PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

MAY 2020: IN SEMESTER ASSESSMENT (ISA) B.TECH. IV SEMESTER

UE18MA251- LINEAR ALGEBRA

MINI PROJECT REPORT

ON

COVID 19 Data Modelling and Predictions in India

Submitted by

1. Name: Rohit Kumar A. S SRN: PES1201801890

2. Name: Hemanth Alva R SRN: PES1201801937

Branch & Section : CSE, 'J' section

PROJECT EVALUATION

(For Official Use Only)

Sl.No.	Parameter	Max Marks	Marks Awarded
1	Background & Framing of the problem	4	
2	Approach and Solution	4	
3	References	4	
4	Clarity of the concepts & Creativity	4	
5	Choice of examples and understanding of the topic	4	
6	Presentation of the work	5	
	Total	25	

Name of the Course Instructor :

Signature of the Course Instructor :

COVID-19 Data Modelling and Predictions in India

Rohit Kumar A. S
Department of Computer Science and Engineering
PES University
Bengaluru, India
rohitkumaras24@gmail.com

Hemanth Alva R
Department of Computer Science and Engineering
PES University
Bengaluru, India
hemanthalva2708@gmail.com

Abstract - On 31 December 2019, the first reported case of COVID-19 was recorded in Wuhan,China and has spread to more than 50 countries. The number of COVID-19 cases are increasing at a rapid pace and the national and local autorities are having a hard time analysing and predicting the spread of COVID-19. The main aim of this paper is to draw statistical modals and find patterns for a better understanding of COVID-19 spread by throughly studying the report. This report describes modeling efforts for evaluating the current level of COVID-19 infections in the world, with a special focus on India. It is noticed that lock-down and isolation are the important techniques to prevent the spreading of the disease. Mathematical modeling of an epidemic like COVID-19 is always useful for strategic decision making, especially to gain understanding of the future of the pandemic in densely populated countries like India. Linear regression, SVM and SIRD (Susceptible-Infected-Recovered-Dead) techinques are all looked and analysed with primary focus on the SIRD model.

I. INTRODUCTION

COVID-19 is a contagion belonging to the "Nidovirus family or "Nidovirales" which includes "Coronaviridae", "Artieviridae" and "Roiniviridae"family, responsible for respiratory illness in humans which may cause common cold to more austere diseases such as "Middle East Respiratory Syndrome(MERS)" and "Severe Acute Respiratory Syndrome(SARS)". The most common trials or symptoms of COVID-19 are fever, tiredness, dry cough, nasal congestion, running nose or sore throat. The spread of this infection is found to be so alarming that the World Health Organization (WHO) declared it as a pandemic disease on 11th March 2020. An infectous disease outbreak is the occurence of a disease that is not usually expected in a perticular community, geographical region or time period. COVID-19 has presented an unprecedented challenge before the world and the number of infected people are increasing exponentially throughout the world (fig.1). Studies till date show that COVID-19 is mainly spread through contact rather than tranmitted through air.

There has been a gradual rise in the number of infection cases as seen in Figure 1. In response, India has implemented international travel bans and a strict lockdown. However, India as a higher risk because of a very large population density, limited infrastructure and healthcare systems on a very large demands, but factors like warm climate as well as humidity may favor India. In the absence of a vaccine, social distancing has emerged as the most widely adopted strategy to control the outbreak of the novel corona virus. Most pandemics follow an exponential curve during the initial spread and eventually flatten out. A classic epidemiological model framework would be capable of describing the dynamics of COVID-19 and predict the impact of the current measures undertaken by the Government of India. In India, the first case of COVID-19 was reported on 30 January 2020 and has since spread to almost every state with a rate of infection approximately 1.9, which is relatively lower than most countries.

