IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Analysis of Post Covid Symptoms Using Machine Learning

[1] Amitha S, [2]Ashoka S, Gagana.K, Kadiyala. Indraja, Manasa M,

[1]Assistant Professor, [2] B.E final Year Students

[1,2]Department of Computer Science and Engineering,

K.S. School of Engineering and Management, Bengaluru, India

Abstract: **Abstract— Over recent months, severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) infection has been confirmed in millions of people around the world, resulting in hospitalization in thousands of cases Multiple symptoms like fever, cough, fatigue, dyspnea, headache, diarrhea, nausea and vomiting, have been reported during the hospital stay. About 60 days after onset of the first COVID-19 symptom, only 13% of the previously hospitalized COVID-19 patients were completely free of any COVID-19-relatedsymptom, while 32% had one or two symptoms and 55% had three or more. Next to the hospitalized patients with "severe" corona virus disease 2019 (COVID-19), millions of people have most probably been infected with SARS-CoV-2 without formal COVID-19 testing and/or medical treatment in the hospital. Indeed, COVID-19 testing capacity was not available for patients who initially were considered to have mild signs and symptoms. These patients are classified as having "mild" COVID-19 as they only require home care and the infection is expected to resolve. Then again, patients with the so-called "mild" COVID-19 may still complain about persistent symptoms, even weeks after the onset of symptoms. To date, however, only anecdotal evidence is available. This study assessed whether or not multiple relevant symptoms recover following the onset of symptoms in hospitalized and non-hospitalized patients with COVID-19.**

Index Terms **-Anecdotal, COVID, Dyspnea, Symptoms**

I. INTRODUCTION

There has recently been a rapid spread of the novel SARS-CoV2 corona virus (designated by the World Health Organization) which gives rise to a respiratory disease COVID-19. The first human corona viruses, 229E and OC43, were identified during the 1960s from human nasal secretions. Other individual virus types classified in this family have been distinguished (such as HCoV NL63 and HKU1) and are thought to arise from zoometric infections as they are endemic in various bat populations. The corona virus infections known were originally viewed as giving rise to innocuous respiratory human conditions that were not life-threatening. The development incidence of serious and deadly respiratory disorders attributed to beta-corona virus subfamily members occurred in the last twenty years with the severe acute respiratory syndrome (SARS) and the Middle East respiratory syndrome (MERS). The SARS-CoV infections arose first in Foshan, China in 2002 and MERS-CoV in 2012 in Saudi Arabia, both causing international alarm and containment efforts due to their rapid spread and high mortality rates. SARS and MERS were associated with mortality rates of 9.6% and 36%, respectively, among those diagnosed patients. These identified corona virus infections as a significant threat to human health with the potential to cause extreme and lethal respiratory tract infections in people, particularly if person-to-person infection occurs easily.

The development and spread of the novel corona virus causing COVID-19 has vastly outpaced the rate of vaccine and therapeutic development. Nevertheless, within weeks of the first observations of COVID-19 disease, the virus was isolated and characterized. One of the most significant SARS-CoV2 protein targets is a 3C-like protease for which the structure is already known. Much effort has been centered around re-purposing known clinically-tested drugs and virtual screening for possible targets using protein structure data). Priority has been given to the identification of infected individuals in order to isolate and (if necessary) treat them. Central to this is the use of clinical symptoms to optimize identification of infected individuals.

II. RELATED WORK

Banda et.al[1] proposed he coronavirus disease (COVID-19) pandemic is a global health emergency with over 6 million cases worldwide as of the beginning of June 2020. The pandemic is historic in scope and precedent given its emergence in an increasingly digital era. Importantly, there have been concerns about the accuracy of COVID-19 case counts due to issues such as lack of access to testing and difficulty in measuring recoveries.

Shen et.al [2] designed Coronavirus disease (COVID-19) has affected more than 200 countries and territories worldwide. This disease poses an extraordinary challenge for public health systems because screening and surveillance capacity is often severely limited, especially during the beginning of the outbreak; this can fuel the outbreak, as many patients can unknowingly infect other people.

