Development of Machine Learning Models for COVID-19 Drug Discovery

A Report Submitted

in Partial Fulfillment of the Requirements

for the Degree of

Bachelor of Technology

in

Computer Science & Engineering

by

Satvik Pandey, Vivek Kumar Meena, Varughese C.V.

to the

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
MOTILAL NEHRU NATIONAL INSTITUTE OF TECHNOLOGY
ALLAHABAD PRAYAGRAJ

October, 2022

UNDERTAKING

I declare that the work presented in this report titled "*Development of Machine Learning Models for COVID-19 Drug Discovery*", submitted to the Computer Science and Engineering Department, Motilal Nehru National Institute of Technology Allahabad, Prayagraj, for the award of the ***Bachelor of Technology*** degree in ***Computer Science & Engineering***, is my original work. I have not plagiarized or submitted the same work for the award of any other degree. In case this undertaking is found incorrect, I accept that my degree may be unconditionally withdrawn.

October, 2022
Allahabad

(Satvik Pandey, Vivek Kumar Meena, Varughese C.V.)

CERTIFICATE

Certified that the work contained in the report titled "*Development of Machine Learning Models for COVID-19 Drug Discovery*", by *Satvik Pandey, Vivek Kumar Meena, Varughese C.V.*, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

(Dr. Pragya Dwivedi)
Computer Science and Engineering Dept.
M.N.N.I.T, Allahabad

October, 2022

Preface

The worldwide effects of COVID-19 spanning more than 200 nations, brought global pressure to speed up treatment options for it. In today's scenario, the drug discovery for COVID-19 is the biggest challenge in front of the whole world. Machine learning has proved itself as a prominent field of study over the last decade by solving many very complex and sophisticated real-world problems. The application areas included almost all the real-world domains such as healthcare, autonomous vehicle, business applications, natural language processing, intelligent robots, gaming, climate modeling, voice, and image processing. Therefore, here we designed a project related to the machine learning based detection system for COVID-19 drug identification. In this project we will develop a robust predictive model for chemical compound screening against COVID-19 disease.

Acknowledgements

Our sincere gratitude goes to our mentor Dr. Pragya Dwivedi for her continuous guidance and support during this process. Her motivation, enthusiasm and insight into the subject has always made us realize and understand the subject in a broader perspective. We would like to thank Prof. D.S. Kushwaha Head of Department of Computer Science, M.N.N.I.T, Allahabad for the providing necessary facility and support.We duly acknowledge the laboratory of computer science department for providing excellent computational facilities of desktops, software's and internet. We express our warm thank to our seniors and our class fellows for their wonderful co-operation and supportive behaviour. It was always helpful to discuss our ideas with them. Our parents and our family deserve a particular note of thanks: your wise counsel and kind words have, as always, served us well.

Contents

Chapter 1

Introduction

In December 2019, the novel SARS-CoV-2 Coronavirus appeared to initiate a pandemic of respiratory disease known as COVID-19, which proved to be a tricky disease that can occur in different forms and degrees of severity ranging from mild to serious with the possibility of organ failure and death. From moderate, self limiting respiratory disease to extreme progressive pneumonia, failure of multiple organs, and death.As the pandemic progresses and the number of confirmed cases and patients suffering serious respiratory failure and cardiovascular problems rises, these are clear reasons on why the effects of this viral infection are of great concern. Significant attention has been paid to define effective approaches to find solutions to the problems associated with COVID-19.

1.1 Motivation

Machine learning is one of the emerging fields in computer science. Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine

develop it's own algorithm, rather than having human programmers specify every needed step. Due to the huge amount of biological and medical data available today, along with well established machine learning algorithms, the design of largely automated drug development pipelines can now be envisioned. These pipelines may guide, or speed up, drug discovery; provide a better understanding of diseases and associated biological phenomena; help planning preclinical wet-lab experiments, and even future clinical trials. This automation of the drug development process might be key to the current issue of low productivity rate that pharmaceutical companies currently face.

Chapter 2

Work Plan

2.1 Methodology:

The Methodology of Proposed Work Include:

2.1.1 Data collection:

The data set used in the study is retrieved from the ChEMBL database.The database exhibits manually curated database of bioactive molecules with drug-like properties including COVID-19 related data. The database provides data set file in machine readable format e.g., tsv and csv formats. The dataset contains COVID-19 active compound structure, its druggable property (pChEMBL value) and several other information regarding its depositor and publication. The structural properties also known as descriptors or features for each compound was calculated by using PaDEL software (http://www.yapcwsoft.com/dd/padeldescriptor/). The software can calculate ¿1000 features as integer-value.

