iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** III **Month of publication:** March 2020

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

Supervised Machine Learning to Recommend Drugs for Testing Against Novel COVID-19

Abhilash V[1], Archana Balasubramanian[2], Bhargavi G[3], Rishith Bhowmick[4]
[1, 3, 4]*Computer Science, PES University, Bangalore, Karnataka, India*
[2]*Bio-Technology, PES University, Bangalore, Karnataka, India*

Abstract: ***Slowly but surely, COVID-19 has taken the world by storm since the end of December, 2019. Procedures are still underway to develop a vaccine for coronavirus. Articles suggest that re-purposing previously known drugs would be a faster approach to developing a vaccine rather than developing the anti-viral from scratch. There are several drugs, as stated in the article, that may be e ective against coronavirus from a biological perspective. In our research, we will use supervised machine learning on a large database of diseases, their respective micro and macro features, and drugs to somehow recommend a list of drugs that can be used to test for e ectiveness against coronavirus. This research, however, is only from a theoretical perspective and may not be from a practical one, unless proven. Any drug recommendations made by our research is only for testing purposes against COVID-19.***

Keywords: COVID-19; Coronavirus; Data; Drugs; Features; Modelling; Recommenda-tion; Testing.

I. INTRODUCTION

With the number of people infected by COVID-19 increasing at an exponential rate, now is the time to come with an e ective drug to combat the illness. COVID-19 is caused by the SARS-CoV-2 virus which is a single stranded RNA virus subject to high rate of mutation. Though it is known to resemble the SARS-related coron- avirus strain, it is still regarded as a one of it's kind virus with no available drug in the market to destroy the virus.

Our objective is to nd the best antiviral available in the existing market by com-paring the similarities to other viral diseases with its corresponding mode of action and available medications using Machine learning.

Machine learning has proved e ective in tackling various issues faced in the eld of medicine. Therefore, our approach is to make use of Supervised Machine Learning to classify the most probable drug for testing(target value) using various features of the virus provided as the input to the algorithm. The outcomes are arranged in descending order based on their frequency values and the rst k outcomes are considered as the most optimum drugs among the others.

We hope to use this information by recommending the said outcomes as valid drugs for clinical testing against COVID-19.

II. CURRENT STATISTICS

As of 27 March, 2020, there are 536,450 con rmed cases with 24,112 deaths globally showing no signs of a slowdown. The Novel COVID-19 was brought to notice, o - cially, to the world on 21 January, 2020 with the release of the rst case report by WHO which con rmed 282 cases from China, Japan, Thailand and The Republic of Korea.

From 21 January, 2020 to 1 February, 2020, there was a 4138% increase (282 to 11,953) in the number of cases worldwide. Italy reported its rst cases, both of which had a travel history to Wuhan City, China. The number of con rmed cases crossed 1,000 on 25 January, 2020 touching 1,320 from 846 that was recorded on 24th January, 2020.

As of 1 February, 2020, 98.89% of the cases reported globally were from China. From 1 February, 2020 to 29 February, 2020, the number of con rmed cases in-creased from 11,793 to 85,403. On 5 February, 2020, Belgium becomes the second European country to report its rst con rmed cases for the COVID-19 disease. On 15 February, 2020, Egypt became the rst country from the African continent to report con rmed cases of COVID 19. From 16 February, 2020 to 17 February, 2020, there has been a noticeable spike in the number of reported cases, spiking from 51,857 to 71,429.

From 1 March, 20201 to 25 March, 2020, the situation becomes intense. The num-ber of reported cases has gone from 87,137 to 372,757. On 11 March 2020, WHO declares the novel COVID-19 disease a pandemic. The number of cases in Italy has been rising exponentially. 10 days were considered from 6 March, 2020 to 16 March, 2020, the cases rose from 3,858 to 24,747 (a 541.44% increase)

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

III. BIOLOGICAL UNDERSTANDING

A. What is COVID-19?

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as stated by the WHO.

B. Why is COVID-19 Regarded as a Pandemic?

The new coronavirus now identified as COVID-19 caused a cluster of cases resulting in acute respiratory disease in Wuhan, Hubei, China in late December, 2019. The outbreak soon spread to more than 190 countries and territories, resulting in more than 19,700 deaths and more than 112,000 recoveries making the WHO to recognise it as a pandemic on 11 March, 2020.

The key strategies in the control of an outbreak are containment, mitigation and suppression to reduce the basic reproduction number (of infected patients) to less than 1 and hence attend the epidemic curve.

