between active and inactive compounds in the main dataset of PubChem. Figure (3.2) shows the distribution of active and inactive compounds in the ChEMBL dataset.

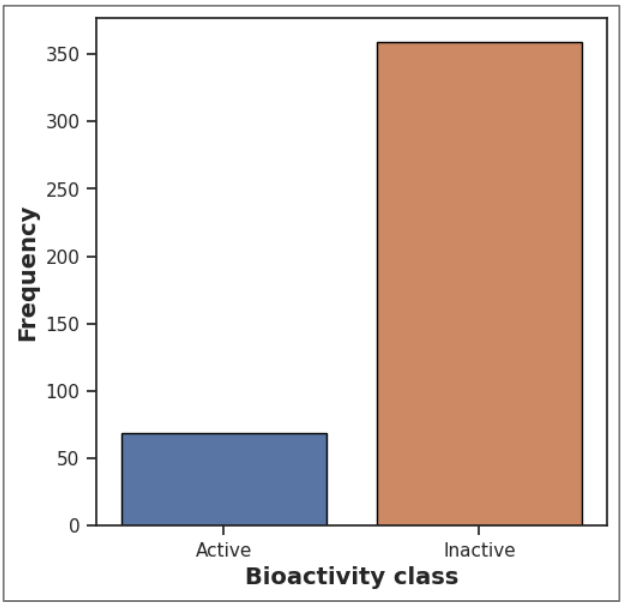


Figure 3.1: Data bioactivity distribution for the PubChem bioassays.

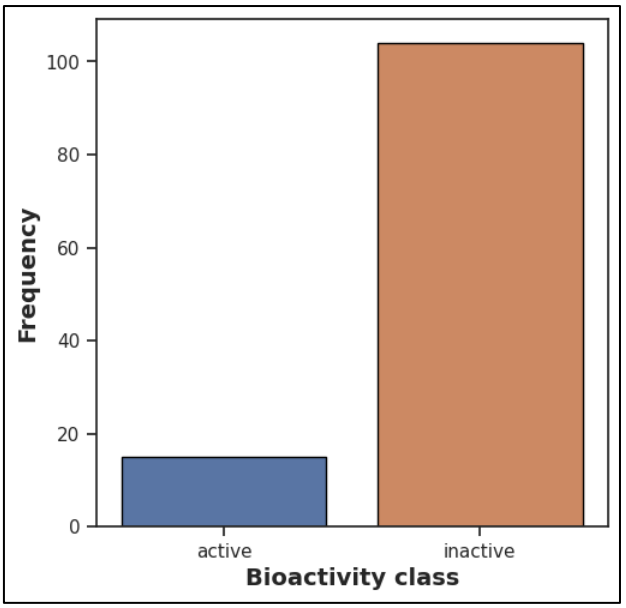


Figure 3.2: Data bioactivity distribution for the ChEMBL.

3.2 Descriptor calculation

The Molecular descriptors are representation of certain structural features of a molecule. And in order to calculate if certain molecules have the functional and structural features that are required to have bioactivity with the target protein, these features must be represented in a mathematical way. For this purpose, PaDEL software [63] was used to calculate the fingerprints for each molecule. PaDEL-Descriptor is a program that calculates molecular

fingerprints and descriptors. The Chemistry Development Kit [64] is primarily used to calculate these descriptors and fingerprints. The presence or absence of a given feature in a molecule is indicated by each element of the fingerprint vector. The fingerprints generated using PaDEL resulted in a matrix of 400 rows × 881 columns. Figure (3.3) shows a sample of the resulting matrix of fingerprints for each molecule.

	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4	PubchemFP5	PubchemFP6	PubchemFP7	PubchemFP8	PubchemFP9	...
0	1	1	0	0	0	0	0	0	0	1	...
1	1	1	0	0	0	0	0	0	0	1	...
2	1	1	1	0	0	0	0	0	0	1	...
3	1	1	0	0	0	0	0	0	0	1	...
4	1	1	1	0	0	0	0	0	0	1	...
...	...	...	...	...	...	...	...	...	...	...	...
395	1	1	1	1	0	0	0	0	0	1	...
396	1	1	1	0	0	0	0	0	0	1	...
397	1	1	1	1	0	0	0	0	0	1	...
398	1	1	1	0	0	0	0	0	0	1	...
399	1	1	1	0	0	0	0	0	0	1	...

400 rows × 882 columns

Figure 3.3: The fingerprints matrix for each molecule resulted from PADEL software.

3.3 Feature Selection

The feature selection method that was used is removing low variance features, this is a basic technique of feature selection. The idea is that if a feature is constant (i.e., has no variance), it can't be used to identify any interesting patterns and should be removed from the dataset. Removing low variance features is seen as appropriate feature selection since these features lack discriminatory power and information, and as a result, do not significantly improve the performance of a ML model. These features may produce noise or increase computational complexity without adding any meaningful information, which might degrade the performance of the model. We may streamline the model and lessen overfitting by eliminating low variance features, which happens when a model is overly complex and tends to fit the noise in the training data instead of the underlying patterns. Additionally, by minimizing the amount of unnecessary or redundant features that may obfuscate the relationship between the input variables and the target variable, the removal of low variance features can enhance the model's interpretability. Removing the low variance features from the dataset resulted in a matrix of 400 rows × 148 columns.

3.4 Dataset division

Model validation is an important step to ensure that if the model is exposed to completely new, unseen data, it will predict with the same accuracy and it will not fail to generalize over the new data, which is the problem of over-fitting. The model should also avoid the problem of underfitting which occurs due to high bias and low variance.

Initially Model validation was made on a subset of the data by using *the Hold-Out* method, which basically split the data between training and testing where the proportion of train to test data is (80,20). The model was trained on 80% of the random split set and then validated on the remaining 20% of the curated dataset, and this was only done for the purpose of selecting the most appropriate ML model.

Although in ML, it's typical to divide the data into 20% for testing and 80% for training. This split, however, could not be enough if the dataset is very small since it can lead to high variance in model performance and incorrect estimations of model performance. When the dataset is small, there may be a lot of randomness and ambiguity in the data, which makes it challenging to determine a model's actual performance. The model may be overfitted or underfitted because the training and testing sets may not have enough data to adequately represent the data's underlying distribution. Alternative methods, like cross-validation may be more applicable in such circumstances.

To avoid overfitting, k-fold cross validation was used to validate the model. Here, training takes place on the training set, followed by validation on the validation set, and finally testing on the test set. However, dividing the original dataset into three sets (train, validation, and test) significantly reduces the amount of data available for training. A process known as k-fold cross-validation was employed as a solution, where k is the number of folds, which was set to 10. The 10-fold cross validation performed the fitting procedure a total of ten times. The model was trained on 9 of the folds set selected at random and with the remaining fold used as a holdout set for validation. The process was repeated ten times, with the performance measure provided after each run. The average was then calculated. The test set was then evaluated after the parameters had been set.

3.5 Model Selection

Choosing the right model is essential since it has a direct impact on the precision and effectiveness of the final model. Drug design can make use of a number of ML algorithms, including DT, RF, SVM, and neural networks. The algorithm selected will rely on the size and complexity of the dataset. Understanding how the model generates its predictions is frequently crucial in drug creation. Moreover, when choosing a model, interpretability is also a crucial factor, and complex models like neural networks are typically more difficult to interpret than linear models, DT, and rule-based models. A variety of ML models were compared, including the Gaussian Process Regressor, DT, and many more, as shown in Table (3.3). Based on the results of the comparison, Extra Trees, DT and Gaussian Process Regressors were the top-performing models, and among them DT was selected and deployed, because of its ability to handle small dataset. DT models are also easy to interpret, don't need a lot of training data, and are computationally efficient.

