International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 5 ǁ 2022 ǁ PP. 78-84

Theoretical approach towards Forecasting Covid-19 Cases using Machine Learning amid suspected cases and their category of admission

Rohan Badajena,*PG Scholar,*
Department of MCA, Dayananda Sagar College of Engineering,
Bengaluru, Affiliated to VTU

Alamma B H, *Assistant Professor, Department of MCA,*
Dayananda Sagar College of Engineering,Bengaluru, Affiliated to VTU

Abstract: *The rapid spread of COVID-19 worldwide has claimed thousands of lives and has put unprecedented pressure on the healthcare systems around the world. The World Health Organization has emphasised the need for comprehensive testing in order to fight the virus [1]. With the lack of testing kits available worldwide, there is a call for novel testing methods that can help arrest the spread faster [18]. Every health care worker exposed to the virus puts additional pressure on an already outstretched infrastructure. Here, in this study we propose a machine learning approach towards predicting Covid-19 cases among a sample population who have undergone other clinical tests and blood spectrum tests. The patient data used in this effort has been donated by Hospital Israelita Albert Einstein, at São Paulo, Brazil [6] for the purpose of research. The problem at hand is divided into two parts: Predict confirmed COVID-19 cases amongst suspected cases based on the laboratory tests of their clinical samples. Predict admission to general, semi-ICU, and ICU wards among those who predicted positive for COVID-19 in the first task. Our approach uses Classification from Supervised Learning techniques to solve this problem. The efficacy of this approach could be used to scale and develop automated systems that could predict the likeliness of Covid-19 based on laboratory tests that are readily accessible. From the features presented to us in the dataset, we are able to predict with 87.0 - 97.4 percent accuracy at a 95 percent confidence level that a patient is suffering from Covid-19 when biomarkers are taken into consideration. Among those that tested positive, we were able to demonstrate that our model could predict with 87.0 - 100 percent accuracy at a 95 percent confidence that whether the patient would be admitted to a particular ward.*

Keyword *- SARS-Cov-2, RT-PCR, Classification, Supervised Learning, Random forest Classifier, Covid-19 Detector, Admission Prediction.*

Date of Submission: 08-05-2022 Date of acceptance: 23-05-2022

I. INTRODUCTION

Blood examination is a rather inexpensive test that is used to diagnose a variety of medical conditions. They are used extensively in diagnosing and monitoring diabetes, cardiovascular issues, forms of cancer, hepatological issues, etc. The complete blood cell (CBC) count is one such frequently obtained blood test whose information content is potentially underused [24]. The CBC is indicative of an individual's overall health and detects a range of potential disorders but is not limited to anaemia, leukaemia, inflammation [4]. The features measured by this test are about Red Blood Cells (carries oxygen), White blood cells (fights infection), Haemoglobin (the oxygen binding molecule in RBC), Haematocrit (the ratio of RBC to fluid) and Platelets (help in clotting).

Apart from CBC, other blood tests obtained routinely are related to liver, glucose, renal (kidney). It is well understood that SARS-Cov, MERS and SARS-Cov-2, the three deadly coronaviruses which share a huge similarity in their genome cause severe respiratory illnesses. They have been found to cause liver impairments in 60% of the patients according to a study published in The Lancet [14]. The study [17] published in the RSNA, Radiological Society of North America, cited a link between thromboembolic complications and Covid-19. It has been observed that the respiratory failure and multi organ failure in Covid-19 is not only caused due to ARDS (acute respiratory distress syndrome) but also due to the complications in thrombotic processes. There is a strong link between D-Dimer levels, disease progression and chest CT scans. D-dimer is a degradation product of cross-linked fibrin and indicates blood clot information. Increased D-dimer levels can be very helpful in linking the severity of the case. Other blood features that can give information related to blood clot

formation in the body could be helpful in diagnosis and management of Covid-19. Partial Thromboplastin Time (PTT) evaluates the ability of an individual to properly form blood clots.