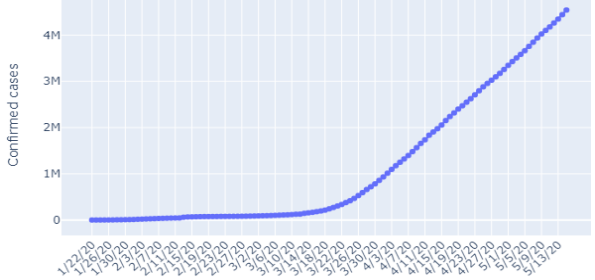

Figure 1:Total Confirmed COVID-19 Cases (globally)

In the present work, we use one of the most effective epidemeological models; SIRD which is simple yet quite effective to predict the future of the epidemic and the effects of lock-down and social distancing via a time-dependent coefficient of the model for India by using the existing data. Exponential and classic susceptible-infected-recovered-dead(SIRD) model based on the avaiable data are used to make short and long-term predictions on a daily basis. The model studies the data with realistic parameters and it shows that even for the relatively small sample size of 600,000 ,the epidemic will be at the peak at around second week of June.

It is reported that there are two important stages of COVID-19; Stage-II and Stage-III. In Stage-II, there is person to person transmission and in Stage-III, there is community transmission, curently India is in Stage-II. According to the stage of COVID-19, a nation can decide on the plan of action to be taken. It is found that the basic reproduction number R_o for India is in the range of 2-5 currently but it varies heavily with location. R_o is the transmission rate given that the population has no immunity from the past exposures or vaccination, nor any deliberate intervention in disease transmission. The number of infections grow and spread in the population if $R_o > 1$.

The current model does not account for factors such as the weather or humidity changes . Several studies have reported that the effects of COVID-19 may change when the weather becomes warmer and other factors,such as differential immunity of Indians due to BCG vaccine are already implicity assumed in the data in the form of basic reproductive number. Curent models predict the transmissions due to Stage-I (individuals with a travel history to high-risk countries) and Stage-II(person-to-person contact). It should be noted that if the reported number of cases begins to exceed the predicted end-state systematically, then the pandemic will enter a new stage, and none of the models described above will be applicable. India having high population density as well as social and demographical issues, it has put India in a high risk for community transmission(Stage-III).In this paper, we propose a mathematical model for constrained scenario, i.e., with lock down, quarantine. This model can approximately predict the number of new Covid19 cases and can provide a concise look at the state of the pandemic in our nation.

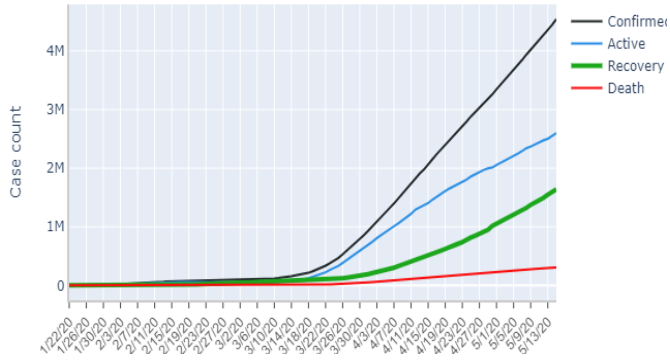

Figure 2: COVID-19 case status(22nd Jan – 13th May)

Figure 2 shows the Infected, Active, Recovered and Dead numbers for 22nd Jan-13th May for India. The rising numbers are alarming and calls for immediate and strict measures against the havoc-wreaking pandemic.

II. LITERATURE SURVEY

In [1], the author used Exploratory Data Analysis (EDA) to assess the spread of the pandemic in India in comparison to other countries, where it was found that India outperformed most countries in slowing the spread of the disease. Age-wise analysis of the Covid-19 confirmed cases was also done; where it was found out that 42% of cases are in people aged 21-40. State wise analysis was also performed on all the states of the country. Maharashtra, Gujarat and Delhi were found to be most affected. Based on Covid-19 test results, it was found that Fever was the most common symptom, found in more than 80% of confirmed cases followed by Dry cough, found in almost 70% of cases. Time-series data (until April 21st) scraped from many sources was passed through the ARIMA model and it predicted that the number of confirmed cases would increase by 25,000 every day and the total number of cases in the country would reach 200,000 by May 7th. A simple analysis of the number of hospital beds vs. number of cases was also performed where it was found that the total number of hospital beds available would run out by August 2020 and further accommodation would be impossible.

