DATA ANALYTICS FOR COVID-19 PANDEMIC USING MACHINE LEARNING

Anjum Sheikh, Dr. Sunil Kumar, Dr. Asha Ambhaikar

Abstract— The whole world is fighting the coronavirus pandemic. The increasing numbers have imposed a challenge on the Governments to initiate necessary steps for combating the effect of coronavirus on the people. Data analytics can be very helpful for knowing the requirements that will arise in the future and at the same time help to know the measures to be taken to deal with the difficult situation. Techniques like machine learning increase the speed of analytics process and therefore be helpful in providing pace to the planning process of pandemic.

Index Terms— Covid-19, Predictive analytics, prescriptive analytics, diagnostic analytics, descriptive analytics, machine learning, big data

1. INTRODUCTION

A pandemic is a term used to describe a contagious disease that spreads from an infected person to healthy person simultaneously in many countries around the world. At present the whole world is fighting with the Coronavirus pandemic which has affected more than 212 countries and territories. The virus gets transferred with the droplets generated by an infected person during sneezing or coughing. Any person who comes in contact with the infected person can get the disease due to accidental inhale of these droplets. The number of infected persons is increasing rapidly and similarly thousands of deaths are being reported everyday at the global level. There is no medicine available and the development of vaccine being in the experimental phase, dealing with this pandemic has become a worldwide challenge. One of the options available with us for facing the situation, is following the guidelines issued by the Government like washing hands, social distancing and wearing masks. The severity of pandemic has forced the researchers to explore solutions for avoiding spread of infections and predict the requirement of medical facilities that will be required to treat the affected patients.

- *Anjum Sheikh is currently pursuing PhD in electronics & communication university at Kalinga University, India, anjnaznus@gmail.com*
- *Dr. Sunil Kumar is head of electrical and electronics engg at Kalinga University, Raipur.*
- *Dr. Asha Ambhaikar is Professor and Dean students welfare at Kalinga University, Raipur.*

Big data analytics along with tools like Machine learning (ML) can be useful in facing the current challenge by analyzing the COVID-19 dataset to predict the disease spread and estimate the growth in number of patients. This analysis will facilitate the arrangement of health care facilities according to the future demands. Machine learning algorithms like Support Vector Machines (SVM), Feature extraction, linear regression etc. can play a critical role in mitigating the impact of Coronavirus.

2. IMPORTANCE OF BIG DATA ANALYTICS FOR COVID-19

Big data is a term that can be used for a large amount of structured and unstructured data that can be mined for gaining meaningful insights. It is a data that exceeds processing capacities of a single machine. The ongoing digital transformations and newly developed algorithms have played a significant role in handling the storage and computing requirements of the voluminous data. According to a research work given in [2] the Big Data consists of three attributes volume, velocity and variety which are generally referred to as 3V's. The concept of 3V's was extended to 9Vs or 3 2 Vs by adding more attributes to it depending on the applications or purpose of analytics. The 3Vs of big data can be related to COVID-19 as follows:

Volume: As the disease has affected billions of people all over the world with numbers increasing every

day, the dataset of the Covid-19 pandemic is huge. The data of the pandemic at the global level and also for some big countries is enormous to be handled by the traditional databases. The volume of data has been increasing everyday with the increase in the number of affected people.

Variety: Big data is able to handle variety of data like the structured and unstructured data. Structured data in the form of texts and values are available in COVID-19 dataset that indicate number of infected patients, active, recovered and deceased along with the details of their country or region and date of observation . The COVID-19 data also consists of unstructured data in the form of X-ray and CT images that are used by the medical practitioners for studying the differences in the pattern of these images as compared to normal people or the ones suffering from diseases like flu and pneumonia. Another source of data is the social media platforms that include responses of the people during pandemics.

Velocity: It is used to measure the speed of data generation. As the virus spread is occurring very fast the real time data of pandemic is arriving continuously. The wide usage of smart phones, laptops, tablets and other digital devices has enabled fast inflow of data.