Developing countries with fewer testing resources have adopted conservative testing methods in order to conserve testing kits on symptomatic and mild to severe cases. As more evidence surfaces around symptoms and associated risks of Covid-19, it has been observed that blood examination can play a key role in the route to extensive testing. It can be used as a first line of defence in the fight against Covid-19 where specialised testing kits are still in short supply (relative to the worldwide population) and every false positive result is a chance lost to screen a true positive. On the other hand, AI and machine learning have played an important role in fighting by processing huge amounts of data and deciphering patterns in this vast spread which would have otherwise been a long-drawn process. Bringing the two together, understanding blood examination patterns with the help of machine learning techniques to predict the disease as well as severity of the disease can help us prioritise which patients should be tested first with a certain degree of confidence.

Supervised learning methods are a subset of machine learning where the model infers about a function from input data mapped onto targeted labels. Each record in the open patient dataset obtained from Kaggle has been categorised into a covid positive and a covid negative patient. Each record consists of measurements of a patient's various clinical parameters derived from blood and other tests. Different patients have been subjected to a varied range of tests. The purpose is to design a model to study the underlying patterns and infer which features are able to classify patients as a covid positive or a covid negative case. A classification algorithm that sorts inputs into two targets or classes is called binary classification. Binary classification is used in the first problem statement. The second problem that classifies patients into their severity of the disease from their admission into the hospital wards has four target classes. Such a classification technique is a multi-class classification.

Procurement of any data poses a challenge of getting reliable unbiased and enough data. This is especially critical for medical and clinical data where the data may always be insufficient to train an algorithm. Through this experiment, we have been able to demonstrate as we capture more features of a patient from their blood examination tests, the overall accuracy of the model improves. On the other hand, the data regarding the category of admission or the severity of their disease is difficult to train as the event rate is less than 1% of the overall data.

Through our experiment we have demonstrated how various classifiers have fared over the data and it has been found random forest ensembles to be well suited for these tasks. During the pre-processing stage it has been observed that we can draw axis parallel boundaries which is well suited for a decision tree-based ensemble. Ensemble methods help in reducing bias variance errors and are able to present us with a best fit model.

A key challenge faced by algorithms in performing well on production data is imbalances in the distribution of classes. Real world data usually occurs in a form where one of the classes is under represented and the other present in majority. Here, in the data used as part of our experiment we have the positive class in the minority and the negative class dominates. The class of interest, the positive class, forms 10% of all the cases in our datasets. Using this data as is to train the model led to poor precision and recall metrics for our class of interest. Hence, we have used sampling techniques to handle the problem of imbalance classification. A mix of both under-sampling and oversampling has been used to form a training set that would be used to train the classifier. The minority class has been over-sampled and the majority has been under- sampled.

II. MATERIALS AND METHODS

A. Data

The patient data used in this effort has been donated by Hospital Israelita Albert Einstein, at São Paulo, Brazil for the purpose of research. The data has been collected between March 28 - April 3, 2020. The blood and clinical samples were collected of patients who visited the hospital for a suspected infection of Covid. The anonymized data was made available in a standardized and normalized form. It has a unit standard deviation and a mean of zero. This data hasn't captured a critical feature D-Dimer and has fewer values of potassium which have been seen to be of critical nature in relation to Covid. The studies showing the importance of these features were published at a later date and at the time, their respective importance was not completely known.

The features which comprise various clinical parameters can be broadly categorized under CBC, liver function, renal analysis, salt tests, blood gas analysis (arterial and venous), influenza tests.

The raw data has 5644 records and 111 attributes. Of these 111 attributes which form features of a model, 1 is a key variable, 73 are of type object and 37 are of type numeric.

B. Methodology

The raw data has a high number of missing values. Exploratory data analysis techniques have been applied in order to clean the data. The cleaned data has been divided into two subsets on the basis of their

completeness. No attempt has been made to impute the data in order to keep it true to its original form. Binary and multi-class classification has been applied in order to predict target variables for the two tasks. Logistic Regressor, Decision Tree and ensemble classifiers were applied on the Analytics Base Table and Random forest classifier was chosen to be well suited for the task at hand. The precision recall and f1- score of the class level metrics reported are that of a Random Forest Classifier.