One of the main reasons for the COVID-19 outbreak to result into a pandemic was due to the fact that infected people who were asymptomatic turned out to be major carriers of the disease. As the incubation period of this disease lies between 2-14 days with some of the only symptoms being fever, dry cough or shortness of breath, it becomes a tedious task to predict the prognosis of the disease. South Korea's CDC reported a woman who became the country's 31st con rmed patient on Feb, 2020 to have infected over 60% of the infected population in South Korea which clearly shows the exponential trend of carrier to carrier transmission of the disease.

C. Mode of Transmission

According to the World Health Organization and the United States Center for Disease Control, it is mainly spread during close contact and via respiratory droplets produced during coughing and sneezing. The virus is not thought to be airborne or to spread over large distances but may survive for up to three hours in aerosol form. The virus can remain infectious for hours to days on surfaces such as door handles making it a widespread disease as it can spread when one touches a contaminated surface to their eyes, nose or mouth.

D. Symptoms

Symptoms of COVID-19 are non-speci c and those infected may exhibit u-like symptoms such as fever, cough,fatigue, shortness of breath, or, muscle pain or re- main asymptomatic.

Symptom	%
Fever	87.9%
Dry cough	67.7%
Fatigue	38.1%
Sputum production	33.4%
Anosmia (loss of smell)	30-66%
Shortness of breath	18.6%
Muscle pain or joint pain	14.8%
Sore throat	13.9%
Headache	13.6%
Chills	11.4%
Nausea or vomiting	5.0%
Nasal congestion	4.8%
Diarrhoea	3.7%
Haemoptysis	0.9%
Conjunctival congestion	0.8%

Fig. 1. Prevalence of Symptoms

The typical signs and symptoms with their prevalence are shown in Fig. 1. CDC lists emergency symptoms including di culty breathing, persistent chest pain or pressure, sudden confusion, di culty waking, and bluish face or lips with compli-cations leading to severe pneumonia, acute respiratory distress syndrome, sepsis, septic shock and death.

E. Coronavirus

The main question that arises is, \Why has no antiviral drug come into e ect to combat the virus?" or \Can an existing drug combat this virus?". The answer to the above mentioned questions lies in the fact that it is a mutated version and is a rst of its kind virus when compared to the previously occurred coronavirus outbreaks i.e., 229E, NL63, OC43, HKU1, MERS-CoV, and the original SARS-CoV. A phylogenetic analysis of the genomes isolated from infected patients showed they were "highly related with at most seven mutations relative to a common ancestor", implying that the rst human infection occurred in November or December, 2019.

F. Structural Biology

Severe acute respiratory syndrome coronavirus 2(SARS-CoV-2 is a positive-sense single-stranded RNA virus). SARS-CoV-2 is a member of the subgenus Sarbecovirus (beta-CoV lineage B) with its RNA sequence approximately 30,000 bases in length. SARS-CoV-2 is unique among known beta-coronaviruses in its incorporation of a polybasic cleavage site, a characteristic known to increase pathogenicity and trans-missibility in viruses.

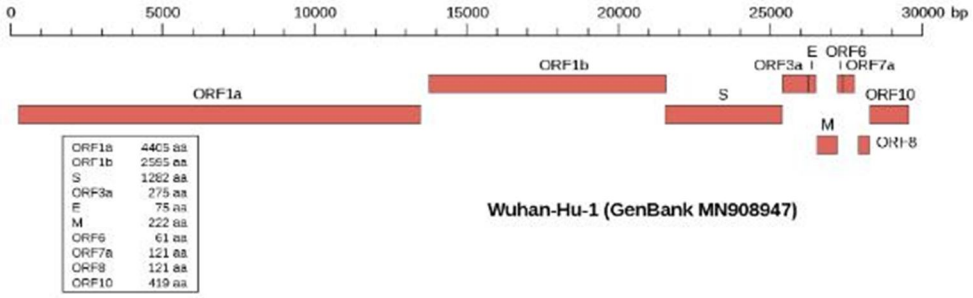


Fig. 2. Genomic Information of COVID-19

G. Mode of Action

SARS-CoV-2 has four structural proteins, known as the S(spike), E(envelope), M(membrane), and N(nucleocapsid) proteins. The spike protein is the protein re-sponsible for allowing the virus to attach to the membrane of a host cell.

