Data Analysis and FuturForecasting of COVID-19

K. Sreesh Reddy
Computer Science and Engineering
PES University
Bengaluru, India
lreddy.sreesh224@gmail.com

Anup Nagareddy
Computer Science and Engineering
PES University
Bengaluru, India
anupnagareddy@gmail.com

Prathvik Nayak
Computer Science and Engineering
PES University
Bengaluru, India
prtvk13@gmail.com

***Abstract*—The outbreak of COVID-19 has affected over 218 countries around the world. The number of infected and deceased patients have been increasing at an alarming rate. It has therefore become necessary for the inculcation of forecasting techniques to develop better strategies and taking preventive decisions. In this study, we focus on in-depth analysis and forecasting techniques to predict the number of cases in the foreseeable future. These predictions might help to prepare against possible threats and consequences.**

***Index Terms*—Analysis and forecasting, COVID-19**

I. Introduction And Background

In the ultimate month of 2019, an outbreak of a virus was announced by the Chinese Government from an undetermined location in Wuhan of the Hubei Province. Since then the virus, now identified as (SARS-CoV-2) or Coronavirus, has spread to over 218 countries and has affected countless lives. As of March this year, the virus has reached all the seven continents. The virus has spread globally, causing thousands of deaths and having a significant impact on the countries' economies and health systems. The WHO (World Health Organisation) had declared the COVID-19 as a *pandemic*. What exactly are coronaviruses? They are a family of viruses that are the reason behind illnesses such as respiratory diseases and gastrointestinal diseases. People usually intertwine the usage of "COVID-19" and "Coronavirus". CoronaVirus is the virus which causes the COVID-19 disease. Coronaviruses are zoonotic, which means that the virus can be transmitted between human beings and animals. Reports show that clinical deterioration occurs quickly, most probably in the 14 days since getting it. Looking at the symptoms of COVID-19, the most common among them are fever, tiredness, and dry cough. Some patients have a bit more severe complications like diarrhoea, sore throat, runny nose, nasal congestion, and aches and pains. Few of the people who have contracted the COVID-19 virus do not develop any of the symptoms and feel perfectly well. These types of cases are known as *asymptomatic* cases. About $4/5^{th}$ of the population who contracted the virus can recover without the need for any specific treatment. But around 1 out of every 6 people becomes severely ill and develops difficulty in breathing. Anyone who experiences such symptoms even at a mild level should consult for medical help with all precautions so as to not affect others if the person does tests positive. There has been a cloud of doubt cast over how the virus can be transmitted. The virus can spread from one person to another through little droplets from the mouth or the nose which exists in the atmosphere when a person having COVID-19 coughs or exhales. These droplets also stick to the surfaces. People can contract it if they touch these surfaces and then touch their eyes, nose or mouth through which it can enter the body. COVID-19 can also be contracted if a person breathes in the droplets exhaled by a person with COVID-19. Hence, in the new normal, it is always recommended to go out with a mask and stay preferably more than 1 meter from others. This pandemic might be the greatest challenge our world has had to overcome since World War II in the 1940s. The virus, apart from its health dangers, has also led the world into a socio-economic crisis. It has the potential to leave a devastating scar in the social, political and economic fields. The World Bank has projected a 110 billion dollar decline in remittances this year, which means 800 million people might not be able to meet their daily needs. Due to an insufficient number of test kits, ventilators, oxygen tanks, hospital beds, and the current unavailability of treatment or vaccine, it is essential to analyze the growth rates of the positive cases, number of recoveries, and several other factors affecting the growth of the disease. For example, if the government had an idea of the number of forecasted cases which might occur the next day, they can make preparations accordingly for the necessary medical equipment. In our project, we analyze and visualize the COVID data since visualizations are easily understandable and to forecast the future cases using the present data at a much more minor level than the experts in the industries of Data Science and AI are doing today. Machine learning and Artificial Intelligence methods have recently made their way into the healthcare field and have had a huge impact and thus, helped medical staff in the long run. We feel Data Analytics is absolutely essential in these recent times to portray the COVID-19 data in an easily understandable way. The graphs and plots show us the steepness in the rise of the cases and alarms us of the damage the virus is doing and hence take necessary precautions. The decisions made during these times by the governments will be of utmost importance and Analytics is an easier way to help those decisions.