Random Forest, developed by Breiman, is a combination of tree-structured predictors (decision trees). Each tree is constructed via a tree classification algorithm and casts a unit vote for the most popular class based on a bootstrap sampling (random sampling with replacement) of the data. The simplest random forest with random features is formed by selecting randomly, at each node, a small group of input variables to split on. The size of the group is fixed throughout the process of growing the forest. Each tree is grown by using the CART (classification and regression tree) methodology without pruning. The number of base estimators of the forest in this study is set to be 200, the number of input variables tried for each node is the square root of the number of total variables, and the minimum size of the terminal nodes is set to be 2. The "score" of Random Forest is the scaled sum of votes derived from the trained trees for out-of-bag samples.

Random Forest includes two methods for measuring the importance of a variable or how much it contributes to predictive accuracy. The default method is the Gini score (the method of this study). For any variable, the measure is the increase in prediction error if the values of that variable are permuted across the out-of-bag observations. This measure is computed for every tree, then averaged over the entire ensemble, and divided by the standard deviation over the entire ensemble. Therefore, the larger the Gini score (ranges from 1 to 100) is, the more important a variable is.

III. EXPERIMENT

A. Data Cleaning

The raw data obtained from public open datasets was combed for quality and inconsistency issues. Records with values populated as 'not_done' and 'Nao realizado' were tagged as null values. Attribute names were cleaned in order to remove extra whitespaces. Categorical variables that were expressed in different forms were represented in binary categories as zeros and ones. The target class represented in the form of positive and negative was replaced with 0 for covid negative and 1 for covid positive. Similar operations were performed for the group of flu variables which were changed from 'detected' and 'not_detected' and 'present' and 'absent'.

The attributes that were identified with more than 98% missing values were eliminated. Over the next pre- processing, a thorough examination of the raw data revealed that instead of looking at all the records as a whole, we should be looking at them as two logical groups that would help address the challenge of completeness columns and row wise. The missing data was not imputed as imputations may render the results compromised and any records with more than 10% of missing column values were dropped.

Following the aforementioned, we were able to derive two base tables and the modelling process on each of them was performed independent of the other.

Case 1: The first subset has been made taking the features related to complete blood count along with age quantile and three categorical variables representing the hospital admission and the target variable ('sars_cov2'). The independent features related to clinical spectrum comprise information related to RBC, WBC and platelets. A total of 20 features have been chosen with 598 records that have no missing values. Few records which had missing values in them were eliminated and no imputations were performed.

Case 2: The second subset has been made using the features related to blood, flu and liver attributes along with age quantile and three categorical variables representing the hospital admission and the target variable ('sars_cov2').In addition to the information, this dataset has information regarding the presence or absence of flu in patients along with liver function parameters. The dataset formed as part of this has 48 attributes with 254 records.

ICU Admission Prediction: For the ward prediction task, the ICU, semi-ICU, regular ward have been merged into one categorical 'Ward' column with ordinal values 0,1,2,3 representing classes 'Not Admitted', 'Regular Ward', 'Semi-ICU' and 'ICU' respectively.

B. Sampling

After Data Cleaning, it was observed that there was an imbalance in class distribution across both target columns 'sars_cov2' and 'Ward' as shown in *Fig. 1* and *Fig. 2* respectively.

The Pie plot (*Fig.1*) clearly indicates that the classes corresponding to the target variable 'sars_cov2' are highly imbalanced. We therefore apply up-sampling and down- sampling for balancing the classes. up-sampling should be done to the Positive Class (minority Class) and down- sampling to the Negative class (majority class). In our dataset of size 254*48, Class 0 (majority class) had 212 samples and Class 1 (minority class) had 42 samples. In order to decide upon the final number of samples of each class, the mean of the number of samples of both classes was taken which came out to be 127. Class 0 was therefore down-sampled and Class 1 up-sampled to meet the requirement of 127 samples each thereby, balancing the number of samples in both classes.