Table 3.3: Classification models comparison table.

Model	Adjusted R-Squared	R-Squared	RMSE
ExtraTreesRegressor	0.84	0.91	0.09
ExtraTreeRegressor	0.84	0.91	0.09
DecisionTreeRegressor	0.84	0.91	0.09
GaussianProcessRegressor	0.84	0.91	0.09
XGBRegressor	0.84	0.91	0.09
MLPRegressor	0.76	0.87	0.11
RandomForestRegressor	0.64	0.81	0.14
HistGradientBoostingRegressor	0.52	0.74	0.16
LGBMRegressor	0.52	0.74	0.16
BaggingRegressor	0.52	0.74	0.16
GradientBoostingRegressor	0.47	0.72	0.17
SVR	0.35	0.65	0.18

3.5.1 Hyperparameter tuning for DT

Underfitting and overfitting conditions should be avoided when developing the optimal model. For the DT hyperparameter tuning is the process of determining the best values for max depth of the DT (one that is neither too small nor too large). Hyperparameters,

which are parameters defined before training the model as they are not learned from the data during training, can significantly affect how well the model performs. Hyperparameters can be done using different methods such as grid search, Bayesian optimization or random search. For our DT model grid search was used. Grid search method is a tuning approach that was used to find the best hyperparameter values. It is an exhaustive search carried out on a model's specific parameter values. For a given k number of folds, it determines all hyperparameter combinations. For the DT regressor model, the Grid search found the optimal values for maximum depth parameter of the model. The GridSearchCV function searches for one point and trains the model using the optimal value along with k-fold cross-validation. The GridSearchCV function returned 20 as the best value for the maximum depth of the DT model.

3.6 Model development

Supervised learning techniques were used to train the model using the labeled data as either active or inactive that was retrieved from PubChem bioassays. The DT regressor, which is in this case doing binary classification, constructs classification models in the shape of a tree structure. It incrementally breaks down a dataset into smaller and smaller subsets while also developing an associated DT. The final result is a tree containing leaf nodes and decision nodes, the predicted output is either active or inactive class in terms of the target protein 3CLpro.
For modeling and evaluation, Python 3.7. was used.

3.6.1 Time Complexity analysis

Building a DT takes O (N * K * log(N)) time, where N is the number of instances in the training dataset and K is the number of features.
When using a DT, k-fold cross-validation entails fitting the model k times on various subsets of the data and validating its performance on the remaining data. As a result, the time complexity of k-fold cross-validation with a DT is O (k * N * K * log(N)) times the time complexity of creating DT.

The time complexity of 10-fold cross-validation with a DT is O (N * K * log(N)), presuming a fixed value of k (in this case, k = 10). Cross-validation therefore has the

same computing cost as creating a single DT on the entire dataset.
The DT's time complexity for prediction is O(log(N)), which is extremely effective and makes DTs acceptable for usage with big datasets.

3.7 Model Evaluation

To evaluate the performance of the model, accuracy, precision, recall and F1-score measure were used as main metrics for evaluating the performance of the model. SK-learn was used to calculate these metrics.
Where accuracy is defined as percentage of accurately predicted values for a given dataset by the model. It is a proportion between the model's total number of predictions and the number of predictions that were correct.

$$Accuracy = (Number\ of\ Correct\ Predictions / Total\ Number\ of\ Predictions) * 100$$

While Precision is the ratio of correct positive predictions made by the model to all positive predictions made; a true positive is a positive prediction made correctly, whereas a false positive is a positive prediction made incorrectly.

$$Precision = TP / (TP + FP)$$

Recall or sensitivity is the ratio of true positives (TP) to the total of true positives and false negatives (FN).

$$Recall = TP / (TP + FN)$$

F1 score combines the model's precision and recall into a single value, which measures the model's overall performance. In our classification tasks, the F1 score is a good evaluation metric, particularly as the dataset is unbalanced, and the cost of false positives and false negatives is not equal.

Confusion matrix which allows for the visualization of the model's performance was also used to display true positives (TPs) and true negatives (TNs), false positives (FPs) and false negatives (FNs). For this context, TPs are the molecules that are active and were predicted as active, TNs are the molecules that are actually inactive and was predicted by the model as

inactive, FPs are the molecules that are actually inactive and were predicted as active by the model, and finally, FNs are the molecules that were predicted to be inactive but were actually active.

3.8 Model Optimization

The goal of model optimization is to enhance the model's performance. We can decrease overfitting and enhance the model's generalization capabilities by optimizing it. To make sure that the ML model works well, generalizes new data, and makes optimum use of computational and resource resources, model optimization is required. There were two model optimization techniques made to the DT model. The first one is Pruning, which is a technique that entails cutting off DT branches that don't help the model perform better on the validation data. The second method was the use of ensemble methods, where multiple models are trained independently and then combined to make a final prediction. The DT model was combined with SVM to create a more robust ML model. DTs can be used to identify key molecular features or substructures that are associated to a target protein. Support vector machines (SVMs) can capture the non-linear relationships between molecular features and the target protein bioactivity. Combining the two algorithms can enhance the model performance.

3.9 Model Deployment

The model was deployed in a web app, where compounds' smiles were passed into the app in the form of text file and the model predicts their activity with the target protein. The deployment of unknown dataset on the predictive model is a crucial step in finding new possible COVID-19 3CLpro enzyme inhibitors. The dataset was virtually screened using the established model, where 748 natural compounds were extracted from ZINC database [68], and then Lipinski's rule of five (RO5) was used to identify drug-like molecules during screening.

Chapter 4

Results and discussion

This chapter represents the results and findings of applying the DT model on the PubChem dataset and using 10-folds cross validation, it also shows the results of using completely new dataset, ChEMBL which was retrieved from different data source than that used to train the model, to validate the model performance. This chapter also shows the result after optimizing the model by applying pruning and ensemble methods to the model. Finally, this chapter will provide a discussion of the results.

4.1 Model Results

The DT model was used to predict the bioactivity of compounds for the inhibition of 3CLpro enzymes in COVID-19 infections. Bioactive chemicals data from five distinct experimental bioassays were extracted from PubChem and curated by removing duplicate compounds and salts, as a result, 400 distinct chemical structures were gathered for model building. To construct an easily reproducible model for the prediction and screening of unknown chemical compounds, compounds fingerprints were generated, where the structural and functional requirements of the 3CLpro enzyme are described by these molecular descriptors. There were 881 descriptors in the model and removing low variance features resulted in 148 descriptors using the from *sklearn* feature selection method *VarianceThreshold*. Based on both internal and external validation methodologies, the results revealed good predictive capacity. For quality and performance measurements different metrics were calculated on the original dataset as well as the external dataset. 10-folds cross validation was used since the data is not large. Figure (4.1) shows accuracy for each of the 10 folds for both training and validation data.