II. Previous Work

Since the last decade, technologies have played an insurmountable role in the field of health, and now the Health Services are seeking for help and support to fight COVID-19. In some papers, the authors stated the possible applications of trending digital technologies like internet of things (IoT), big-data analytics, AI (AI), deep learning and blockchain technology to develop strategies for monitoring, detection and prevention of epidemic and also to spot the impact of the epidemic to the healthcare sector. In an extremely research intensive work proposed by Benvenuto et al., authors had opted for an autoregressive integrated moving average (ARIMA) model to predict the spread of COVID-19. During this paper, the author forecasted the varied parameters for the following 2 days supporting the study about the prevalence and incidence of the COVID-19. This research work also shows the correlogram and graph of the ARIMA forecast for the epidemic prevalence and incidence. Deb et al. proposed a statistical method to analyse incidence patterns and therefore the estimated reproduction number of COVID-19 outbreak. They performed a statistical analysis to view the outbreak trends in the epidemiological period so as to implement various policies during this time to keep the virus in check.

As per the current situation, it's important to know the first spread patterns of the infection to contain and put in effect the various measures in place.In this direction, Kucharski et al. proposed a scientific model of critical SARS-CoV2 transmission by using different datasets to review the COVID-19 outbreak inside and outside the doors of Wuhan. Using that, they had felt that there might be a possible outbreak outside of Wuhan. Recently there are several studies conducted on the epidemiological outbreak of COVID-19 using exploratory data analysis (EDA) supporting various available datasets. The studies mainly center around the occurrence of confirmed, death, and recovered cases in and around Wuhan and also the remainder of the globe to understand the suspected threats and subsequent planning of containment activities. Lauer et al., in their research work raised the problem of the severity of the time period for COVID-19. They studied 181 confirmed cases and identified that the time period may vary from 5 days to 14 days and supported this better surveillance and control activities are often planned.

In recent research work, Singer analysed data of 25 of the infected countries to follow predictions in a short period of time about the COVID 2019 outbreak. This research pointed out that the spread of the virus is either steady or it is explosive depending upon the different parameters. With this understanding, the authors analysed the impact of lockdown in various parts of the globe.

Based on the above literature, it's evident that there is sufficient work out there on exploratory data analysis to know the present trend of the epidemic but there is still plenty of scope to develop and test efficient machine learning based prediction models in order that proactive strategies can be identified to cater the immediate needs. The precise issue we would like to handle and help with is that the un-sureness of the cases within the future and helping everyone have an inspiration of how the graph might be taken care of in time. Machine learning and Artificial Intelligence algorithms play a very important role, especially in this day and age, in epidemic analysis and forecasting. Within the presence of massive epidemic data, machine learning techniques help to seek out the epidemic patterns in order that early action may be planned to prevent the spread of the virus.

III. Proposed Solution

Since this is a pandemic which has occurred rarely in our history, a model which is specific to such kinds of outbreaks have been developed for us to use. The model or solution we try to apply here is the ***SIR Model***, a Differential Equation Model which is an optimal model for recording the spread of a disease. In the first step of the modelling process, we try to identify the independent and the dependent variables. Here, the variable which is independent is time *t*. It is one of the most primitive compartmental models, and many other models are derivatives of this model. This model consists of three parts:

S: This represents the number of **susceptible** individuals. Whenever a susceptible person comes in contact with an infectious person, the person who is susceptible contracts the disease and will be converted into an infectious compartment.

I: This shows the number of individuals who are **infectious**. These kinds of people are those who can infect a susceptible person.

R: Here, R represents the people who are **removed** from the infections. i.e. they have already contracted the disease and are now immune to it, or they are currently deceased. It is usually assumed that the number of people who die due to the disease are not comparable to the size of the population.

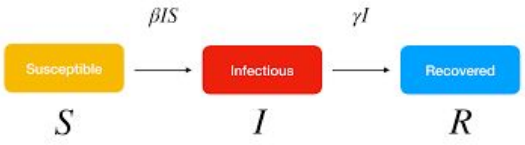

Fig: The SIR Model's Workflow

The SIR Model allows us to describe the number of people in each compartment with an ordinary differential equation. **β** is a parameter which controls how much the disease can be transported through exposure. The factors that influence this are the chance of coming into contact and the probability of transmission of the disease. **γ** is a parameter which expresses how much the disease can be recovered from in a specific amount of time. Since, once the people contract the disease and then heal back, there is no way for them to contract the same disease again.