In [2], the researchers used the SIR (D) model to predict the future of the pandemic in India. They also estimated the effect of lockdown/social isolation using a time-dependent co-efficient of the model. Two cases; optimistic and pessimistic were considered with different values of parameters. Without lockdown, in the optimistic scenario, it was found out that the peak would come at around September 2020 and that almost 200 million people would be infected. In the pessimistic cases it was found that the peak

would come at the end of June, where almost all of the population was infected. A 40-day lockdown delayed the appearance of the peak by approximately forty days with more than 100 million deaths while a 100-day lockdown delayed the peak by the same number of days but the total deaths would be in millions. Based on the data available during the first lockdown period, the researchers found that even with a 100-day lockdown, the peak would appear by the end of June.

In [3], the researcher used the SIR model and exponential model to compare the growth of the pandemic in India with various affected cities in the US. It was found that the pattern of Washington DC matched closely with India. Since the outbreak happened earlier in Washington DC, the paper proposed that study of the pandemic in Washington DC might help us predict the course of the disease in India. Long term predictions based on this theory predicted the number of cases to touch 0.5 million by May 2020, with single day increases as large as 90,000. The paper also studied the effect of lockdown in Hubei, China. It was found effective with the curve flattening after 30-40 days after imposing the lockdown. The paper also analysed that since China imposed the lockdown only after reaching Stage 3, i.e. community transmission, it might've been less effective. Since India imposed the 'Janata Curfew' in stage 2, the paper predicted that the lockdown would have greater effect in controlling the spread.

In [4], the researchers used a "Bats-Hosts-Reservoir-People transmission network" model for simulating the potential transmission from the infection source (assumed bats) to the human infection. Since the Bats-Hosts-Reservoir network was hard to explore clearly and public concerns were focusing on the transmission from Huanan Seafood Wholesale Market (reservoir) to people, it was simplified as Reservoir-People (RP) transmission network model. The next generation matrix approach was adopted to calculate the basic reproduction number (R0) from the RP model to assess the transmissibility of the SARS-CoV-2. The value of R0 was estimated of 2.30 from reservoir to person and 3.58 from person to person. It's expected that the number of secondary infections that result from introducing a single infected individual into an otherwise susceptible population was 3.58.

In [5], the authors assess the effectiveness of various stages of government intervention on the effective reproductive number ('R') in China. Using the parameterized SIRD model, they simulated the spread dynamics of corona virus disease 2019 (COVID-19) outbreak and impact of different control measures, conducted the sensitivity analysis to identify the key factor, plotted the trend curve of effective reproductive number (R), and performed data fitting after the simulation. By simulation and data fitting, the model showed the peak existing confirmed cases of 59 769 arriving on 15 February 2020, with the coefficient of determination close to 1 and the fitting bias 3.02%, suggesting high precision of the data-fitting results. More rigorous government control policies were found to be associated with a slower increase in the infected population. There was an upward trend of R in the beginning, followed by a downward trend, a temporary rebound, and another continuous decline. The feature of high infectiousness for severe acute respiratory syndrome corona virus 2 (SARS-CoV2) led to an upward trend, and it was predicted that the government measures contributed to the temporary rebound and declines. The declines of R could be exploited as strong evidence for the effectiveness of the interventions. The researchers used a parameter ('k') which denoted the ratio of suspected to confirmed cases to analyse the effect of varying degrees of governmental measures. The faster and more the intervention, the susceptible population decreased lower and sooner.

III. MODEL DESCRIPTION / REPORT ON THE PRESENT INVESTIGATION

Before analysing the epidemiologically more effective SIRD model, we analysed the data using Linear Regression and Support Vector Machine (SVM) models.

1) Least squares Linear Regression

Linear regression [6] is a method for modelling the relationship between two scalar values: the input variable x and the output variable y. The model assumes that y is a linear function or a weighted sum of the input variable: $y = b0 + b1.\ x1$. The objective of creating a linear regression model is to find the values for the coefficient values (b) that minimize the error in the prediction of the output variable y.