Prakash et.al [3] identifying potential Covid-19 patients in the general population is a huge challenge at the moment. Given the low availability of infected Covid-19 patients clinical data, it is challenging to understand and comprehend similar and complex patterns in these symptomatic patients.

Banda et.al [4] in order to get a complete picture, there is a need to use patient generated data to track the long-term impact of COVID-19 on recovered patients in real time. There is a growing need to meticulously characterize these patients' experiences, from infection to months post-infection, and with highly granular patient generated data rather than clinician narratives.

Kwock et.al [5] COVID-19 is one of the greatest threats to human beings in terms of health care, economy, and society in recent history. Up to this moment, there have been no signs of remission, and there is no proven effective cure. Vaccination is the primary biomedical preventive measure against the novel coronavirus. However, public bias or sentiments, as reflected on social media, may have a significant impact on the progression toward achieving herd immunity. Application of Internet of things and machine learning are proposed in [6][7][8].

Extraction of features from unstructured raw data (hospitalized patient information in text format) using string matching algorithms and use of this data to construct a processed dataset. Identification of the significant symptoms of COVID-19 patients by analyzing their association using five different machine learning approaches. Developing a comprehensive predictive model to predictCOVID-19 positive patients among suspected and confirmed individuals. Analyzing the relationship between patient age and COVID-19 confirmation. Identifying patient travel history and measure how it influences disease progression. Use statistical analysis to calculate the impact and contribution of particular patient features to COVID-19 diagnosis.

For this study secondary data has been collected. From the website of KSE the monthly stock prices for the sample firms are obtained from Jan 2010 to Dec 2014. And from the website of SBP the data for the macroeconomic variables are collected for the period of five years. The time series monthly data is collected on stock prices for sample firm sand relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE -100 Index is taken from yahoo finance.

III. HELPFUL HINTS

We collected raw hospital data, obtained through GitHub repository. A record of their information is made available in anonymized form when a person has presented to hospitals and clinics for diagnosis and treatment. In our datasets, there were data from 6,512 patients from seven different provinces (Anhui, Guangdong, Henan, Jiangsu, Shandong, Shanxi, and Zhejiang) in China. The original dataset was written in Mandarin Chinese, which was translated by Google Translator, and was checked and validated by a native Chinese speaker and researcher (Haoming Xu) to confirm its accuracy. With the spread of the novel corona virus, the accumulation of related national epidemiology data, and its availability can be used for ML studies. However, much of this data was in the form of unstructured text information which can be difficult to process. The data used here were collected from a study by a group at Beijing University's Big Data High-accuracy Centre. They collected these datasets from the official channels of the national government websites.

The detail of the dataset is as follows – basic information regarding gender, age, habitual residence, work and Wuhan/Hubei contact history; trajectory information is time, place, transportation and event up to February 20, 2020. We extracted important features of basic information (age, gender), symptoms (fever, cough, muscle soreness), diagnostic results (lung infection, radiographic imaging), prior disease/symptom history (pneumonia, diarrhea, runny nose) and some trajectory information (isolation treatment status, travel history) that are directly or indirectly related to COVID-19 disease. The original Chinese datasets did not include information about which patients were suspected positive and which were confirmed for all patients. The definition of a suspected case is the patients who develop symptoms and have communication with confirmed COVID-19 patients but didn't confirm as COVID-19 after diagnosis. Moreover, confirmed cases defined as, the patients who are confirmed as positive for COVID-19 in the CDC approved test report or the doctors mentioned confirmed cases after diagnosis in the root dataset. The data contain patient symptoms in a text format. For this reason, we find symptoms of every individual patient and some trajectory information applying various string-matching algorithms. In detail, we selected some keywords for each feature then we matched those keywords to text data and extract the features individually.

Lastly, we generated our final dataset which contained the following features (described in the gender, age, fever, tussis (cough), rhinorrhea (runny nose), pneumonia, lung infection, muscle soreness, diarrhea, and travel history and isolation treatment status. This dataset consists of 1,572 cases of confirmed COVID-19 and 4,940 suspected cases. All the patients did not develop the same symptoms, although, diarrhea and, muscle soreness occurred only rarely. Then we preprocessed the dataset, firstly cleaned the dataset and eliminated unwanted fields. One of the important issues with missing value is the missing value mechanism. It's important because it affects how much the missing value biases our results, so we took it into account when choosing a method to deal with the missing value. Our dataset contained 2.1% missing values only in the gender and age fields, and the propensity for the data point to be missing gender and age fields were completely random, i.e., Missing Completely at Random (MCAR) types of missing data.