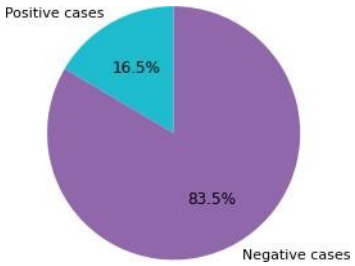

Fig 1: sars_cov2 Class Distribution

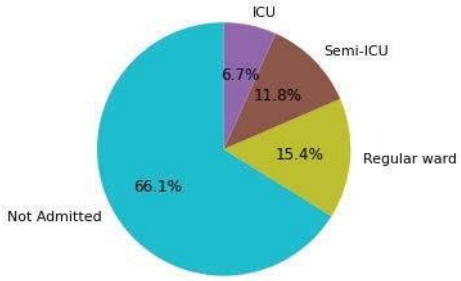

Fig 2: Ward Class Distribution

The Pie plot (*Fig. 2*) for ward distribution shows that the classes corresponding to the target variable 'Ward' are also imbalanced. We therefore, apply up-sampling and down- sampling for balancing the classes. 'Not Admitted' class needs down-sampling while the remaining three classes need up-sampling. In our dataset of size 254*48, Class 0 (Not admitted) had 168 samples, Class 1 (Regular ward) had 39 samples, Class 2 (Semi-ICU) had 30 samples and Class 4 (ICU) had 17 samples. The mean of these values, 64 was chosen to be the number of final samples for each Class. Up-sampling was done on Classes 1, 2 and 3. Whereas. Class 0 (the majority class) was down sampled. The up-sampling and down-sampling techniques were done to ensure good accuracy of the machine learning models.

C. Imputation

Data Cleaning helped remove the majority of null values in the initial data set. After sub-setting the data we were still left with a few records that had missing values. These missing values were then imputed with a randomly chosen value of '-25' for the Classifier to perform satisfactorily.

C. Classification

Before performing classification, the data was split into train and test data on a split percentage of 70 and 30 respectively. We performed classification using different machine learning algorithms and compared the testing and training accuracy of the different classifiers used (*Table 1*). The best accuracy was obtained using Random Forest Classifier with a training accuracy of 97.17% and testing accuracy of 94.80 % .

Scenario-1*:* Area under the ROC-curve (*Fig. 3a*) for Covid- 19 Predictions was observed to be 88.58%.

Scenario-2: Area under the ROC-curve (*Fig. 3b*) for Covid- 19 Predictions was observed to be 97.82%.

Prediction accuracy of the model is also an important factor in determining the efficiency of the model. Our model performs well on the test data as shown in *Fig. 4.* There is 95% confidence that the Test Accuracy is between 87 and
97.4 percent.

B. Admission Predictions

As per *Fig. 5*, There is 95% confidence that the test Accuracy is between 87 and 100%. Additionally, that most frequent accuracies lie between 95 to 96 percent in predicting the admission to wards based on severity.

TABLE.1. COVID-19 PREDICTION ACCURACIES WITH DIFFERENT ALGORITHMS

Training and Test Accuracies		
Model	***Training Accuracy (%)***	***Testing Accuracy (%)***
Logistic Regression	91.52	89.61
Decision Tree	90.39	83.11
Random Forest	97.17	94.80
Bagging	100	96.10
Gradient Boost	100	94.80
ADA Boosting	98.30	93.50