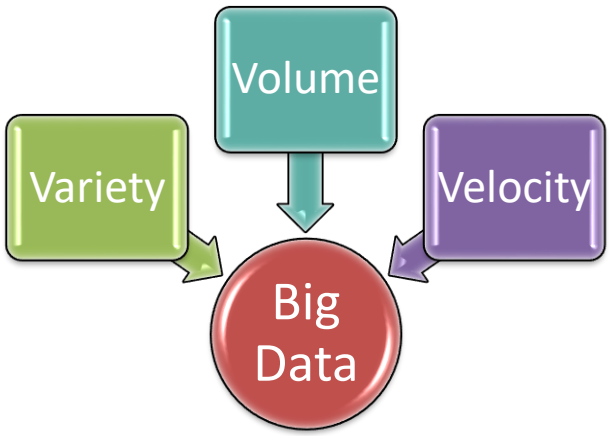

Fig. 1. 3Vs of Big Data

Researchers all over the world have adopted different techniques of Big data analytics to know the important measures for facing the challenging situation. The four types of Big data analytical techniques are descriptive, diagnostic, predictive and prescriptive. Descriptive analytics uses statistical analytics for the given database to derive possible opportunities while diagnostic uses the past data to find the reasons for certain events. Predictive analytics can be used to forecast certain events which can then be used to plan accordingly to achieve our targets. Prescriptive analytics uses the power of decision science to select the best among all the available models to maximize chances of success for the targets.

The color coded world and country maps that display aggregate number of patients in various categories like active , recovered or deceased is an example of descriptive analytics. It uses statistical charts to display the changes in numbers of patients in various categories which can be plotted on the basis of data obtained daily, weekly and monthly. Similarly the spread of the virus, their peak periods and impact can be indicated according to three levels that include regional, country and continent. These visualization patterns only reflect changes in the current trend of infection that can be utilized by the analysts to perform diagnostic or predictive analysis. Diagnostics data analytics have efficiently been used by some countries like Taiwan to control the spread of pandemic. The team of Taiwenese officials carried out a detailed mapping of the infected persons to determine their source of infection. This database was integrated with the immigration and customs information to determine their travel history for the last two weeks. The health information collected from all the international travelers was helpful in knowing about their contacts after the travel and also about any symptoms of the infection observed by them. Data processing and analytics facilitated better information, enabled faster and accurate decisions to combat the spread of virus in Taiwan.

Predictive analysis is one of the analytical techniques that can play a significant role in knowing the demand of medical facilities by studying the rate of growth of infection in a particular area. It can be used to develop models for different scenarios like best case and worst case that can be adjusted according to the real time situations or change in data. This information can be utilized to envisage the demand for health care facilities like personal protective equipment (PPE) kits, ICU beds, ventilators, medicines, quarantine facilities, ambulances services, testing labs and number of medical as well as paramedical staff. Predictive analytics play a crucial role in identifying Coronavirus patterns by observing health records of patients and determination of hotspot areas. The outcomes of analytics can be used

to suggest plan of action for the quarantine and preventive measures. This type of analytics can be powerful while making efforts to handle the pandemic situations by identification of factors responsible for the quick spread of infection. It can thus play a great role in recognizing treatment patterns, therapies and development of vaccines.

The prescriptive analysis that helps in selecting the best possible solutions can be helpful in deciding policies for social distancing, quarantine and lockdown. The incubation period of the virus being 5-14 days the infected person does not show any symptoms and leads a normal life. Without being aware that he/she has become a potential carrier of the disease the infection spreads to many people. Researchers and data scientists after analyzing the rate of virus spread in severely affected countries found that if people are allowed to stay home they would not come in contact with others and this can be helpful in combating the infection. The suspects or the close contacts were therefore sent to quarantine in which they would be away from the social life thereby reducing the chances of spreading infection to others. National lockdown was another preventive step enforced by many countries to handle the difficult situation. The imposition of lockdown put limitations on movement of people outside their houses, allowing access only to essential commodities. Though the lockdown phase hit the economic development of the countries but the favorable part of is that rate of virus spread could be controlled that has saved many lives.

3. Role of Machine Learning in data analysis of Coronavirus Pandemic

Machine learning algorithms allow the machine to acquire knowledge from the received data and perform analysis to derive solutions for the given problems. Pandemics have been always a serious threat to the world. The world has witnessed a number of pandemic and the current Covid-19 is not the last one. Data scientists are using Machine Learning approaches for the big data analytics to fight the pandemic. This section will discuss some of the solutions provided by Machine Learning for facing the pandemic scenario.