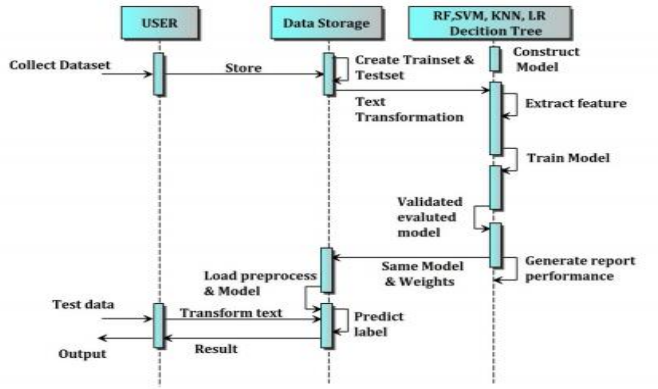

Figure 1. Sequence Diagram of Data Processing

Random Forest is an ensemble of regression and classification trees, which can train a similar size of training datasets called bootstraps, and at the end combine them for a more accurate result. Figure 1 depicts about the sequence of data processing. The bootstraps are created by random resampling from the training dataset. Random Forests perform far better than a single tree. This approach can work with higher dimensional large datasets with comparatively greater accuracy. The model will be built with the following equations.

Calculate the constant value and initialize the model

$$F0=\gamma \operatorname{argmin} \sum_{j=1} mL(yj,\gamma) \quad (1) \qquad \text{------(1)}$$

Compute the pseudo-residuals r for i=1…n ------(2)

$$rjm=[\delta L(yj,F(xj))\delta F(xj)]F(x)-Fm-1(x)) \text{-------(3)}$$

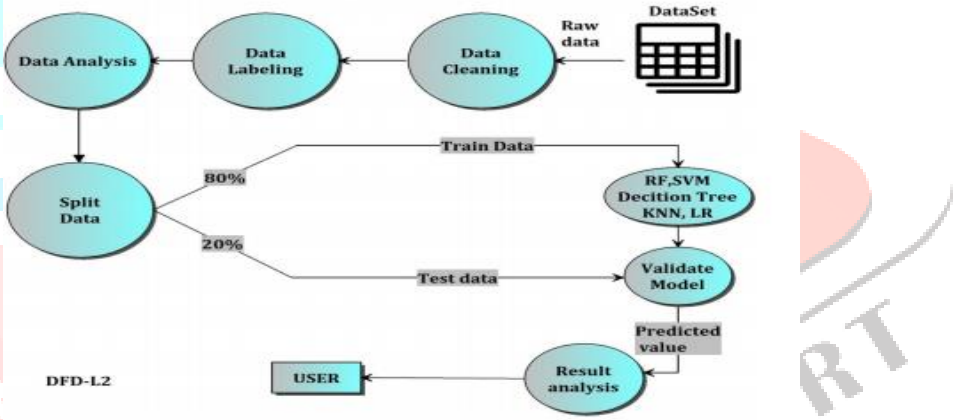

Figure 2. Processing of collected data

According to the statistics, the median age was 43 years with IQR 32–55, composed of approximately half males and half females. Most of the patients presented with fever, cough and radio-graphic chest imaging results that indicated that around 50% of confirmed patients had one or both lungs affected by the infection. Figure 2 shows data processing. In suspected patients, 29.01% were affected with fever, whereas 79.01% confirmed patients have fever & 75.57% have a cough. Travel history was notable for being one of the major associated features to COVID-19 infection, as would be expected with 65.1% of patients having recently travelled a long distance. Some other symptoms were also related to COVID-19 status but were less commonly seen, including muscle soreness and diarrhea; these features, particularly diarrhea, were much more prominent in the earlier SARS epidemic. However, it is striking that 6.74% of the confirmed COVID-19 positive and 69.53% of the suspected patients did not develop any type of symptoms. As these patients cannot be detected or predicted by symptoms alone, our machine learning approach is of no use for assessing these people, although it is possible that they may have other factors that may lend themselves to detection in this way. However, the importance of particular social factors are likely to vary over time; notably, foreign travel may come to be less critical as local community transmission becomes the most common form of infection.