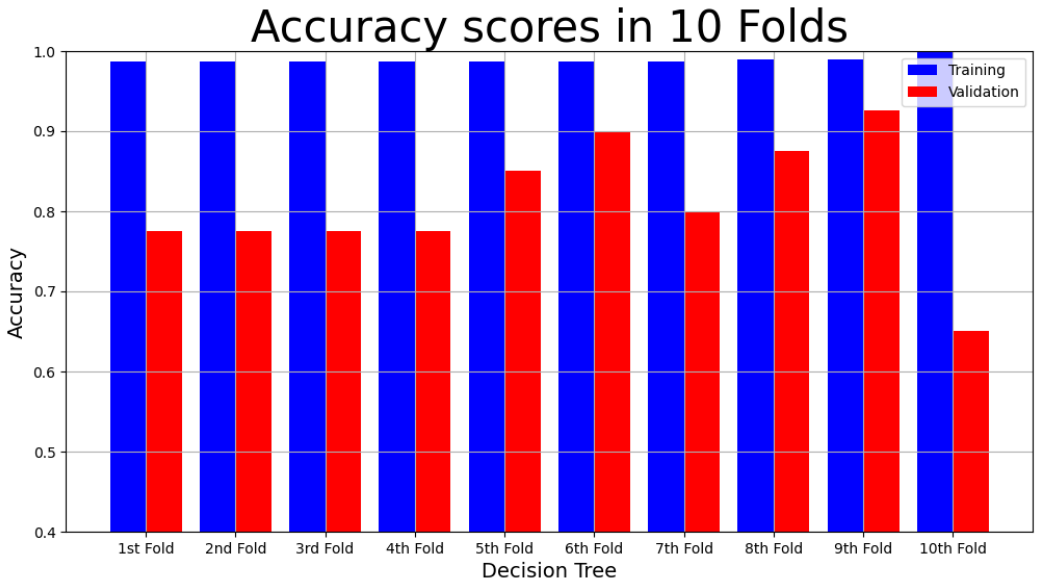


Figure 4.1: Accuracy scores in 10 folds.

As can be shown from Figure (4.1), the accuracy for the training data was very high, and for the validation data was also close to the accuracy of the training in most of the folds. Moreover, the model performance was evaluated after each run of the 10 folds, where accuracy, precision, recall and F1 score were measured for both training data and validation data.

Table 4.1: Model performance metrics for training data for each of the 10 folds.

	Training										**Mean**
Accuracy	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.989	0.989	1	**0.988**
Precision	0.928	0.928	0.928	0.928	0.928	0.928	0.930	0.952	0.952	1	**0.94**
Recall	0.951	0.951	0.951	0.951	0.951	0.951	0.952	0.952	0.952	1	**0.956**
F1-Score	0.9397	0.939	0.939	0.939	0.939	0.939	0.941	0.952	0.952	1	**0.948**

Table 4.2: Model performance metrics for validation data for each of the 10 folds

	Validation										**Mean**
Accuracy	0.8	0.825	0.85	0.8	0.825	0.925	0.825	0.875	0.925	0.725	**0.8375**
Precision	0.2857	0.333	0.333	0.2	0.4	0.67	0.28	0.43	0.67	0.11	**0.377**
Recall	0.4	0.4	0.2	0.2	0.8	0.8	0.5	0.75	0.5	0.25	**0.48**
F1-Score	0.333	0.364	0.25	0.2	0.533	0.727	0.3636	0.545	0.545	0.154	**0.41**

Table (4.1) shows the performance metrics for the training data, while Table (4.2) shows the performance metrics for the validation data in each run. The mean accuracy value for the training data is 0.988 and 0.837 for the validation data, while the mean value for the precision was 0.94 for training and 0.37 for validation and for recall it was 0.956 for training and 0.48 for validation.

4.2 Results after optimizing the Model

Figure (4.2) shows accuracy for each of the 10 folds for both training and validation data after applying ensemble methods to the model and Figure (4.3) shows accuracy for each of the 10 folds for both training and validation data after applying pruning. Table (4.3) shows a comparison of the result of the model performance after applying the pruning and ensemble methods to the model.

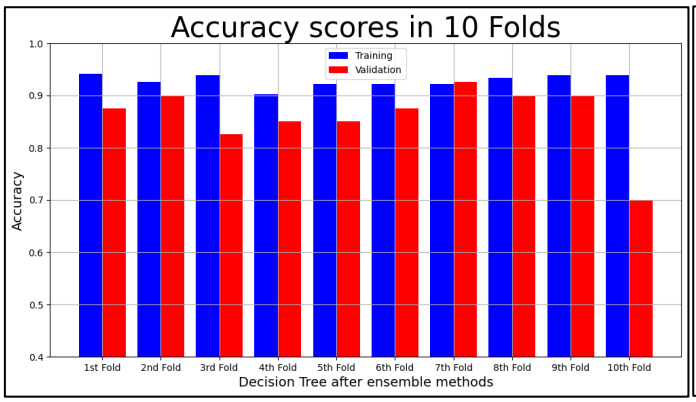


Figure 4.2: Accuracy scores in 10 folds after ensemble methods.

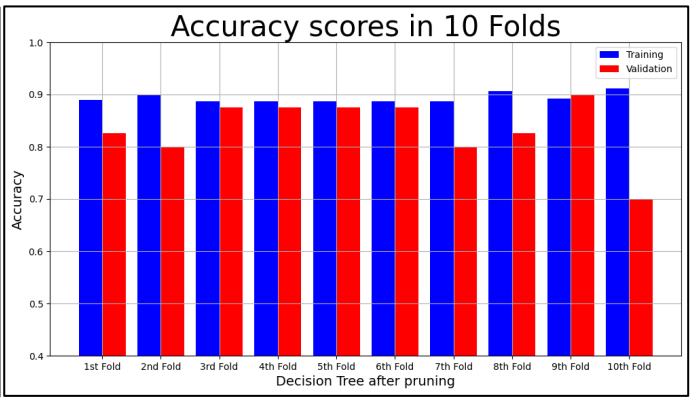


Figure 4.3: Accuracy scores in 10 folds after pruning.

Table 4.3: Ccomparison of the result of the model performance after applying the pruning and ensemble methods to the model.

	Original	After Pruning	After Ensemble Method
Accuracy	84%	89%	89%
Precision	0.38	0.43	0.75
Recall	0.48	0.67	0.6
F1-Score	0.41	0.52	0.67

4.3 External Validation Results

After performing the 10 folds cross validation, external dataset was retrieved from new data source ChEMBL which is different than the data source from which that data was retrieved for

model training (PubChem). The external data set was passed to the model and the model evaluation metrics were calculated. The classification report is shown in Figure (4.4), and confusion matrix is shown in Figure (4.5). As shown from the results, the model is predicting the bioactivity of the compounds with high accuracy.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	104
1	1.00	0.87	0.93	15
accuracy			0.98	119
macro avg	0.99	0.93	0.96	119
weighted avg	0.98	0.98	0.98	119

Figure 4.4: Classification report for the external (ChEMBL dataset).

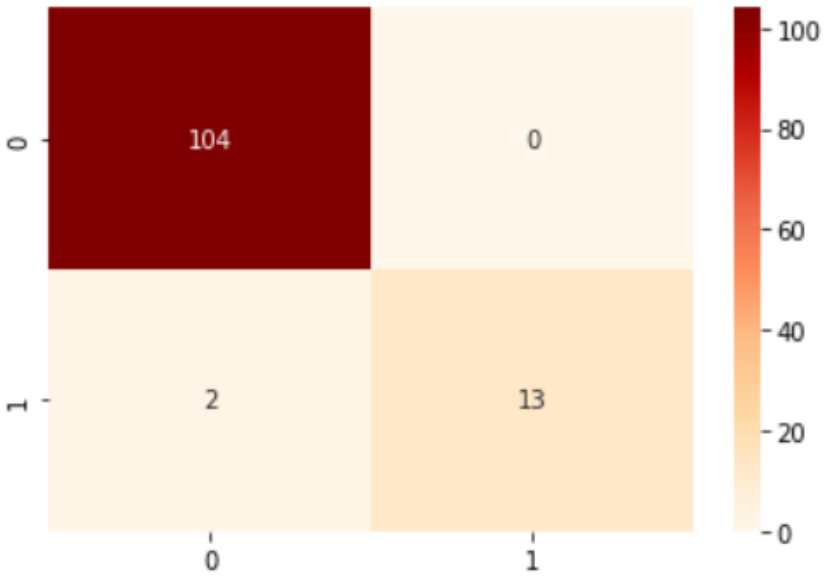


Figure 4.5: Confusion Matrix for the external (ChEMBL) dataset.