The authors in [7] during their research have mentioned that the large data obtained from the patients can be analyzed using machine learning. As shown in fig 2 the data to be collected can be of various types. The dataset will consist of geographic information like country, state, province or city.
Another set of important information for analytics includes kind of symptoms seen in the patients along with its severity levels observed depending on their age and gender. Apart from this information of dates of onset of disease, testing, domestic or international travel and hospitalization can play a key role in the analytics process. The advantage of using Machine Learning is its high speed, accuracy and simple operations. The data scientists working with ML models feed data to the machine and an appropriate algorithm is developed by the machine. The accuracy and high speed of the ML models is beneficial in finding preventive measures; accelerate the process of research for finding drugs to cure the disease and making arrangements for the medical facilities according to the outcomes of the analytics.

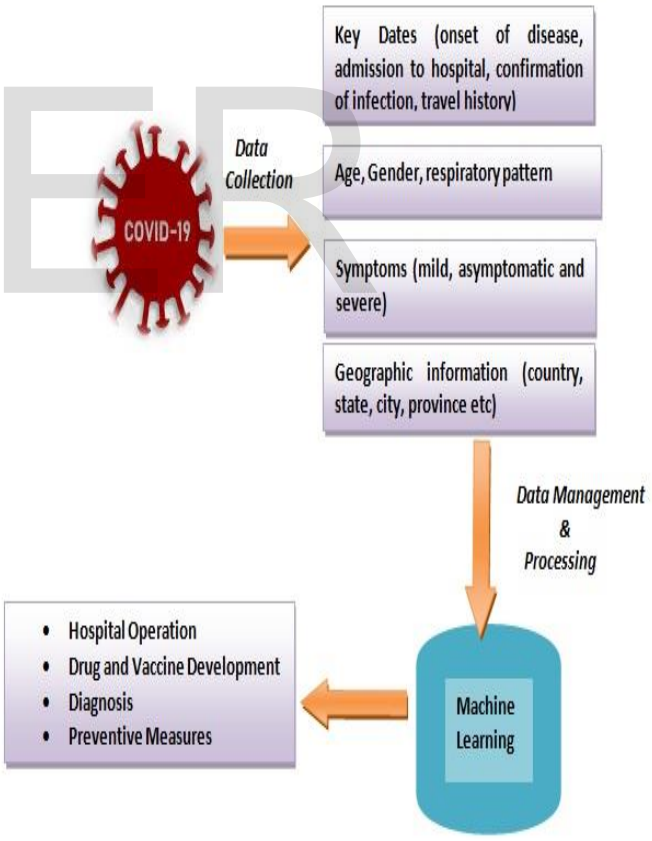

Fig 2: Application of Machine Learning for Fighting Coronavirus Pandemic

ML approaches provide valuable support in identifying the risks of infection and ineffectiveness of treatment method. The available data set can be used to predict the causes of infection like unhygienic practices, social interaction and climatic changes. Similarly the classification on the basis of age or health conditions allows identifying the group of elderly people or the people with ailments like diabetes, hypertension etc., by which special methods can be adopted to protect them from infection. ML has been used earlier to predict the outcomes of treatment for diseases like epilepsy and cancer. As there are no medicines available for the treatment of Covid-19, the doctors are trying different type of medicines to treat the disease. The data of various treatment methods can be analyzed using machine learning to know the effect of medicines, any side effects of those medicines and response of the patients based on their age as well as health condition.

The novel coronavirus outbreak left the whole world wondering with its intensity of transmission. Thousands of death has been reported all over the world. Diagnosis of the disease is essential to provide medical assistance to the infected persons and isolate them to avoid further transmission to others. Collection of medical samples for large populated countries is difficult. To avoid long waiting period for the medical reports due to increased burden on the testing labs, Machine learning can be used for the development of some faster and cheaper methods for diagnosing the disease by analysis of data related to the symptoms of coronavirus infection. For example a face scanning machine can be used to identify symptoms like fever. A similar system was launched at a hospital in Florida that uses thermal scanning face camera to detect fevers and sweating. Such cameras can be ideal to be used at grocery stores, hospitals and similar places which are visited by large number of people. Machine learning based smart watches and wearable are used for monitoring health parameters like body temperature, heart beat and breathing rate. These smart watches with some improvements can be used for detecting onset of infection and tracking recovery process. Both the given examples are under research and no effective results are available but using them can be beneficial in handling the increasing number of patients.

The medical treatments provided to the affected patients are taking place on trial and error basis everywhere. A lot of research is going on to develop a vaccine but most of the health organizations have predicted that arrival of vaccine will take a long time. Machine learning has been used in the past for handling the outbreak of Ebola virus in past. Similar experiments can be helpful in identification of drug and development of vaccine to face the current pandemics by analyzing review of drugs.