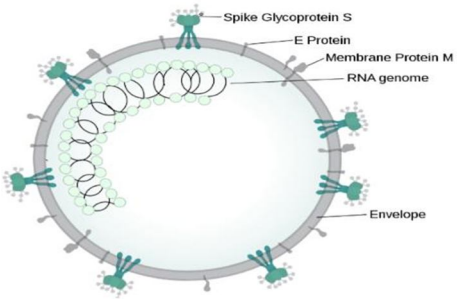


Fig. 3. SARSr-COV Virion

Protein modeling experiments and reverse genetics have depicted that the SARS-CoV-2 has su cient a nity to the angiotensin converting enzyme 2(ACE2) recep-tors of human cells to use them as a mechanism of cell entry.
Initial spike protein priming by transmembrane protease, serine 2(TMPRSS2) is essential for entry of SARS-CoV-2 into the host cell to release it's RNA into the cell and take over the host cell's machinery. SARS-CoV-2 produces at least three virulence factors that promote shedding of new virions from host cells and inhibit immune response.
The above mentioned sites can serve as target sites for a drug or a vaccine to act upon.

H. Similarity to Existing Diseases

The International Committee on Taxonomy of Viruses(ICTV) identi ed 2019-nCoV as a strain of SARS-related coronavirus. One theory states that the previous SARS Cov1 and the present SARS CoV2 use the same \receptor" for entry in cells and the suggestion is that the \envelope proteins" of the present SARS CoV2 are just more e cient. No existing research shows the genetic similarity of this virus to previously existing ones which explains the reason for the non-availability of drugs in the market.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

I. Exisiting Anti-Virals

Scientists have suggested dozens of existing compounds for testing, but WHO is focusing on what it says are the four most promising therapies: an experimental antiviral compound called remdesivir; the malaria medications chloroquine and hydroxychloroquine; a combination of two HIV drugs, lopinavir and ritonavir; and that same combination plus interferon-beta, an immune system messenger that can help cripple viruses.

The one good news is that the above mentioned drugs cater to the broad-spectrum aspect of antiviral drugs but to be proven e ective is required in high dosages which may lead to drug toxicity in one's body.

Few other downsides are the expense factor and the fact that these can't be used for milder symptoms. The other possibility is that they may fail in a particular testing batch but might prove e ective on a large scale.

Our basis is to now nd the best antiviral in the existing market that caters to the above speci ed factors without having many downfalls.

IV. FEATURES FOR MODELLING

The database that will be used for training the model as accurately as possible would be to provide the features that have the most signi cant impact on the drug used against the virus at a micro/macro level

Few of the important features, for each disease, in predicting the most e ective drug in the market for testing against COVID-19 are as follows - Structure of the Virus Available Anti-Viral Drug Mode of Action of the Drug Toxicity of the Drug

V. DATA PRE-PROCESSING

Data Pre-Processing is one of the rst and most important steps performed in Ma-chine Learning. It is the process of preparing the data in a format that is recognised by the Machine Learning model. It is also one of the crucial steps that needs to be performed as data, if not processed correctly, can lead to misleading results, which is undesirable.

Before any processing is done on the given data, it needs to be cleaned. Cleaning is the process of removing or modifying incorrect, incomplete, irrelevant, duplicated, or improperly formatted data. Cleaning is essentially nding a way to maximise the accuracy of the given data without necessarily deleting information. It also involves xing spelling and syntax errors, standardising data sets, and correcting empty elds, missing data, and identifying duplicates.

The next step of Data Pre-Processing is converting the features of the data into a form that is understood by the model. Some of the features and targets could be classi ed into categories, so they need to be encoded in a format understood by the model. There are various encoding techniques such as Label Encoder, One Hot Encoder.

Another step that is involved in Data Pre-Processing is Feature Scaling. It is the process of bringing all the features to a common scale. This is to prevent the model from giving more preference to features that have a large range of values. This is done to models that involve Euclidean distances.

VI. SML MODELLING ALGORITHM

A. Tree-Based Modelling

Decision Tree is a supervised predictive model that draws conclusions about the target value from observations about the features. The features contain continuous as well as categorical data. The feature that best classi es the drugs(target values) is considered as the root node which is measured on basis of the gini impurity value. Lower the impurity value, better the classi cation. GI is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

Considering C as the total classes and p(i) is the probability of picking a data-point with class i, then the GI is calculated by plugging in the values into the GI Equation as depicted in below. CART (Classi cation and Regression Trees) makes use of GI as metric.