2.1.2 Dataset preprocessing:

Here after feature calculation the data get transformed so that the machine can easily parse it. The methods include data quality assessment and dimensionality reduction. For data quality assessment the dataset was checked for missing values, inconsistent data and duplicate data points.

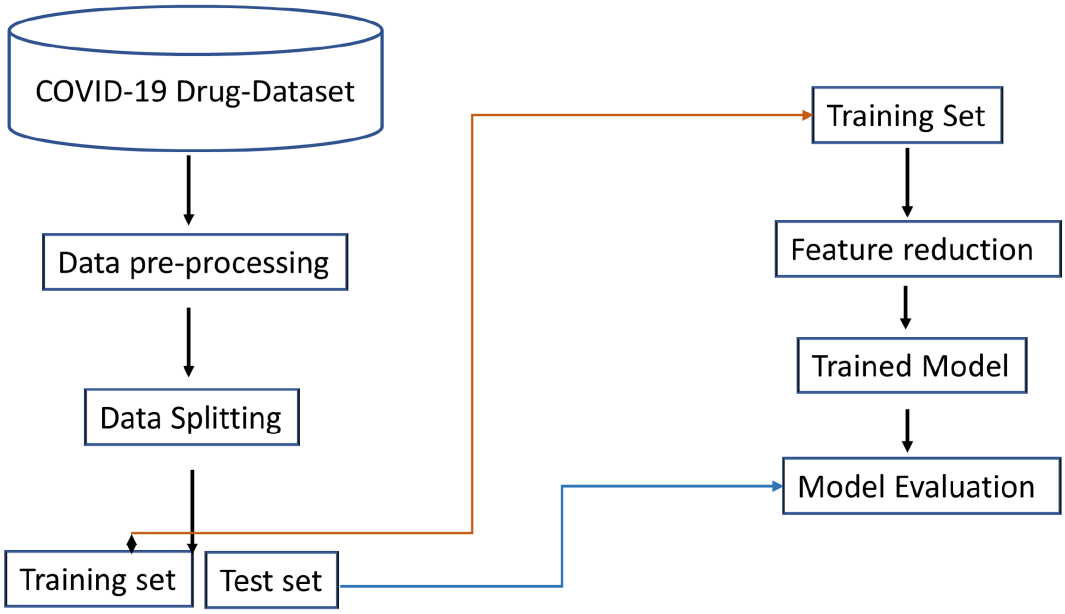

Figure 1: Workflow of Proposed Project

2.1.3 Feature reduction:

- *Missing Values Ratio* -

Data columns with too many missing values are unlikely to carry much useful information. Thus, data columns with number of missing values greater than a given threshold can be removed. The higher the threshold, the more aggressive the reduction.

- *Low Variance Filter* -

Similarly, to the previous technique, data columns with little changes in the data carry little information. Thus, all data columns with variance lower than a given threshold are removed. A word of caution: variance is range dependent; therefore, normalization is required before applying this technique.

- *High Correlation Filter -*

Data columns with very similar trends are also likely to carry very similar information. In this case, only one of them will suffice to feed the machine learning model. Here we calculate the correlation coefficient between numerical columns and between nominal columns as the Pearson's Product Moment Coefficient and the Pearson's chi square value respectively. Pairs of columns with correlation coefficient higher than a threshold are reduced to only one. A word of caution: correlation is scale sensitive; therefore, column normalization is required for a meaningful correlation comparison.

- *Principal Component Analysis (PCA) -*

Principal Component Analysis (PCA) is a statistical procedure that orthogonally transforms the original n coordinates of a data set into a new set of n coordinates called principal components. As a result of the transformation, the first principal component has the largest possible variance; each succeeding component has the highest possible variance under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components. Keeping only the first m ¡ n components reduces the data dimensionality while retaining most of the data information, i.e. the variation in the data.Notice that the PCA transformation is sensitive to the relative scaling of the original variables. Data column ranges need to be normalized before applying PCA. Also notice that the new coordinates (PCs) are not real system-produced variables anymore. Applying PCA to your data set loses its interpretability.If interpretability of the results is important for your analysis, PCA is not the transformation for your project.