Machine learning models can be used to evaluate the possibilities of virus contamination by interpreting the interactions of people through social media and other such communication platforms. These models cannot help in diagnosing but can be used by people to locate the nearby hospital. It can also be used to estimate the spread of infection and predict the total number of patients in the upcoming weeks. It can used to track the travel history of a person for last few days to gather information of the places travelled by the infected person and know the people who came in contact with him. These models can be used like an app on mobile that will be capable of sending alerts of contaminated zones near you. Some of the Governments have been using these apps to collect data of the users like their name, gender, age, travel history and any symptoms of Covid-19 observed by the user. This facilitates tracing of the suspects so that they reach for testing. The user of app will receive notifications if any Coronavirus positive patient is near to him for enabling the users either to take precautionary measure or leave the place immediately.

The research work for the Covid-19 pandemic is going through a developmental phase. Though Machine Learning is playing an important role in containing the pandemic situation but it has some limitations. Machine Learning requires large and good quality of data to produce the desired results. Unavailability of data has become one of the causes for the slow development of medical research. As most of the hospitals and the health organizations hesitate to share the patient's data due to security or privacy concerns, a large amount of data is being underutilized. Machine learning does not work well with the few data points and the decisions or predictions given by the system tend to be biased in such cases. Moreover incomplete data and presence of errors will produce inappropriate results. In this situation the researchers may tend to take wrong decisions based on the obtained results which can prove to be fatal for the people. To enable sharing of data the concerned organizations can remove personal information of the patients and at the same time the research groups should clearly specify their objectives of research along with an assurance of avoiding misuse of data. Another risk while using

Machine Learning program is that immediate detection of error may not be possible always. Once the problem is identified it requires lot of time and efforts to find the root cause and make corrections for it. The number of countries and people affected with the Covid-19 pandemic is increasing exponentially. Machine learning can be used for the data from different geographic locations but standardization of the collected data at the global level is very difficult.

4. MACHINE LEARNING MODELS FOR COVID-19

A research by MIT Sloan School of Management used ML and data analytics to figure out some solutions for the Covid-19 pandemic. The research focused on some of the primary areas of the pandemic like prediction of mortality rate, infection spread, ventilator requirements and testing facilities. The team developed a mortality risk calculator by using the data published by countries like Italy, United States of America and China. It used factors like age, gender, blood pressure, body temperature to analyze the data related to infection, death rate, effects of isolation and health conditions of patients in ICU. Another model called as DELPHI was developed by the team for analyzing the spread of infection. DELPHI was based on the standard SEIR model that classifies the population into four categories: Susceptible, Exposed, Infected and Recovered [10]. This model was helpful in estimating the virus infections, changes in the requirements of medical facilities and deaths. An important concern for facing the challenging pandemic was availability of ventilators for the critical patients. The patients with worse affects of infection need ventilators to increase the oxygen levels of the lungs. In the countries where the number of infected patients was increasing rapidly it was difficult to meet the fluctuating demand of ventilators. A model developed by this research team optimized allocation of ventilators to meet the shortage by allowing sharing of resources among the states. The team collected set of samples and fed it to a ML algorithm for improving accuracy of Covid-19 testing procedures.

Coronavirus affects the lungs that causes respiratory problems and may become cause of death for some critical patients. Radiologists can detect the symptoms of Covid-19 by using computed tomography and X-ray images of chest. Most of the emergency clinics have x-ray imaging machines. Therefore using X-ray images to detect coivd-19 can reduce the burden of testing labs. Some of the researchers have used supervised machine algorithm called as Support Vector Machine (SVM) for X-ray images of lungs to differentiate between the healthy persons, pneumonia patients and covid-19 patients [8]. SVM is generally used for classification problems. It uses hyperplane to classify the data points into two classes. The automated techniques like SVM provide better accuracies as compared to the traditional image classification methods. It can be used for large datasets, saves time of the medical professionals due to the high speed of the algorithm and thus facilitates obtaining the results of the screening in less time. Radiologists can therefore use SVM on the images to classify infected lungs of Covid-19 so that an early detection of lung damage signs can be done which can reduce the delay in providing medical attention to the affected people. The technique will also be useful in observing the changes during the recovery process [9].

Regression analysis using Machine Learning is a prominent method to make predictions for the dangerous circumstances that may arise in the future due to the pandemic. Logistic and linear regression techniques can play a crucial role in forecasting the effects of Covid-19 pandemic. The linear regression model describes relation between a dependent variable and one or more independent variable using a regression line. On the other hand logistic regression is used to determine probabilities of all the possible outcomes for an event using dependent variable that are binary in nature like 0/1 and yes/no. Linear regression can be used to predict the probability of quantitative parameter while the logistic method can be used for quantitative parameters. Linear regression can e used to predict the number of confirmed, recovered and deceased cases after a given interval of time . The accuracy of outcomes while using this model can be improved by making the dataset more informative by inclusion of using some more parameters like date, gender, immune system of patients and preventive measures. The predictions obtained by the linear regression enabled handling of the risky situation through formulation of policies to protect the community from the deadly virus. The predictions done for short term interval can be modified for the long term predictions. The health care ministry and organizations of a country can use these predictions to plan for the essential services. Some of the steps taken by the countries all over the world to prevent transmission are social distancing, quarantine and national lockdown. Logistic regression can be used for sentiment analysis of people using the data from the social media platforms. Most of the people use

social media to express their fear for fast spread of coronavirus infection. Some of them may talk about the impact of the pandemic on the economy of the country and how it would affect the common lives of the people. The sentiments or expressions of the person will vary according to time and place with both positive and negative emotions being shared through the social media like Facebook and twitter. The sentiment analysis using logistic regression can be used to predict the impact on hospitality and tourism industry due to changes in behavior of people for spending on travel and entertainment. This situation can arise for some of the people due to fear of infection. Another reason for the change in behavior will be financial instability due to the increase in unemployment and loss in business during the Coronavirus pandemic. The sentiment analysis can be done for making much other kind of predictions by creating a subset of data. This gathered data can be classified into positive and negative emotions using logistic regression model.

5. CONCLUSION

Big data analytics and Machine Learning can be very effective in finding solutions for the Covid-19 pandemic. The results of analytics help in predicting the growth of affected patients in the future which can thus be utilized for forecasting the health care demands. An advantage of using Machine Learning for analytics is high speed and accuracy. This article has described the role of Machine Learning and some of the models that can be helpful in facing the challenging situation. As the pandemic continues, new challenges will keep on arising but the technological advancements will surely try to find a way through this devastating phase and help the world to mitigate the impact of virus.

REFERENCES

[1] Douglas Laney, 3D Data Management: Controlling Data Volume, Velocity and Variety, Application Delivery Strategies, Meta Group, 6 Feb 2001, pp 1-4.
[2] https://www.dqindia.com/data-analytics-powers-fight-coronavirus/
[3] Ifeyinwa Angela Ajah ,Henry Friday Nweke, Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications, Big Data and Cognitive Computing, MDPI,2019
[4] Anis Koubaa, Understanding the COVID19 Outbreak: A Comparative Data Analytics and Study, Arxiv:2003.14150v1,March 2020
[5] https://www.datarevenue.com/en-blog/machine-learning-covid-19
[6] https://www.datanami.com/2020/04/16/new-mit-analytics-tools-predict-covid-19-patient-outcomes-and-more/
[7] Ahmad Alimadadi, Sachin Aryal, Ishan Manandhar, Patricia B. Munroe, Bina Joe, Xi Cheng, Artificial Intelligence And Machine Learning To Fight COVID-19, Physiol Genomics 52: 200–202, 2020
[8] Prabira Sethy, Santi Kumari Behera, Pradyumna Kumar, Preesat Biswas, (2020). Detection of coronavirus Disease (COVID-19) based on Deep Features and Support Vector Machine. 643-651. 10.33889/IJMEMS.2020.5.4.052.
[9] Lamia Nabil Mahdy, Kadry Ali Ezzat, Haytham H. Elmousalami, Hassan Aboul Ella, Aboul Ella Hassanien, Automatic X-ray COVID-19 Lung Image Classification System based on Multi-Level Thresholding and Support Vector Machine, 2020, medRxiv preprint
[10] Gaurav Pandey, Poonam Chaudhary, Rajan Gupta, Saibal Pal, SEIR and Regression Model based COVID-19 outbreak predictions in India, arXiv:2004.00958
[11] https://analyticsindiamag.com/a-machine-learning-approach-for-monitoring-covid19-indicators/