Linear regression can be stated using Matrix notation; for example: Ax= b, where

$$A = \begin{bmatrix} a11, a12 \\ a21, a22 \\ a31, a32 \end{bmatrix} \quad x = \begin{bmatrix} x1 \\ x2 \end{bmatrix}$$

[b1
b = b2
b3]

Ideally, the vector 'b' should lie in the column space of A but in many real-world cases, it doesn't (Fig 3)

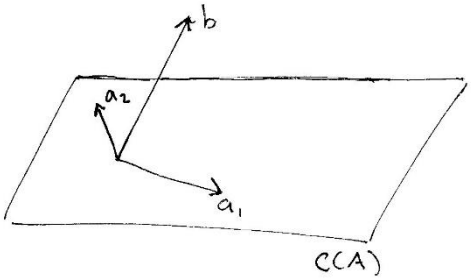

Figure 3: b lies outside the column spcace of A

We swap out b for another vector that's pretty close to it but that fits our model. Specifically, we want to pick a vector p that's in the column space of A i.e. C(A) but is also as close as possible to b. Hence, we project 'b' onto C(A) and this projection is called vector 'p'.

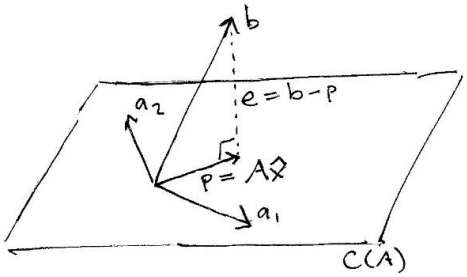

Figure 4: 'p' is projection of 'b' on the column space of A

The line marked e is the "error" between our observed vector b and the projected vector p that we're planning to use instead. The goal is to choose the vector p to make e as small as possible. That is, we want to minimize the error between the vector p used in the model and the observed vector b.

Since the vector e is perpendicular to the plane of A's column space, that means the dot product between them must be zero i.e.

$$A \cdot e = 0$$

But since e = b - p, and p = A times x-hat, we get,

$$A^T(b - A\hat{x}) = 0$$

Solving for x-hat, we get,

$$A^T b = A^T A\hat{x}$$

$$\Rightarrow \hat{x} = (A^T A)^{-1} A^T b$$

This x hat is our linear regression parameter that helps us draw the best fit line with least mean-square error.

2) Support Vector Machine (SVM):

Support vector regression (SVR) [7] is a statistical method that examines the linear or non-linear relationship between two continuous variables.

The equation of the line in its simplest form is described as below **y=mx +c.** In the case of regression using a support vector machine, we do something similar but with a slight change. Here we define a small error value ***e*** (error = prediction - actual). The value of ***e*** determines the width of the error tube (also called insensitive tube). The value of ***e*** determines the number of support vectors, and a smaller ***e*** value indicates a lower tolerance for error. Thus, we try to find the line's best fit in such a way that:

(mx+c)-y ≤ *e* and **y-(mx+c) ≤ *e***

Also, we do not care about errors as long as they are less than ***e.*** So, in this case, only those data points that are outside the ***e*** error region will be contributing to the final cost calculation. Hence, the support vector regression model depends only on a subset of the training data points, as the cost function of the model ignores any training data close to the model prediction when the error is less than **e.**

SVM is memory efficient, which means it takes a relatively lower amount of calculation resources to train the model. This is because presenting the solution by means of a small subset of training points gives enormous computational advantages.

3) SIRD Model:

SIRD [8] model is a compartmental model describing the dynamics of infectious disease. The model divides the population into compartments. Each compartment is expected to have the same characteristics. SIR represents the three compartments segmented by the model;

S(t) – Susceptible to disease but not infected yet

I(t) – Infected by the disease

R(t) – Recovered from the disease and immune

D(t) - Dead from the disease

The model is built on the fact that on any given day the total population can be divided into mainly four different categories namely those who are susceptible to the infection (S), already infected people or active number of patients (I), number of recovered people (R), and the number of deaths (D). It is customary to neglect the usual daily birth and death rate which is also known as demographic. The SIRD model also assumes that the quantities S, I, R, and D are dependent only on

time and not on the location. SIR model is a framework describing how the number of people in each group can change over time (Fig. 5)