Usually, the effect of the natural birth or the death rate is

not considered or considered negligible since the SIR model assumes that the long period of the disease is much shorter than the lifetime of the average human being. Due to this, we know the importance of knowing parameters such as beta and gamma. When we are able to estimate both the values, there are several insights which can be derived from them. D is generally the period of time a person stays in the *infectious* period.

$$\frac{dS}{dt} = -\beta IS$$

$$\frac{dI}{dt} = \beta IS - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

Fig: The SIR Differential Equations

The model was implemented by first defining the equations for Susceptibility, Infected and Recovered/Deceased. Then the ***Runge-Kutta method*** is used for the three dimensions. We made use of the *odeint* function in Python where the differential equations are solved given a model, the initial conditions for the differential states and the time at which the solution should be reported.

We also use a newer model which is the SIR Model but with more components since the model has to be as close to the real world as possible. Hence, we add three more parameters to the model making it an ***SEIHRD Model.*** To make the model more accurate, we add parameters such as:
E(t) : They are the number of people who are exposed on day t.
D(t) : Number of people who are dead on day t.
H(t): No. of people who are hospitalized on day t.

For preprocessing, we make an array of columns which we need to clean such as *'Deaths', 'Confirmed', 'Recovered', 'Active'*. We filled the missing values in the *'Country_Region'* with an empty string ‘’ and the columns whose values were numerical in nature had their missing values replaced with 0 using the fillna() method. Since the columns we arranged in an array were the only ones we needed to clean, we dropped all the unnecessary columns which had nothing to do with the prediction or wouldn't help the prediction in any way such as *'Last_Update'*.

The dataset was collected from the official John Hopkins Repository for the COVID-19 data which automatically updates to the most recent numbers. Some interesting visualizations were carried out using the data where it was observed that the rise of the cases were very steep in recent times and the United States of America and India were almost at the forefront in most of the cases. Another well known source for collection of COVID-19 was the website ourworldindata, which was majorly used in the building of the SIR and SEIHRD Models.

Various types of Graphs were visualized based on the COVID-19 data. Bar Graphs were made for comparing Cases with Countries and Continents alike.

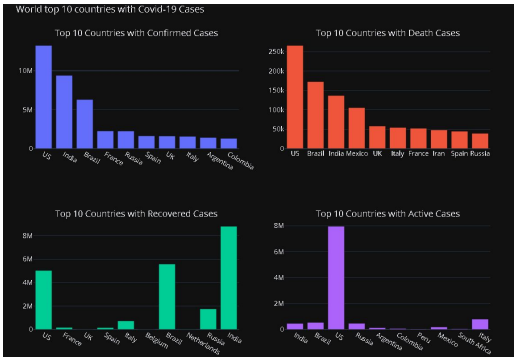

Fig: Plots between countries for various types of cases

Bubble Graphs were also visualized for a better understanding of the difference between severely affected countries like America, Brazil, India and not so severely affected. But the most fascinating part about the dataset was that we were able to visualize the data on a globe. The animations using choropleth helped us get the idea of the spread of the virus around the world in a matter of months.

Fig: Animating the spread of COVID-19

To get an idea about how the intensity of the cases of people who have recovered is distributed worldwide, we have plotted another choropleth basemap:

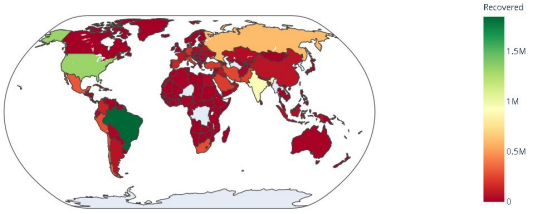

Fig: Intensity of Recovery Cases around the globe

IV. Experimental Results

The SIR Model has many versions, one which considers the births and the deaths (known as SIRD Model with

demography), with states in between. But our interest isn't long term. We will assume that people develop immunity and once the patient has reached the *recovered* stage then there is no chance of him going back to the other 2 stages, i.e. susceptible and *infectious*. This assumption would not work in the long term because the immunity of a person after developing antibodies against the virus may be lost from 3 to 6 months and COVID-19 may come back at a certain season.