IV. RESULT

A. Covid-19 Predictions

(a) Blood subset (b) Complete Dataset

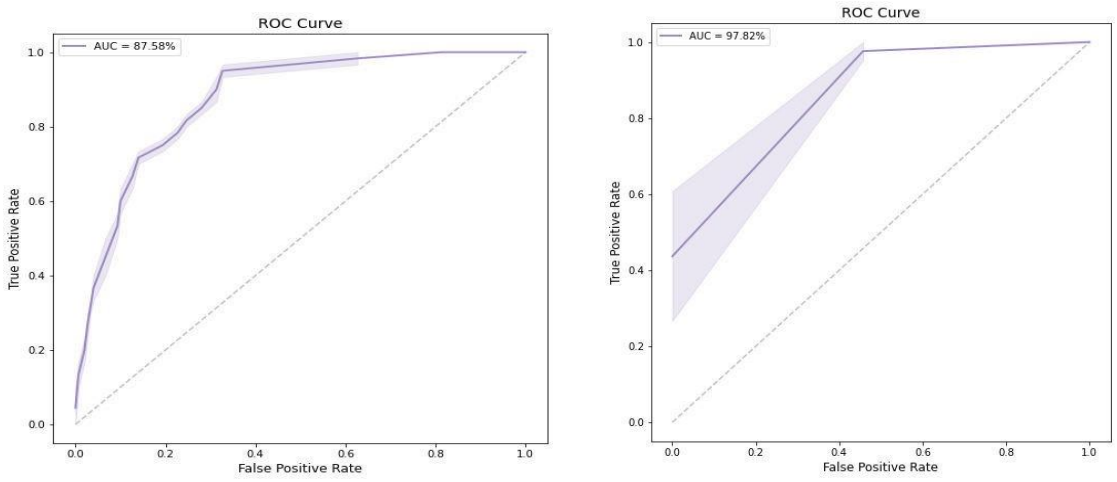

Fig 3: ROC Curve for Covid-19 Prediction on (a) Blood subset and on (b) Complete Dataset

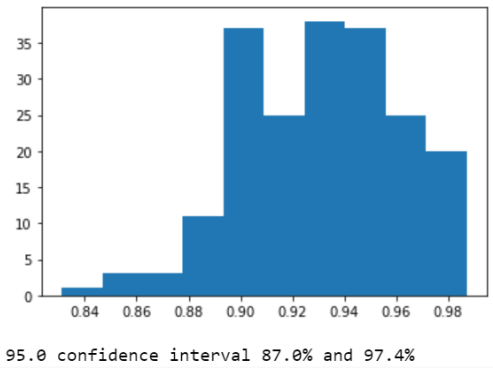

Fig 4: 95% Confidence Interval for Covid-19 Prediction

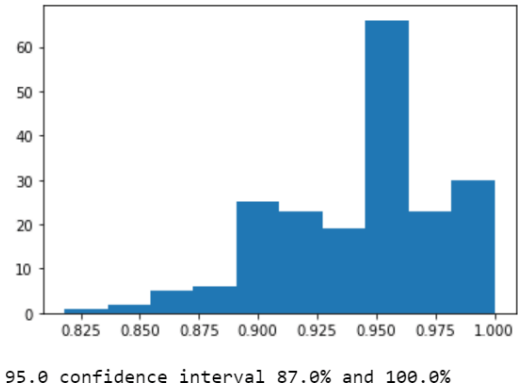

Fig 5: 95% Confidence Interval for Admission Prediction

V. CONCLUSION AND DISCUSSION

Through this study, we have demonstrated a machine learning approach towards predicting occurrence of Covid- 19 in a patient using clinical spectrum data. We achieve this by showcasing that as we add more bio-markers to our training data, the precision, recall and f1-score of the model significantly improves. Adding attributes related to flu and liver tests to the existing blood count spectrum helps this increase. As countries grapple with a shortage of testing kits, adding these tests in the testing protocol can help narrow down the pool of patients requiring the specialized testing kit. This can further enable minimizing the contact for healthcare workers as the number of patients visiting the hospital for a check-up for a suspected case can be reduced. The clinical tests (good indicators/attributes) suggested as part of the study are the ones that can be conducted at the convenience of home and minimize delay.

For implementing an enhanced version of this study to help a larger population, we would propose the model to be trained on more data and attributes significant for coagulation (D-Dimer), Salts (Potassium), information about gender, in order to get the best-fit model for our class of interest. We also propose an extension of this study for health workers and hospitals. They can use the proposed model to define a policy for a minimum acceptable sensitivity level to be prioritized as well as specify admission prediction (to the ward). These attributes can act as an input for the model to determine patients that would be prioritized as likely to test positive. The model's output can be used as a tool for prioritization and to support further medical decision-making processes. The model would then output a binary indicator for SARS-CoV-2 infection, likelihood measure and accuracy. On a periodic basis, input parameters and hospital's policy can be updated depending on health system conditions. The model would then be re- trained also incorporating newly available data.