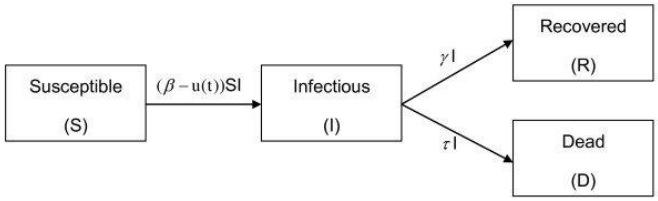

Figure 5: SIRD model mechanism

The time rates of change of these quantities are given by the following coupled ordinary differential equations:

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

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

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

$$\frac{dD}{dt} = \mu I,$$

Figure 6: Differential equations of SIRD model

where β is the transmission rate, μ is the mortality rate and γ is the average recovery rate. The constant N is not the population of the country but population composed of susceptible (S), infected (I), recovered (R) and dead(D).

In the present model, at any time the total population N = S(t) + I(t) + R(t) + D(t) remains constant as **dS(t)/dt + dI(t)/dt + dR(t)/dt = 0**

We note that the dynamics of the infectious class depends on the ratio β/γ, called the reproduction number (R_0). R_0 is the average number of people infected from one other person. If it is high, the probability of pandemic is also higher.

IV. RESULTS

The data for the models have been taken from "John Hopkins University Centre for Systems Science and Engineering (JHU CSSE)" [9] and "Our World in Data" website which scrapes data from the "European CDC" [10]. The dataset undergoes the process of choosing of essential columns using filtering and visualizing the data in the graphical format. This paper used "Python" for data processing, "data visualisation" and "data prediction". It uses "pandas" and "NumPy" libraries to process and extract information from the available dataset. Appropriate graphs are created for better visualization using "Matplotlib" and "Plotly" along with the above-mentioned libraries. "Sklearn" was used for data prediction.

A. Comparison of India with the world

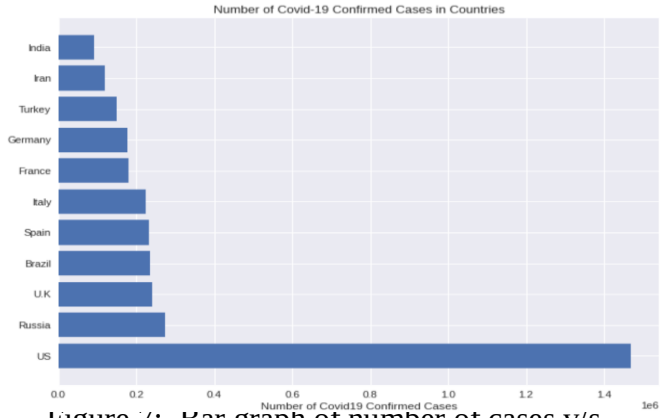

Figure 7: Bar graph of number of cases v/s countries

Inferences from figures 7 and 8:

- US are the worst affected country in the world with more than 1.4 million cases.
- Contrary to popular expectations, the developed Western countries have been hit worse than the developing countries.
- Even though 'China' was the origin and initially the worst hit country, it does not feature in the top 10 (as of May 15,2020)
- India comes in 11th position with almost 0.1 million cases and it has performed better than other world powers.
- US accounts for more than a quarter of the pie chart, owing to their 1.4 million cases out of the world total of 4.5 million.

B. Covid-19 spread timeline in India

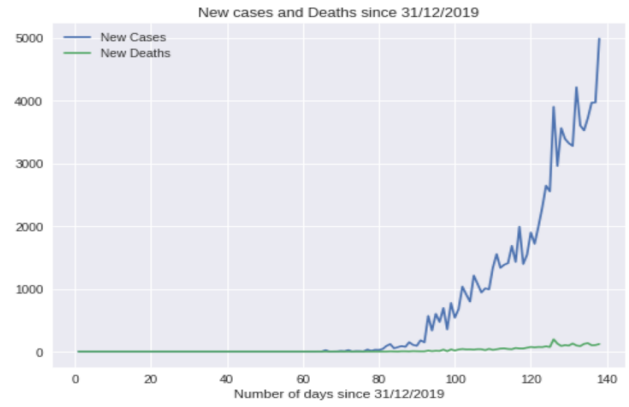

Figure 8: New cases and deaths per day in India

The graphs are plotted for 130 days starting from 31/12/2019. In both graphs, blue line represents cases and the green line represents deaths.