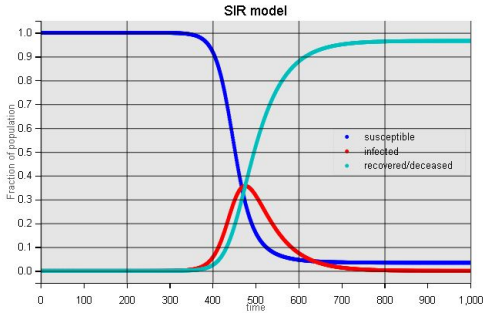

Fig: Graph of SIR Model

The implementation of the SIR Model can be done in many ways. It can vary from using the dynamics on a graph (Social Networking purposes) or it can be done using the differential equations - with mean field approximation.

To not go into many complexities, we chose the first option. To solve the differential equations, we use a numerical method known as runge-kutta of the *4th* order.

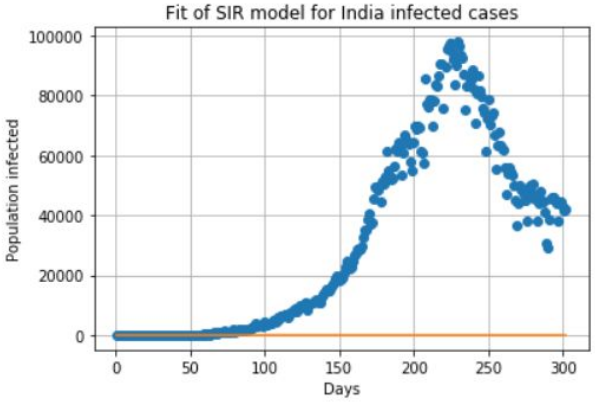

Fig: The SIR Model fitting for the India Infected Cases.

Which is similar to the actual graph:

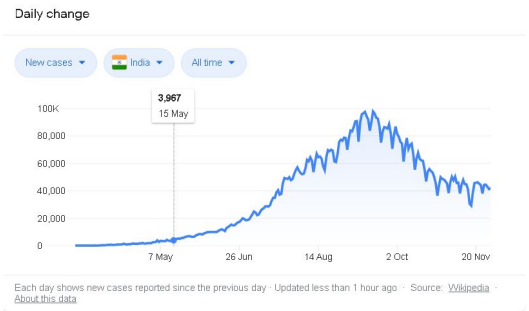

Fig: Actual graph of Indian cases from Google

But this can be improved using the SEIHRD Model. Some new parameters are added such as **α** (the fatality rate), **ρ** (Rate at which people are dying) etc. for the new SEIHRD Model. We use the US data here:

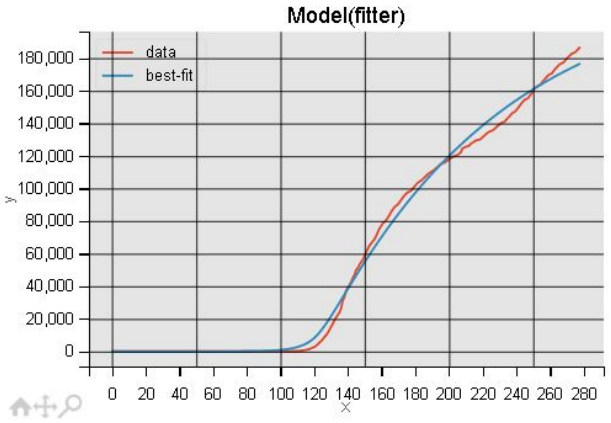

Fig: The fitter after applying the SEIHRD Model

Another modelling we did was on China's data where we fit the model into China's data and further improved the model with the addition of a two-time delay for the individuals who are infected. To find the rates of contact and recovery, the Monte Carlo simulation was performed, to run the simulation in China since there is data available for all three stages, early infection growth, rapid infection growth and near zero infected cases.

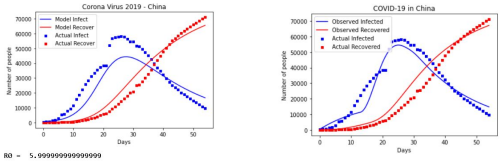

Fig: SIR Model in China (after modelling)

Since this model is apt for the current situation we all are in, we felt that this is an appropriate time to learn about the compartmental models. Although these models are very efficient for forecasting the future of the outbreak of a pandemic, this model however cannot be used for many of the other fields machine learning and artificial intelligence is required in.