- *Backward Feature Elimination -*

In this technique, at a given iteration, the selected classification algorithm is trained on n input features. Then we remove one input feature at a time and train the same model on n-1 input features n times. The input features whose removal has produced the smallest increase in the error rate is removed, leaving us with n-1 input features.

The classification is then repeated using n-2 features, and so on. Each iteration k produces a model trained on n-k features and an error rate e(k). Selecting the maximum tolerable error rate, we define the smallest number of features necessary to reach that classification performance with the selected machine learning algorithm.

- *Forward Feature Construction -*

This is the inverse process to the Backward Feature Elimination. We start with 1 feature only, progressively adding 1 feature at a time, i.e. the feature that produces the highest increase in performance. Both algorithms, Backward Feature Elimination and Forward Feature Construction, are quite time and computationally expensive. They are practically only applicable to a data set with an already relatively low number of input columns.

2.1.4 Data set splitting:

Further the data set was divided into training and test in 80:20 ratio as standard splitting ration. The training set is the part on which machine learning algorithms are applied to build the model. The test set is used to test the developed model robustness.

2.1.5 Model development:

Multiple Linear Regression: Multiple linear regression is the most usable statistical technique for predictive analysis in machine learning. Linear regression determines a linear relationship between dependent and independent variables. Say for example x is independent and y is dependent variable the equation below represents how y is related to x.

$$minimize = \frac{1}{n}\sum_{k=1}^{n}(pred_i - y_i)^2$$

It can also be represented as:

$$g = \frac{1}{n}\sum_{k=1}^{n}(pred_i - y_i)^2$$

Where, g is called as cost function, which is the root mean square of the predicted value of y ($pred_i$) and actual y (y_i), n is the total number of data points.

2.1.6 Model Evaluation parameters:

Here in this project we will evaluate the performance of model in terms pf R2, mean square error (MSE), mean absolute error (MAE) and root mean square error (RMSE). R2 score finds the scatteredness of data points around the regression line which can also be referred to as the coefficient of determination.

$$R^2 = \frac{Variance..explained..by..model}{Total..varience}$$

Mean absolute error (MAE): MAE is the average magnitude of the errors in the model predictions which is calculated as:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_i - \hat{y}_j|$$

Root Mean Square Error (RMSE): It is a standard deviation of the prediction error. It is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_i - \hat{y}_j)^2}$$

Chapter 3

Conclusion

In this project we intend to develop a mathematical model which could predict the effectiveness of drug against COVID-19 virus. This could speed up the drug discovery as preclinical wet lab experiments would be carried out on those drugs which showed promising results upon prediction from the model.

References

1. *Foertter, F., Gaither, K., Hinsen, K., West, J. (2020). Computational Science in the Battle Against COVID-19. Computing in Science Engineering, 22(6), 9-10.*

2. *Bagula, A., Maluleke, H., Ajayi, O., Bagula, A., Bagula, N., Bagula, M. (2020,July). Predictive Models for Mitigating COVID-19 Outbreak. In 2020 IEEE Symposium on Computers and Communications (ISCC) (pp. 1-7). IEEE.*

3. *Zamzami, N., Koochemeshkian, P., Bouguila, N. (2020, August). A Distribution-based Regression for Real-time COVID-19 Cases Detection from Chest X-ray and CT Images. In 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI) (pp. 104-111). IEEE.*

4. *Li, B., Dai, C., Wang, L., Deng, H., Li, Y., Guan, Z., Ni, H. (2020). A novel drug repurposing approach for non-small cell lung cancer using deep learning. Plos one, 15(6), e0233112.*

5. *Issa, N. T., Stathias, V., Schürer, S., Dakshanamurthy, S. (2020, January).Machine and deep learning approaches for cancer drug repurposing. In Seminars in Cancer Biology. Academic Press.*

6. *Feng, Z., Chen, M., Xue, Y., Liang, T., Chen, H., Zhou, Y., ... Xie, X. Q. (2020). MCCS: a novel recognition pattern-based method for fast track discovery of anti-SARS-CoV-2 drugs. Briefings in Bioinformatics.*

7. *Fleming, N. (2018). How artificial intelligence is changing drug discovery. Nature,557(7706), S55-S55.*