In conclusion, we are optimistic that using this process flow, training it on an ethnically diverse group of people with rich data, the predictions of this study can have a wider impact and achieve greater accuracy at a less risk tohuman lives.

REFERENCES

[1]. A guide to WHO's guidance on Covid-19. (n.d.). Retrieved from https://www.who.int/news-room/feature-stories/detail/a-guide-to- who-s-guidance

[2]. Andre Filipe de Moraes Batista, J. L. (April-14-2020). COVID-19 diagnosis prediction in emergency care patients: a machine learning approach.

[3]. Chih-Cheng Lai, T.-P. S.-C.-J.-R. (2020). Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and Coronavirus disease- 2019 (COVID-19): The Epidemic and the Challenges. Int J Antimicrob Agents.

[4]. Complete Blood Count. (n.d.). Retrieved from https://www.mayoclinic.org/tests-procedures/complete-blood-count/about/pac-20384919

[5]. Coronavirus disease (COVID-19) outbreak situation. (n.d.). Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[6]. Diagnosis of COVID-19 and it's clinical spectrum. (n.d.). Retrieved from Kaggle: https://www.kaggle.com/einsteindata4u/covid19

[7]. Diagnosis of COVID-19 and its clinical spectrum. (n.d.). Retrieved from Kaggle: https://www.kaggle.com/einsteindata4u/covid19

[8]. Dong Chen, M., Xiaokun Li, M., & Qifa Song, M. (n.d.). Assessment of Hypokalemia and Clinical Characteristics in Patients With Coronavirus Disease 2019 in Wenzhou, China.

[9]. Fei Zhou (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet.

[10]. How to Protect yourself and others. (n.d.). Retrieved from Centerfor Disease control and Prevention: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting- sick/prevention.html

[11]. Imbalanced-learn API. (n.d.). Retrieved from https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.S MOTE.html

[12]. John Willan, c. h. (2020). Challenges for NHS hospitals during covid-19 epidemic. BMJ. Retrieved from https://doi.org/10.1136/bmj.m1117

[13]. LabCorp COVID-19 RT-PCR test EUA Summary. (n.d.). Retrievedfrom Food and Drug Administration:

[14]. Liver injury in Covid-19 - Management and Challenges. (n.d.).Retrieved from The Lancet: https://www.thelancet.com/journals/langas/article/PIIS2468- 1253(20)30057-1/fulltext

[15]. M. CECCARELLI, M. B. (2020). Editorial – Differences and similarities between Severe Acute Respiratory Syndrome. EuropeanReview for Medical and Pharmacological Sciences

[16]. Machine Learning Classifiers. (n.d.). Retrieved from Towards Data Science: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

[17]. Matthijs Oudkerk, H. R. (n.d.). Diagnosis, Prevention, and Treatment of Thromboembolic Complications in COVID-19: Reportof the National Institute for Public Health of the Netherlands.

[18]. Meagan N Esbin, O. N. (n.d.). Overcoming the bottleneck to widespread testing: A rapid review of nucleic acid testing approaches for COVID-19 detection.

[19]. Nervous system involvement after infection with COVID-19 and other coronaviruses. (n.d.). Retrieved from Science Direct: https://www.sciencedirect.com/science/article/pii/S08891591203035 73

[20]. Real-Time Reverse Transcription–Polymerase Chain Reaction Assay for SARS-associated Coronavirus. (n.d.). Retrieved from National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322901/

[21]. Stephen A. Lauer, M. P. (5-5-2020). The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application Annals of Internal Medicine. Retrieved from https://doi.org/10.7326/M20-0504

[22]. Symptoms of Coronavirus. (n.d.). Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/symptoms- testing/symptoms.html
[23]. Understanding AUC-ROC Curve. (n.d.). Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-auc- roc-curve-68b2303cc9c5
[24]. Usefulness of a Complete Blood Count-Derived Risk Score to Predict Incident Mortality in Patients With Suspected Cardiovascular Disease. (n.d.). Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/abs/pii/S0002914906 020005