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

The leaf nodes of the tree are the target values. Any disease we consider are iden-ti ed by many factors which includes speci c symptoms pertaining to it, list of retroviral drugs, toxicity of drugs, structure of the said viruses, etc., which are the features that the model will be trained on. As the factors pertaining to COVID-19 are of main focus here, the features that best classify COVID-19 to the most probable vaccine are given higher priority from the root to the leaf nodes.

1) *Why Tree-Based Modelling?:* The Tree based approach requires less cleaning when compared to other modelling techniques. It is neither in uenced by outliers nor missing values. Both categorical as well as continuous values can be handled by this approach. It is the most optimum in classi cation of the testing data(COVID-19) into categorical target values(the drugs) based on the similarity of the features of the data.

B. Decision Tree vs. Random Forest

Decision Trees are more prone to over tting as decision trees are highly speci c while leading to smaller samples of events. The error due to the bias as well as the variance must be reduced and hence we use the Random Forest Classi cation instead.

Random Forest consists of many individual decision trees that operate as an en-semble, i.e multiple machine learning algorithms that obtain better predictive per-formance. The predictions made by individual decision trees have low correlation with each other. If some trees provide wrong predictions, many others provide the correct predictions. Therefore the group of trees move towards the right direction. The features mentioned above of each disease are uncorrelated or have very low correlation with each other.

Random Forest tackles over tting using two di erent ways. One way is to train on di erent train samples of the data. The second way is to train on random sub-sets of the features. Out of n features of the diseases, we take k random samples as the features to one of the decision trees and other k random samples to the other trees. The features of our data include both micro level and macro level descriptions which make them much less susceptible to any changes.

C. Random Forest Algorithm

Random Forest is a modelling technique comprising of multiple decision trees. This operates like an ensemble. Each individual decision tree gives out its class predic-tion. The class with the highest tally is the output of the model. This model uses two key concepts that gives it the name random:

Random sampling of training data points when building trees Random subsets of features considered when splitting nodes

The reason random forest performs so well is "A large number of relatively uncorre-lated models (trees) operating as a committee will outperform any of the individual constituent models."

The feature importances in a random forest indicate the sum of the reduction in Gini Impurity, as mentioned in Section 6.1, over all the nodes that are split on that feature. Considering N1 and N2 are the number of samples that go into T1 and T2 branches respectively, then -

$$Gini_{split}(T) = \frac{N_1}{N} Gini(T_1) + \frac{N_2}{N} Gini(T_2),$$

With respect to COVID-19, the features that should be used for modelling are of a large number and feature extraction plays a crucial role in nding out what fea-tures have a sign cant impact on the target variable, i.e the drug used. Feature importances can give us insights into the problem by telling us what features of the existing viruses are the most discerning between classes, i.e drug used.The distinguishing features of the existing viruses acts as the nodes for the random forest algorithm, as seen in Fig. 3. In Fig. 3, there are only two classes, but in our application we will have 'n' classes, where n = Number of Distinct Drugs. The classes refers to the drugs used for the respective viruses. The nal class, based on majority voting, will be the drug used.

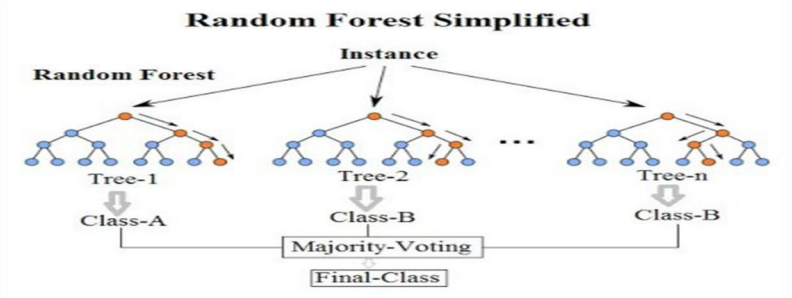


Fig. 4. Simple Representation of Random Forest Algorithm of 2 Classes

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

D. Why not any Other ML Techniques?

Machine Learning has four types of prediction algorithms - supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

Unsupervised learning is the machine learning technique which draws inferences from data points without the target values or labelled responses. K means clus-tering is an example of unsupervised learning that is mostly used for exploratory data analysis and grouping of similar data points together. The objective is to pre-dict a set of most probable vaccines for the disease which learns from the training dataset that includes the di erent vaccines required for each of the data points in the dataset.