Contact with infected individuals would be and remains an excellent predictor, but this relies on rigorous contact tracing and social network analysis. Mann– Whitney U test and chi-square tests indicated that all the features were impacted except muscle soreness and diarrhea. These significant symptoms matched with findings from our machine learning analysis.

IV Conclusion

The development of the COVID-19 pandemic currently represents a dangerous threat to global health. The key to stopping this spread is the development of methods to identify infected individuals as early as possible. This can be challenging given the delay in symptom presentation; however, machine learning algorithms provide a promising approach to address this problem that can be rapidly and cheaply applied in a pandemic situation. In our study, we developed and tested a range of machine learning approaches and found the most significant clinical COVID-19 predictive features were (in descending order): lung infection, cough, pneumonia, runny nose, travel history, fever, isolation, age, muscle soreness, diarrhea, and gender. Our models were able to predict the stage of COVID-19 based on basic patient information (age and gender), travel and isolation, and clinical symptoms (including

fever, cough and runny nose and pneumonia). The accuracy of our algorithms was highest for the age range 0– 20 years, with the SVM algorithm with 93% accuracy, but it was notable that the other algorithms performed almost as well with greater than 85% accuracy.

REFERENCES

[1] Banda, Juan M., Nicola Adderley, Heba AlGhoul, Osaid Alser, Muath Alser, Carlos Areia, Mikail Cogenur et al. "Characterization of long-term patient-reported symptoms of COVID-19: an analysis of social media data." *medRxiv* (2021).

[2] Ahamad, Md Martuza, Sakifa Aktar, Md Jamal Uddin, Md Rashed-Al-Mahfuz, A. K. M. Azad, Shahadat Uddin, Salem A. Alyami et al. "Adverse effects of COVID-19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and post-vaccination reactogenicity." *medRxiv* (2021).

[3] Aktar, Sakifa, Md Martuza Ahamad, Md Rashed-Al-Mahfuz, A. K. M. Azad, Shahadat Uddin, A. H. M. Kamal, Salem A. Alyami et al. "Machine Learning Approach to Predicting COVID-19 Disease Severity Based on Clinical Blood Test Data: Statistical Analysis and Model Development." *JMIR Medical Informatics* 9, no. 4 (2021): e25884.

[4] Quiroz, Juan Carlos, You-Zhen Feng, Zhong-Yuan Cheng, Dana Rezazadegan, Ping-Kang Chen, Qi-Ting Lin, Long Qian et al. "Development and validation of a machine learning approach for automated severity assessment of COVID-19 based on clinical and imaging data: Retrospective study." *JMIR Medical Informatics* 9, no. 2 (2021): e24572.

[5] Kwok, Stephen Wai Hang, Sai Kumar Vadde, and Guanjin Wang. "Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: Machine learning analysis." *Journal of medical Internet research* 23, no. 5 (2021): e26953.

[6] Kulkarni, Sandhya A., C. S. Sowmya, P. Subhalakshmi, S. A. Tejashwini, V. R. Sanusha, S. Amitha, and Vandana Jha. "Design and Development of Smart Helmet to Avoid Road Hazards Using IoT." In *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, pp. 1-6. IEEE, 2020.

[7] Amitha, S., Pooja N. Raj, H. P. Sonika, Sushma Urs, B. Tejashwini, Sandhya A. Kulkarni, and Vandana Jha. "Segregated Waste Collector with Robotic Vacuum Cleaner using Internet of Things." In *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, pp. 1-5. IEEE, 2020.

[8] Kulkarni, Sandhya A., Vishal D. Raikar, B. K. Rahul, L. V. Rakshitha, K. Sharanya, and Vandana Jha. "Intelligent Water Level Monitoring System Using IoT." In *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, pp. 1-5. IEEE, 2020.