The classification model's accuracy in classifying data from different classes is summarized in a table like form which is the confusion matrix. The model's predicated label is on one axis of the confusion matrix, and the actual label is on the other. We can use confusion matrix to check how well the model predicted TPs and TNs. The model can be considered a high-performance model if it accurately predicted TPs and TNs which is shown clearly in the confusion matrix. The model predicted all inactive compounds correctly, and only predicted two of the active compounds as inactive as shown in Figure (4.5), where it shows that it has predicted all the 104

inactive compounds as inactive which represents the TNs, and from the 15 active compounds, it predicted 13 as active which represent the TPs, while miss predicting only 2 out of the 15 as inactive which represent FPs.

4.4 Screening Results

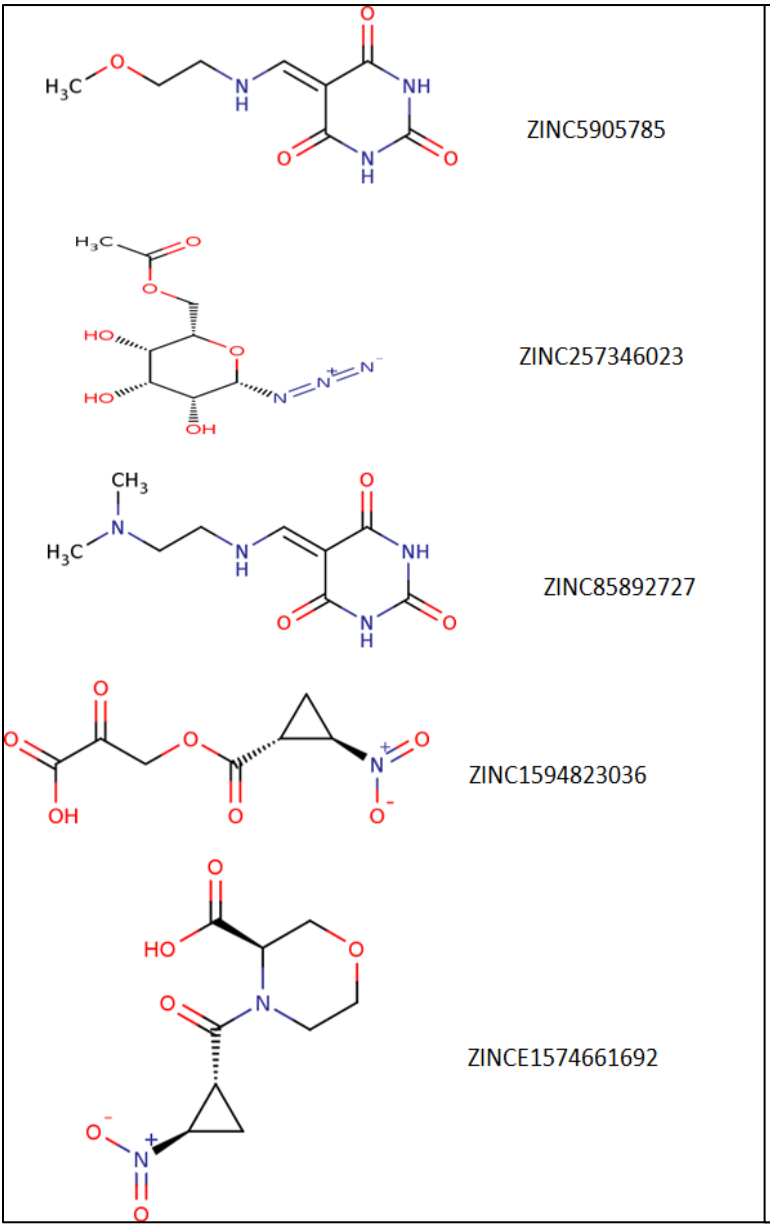
The model was deployed in a web application where the application takes molecules IDs and their SMILES, and it will return descriptors for these molecules as well as classification to whether or not the molecule has bioactivity to the 3CLpro. New unknown datasets from ZINC [70] were passed to it, and it has found anti-COVID-19 agents in 26 of 748 natural compounds from the ZINC natural product database. Moreover, the possibility of drug repurposing was also checked by passing FDA approved drugs data set to the model where it has found anti-COVID-19 agents in 1 out of 63 from the FDA approved drugs, as Acebutolol was classified as active to 3CLpro. Furthermore, Lipinski's RO5 was applied to all of the screened anti-COVID-19 compounds except for the FDA approved ones in order to prioritize drug-like compounds.

After applying Lipinski RO5 to the screened molecules, 25 out 26 compounds have the properties of drug like compounds, where these compounds can be used in clinical trials. The screened compounds that have drug like properties of Lipinski RO5 and their SMILES, molecular weight, log P, number of Hydrogen donors, and number of Hydrogen acceptors shown in Table (4.4). Figure (4.6-a) and Figure (4.6-b) shows each of the 25 screened molecules has their chemical structure.

Table 4.4: The screening of the active compounds with 3CLpro.

SMILES	Molecule ID	MW	LogP	NumHDonors	NumHAcceptors
COCCNC=C1C(=O)NC(=O)NC1=O	ZINC5905785	213.193	-1.5277	3	5
CC(=O)OC[C@@H]1O[C@H](N=[N+]=[N-])[C@H](O)[C@H](	ZINC257346023	247.207	-1.3326	3	7
CN(C)CCNC=C1C(=O)NC(=O)NC1=O	ZINC85892727	226.236	-1.6125	3	5
O=C(O)C(=O)COC(=O)[C@@H]1C[C@H]1[N+](=O)[O-]	ZINC1594823036	217.133	-1.1515	1	6
O=C(O)[C@H]1COCCN1C(=O)[C@@H]1C[C@H]1[N+](=O)[O	ZINC1574661692	244.203	-1.0364	1	5
CN1C(=O)CN(C(=O)[C@@H]2C[C@H]2[N+](=O)[O-])CC1=O	ZINC848048479	241.203	-1.5212	0	5
Cn1cnn(CC(=O)OCCC[N+](=O)[O-])c1=O	ZINC860920224	244.207	-1.2082	0	8
O=C(Cn1cc([N+](=O)[O-])cn1)N[C@H]1CNOC1	ZINC722472253	241.207	-1.1891	2	7
O=C(O)COC(=O)Cn1cnc([N+](=O)[O-])n1	ZINC1610784971	230.136	-1.1859	1	8
O=C(O)CNC(=O)CNC(=O)[C@@H]1C[C@H]1[N+](=O)[O-]	ZINC1602069684	245.191	-2.0314	3	5
O=[N+]([O-])c1cnc(NCCn2ncnn2)nc1	ZINC354596249	236.195	-0.5166	1	9
O=C(Cn1cnc([N+](=O)[O-])n1)O[C@H]1CNOC1	ZINC1326769379	243.179	-1.3671	1	9
O=C(O)CNC(=O)COC(=O)[C@@H]1C[C@H]1[N+](=O)[O-]	ZINC1606264945	246.175	-1.6044	2	6
NCC(=O)NCCn1cc([N+](=O)[O-])cn1	ZINC1118269651	213.197	-1.1338	2	6
CNC(=O)[C@H](O)CNc1ncc([N+](=O)[O-])cn1	ZINC758115244	241.207	-1.0964	3	7
O=C(O)[C@H]1COCCN1C(=O)[C@@H]1C[C@H]1[N+](=O)[O	ZINC1574661692	244.203	-1.0364	1	5
COC(=O)COCCNC(=O)[C@@H]1C[C@H]1[N+](=O)[O-]	ZINC862950467	246.219	-1.0426	1	6
Cn1cnn(CC(=O)OCCC[N+](=O)[O-])c1=O	ZINC860920224	244.207	-1.2082	0	8
O=C(Cn1cc([N+](=O)[O-])cn1)N[C@H]1CNOC1	ZINC722472253	241.207	-1.1891	2	7
CN(CCNC(=O)[C@@H]1C[C@H]1[N+](=O)[O-])CC(=O)O	ZINC1598492128	245.235	-1.2158	2	5
O=C(O)[C@@H]1CN(C(=O)[C@@H]2C[C@H]2[N+](=O)[O-])	ZINC1594947991	244.203	-1.0364	1	5
O=C(O)[C@@H]1COCCN1C(=O)[C@@H]1C[C@H]1[N+](=O)	ZINC378180195	244.203	-1.0364	1	5
CNCCNC(=O)Cn1cnc([N+](=O)[O-])n1	ZINC1326581486	228.212	-1.4781	2	7
O=C(O)CNC(=O)COC(=O)[C@@H]1C[C@H]1[N+](=O)[O-]	ZINC1606264945	246.175	-1.6044	2	6
O=C(O)[C@H]1CN(C(=O)[C@@H]2C[C@H]2[N+](=O)[O-])CC	ZINC1594947990	244.203	-1.0364	1	5

In the above table if we take a look at any of the rows, we can see that the molecular weight (MW) is less than 500 Da, the hydrogen bond acceptors are all less than 10, all the hydrogen bond donors are less than 5 and partition coefficients are less than 5. These values are all in accordance to Lipinski's RO5, so all of these molecules can be drugs and have the drug-like properties.



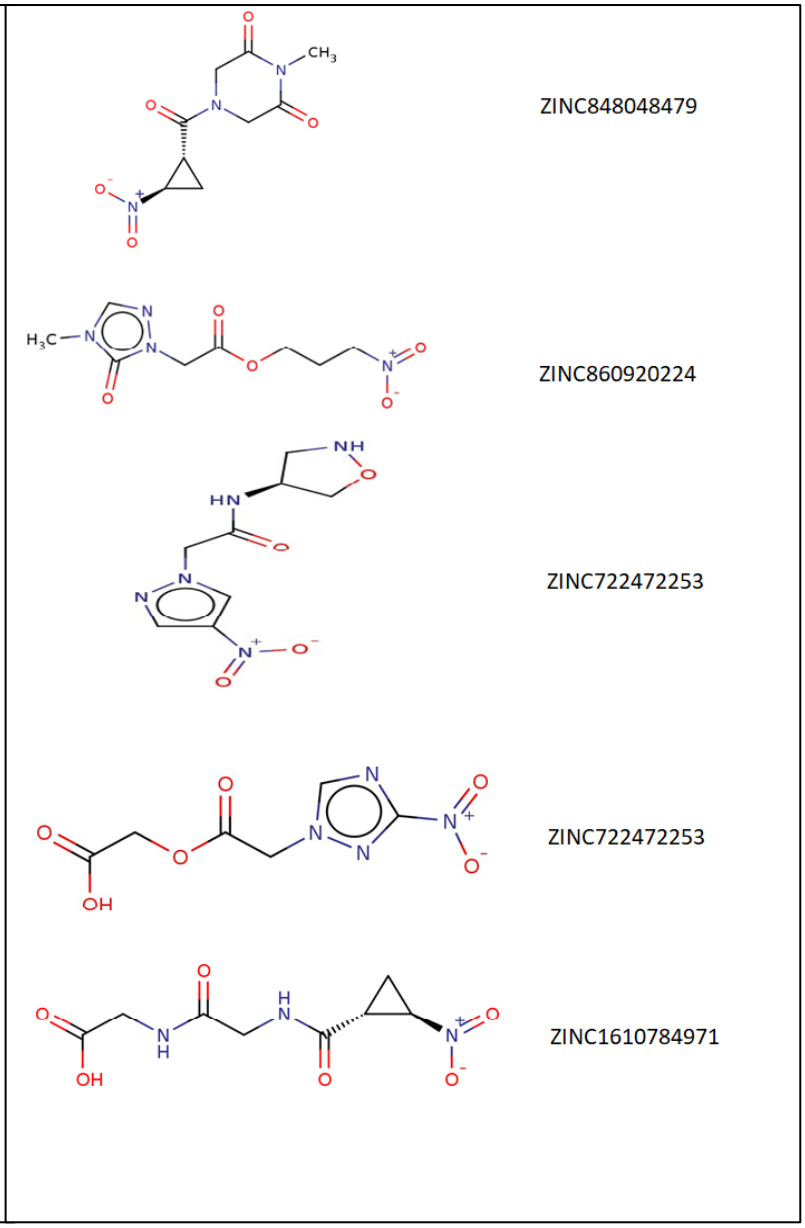


Figure 4.6-a: Screened compounds and their chemical structure.

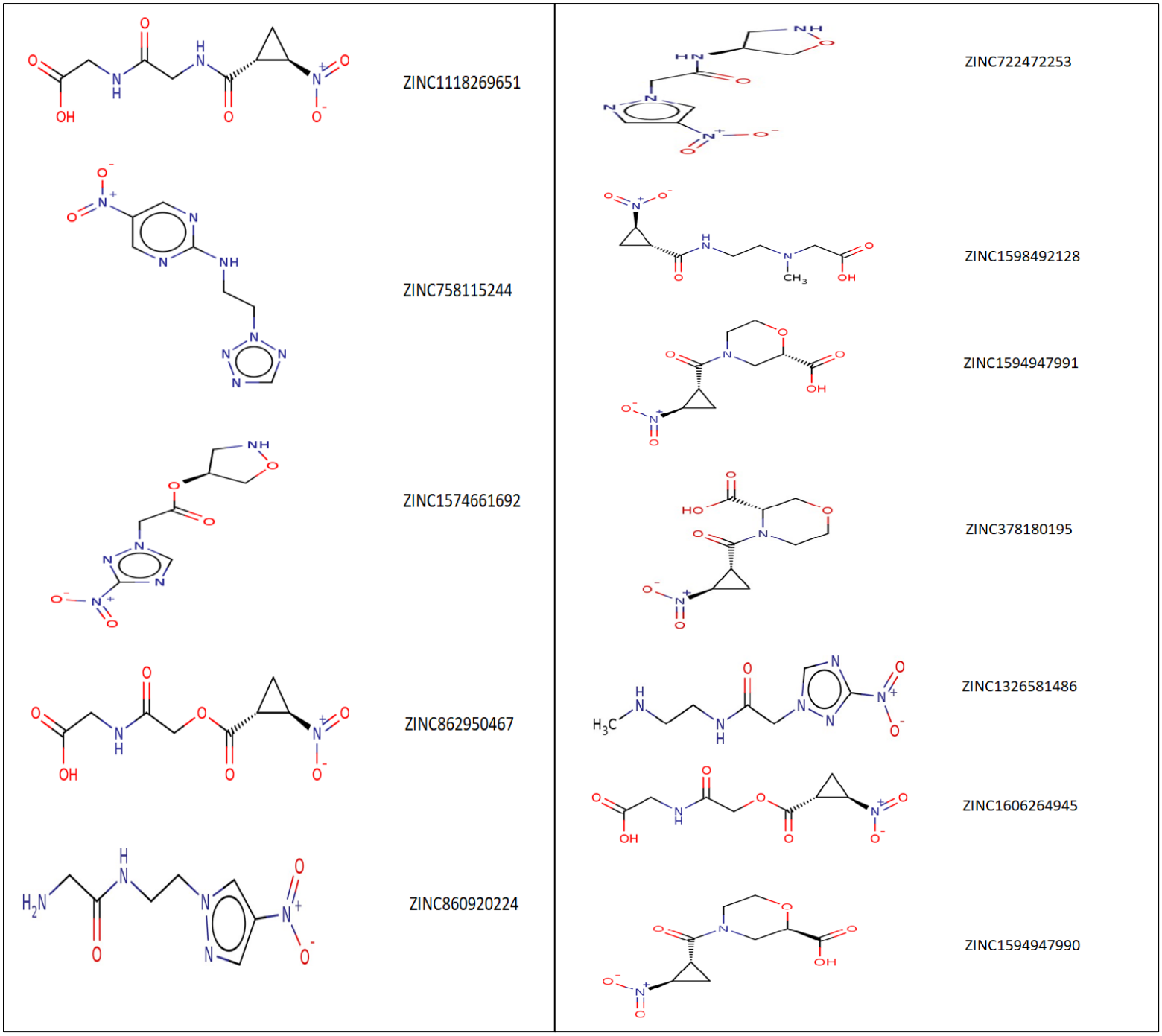


Figure 4.6-b: Screened compounds and their chemical structure

4.5 Discussion

This study was performed to deploy ML model, which in this case based on DT regressor in order to discover drugs for COVID-19 virus by predicting the bioactivity of unknown molecules with the 3CLpro enzyme. The seed values that were used is 20 for the max depth of

the DT tree, 10 for the number of folds for cross validation.
The model has shown good results with high accuracy for training and validation set, while the precision and recall for the training set was much higher than that of the validation set, which indicates that it is likely that the model is overfitting to the training data or due to class imbalance, as the dataset had much more samples of the inactive class than the active class.

As for the external validation set the results show very high performance on external data that was never exposed to the model. It is interesting to see that the model performed much better on external validation more than internal validation and this could be due to many reasons, such as that the data sets are not equally sampled and distributed, differences in the quality of the validation datasets, as in some cases, the DT model may be able to fit the noise or outliers in one dataset but not in another, leading to different performance levels, and differences in the pre-processing of the two dataset, as in the external validation dataset, we used the negative logarithmic scale which is -log10(IC50) which gave us even distribution of the data. So, although 10 folds cross validation was used to avoid the problem of overfitting, the model was still facing this due to the data. The DT model may be overfitting if it is too sophisticated for the quantity of training data provided, which is one cause. The model might be able to remember the training data in this situation and fit the data noise, but it won't be able to generalize to new data that it hasn't seen before. The use pruning and ensemble methods combining the DT with SVM enhanced the model performance and reduced the overfitting of the model.

When comparing this study findings with previous studies' findings, we could see similar results in terms of discovering new compounds although different machine learning models were used, which leads us to conclude that regardless of the machine learning model deployed, AI can proof to be effective in discovering bioactive unknown molecules with the specified target protein. However, deploying different machine learning models than other studies has resulted in different performance metrics, and different and new compounds discovered as each study has different data source.

When comparing the finding with Kumari M, Subbarao N study that developed CNN model to predict bioactivity for the same target protein 3CLpro, their model was trained on 282 compounds and predicted an external validation test set of 141 compounds with an accuracy of 0.86, a precision of 0.73, a recall of 0.45, an F1-score of 0.55. while our model showed 0.89

accuracy, precision of 0.75, recall of 0.6 and f1 score of 0.67 for internal validation. While the finding of previous study by Mody, V., Ho, J. [69] which has investigated 47 FDA approved drugs that inhibit the SARS-COV-2 3CLpro enzymatic activity were used, as that study performed an in vitro enzymatic inhibitory assay using commercially available assay kits, and has identified that boceprevir, paritaprevir, and tipranavir were able to partially inhibit the 3CLpro enzymatic activity at 50 μM drug concentration while PIs lopinavir and ritonavir did not exhibit any 3CLpro inhibitory activity. These drugs were passed to our model and among the three of them the paritaprevir was predicted to be active, the PIs lopinavir and ritonavir were predicted to be in-active using our model which agrees with the previous study findings as well.

Chapter 5

Conclusion and future work

5.1 Conclusion

This thesis developed ML model that is based on DT regressor to predict the bioactivity of unknown compounds with 3CL-Protease, it has successfully identified new compounds that could be candidates for the inhibition of 3CLpro enzyme that could be used in clinical trials. The DT model's performance was compared to various classification approaches before deployment such as RF, MLP, SVR and other regression models and has shown better performance than others. The DT model was trained and validated on 400 compounds from PubChem bioassays and 119 compounds from ChEMBL database were used for external validation. The model implemented 10 folds cross validation, and although the model showed very good accuracy of 0.84 for internal validation, it had poor precision, recall and f1-score with values of 0.37, 0.48 and 0.41 respectively, which indicated overfitting, the performance was enhanced by applying DT pruning and ensembled methods combing the DT with SVM, this resulted in 0.89 accuracy, precision of 0.75, recall of 0.6 and f1- score of 0.67 for internal validation. On the other hand, the model should an excellent performance for the external dataset with 0.98 accuracy, 0.99 precision, recall of 0.93 and f1-score of 0.96. The model was deployed in a web application where new unknown datasets were passed to it, resulting in finding 26 bioactive molecules out of 748 unknown compounds, in which 25 molecules could be used as drug as they have the drug-like properties of Lipinski rule. These compounds could be used in clinical trials to test their effectiveness on the virus. The number of discovered compounds could certainly increase if more unknown molecules are passed to the model. The chemical structures of FDA approved drugs were also passed to the model, where their bioactivity with the 3CLpro was observed in previous studies and the result of model has agreed with this observation to some extent.

The limitation of this work was in the small amount of data that was used for training the model, and since the model is based on supervised learning, the amount of data might have caused overfitting in the model, and this was due to the limited amount of data available in the published bioassays in relation to the target protein.

In conclusion, quality, size, preprocessing and mostly the distribution of the validation datasets

can affect the model performance greatly. DTs are a useful tool in drug discovery as they are relatively easy to interpret and visualize, However, they are not always the most accurate or effective method for every task, and by combining DTs and SVMs, it is possible to create a more robust and accurate machine learning model for drug discovery. The DT can capture important features of the compounds, while the SVM can learn more complex relationships between these features and biological activity.

This thesis shows that drug development can benefit greatly from the use ML models that can dramatically improve drug development while reducing human engagement in medical practice, however, the developed ML models still must be deployed in real world to study their effectiveness.

5.2 Future Work

There are several potential directions for future work on the model, such as increasing the data size by searching for new data sources or bioassays in regard to the target protein. Another method of improving the model is by doing feature engineering to identify the features that might be crucial for drug discovery. This entails combining feature importance analysis and domain knowledge to identify the features that capture important chemical properties. Exploring deep learning techniques like CNN in combination with the current model is another way to improve the model and increase the model's accuracy and predictive potential. In addition, it's critical to make sure the model is easily understood and able to provide insight on how it generates predictions, to better understand how various features and models contribute to the final prediction, this could entail implementing approaches like DT visualization.

References

[1] Song, C.M.; Lim, S.J.; Tong, J.C. Recent advances in computer-aided drug design. Brief. Bioinform, 2009.

[2] Lavecchia, A.; di Giovanni, C. Virtual screening strategies in drug discovery: A critical review. Curr. Med. Chem, 2013.

[3] Moore, T.J.; Zhang, H.; Anderson, G.; Alexander, G.C. Estimated costs of pivotal trials for novel therapeutic agents approved by the us food and drug administration, 2015–2016, 2018.

[4] A. Chandra Kaushik and U. Raj, "AI-driven drug discovery: A boon against COVID-19?", 2020.

[5] 6 Things we learned about artificial intelligence in drug discovery from 330 scientists (n.d.). https://blog.benchsci.com/6-things-we-learned-about-artificial-intelligence-in-drug-discovery-from-330-scientists. (Accessed 4 March 2022).

[6] Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S.H. Structure-based virtual screening for drug discovery: A problem-centric review. AAPS J, 2012.

[7] Swinney, D.C.; Anthony, J., "How were new medicines discovered?", 2011.

[8] Lionta, E.; Spyrou, G.; Vassilatis, D.K.; Cournia, Z. Structure-based virtual screening for drug discovery: Principles, applications and recent advances, 2014.

[9] Batool, M.; Choi, S. Identification of druggable genome in staphylococcus aureus multidrug resistant strain, 2017.

[10] Walters, W.P., Murcko, M. Assessing the impact of generative AI on medicinal chemistry. Nat. Biotechnol. 38, 143–145, 2020.

[11] C.W. Lin, F.J. Tsai, C.H. Tsai, C.C. Lai, L. Wan, T.Y. Ho, C.C. Hsieh, P.D. Chao, AntiSARS coronavirus 3C-like protease effects of lsatis indigotica root and plant derived phenolic compounds, Antivir. Res. 68, 2005.

[12] S. Chen, L.L. Chen, H.B. Luo, T. Sun, J. Chen, F. Ye, J.H. Cai, J.K. Shen, X. Shen, H.L. Jiang, Enzymatic activity characterization of SARS coronavirus 3C-like protease by fluorescence resonance energy transfer technique, 2005.

[13] R. Ramajayam, K.P. Tan, H.G. Liu, P.H. Liang, Synthesis and evaluation of pyrazolone compounds as SARS-coronavirus 3C-like protease inhibitors, 2010.

[14] V. Kumar, K. Roy, Development of a simple, interpretable and easily transferable QSAR model for quick screening antiviral databases in search of novel

3C-like protease (3CLpro) enzyme inhibitors against COVID-19 diseases, SAR QSAR Environ, 2020.

[15] Álvarez-Machancoses, Óscar, and Juan Luis Fernández-Martínez. "Using artificial intelligence methods to speed up drug discovery." Expert opinion on drug discovery 14.8, 2019.

[16] Swinney, D.C.; Anthony, J. How were new medicines discovered?, 2011.

[17] C.W. Lin, F.J. Tsai, C.H. Tsai, C.C. Lai, L. Wan, T.Y. Ho, C.C. Hsieh, P.D. Chao, AntiSARS coronavirus 3C-like protease effects of lsatis indigotica root and plant derived phenolic compounds, Antivir, 2005.

[18] Chan, H.C.S., Shan, H., Dahoun, T., Vogel, H., Yuan, S., Advancing drug discovery via artificial intelligence. Trends Pharmacol, 2019.

[19] Duch, W.; Swaminathan, K.; Meller, J. Artificial intelligence approaches for rational drug design and discovery. Curr. Pharm. Des,2007.

[20] Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vectormachine and artificial neural network systems for drug/nondrug classification. J. Chem. Inf. Comput. Sci, 2003.

[21] Zernov, V.V.; Balakin, K.V.; Ivaschenko, A.A.; Savchuk, N.P.; Pletnev, I.V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. J. Chem. Inf. Comput,2003.

[22] Warmuth, M.K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. J. Chem. Inf. Comput,2003.

[23] Jorissen, R.N.; Gilson,M.K. Virtual screening of molecular databases using a support vector machine. J. Chem. Inf.Model,2005.

[24] Koohy, H. The rise and fall of machine learning methods in biomedical research, 2012.

[25] Young, J.D.; Cai, C.; Lu, X. Unsupervised deep learning reveals prognostically relevant subtypes of glioblastoma. BMC Bioinform, 2017.

[26] Popova,M.; Isayev,O.; Tropsha,A.Deep reinforcement learning for de novo drug design. SciAdv, eaap7885, 2018.

[27] Lo, Y.C.; Rensi, S.E.; Torng, W.; Altman, R.B. Machine learning in chemoinformatics and drug discovery.Drug Discovy, 2018.

[28] Ma, X.H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z.R.; Chen, Y.Z. Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. COMB Chem. High Throughput Screen, 12, 344–357, 2009.

[29] Han, L.Y.; Ma, X.H.; Lin, H.H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z.R.; Cao, Z.W.; Ji, Z.L.; Chen, Y.Z. A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. J. Mol. Graph. Model, 26, 1276–1286, 2008.

[30] Liu, X.H.; Ma, X.H.; Tan, C.Y.; Jiang, Y.Y.; Go, M.L.; Low, B.C.; Chen, Y.Z. Virtual screening of abl inhibitors from large compound libraries by support vector machines, 2009.

[31] Prieto-Martínez, F.D.; López-López, E.; Eurídice Juárez-Mercado, K.; Medina-Franco, J.L. Chapter2—computational drug design methods—current and future perspectives, 2019.

[32] Cox, D.R.; Kartsonaki, C.; Keogh, R.H. Big data: Some statistical issues. Stat. Probab, 2018.

[33] P. Hop, B. Allgood, J. Yu, Geometric deep learning autonomously learns chemical features that outperform those engineered by domain experts, 2018.

[34] Hartenfeller, M.; Schneider, G. De novo drug design,2011.

[35] Richardson, J.S.; Richardson, D.C. The de novo design of protein structures, 2011.

[36] Lameijer, E.W.; Tromp, R.A.; Spanjersberg, R.F.; Brussee, J.; Ijzerman, A.P. Designing active template molecules by combining computational de novo design and human chemist's expertise. J. Med. Chem, 2007.

[37] Gillet, V.J. New directions in library design and analysis. Curr. Opin. Chem. Biol., 372–378, 2008.

[38] Prada-Gracia, D.; Huerta-Yepez, S.; Moreno-Vargas, L.M. Application of computational methods for anticancer drug discovery, design, and optimization. Bol. Med. Hosp. Infan.t Mex, 2016.

[39] Meng, X.Y.; Zhang, H.X.; Mezei, M.; Cui, M. Molecular docking: A powerful approach for structure-based drug discovery. Curr. Comput. Aided Drug Des. 146–157, 2011.

[40] Kitchen, D., Decornez, H., Furr, J. et al. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3, 935–949, 2004.

[41] Huang, S.Y.; Zou, X. Advances and challenges in protein-ligand docking. Int. J. Mol. Sci 2010, 11, 3016–3034, 2004.

[42] Lopez-Vallejo, F.; Caulfield, T.; Martinez-Mayorga, K.; Giulianotti, M.A.; Nefzi, A.; Houghten, R.A.; Medina-Franco, J.L. Integrating virtual screening and combinatorial chemistry for accelerated drug discovery. Comb. Chem. High. Throughput Screen, 2011.

[43] Kapetanovic, I.M.Computer-aided drug discovery and development (caddd): Insilico-chemico-biological approach. Chem. Biol. Interact, 2008.

[44] Ain, Q.U.; Aleksandrova, A.; Roessler, F.D.; Ballester, P.J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdiscip. Rev. Comput Mol, 405–424, 2015.

[45] Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C.R. Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. Br. J. Pharm. 153, 7–26, 2008.

[46] Huang, S.Y.; Grinter, S.Z.; Zou, X. Scoring functions and their evaluation methods for protein-ligand docking: Recent advances and future directions. Phys. Chem. Chem. Phys, 2010.

[47] David, H.; Gary, B.F. Computational intelligence methods for docking scores. Curr. Comput. Aided Drug Des., 2009.

[48] Sousa, S.F.; Fernandes, P.A.; Ramos, M.J. Protein-ligand docking: Current status and future challenges. Proteins, 2006.

[49] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: A review," Chaos, Solitons and Fractals, vol. 139, 2020.

[50] E. N. Muratov, "A critical overview of computational approaches employed for COVID-19 drug discovery," Chem. Soc. Rev., vol. 50, no. 16, pp. 9121–9151, 2021.

[51] M. Batool, B. Ahmad, and S. Choi, "A structure-based drug discovery paradigm," Int. J. Mol. Sci., vol. 20, no. 11, 2019.

[52] N. Gianchandani, A. Jaiswal, D. Singh, V. Kumar, M. Kaur, Rapid COVID-19 diagnosis using ensemble deep transfer learning models from chest radiographic images, 2020.

[53] Singh D, Kumar V, Kaur M. Densely connected convolutional networks-based COVID-19 screening model, 2021.

[54] kumari, Madhulata. "Evaluation of Predictive Models Based on Random Forest, and Support Vector Machine Classifiers and Virtual Screening of Anti-Mycobacterial Compounds." International Journal of Computational Biology and Drug Design, 2017.

[55] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, Q. Chen, S. Huang, M. Yang, X. Yang, S. Hu, Y. Wang, X. Hu, B. Zheng, K. Zhang, H. Wu, Z. Dong, Y. Xu, Y. Zhu, X. Chen, H. Yu, Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography, Sci. Rep. 10 19196, 2020.

[56] J. Peng, J. Li, X. Shang, A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network, BMC Bioinf. 21, 2020.

[57] S. Hu, C.P. Chen, J. Zhang, B. Wang, Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks, BMC Bioinf. 20- 689, 2019.

[58] J.G. Meyer, S. Liu, l.J. Miller, J.J. Coon, A. Gitter, Leaming drug functions from chemical structures with convolutional neural networks and random forests, J. Chem. Inf. Model. 59- 4438-4449, 2019.

[59] Madhulata Kumari, Naidu Subbarao, Deep learning model for virtual screening of novel 3C-like protease enzyme inhibitors against SARS coronavirus diseases, 2021.

[60] Kim, Sunghwan, "PubChem in 2021: new data content and improved web interfaces.", 2021.

[61] Kim S, Thiessen PA, Cheng T, Yu B, Bolton EE. An update on PUG-REST: RESTful interface for programmatic access to PubChem. Nucleic Acids Res. 2018, https://pubchempy.readthedocs.io/en/latest/guide/contribute.html (Accessed 22 March 2022).

[62] Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR. (2017) 'The ChEMBL database in 2017.', 2017.

[63] Y. Zhou, F. Wang, J. Tang, R. Nussinov, and F. Cheng, "Artificial intelligence in COVID-19 drug repurposing," Lancet Digit. Heal., vol. 2, no. 12, pp. e667–e676, 2020.

[64] J. Cheminform Willighagen, The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching, 2017.

[65] Benet LZ, Hosey CM, Ursu O, Oprea TI: BDDCS, the Rule of 5 and drugability,2016.

[66] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, 2001.

[67] Chava L Ramspek, Kitty J Jager, Friedo W Dekker, Carmine Zoccali, Merel van Ddiepen, "External validation of prognostic models: what, why, how, when and where?" 2021.

[68] Irwin, J. J., & Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening, 2005.

[69] Mody, V., Ho, J., Wills, S., Identification of 3-chymotrypsin like protease (3CLpro) inhibitors as potential anti-SARS-CoV-2 agents. Commun Biol 4, 93 (2021). https://doi.org/10.1038/s42003-020-01577-x (Accessed 28 December 2022).

[70] Sterling T, Irwin JJ ZINC 15—ligand discovery for everyone. J Chem Inf Model 55:2324–2337. https://doi.org/10.1021/acs.jcim.5b00559, 2015.

نموذج التعلم الآلي لاكتشاف عقار كوفيد -19 باستخدام الذكاء الاصطناعي

اعداد : كلاوديا الياس رأفت علاوي

اشراف: د. رشيد جيوسي و د. يوسف نجاجرة

ملخص

كان فيروس كورونا مشكلة كبيرة تواجه العالم، ولا يزال تطوير دواء فعال للفيروس قيد البحث. ومع ذلك، فإن تطوير دواء جديد عملية طويلة ومكلفة وقد تستغرق سنوات عديدة. يمكن أن يلعب الذكاء الاصطناعي دورًا حيويًا في اكتشاف الأدوية بشكل أسرع وأكثر فعالية من حيث التكلفة. إن البروتياز الأساسي الضروري الازم لتضاعف و نمو الفيروس هو إنزيم البروتيز الرئيسي (3CLpro). في هذه الأطروحة، تم تطوير نموذج التعلم الآلي الذي يمكن استخدامه للتنبؤ بالنشاط المثبط للبروتيز الرئيسي من خلال تطبيق شجرة القرار. تم الحصول على الواصفات التي تمثل الجزيئات الكيميائية باستخدام برنامج واصف PADEL، وتم إدخال هذه الواصفات في نموذج شجرة القرار لتدريبها والتنبؤ بالنشاط الحيوي لمركبات غير معروفة مع البروتين المستهدف. تم تحسين النموذج باستخدام طرق التقليم والتجميع، حيث تم دمج شجرة القرار مع آلة المتجه الداعم لتحسين أداء النموذج. ركز البحث على كلا النهجين الخارجي والداخلي للتحقق من أداء النموذج. اكتشف النموذج بنجاح 26 مركبًا غير معروف من مصدر بيانات منتج الزنك الطبيعي الذي أظهر نشاطًا حيويًا مع البروتين المستهدف. علاوة على ذلك ، تم تطبيق قاعدة (RO5) Lipinski لتحديد المركبات التي تصلح ان تكون عقاقير و لها خصائص العقاقير مما انتج عن 25 من المركبات المكتشفة التي لها خصائص شبيهة بالعقاقير ويمكن استخدامها في التجارب السريرية. تم التحقق من صحة النموذج باستخدام التحقق المتقاطع المتكون من 10 تقاطعات وتم التحقق من صحته أيضًا باستخدام مجموعة بيانات خارجية من مصدر بيانات مختلف عن مصدر البيانات المستخدم في تدريب النموذج ، وقد أثبت النموذج المقترح فعاليته على كل من مجموعات البيانات الخارجية والداخلية و لكن أظهر النموذج أداءً أعلى في البيانات الخارجية. فقد كانت النتائج بدقة 0.89 ، و إحكام 0.75 ، واسترجاع 0.6 و النتيجة الكاملة للكفائة ودقة النموذج 00.67 للتحقق الداخلي ، بينما بالنسبة للتحقق الخارجي 0.98 دقة، و إحكام 0.99، واسترجاع 0.93 و النتيجة الكاملة للكفائة ودقة النموذج 0.96. مقارنة بالدراسات المماثلة التي تستخدم التعلم العميق ، أظهر نموذج التعلم الآلي لدينا أداءً أفضل. في الختام، يمكن أن يكون النموذج المقترح مفيدًا في اكتشاف الأدوية لمركبات جديدة لفيروس كورونا.