Reinforcement learning is a technique used widely in Deep Neural Networks. Ex. Image captioning . Our algorithm is supposed to predict the vaccines which are present in the dataset, i.e the outcome given by the model is compulsorily present in the data. Reinforcement learning is mainly based on prediction of a value that is not present in the data and how well the model performs when it encounters an unseen situation. Hence, reinforcement learning cannot be adopted to get the

desirable outcome. Semi-Supervised Learning is a combination of unsupervised and supervised learn-ing. It deals with a large number of unlabelled data and few labelled data. This algorithm is not suited for the approach taken up.

VII. LIMITATIONS OF THE MODELLING ALGORITHM

Random Forest creates a lot of trees which requires much more computational power and resources. By default, Random Forest creates 100 trees in Python. It makes a decision based on the majority of votes. As the number of trees generated becomes larger, the time taken for training increases. For data including categorical values with di erent levels, Random Forest may be biased towards those attributes with more number of levels. Random Forests operate as an ensemble which are inherently less interpretable than individual decision trees. Predictions are slower which may create slower applications. They are capable of taking up large amounts of memory and are slow to evaluate. Looking from this perspective, there won't be much impact on the outcome, as such.

VIII. MODEL EVALUATION METRIC

The model performance will be evaluated on basis of the micro-averaging precision obtained from the confusion matrix. In \micro averaging", we calculate the perfor-mance from the individual true positives, true negatives, false positives, and false negatives of the the k-class model(k = Number of Distinct Drugs):

$$PRE_{micro} = \frac{TP_1 + \cdots + TP_k}{TP_1 + \cdots + TP_k + FP_1 + \cdots + FP_k}$$

Where TP = True Positive, FP = False Positive.

IX. MODEL IMPLEMENTATION

After training the model on the provided data, the model will be tested on the COVID-19 data with the same features used for modelling, where the target variable is the drug used. Once the model completes its job, we will extract the predicted variable, i.e the drug to be recommended for testing. Once we have the extracted values, the drugs can be ordered in descending order on basis of the statistical mode of the predicted variable. Once extracted and ordered, the encoded values for the drugs will be converted back to the original drug name given before Data Pre-Processing. The top drugs will be recommended for testing purposes against the COVID-19 virus.

X. CONCLUSION

As we can see from a biological understanding, the features mentioned in Section 4 can help di erentiate one virus from another and identify the drug that can be used against it. The data for the model to be trained on comprises of the features of the previously known viruses.