V. CONCLUSIONS

The paper contains the data analysis and the forecasting of COVID-19 cases, with data collected from the official John Hopkins Github repository and OWID (Our Word in Data) to carry out different types of EDA and visualizations. The forecasting of cases was done using the SIR Model, and improved upon using the SEIHRD Model. This was the first

time we encountered such a model which exists to model the outbreak of the disease, and found it very interesting. Sreesh's contributions were to find the compartmental models, Prathvik's contributions were to gather and visualize the data , and Anup's contribution was to perform EDA on the data. We all worked together and helped each other whenever we were stuck. The main learnings we obtained from our research for the papers and the models was the impact which analytics had on the current generation and the situation. The visualizations we did were the ones which were seen on the news channel almost daily to understand how successfully the governments and their countries are dealing with the containment of the virus. The main slogan during the lockdown was to 'flatten the curve' which was inspired from the peaks countries were observing through analysis of data. Since our topic is one which has impacted people all around the world and is austere in nature, we would like to give a few precautions and how to contain this pandemic.

To prevent the spread of COVID-19:

-> Always keeps your hands clean. Use alcohol-based sanitizers to clean.

-> Always keep a safe distance from any person who is coughing or sneezing.

-> Whenever social distancing isn't viable, always keep a mask on you.

-> Do not touch your eyes, nose or mouth after touching any surface.

-> Whenever you cough or sneeze, always use the inside part of your elbow.

-> If you feel uneasy, it is better to not step out and stay at home..

-> If you feel any of the symptoms of the virus, go seek medical help.

This pandemic is far from over and the only way to conquer it is by working together and keeping in mind yours' and others' health and safety. Stay well! Stay safe!

VI. References

Some of the references that are not cited here were used to gain a deeper insight of the respective topic. The information provided by these references have been used in the document as a whole and cannot be cited at specific locations.

Ting, Daniel Shu Wei, Lawrence Carin, Victor Dzau, and Tien Y. Wong. "*Digital technology and COVID-19*." Nature Medicine (2020).

Benvenuto, Domenico, Marta Giovanetti, Lazzaro Vassallo, Silvia Angeletti, and Massimo Ciccozzi. "*Application of the ARIMA model on the COVID-2019 epidemic dataset*." Data in brief (2020)

Deb, Soudeep, and Manidipa Majumdar. "A time series method to analyze incidence pattern and estimate reproduction number of COVID-19." (2020).

Kucharski, Adam J., Timothy W. Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M. Eggo et al. "*Early dynamics of transmission and control of COVID-19: a mathematical modelling study*." The lancet infectious diseases (2020).

Dey, Samrat Kumar, Md Mahbubur Rahman, Umme Raihan Siddiqi, and Arpita Howlader. "*Analyzing the Epidemiological Outbreak of COVID-19: A Visual Exploratory Data Analysis (EDA) Approach.*" Journal of Medical Virology (2020).

Lauer, Stephen A., Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. "*The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application.*" Annals of internal medicine (2020).

Singer, H. M. "*Short-term predictions of country-specific Covid-19 infection rates based on power law scaling exponents.*" (2020).

https://www.undp.org/content/undp/en/home/coronavirus.html

https://www.who.int/health-topics/coronavirus#tab=tab_1

https://www.physio-pedia.com/Coronavirus_Disease_(COVID-19)

http://cajgh.pitt.edu/ojs/index.php/cajgh/article/view/466

https://en.wikipedia.org/wiki/Runge%E2%80%93Kutta_methods

https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology

https://www.kaggle.com/saga21/covid-global-forecast-sir-model-ml-regressions#2.-SIR-model-

https://www.lewuathe.com/covid-19-dynamics-with-sir-model.html

https://www.youtube.com/watch?v=NKMHhm2Zbkw

https://www.lewuathe.com/covid-19-dynamics-with-sir-model.html?fbclid=IwAR2t1lhNhmZOHB_DHf2Tvy3rDOjsr-W7uCJE8GaodRjmLv1oxjV5cKpZ4o8