Inferences from figures 9 and 10 :

- Number of new cases per day has seen a drastic increase day by day and is reaching worrying levels.
- Number of new deaths has been very less, which indicates that the recovery rate is high in India.
- The plot of total cases shows an exponenetial growth. This is expected for a pandemic of this magnitude.

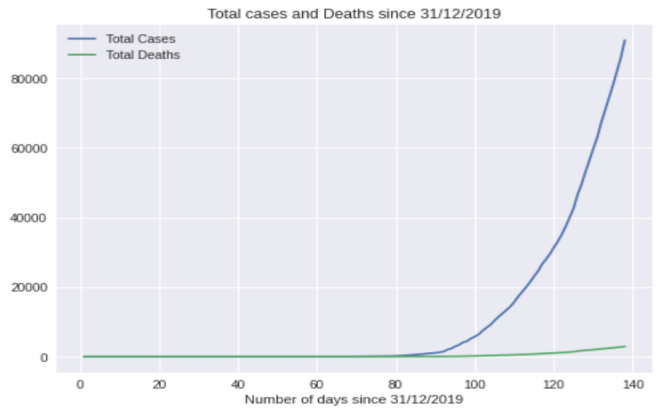

Figure 9: Total cases and deaths in India

C. Linear Regression And Exponential Growth

Best fit and natural log line was drawn starting from Day 65 onwards, this is because India only experienced new case growth after Day 65.

As seen from the figures 11 and 12, the exponential growth goes haywire and predicts that by Day 150 (May end) there will be more than 45,000 new cases every day !!

While exponential growth is useful to predict how a disease would take course if left unattended, it assumes that people who are ill will never get better or attain immunity !! Hence, we next try a better model : SVM regression.

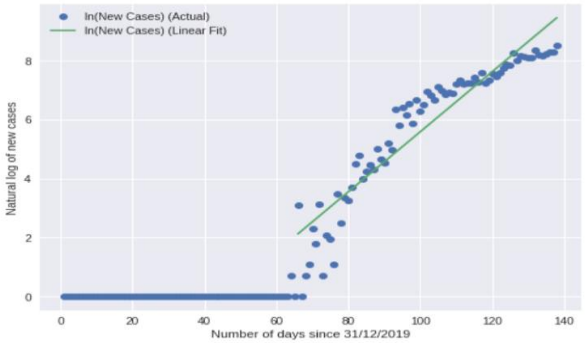

Figure 10: Natural log of new cases and the linear model fit to that data

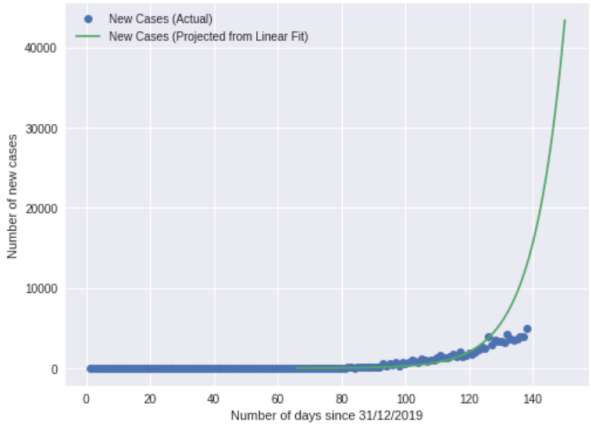

Figure 11: Actual data and prediction of next 10 days (Day 140-Day150)

D. Support Vector Machine(SVM) Regression

SVM Regression was used to plot the non-linear regression best fit model seen in Figure 13. The solid blue line is the approximate curve for the number of confirmed cases over time while the dotted purple line represents the SVM predicted curve.

As it can be seen from the comparison of figures 12 and 13, the SVM model gives better prediction than the linear regression model, but it is also non-terminating and shows no peak or flattening of the curve. Hence we next look at a very useful epidimealogical model for pandemics, the SIRD model.

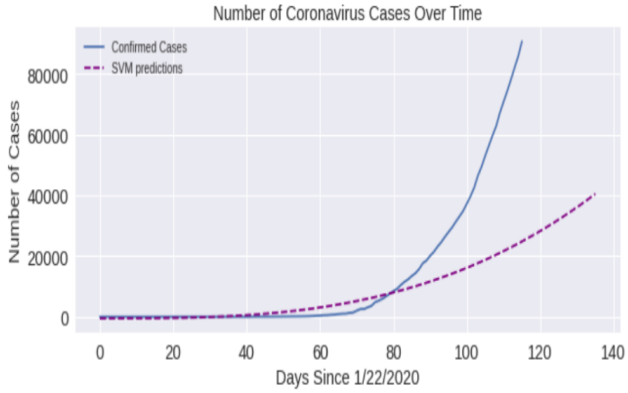

Figure 12: SVM model prediction and confirmed cases plot

E. SIRD Model

The differential equation in Figure 6 were solved by “solve_ivp” function of scipy module. The model is trained and loss i.e Root Mean Square Error between predicted and actual data, is minimised .

The model requires the population/sample size i.e size of susceptible population and the basic reproduction number(R_0) [11]. Since R_0 varies from place to place, the initial R_0 of the worst hit

places such as U.K and certain citites in the U.S was taken for our model as well. This number was in between 4 and 7.

Since no data is available on the susceptible count in the country, we ran tests with various intial conditions with different values of the parameters. Based on eye optimization, a good fit between the prediction and actual data was found .

The susceptible population was taken as 6 lakh and the initial reproduction number was taken as 5. The model was simulated for 200 days starting from 1/22/2020 and the predictions were plotted along with the actual data.

In Figure 14, the brown line represents the number of infected cases and the light blue line represents the recovered cases while the purple lines represents the susceptible cases.

As observed from the figures 14 and 15, there is a peak at mid June where the number of infected cases reaches around 1,74,000 cases. The curve eventually reduces and flattens out after August. The number of susceptible people also reduces uniformly and comes down to less than 50,000 after August.

It is noteworthy to point out that the infected curve doesn't really reach zero even at the end of September. Hence, there is a good chance that we might have to be wary and alert even when the pandemic has died down , for asymptomatic and isolated cases might cause another outbreak. This is also a testament to the high infection rate of the coronavirus and the lack of a vaccine.

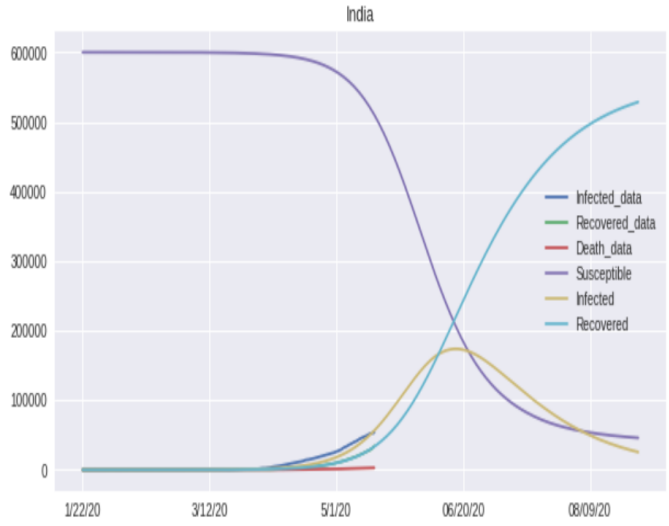

Figure 13: SIRD Model prediction for India.

The peak number of infected cases (brown line) comes at around the end of June. The curve eventually starts flattening out at the end of August but not completely. This is due to the high infection rate of the coronavirus and the lack of a viable vaccine.

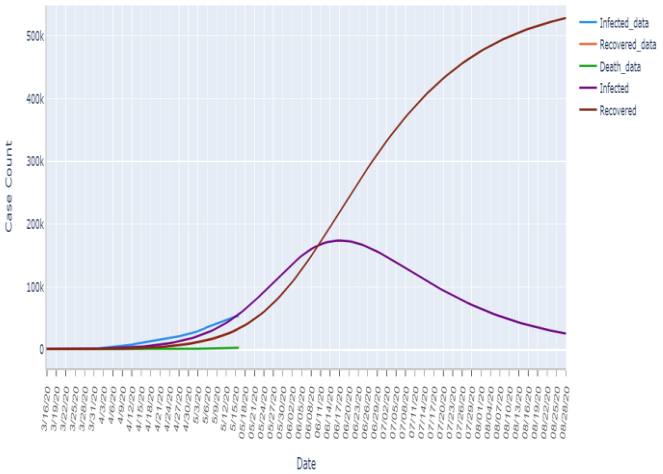

Figure 14: SIRD model without the suscpetible curve.

The peak number of infected cases (purple line) comes at around the end of June. It can be observed that the total number of recoveries doesn't reach the whole population/sample size of 600,000 even at the end of August.

V. CONCLUSION AND FUTURE WORK

The main aim of this paper is to study and analyse the COVID-19 spread and the pattern of outbreak of the virus in India. We get a detailed understanding of how this epidemic is spreading and also the measures to be taken by the Healthcare sector of India and lastly predict the future of this epidemic in India. This Paper work can be extended to higher level in the future where the results can be taken as independent criteria and moreover the results can be extended to be more accurate and predict more accurate number of cases in India.

The present analysis of the SIR(D) model indicates a scary high number of peak of infection for India by comparing the existing data to the model simulation. Moreover, India is on high risk to enter into community transmission due to reported violation of quarantine norms by individuals as well as other social and demographic issues. The predictions made using the current epidemiological model in the current work will be invalid if such an event occurs. The results from this paper should be used only for qualitative understanding and reasonable estimation of the outbreak, but are not advisable for any decision making or policy change. Human civilization has witnessed and survived many devastating pandemics during the past few millenniums and will do the same this time as well. Fortunately, we have advanced technology and medical facilities which will help us fight better against the new pandemic.

VI. REFERENCES

[1] Sarvam Mittal, "An Exploratory Data Analysis of COVID-19 in India", *Article in International Journal of Engineering and Technical Research · April 2020.*

[2] Rajesh, Aditya, et al. "CoVID-19 prediction for India from the existing data and SIR (D) model study." *medRxiv* (2020). DOI: 10.1101/2020.05.05.20085902

[3] Rajesh Ranjan, "Predictions for COVID-19 outbreak in India using Epidemiological models" medRxiv (2020). DOI: 10.1101/2020.04.02.20051466

[4] Chen, T., Rui, J., Wang, Q. *et al.* A mathematical model for simulating the phase-based transmissibility of a novel corona virus. *Infect Dis Poverty* **9,** 24 (2020). https://doi.org/10.1186/s40249-020-00640-3

[5] Fang, Yaqing, Yiting Nie, and Marshare Penny. "Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis." *Journal of medical virology* 92.6 (2020): 645-659.

[6] Linear algebra view of least square regression: https://medium.com/@andrew.chamberlain/the-linear-algebra-view-of-least-squares-regression-f67044b7f39b

[7] Support Vector regression: https://heartbeat.fritz.ai/support-vector-regression-in-python-using-scikit-learn-89cc18e933b7

[8] SEIR model- Cessi : http://www.cessi.in/coronavirus/page1.php

[9] Dataset used for number of confirmed COVID-19 cases : https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv
Dataset used for number of recovered COVID-19 cases : https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv
Dataset used for number of dead COVID-19 cases : https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_dead_global.csv
Country dataset : https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-data/data/cases_country.csv

[10] European CDC Dataset : https://covid.ourworldindata.org/data/ecdc/full_data.csv

[11] Basic reproductive number- Wikipedia: https://en.m.wikipedia.org/wiki/Basic_reproduction_number

[12] SIR model analysis example https://scipython.com/book/chapter-8-scipy/additional-examples/the-sir-epidemic-model/

[13] Data Modelling & Analysing Corona virus (COVID19) Spread using Data Science & Data Analytics in Python Code, blog: "https://in.springboard.com/blog/data-modelling-covid/#"