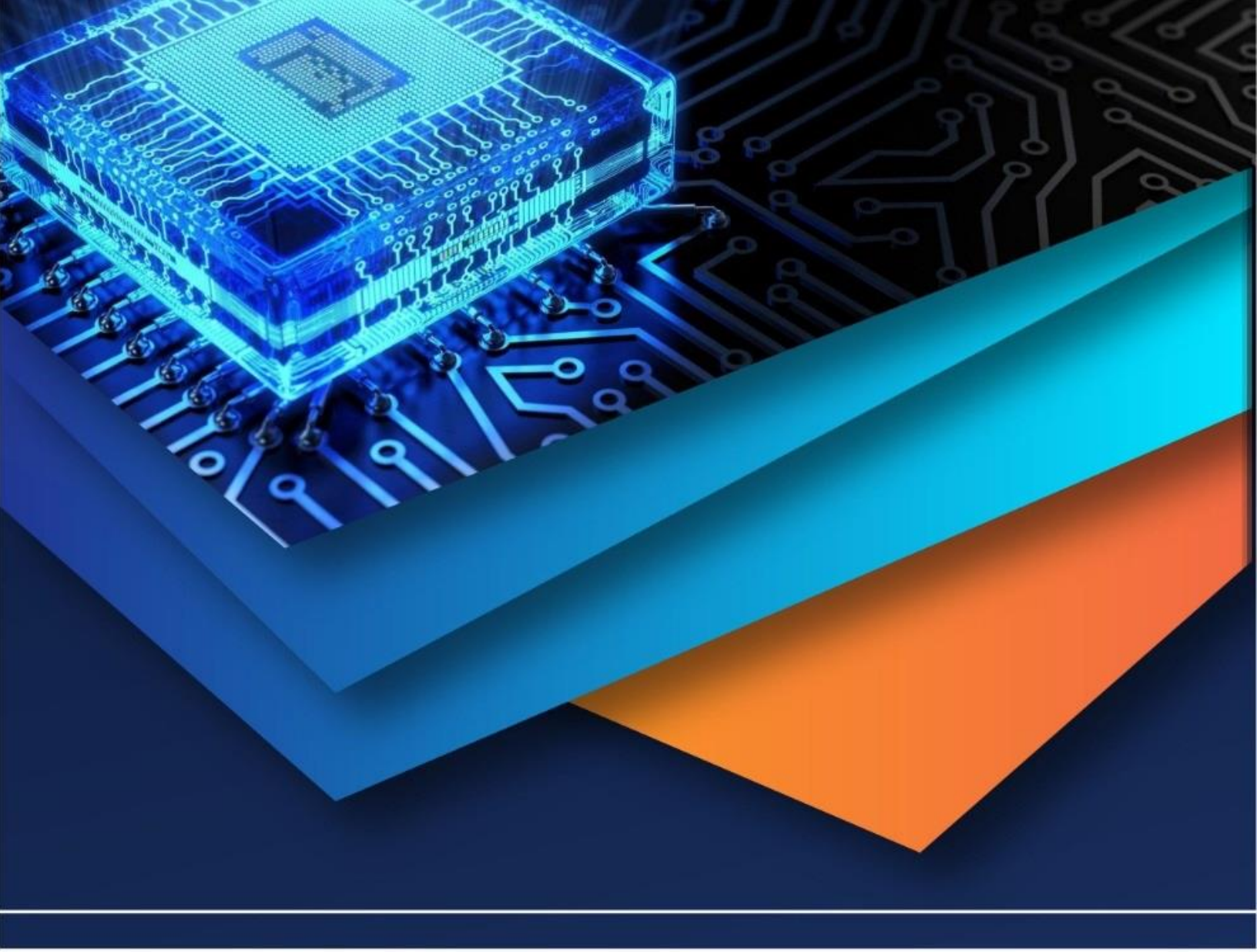
The model will be used on the COVID-19 data with the same features to recommend drugs that can be tested to check for e ectiveness against the COVID-19 virus. As of now, the random forest algorithm seems like the best SML approach considering the type and features of the data we are dealing with. Further changes may be made concerning the model once the performance of the model is known on testing.

XI. ABBREVIATIONS

COVID, Coronavirus Disease; SML, Supervised Machine Learning; ML, Machine Learning; GI, Gini Impurity; CDC, Centre for Disease Control

REFERENCES

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
[2] https://www.worldometers.info/coronavirus/
[3] https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/
[4] https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f
[5] https://hackernoon.com/reinforcement-learning-and-supervised-learning-a-brief-comparison-1b6d68c45 a
[6] https://geohackweek.github.io/machine-learning/01-tree-based/
[7] https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76
[8] Grus, Joel. (2015). Data Science from Scratch: First Principles with Python. O'Reilly Publications
[9] https://towardsdatascience.com/understanding-random-forest-58381e0602d2
[10] https://victorzhou.com/blog/gini-impurity/
[11] http://danielhomola.com/wp-content/uploads/2018/03/DanielHomola PhdThesis nal.pdf
[12] https://medium.com/datadriveninvestor/decision-tree-algorithm-with-hands-on-example-e6c2afb40d38
[13] https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d
[14] https://machinelearningmastery.com/confusion-matrix-machine-learning/
[15] https://sebastianraschka.com/faq/docs/multiclass-metric.html https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/
[16] https://en.wikipedia.org/wiki/Decision tree learning
[17] https://towardsdatascience.com/understanding-random-forest-58381e0602d
[18] https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991
[19] https://www.geeksforgeeks.org/ml-semi-supervised-learning/
[20] https://www.oreilly.com/library/view/hands-on-machine-learning/9781789346411/e17de38e-421e-4577-afc3-efdd4e02a468.xhtml
[21] https://www.sisense.com/glossary/data-cleaning/
[22] https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19- nal-report.pdf
[23] https://www.cdc.gov/coronavirus/2019-ncov/prepare/cleaning-disinfection.html
[24] https://www.nih.gov/news-events/news-releases/new-coronavirus-stable-hours-surfaces
[25] https://www.nature.com/articles/s41591-020-0820-9
[26] https://linkinghub.elsevier.com/retrieve/pii/S0166354220300528
[27] https://www.nature.com/articles/s41564-020-0695-z
[28] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7089049/
[29] https://gizmodo.com/scientists-create-atomic-level-image-of-the-new-coronav-1841795715
[30] https://en.wikipedia.org/wiki/Severe acute respiratory syndrome coronavirus 2



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY