Data Analysis, Modeling and Forecasting of COVID-19 using Machine Learning

Toshaniwali Bhargav
Dept. of Computer Science
Rungta College of Engg. & Tech.
Raipur, Chhattisgarh, India

Toran Verma
Dept. of Computer Science
Rungta College of Engg. & Tech.
Raipur, Chhattisgarh, India

ABSTRACT

In this work, we present a global analysis and exploring the World Wide data of Covid-19. SIR Model and Mathematical Curve Fitting Method have been used to predict the future spread of the pandemic in India. Odisha, Madhya Pradesh, and Chhattisgarh three states of INDIA are selected based on the pattern of the disease spread in INDIA. The parameters of the models are estimated by utilizing real-time data. The models predict the ending of the pandemic in these states and estimate the number of people that would be affected under the prevailing conditions.

For analyzing and designing this model available datasets have been used. These consist of a record of cases globally from March 21st to June 1st, 2020. Hence we will make two predictions from our model. The first model will analyze the COVID confirmed cases globally. And, the second model fetches real-time data through which we will predict total confirmed cases, total deaths, and total recovery in India.

This model proposes the aim for understanding its everyday exponential behaviour along with the prediction of future reach ability of the COVID-2019 across the nations by utilizing real-time. With lockdown continuing even after May 2020, we expect our model to reflect the peak cases either in the month of September or October 2020.

General Terms

Data Mining, Machine Learning, Mathematical Modeling

Keywords

COVID-19, Data Analysis, Modelling, Forecasting, SIR, Mathematical Curve Fitting, SARS-CoV-2, WHO, Corona Virus

1. INTRODUCTION

The Corona virus disease 2019 is infectious and results from Severe Acute Respiratory Syndrome Coronavirus2. The first case was reported on 17 November 2019 in Wuhan, China and since then it has spread exponentially. The first confirmed case of COVID 19 in INDIA was on 30 January 2019 in the state of Kerala. The concerned person had been to Wuhan before returning to India. The status of confirmed cases in India till 25th June 2020 happens to be 4.3 Lakhs. This has now resulted in the deaths of approximately 14000 people. The worldwide (188 countries) data shows that till 14th July 2020, 13.1 million cases have been reported. This has resulted in the deaths of more than 573000 people. Over 7.26 million people have also recovered after being infected.

Cough, Loss of smell, Fever, and shortness of breath are some of the common symptoms of this virus. It has been observed that the majority of the cases have begun with mild symptoms. The usual time from exposure to the beginning of symptoms is approximately 5 days. On a general level, it may vary between 2 to 14 days.

Coughing, Speaking to one another and sneezing are some of the modes of COVID-19 transmission. Close proximity with an infected person increases the risk even more. Sneezing results in showering droplets in the space which falls on the ground. Smaller Droplets (technically aerosol) remain suspended in the air for a longer duration of time.

A. Source and Modes Of Transmission

Stage 1 indicates the beginning of the disease in people who had some travel history. Infections are very low at this stage.

Stage 2 marks local transmission. The people who got infected via traveling spread the virus to their close friends and family. Everybody who came in close proximity with the infected is usually traced and then isolated.

Stage 3 is Community Transmission. This is when the source of the virus is untraceable and the virus spreads to the public. It is suggested that large scale geographical lockdowns are implemented to reduce virus propagation.

Stage 4 is alarming indicating the impact of a virus as an epidemic. China is the perfect example as it had a large number of infections and unabated deaths.

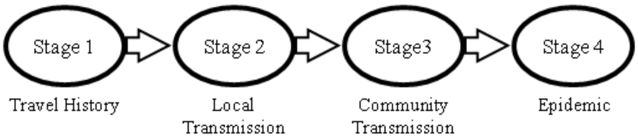


Fig 1 Transmission Stage of COVID-19

2. RELATED WORK

He, K., Zhang et al [4] discussed three-dimensional deep learning, that is COVID-19 detection Neural Network (COVNet) based on volumetric Chest CT images. Here three types of CT images are COVID-19 Community-Acquired Pneumonia (CAP) and two other non-pneumonia cases have been observed and modeled to test the strength of the future Model. The accuracy obtained from AUC is 0.96 to detect the future model of COVID-19 cases. Deep learning [8] process based on the -location attention mechanism and the three-dimensional CNN ResNet-18 network using Pulmonary CT images has also been employed to sense this pandemic The overall accuracy obtained from this method is 86%. It is helpful for doctors to analyze Chest CT images and thereby perform a premature screening of COVID-19 patients. Roosa & Chowell et al [3] Presented Phenomenological Models to forecast the outbreak of Coronavirus. These models help in evaluating short-term forecasts to determine total confirmed cases in the Hubei region. The prediction indicates continuously increasing confirmed cases in Hubei and other regions with 7409-7496 cases in the first 5 days. 1128-1929 cases in the next five days. Till February 24, 2020, the total confirmed cases are 37,415 – 38,028 in Hubei. In other regions of CHINA, the total confirmed cases range between 11,588 -13,499. In this paper [7], the study of the Corona Tracker website is used to access the outbreaks of COVID-19. The Data have been taken from Jan 2020 to 3rd March 2020. A Real-time query has been used and implemented in SEIR modeling to estimate the global impact of this pandemic. The result obtained from this indicates that the peak month of this disease is May 2020 and started dropping from early July 2020. Sheng Zhang et al [6] discuss the spread of COVID-19 in Diamond Princess Cruise Ship. A total of 355 confirmed cases has been recorded till February 16, 2020. The objective of this paper was to predict daily new cases on the ship. Reported Serial Interval in the form of Mean and Standard Deviation were fitted with Gamma Distribution and applied “earlyR” package in R The estimated total number of cumulative cases decreased to 1081 (981-1177) when the R0 was reduced by 25 %. It was also observed that the estimated total number of cumulative cases decreased to 758 (697-817) when the R0 was reduced by 50 %. Joby and Mahanthesh [12] put forward a Mathematical Model that is a blend of SIR & Logistic Growth Models. The COVID-19 prediction has been carried out in three states namely Maharashtra, Karnataka & Kerala. In the case of Maharashtra, the highest removal rate predicted is 0.914 using SIR Model. This was followed by Karnataka 0.76 & Kerala 0.114.

3. PRELIMINARIES

3.1 SIR Model

Parameters within this model are:

1. S(t) is the number of susceptible individuals,
2. I(t) is the number of infected individuals
3. R(t) is the number of recovered individuals
4. N is the measured constant population size involved in the disease
5. β is the contact rate of the disease
6. γ is the mean recovery/removal rate.

We can explain the virus communication by the nonlinear ordinary differential equation as shown in equation (2) to (4)

$$S(t) + I(t)+ R(t) = N \quad (1)$$

$$\frac{dS}{dt} = -\beta SI \quad (2)$$

$$\frac{dI}{dt} = -\beta SI - \gamma I \quad (3)$$

$$\frac{dR}{dt} = \gamma I \quad (4)$$

As seen when the three initial conditions for S(t), I(t), R(t) should be defined. So:

- Infected I(t) persons are usually set as 1, i.e. to say, the model begins on its first day with one individual infected.
- By this way, the recovered persons R(t) starts as zero.
- Finally, susceptible people S(t) is given to be the resulting value between N – I(t) – R(t) since all, i.e. S(t)+I(t)+R(t) should always be equal to 100% of the population which is N.

3.2 Curve Fitting

Curve fitting is the process of constructing a curve, or Mathematical Function, that has the best fit to a series of data points, possibly subject to constraints.

3.2.1 Exponential Models

Exponentials are often used when the rate of change of a quantity is proportional to the initial amount of the quantity. If the coefficient associated with *b* and/or *d* is negative, *y* represents exponential decay.

y = aebx

y = aebx + cedx

3.2.2 Polynomial Models:

Polynomial models for curves are given by

$y = \Sigma i = 1n + 1pixn + 1 - i$

Where n + 1 is the order of the polynomial, *n* is the *degree* of the polynomial, and $1 \leq n \leq 9$. The order gives the number of coefficients to be fit, and the degree gives the highest power of the predictor variable. In this guide, polynomials are described in terms of their degree.

4. METHODOLOGY

In this section, we explain our approach to modeling diseases such as COVID-19 through MATLAB. Including the definitions and the mechanism of various algorithms used to predict the COVID-19 pandemic and to find out the best model.

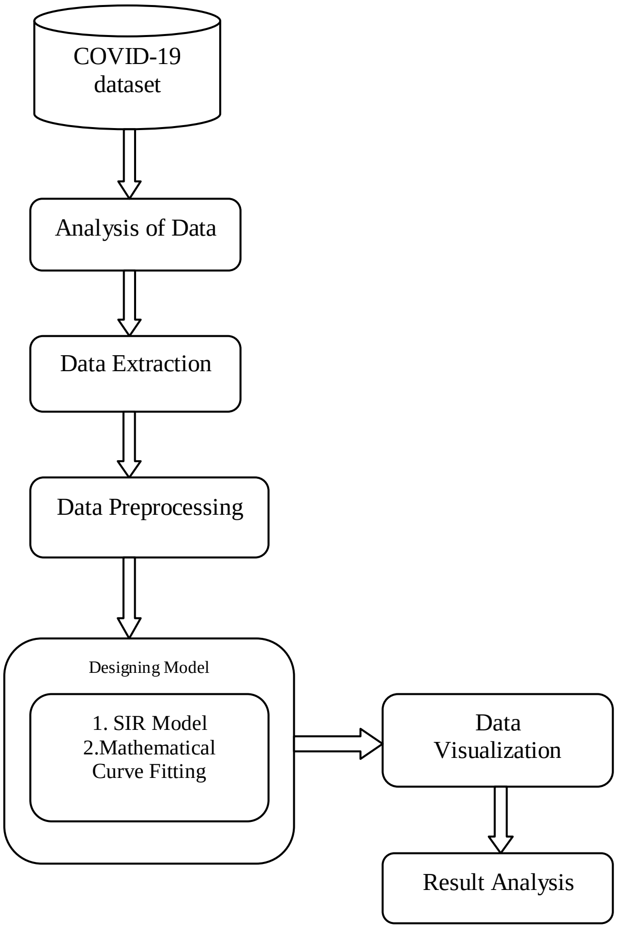


Fig 2 Architectural Diagram

4.1 Data Source

Global data were collected from https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide for different countries and continents on cases over the last 63 days from 21st March 2020 to June 1st, 2020 we only considered datasets containing total confirmed cases, total death and daily increasing cases globally and the next step is to create two data sets of India hosted on website https://www.covid19india.org/. As stated earlier, the aim of this work to forecast the outbreak of this epidemic state-wise through the SIR model and second mathematical curve fitting.

4.2 Data Analysis

After the collection of datasets of COVID-19, now we can analyze that collected data's which automates analytical model building, that systems can learn from data, identify patterns and make decisions with minimal human intervention.

4.3 Data Extraction

In this process now we can extract our data or process of retrieving the COVID-19 data out of (usually unstructured or poorly structured) data sources for further data processing or data storage (data migration).

4.4 Data pre-processing

After extracting the data, we can pre-process the raw data of COVID-19 and make it suitable for a machine learning model. In this process our COVID-19 data gets transformed or *encoded*, to bring it to such a state that now the machine can easily parse it.

4.5 Designing *Model*

Data Modelling with model construction, i.e. finding the best model that fits the COVID-19 pandemic

4.5.1 SIR Model

The Susceptible Infected Recovered (SIR) model is a Mathematical Model, which is also known as the compartmental disease model to depict the disease increase in a population. This disease is divided into one of several different compartments, which shows their health status with respect to the infection. The dynamic of this pandemic can be analyzed as the rate of transfer of this disease between these compartments. The most conceptual model is the SIR model, which forms the origin of infectious disease modeling as shown in figure 4.

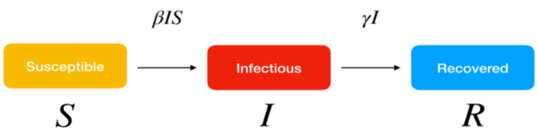


Fig 3 SIR Model

Susceptible**:** Susceptible persons have never been infected, but are susceptible to infection. If they turn into infected they come into the Infected compartment.

Infected**:** Infected individuals can infect susceptible individuals. After a period of time, they move into the compartment.

Removed: Removed individuals have either recovered from the infection and are immune to further infection, or have died.

4.5.2 Curve Fitting

Curve fitting is the process of constructing a curve or Mathematical Function, that has the best fit to a series of data points, possibly subject to constraints. Curve fitting can involve either interpolation, where an exact fit to the data is required, or smoothing, in which a "smooth" function is constructed that approximately fits the data. Fitted curves can be used as an aid for data visualization, to infer values of a function where no data are available, and to summarize the relationships among two or more variables.

4.6 Data Visualization

After designing the model, now we had done data visualization to see the graphical representation of the designing model of COVID-19. This will provide an accessible way to see and understand trends, outliers, and patterns in data.

4.7 Result Analysis

Finally we have done the result analysis and predict the outcome of COVID-19 with this modeling.

5. RESULT AND DISCUSSION

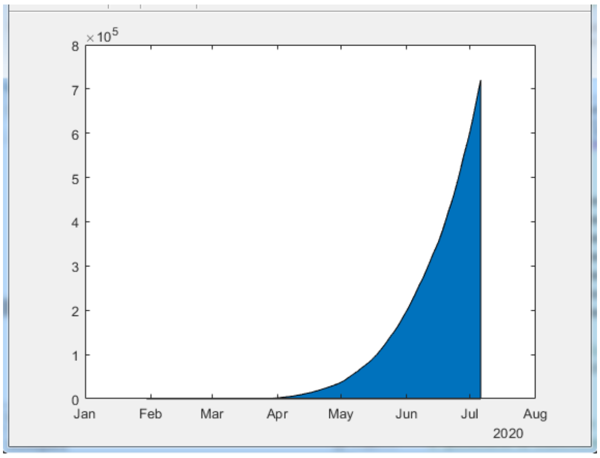


Fig 4 Total confirmed cases in India

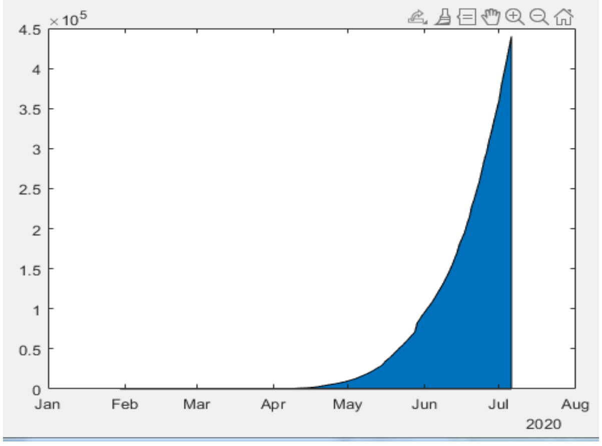


Fig 5 Total Recovered cases in India

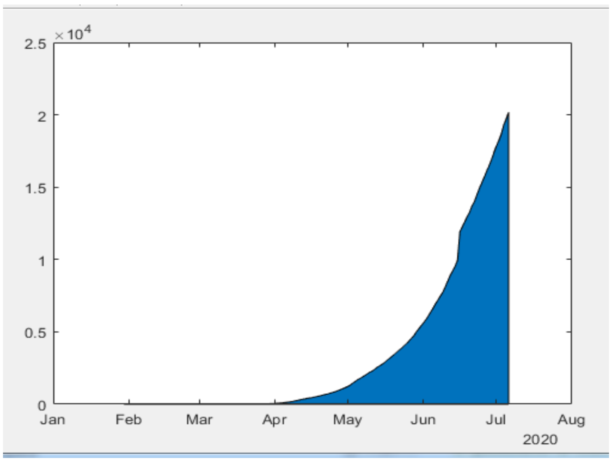


Fig 6Total Decreases cases in India

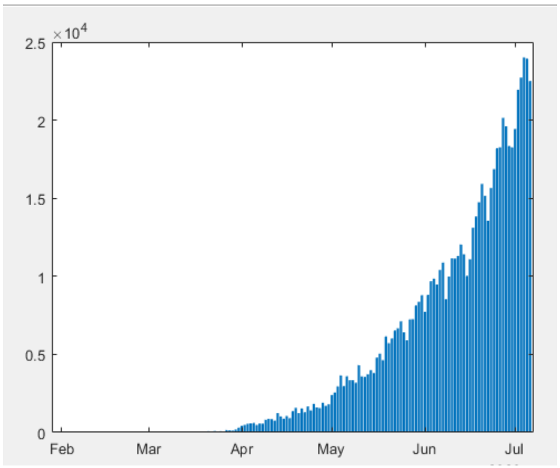


Fig 7 Daily Confirmed cases in India

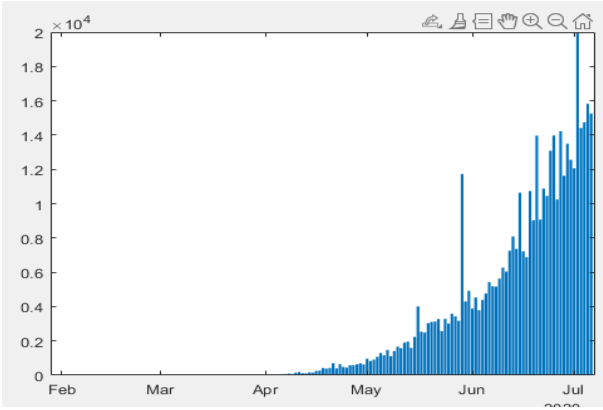


Fig 8 Daily Recovered Cases in India

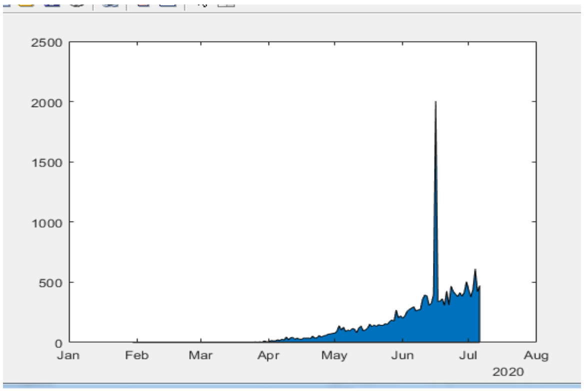


Fig 9 Daily Decreased cases in India

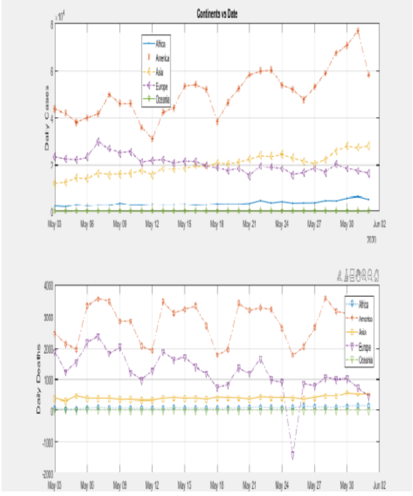


Fig10 Daily cases and daily death of Covid-19 in each continent v/s date

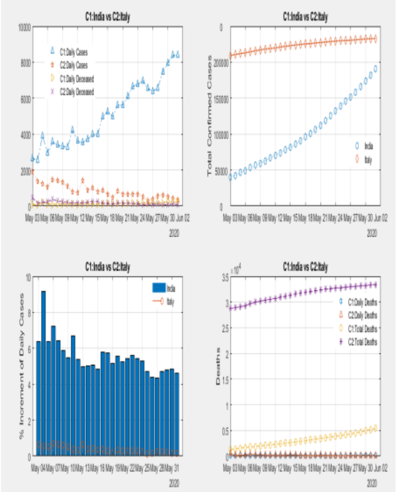


Fig 11 TotalConfirmed, Daily cases and deaths of Covid-19 in India v/s Italy

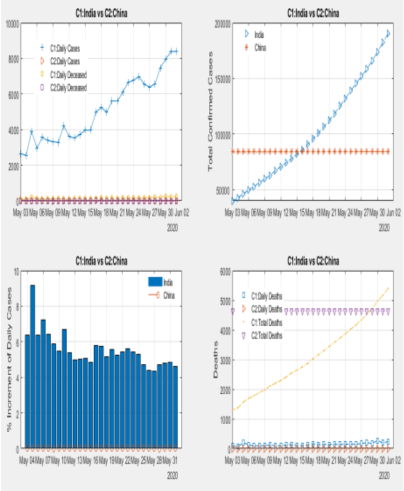


Fig 12 TotalConfirmed, Daily cases and deaths of Covid-19 in India v/s China

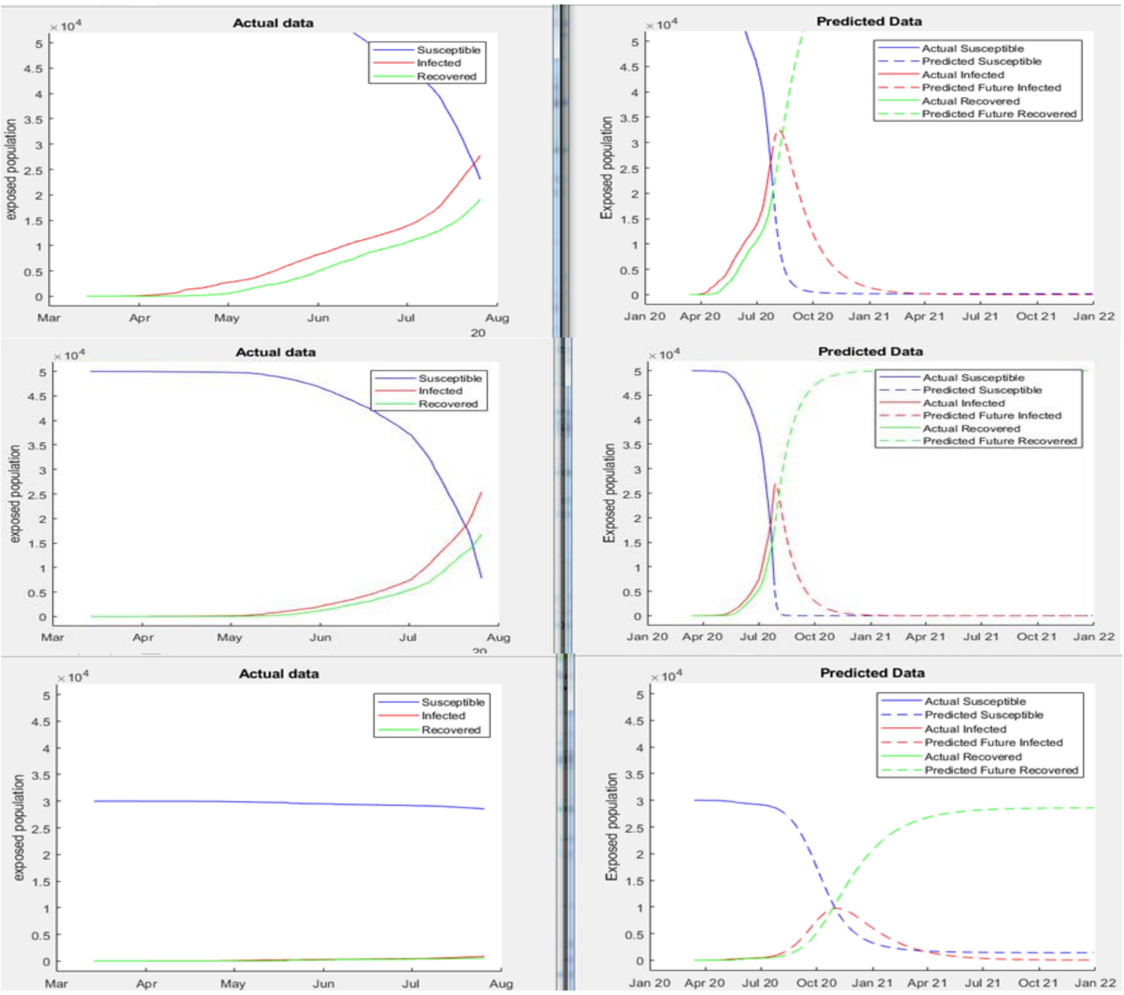
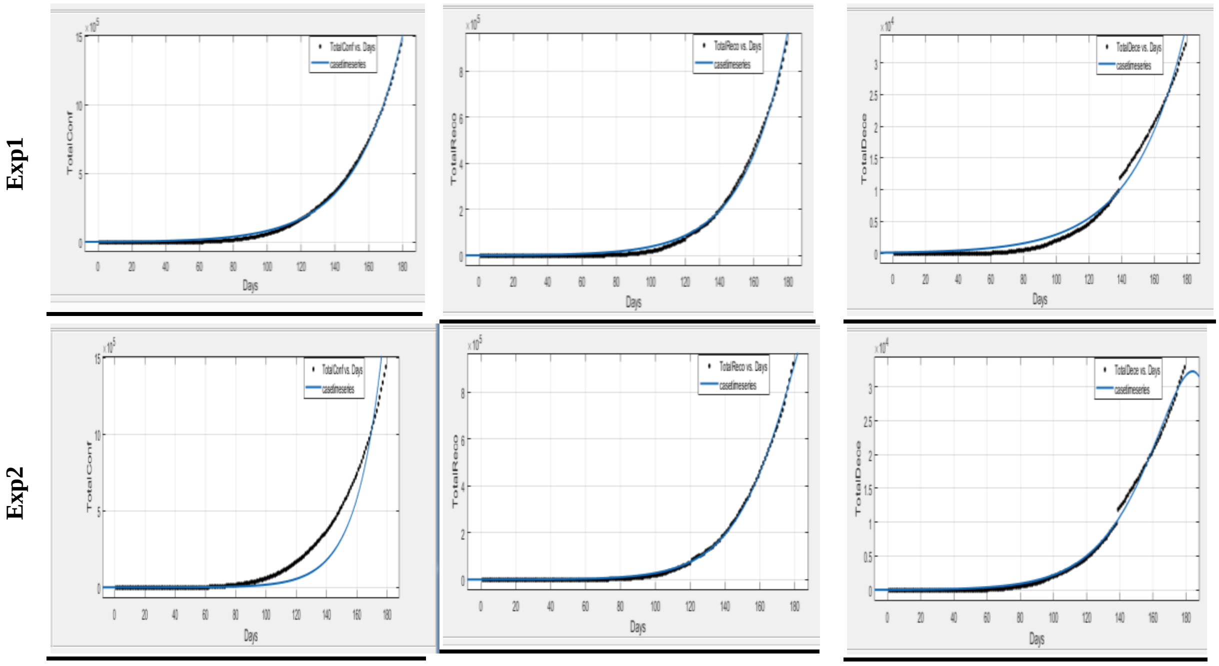


Figure 13 Actual and Predicted Data of Odisha, Madhya Pradesh & Chhattisgarh

Total Confirmed Cases **Total Recovered Cases** **Total Decreased Cases**



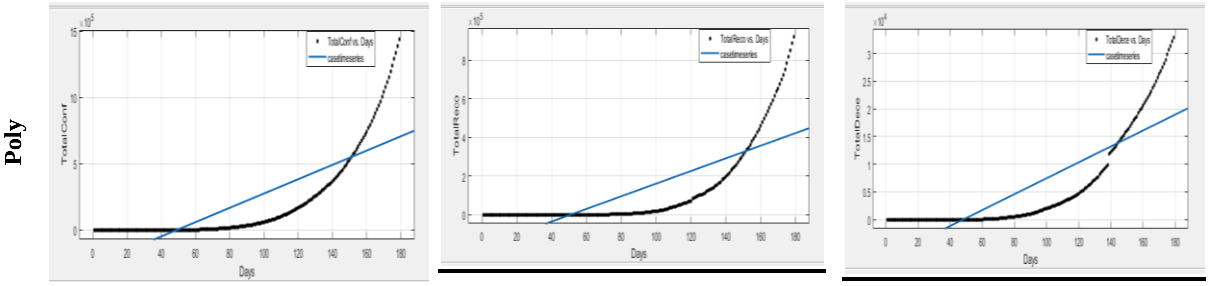


Figure 14 Mathematical Curve fitting of total confirmed, recovered and decreased cases in India

The analysis of Covid-19 from 21st March to 1st June 2020 country -wise total cases have been estimated for Africa-146996, America- 2904566, Asia-1123892, Europe-1951284, Oceania-8651 and, other-696 cases have been recorded.

Daily increasing cases and daily death of Covid-19 in each continent concerning time has been shown in Fig 10. The outcome with this model reflects that in America-2904566 total confirmed cases and the daily death rate is comparatively highest when compared to others and we can figure out the pattern of this pandemic.

Regarding Fig 11 and Fig: 12, we have analyzed and evaluated all the countries through this model. Along with India, we have compared the two most infected countries from 21st March to 1st June 2020 during this pandemic respectively. The total confirmed cases of

India, Italy, and China have been recorded as190500, 233000, and 84150. The percentage of daily increase in cases for India, Italy, and China are 0.1526, 4.607, and 0.0225.

The death cases till the 1st of June have been recorded as 5394, 33420, and 463800 in India, Italy & China.

In Fig 13, the outcome of this experiment revealed the situation of virus spread in India. SIR model indicates that in Madhya Pradesh the number of infected cases is highest compared to the other two states. The cases in MP and Odisha are expected to peak by August. As far as Chhattisgarh is concerned, it is expected to see the peak by October. The model also allows us to predict that the pandemic shall end by the mid of December in all three states as shown in Table I.

With reference to Fig 14 represents the forecast of the future outbreak of infected persons in India of total confirmed, Recovery, and decreased cases. And Table II reflect the results of the accuracy of mathematical modeling for the exponential 1, exponential 2 and polynomial equations.

Table-1 Result of Estimated Susceptible-Infected-Recovered (SIR) model

		Actual Data (Till July 2020)			Forecast(Till December 2020)		
S.N.	State	Susceptible	Infected	Recovered	Susceptible	Infected	Recovered
1	Madhya Pradesh	23070	27800	19130	1	1775	51090
2	Odisha	7818	25390	7818	296	0	49880
3	Chhattisgarh	28540	852	887	1396	18	28590

Table 2 Estimate of Case time series in India.

Fit Name	Data	Fit Type	SSE	R Sq.	DFE	Adj. R Sq	RMSE	Coeff.
Casetimeseries	Tot Conf. v/s Days	exp 1	4.4949e+10	0.9979	177	0.9979	1.5936e+04	2
		exp 2	1.8481e+12	0.9143	175	0.9143	1.0277e+05	4
		Poly1	7.6286e+12	0.6461	177	0.6461	2.0760e+05	2
	Tot Rec. v/s Days	exp1	2.8261e+10	0.9967	177	0.9967	1.2636e+04	2
		exp2	6.6773e+02	0.9992	175	0.9992	6.1771e+03	4
		Poly1	3.4705e+12	0.5952	177	0.5952	1.0403e+05	2
	Tot Dec. v/s Days	exp1	1.6773e+08	0.9881	177	0.9881	973.4532	2
		exp2	3.8989e+07	0.9972	175	0.9972	472.0126	4
		Poly1	4.2587e+09	0.6987	177	0.6970	4.9051e+03	2

6. 6. CONCLUSION AND SCOPE OF FURTHER WORK

Information and communication technology helps in the decision-making process. It is based on a large amount of historical data. Gathering information and getting an interesting pattern out of the accumulated data is a challenging task. With the existing data on confirmed, recovered, and death cases all across India, we have predicted the approximate month when this pandemic will cease.

The study done in this work indicated that India is yet to achieve a peak in the spread of Coronavirus disease. The predictive model proposed, revealed that it is unlikely to get rid of Covid-19 before the end of December 2020 in India. It's unfortunate but has to be tackled with bravery and scientific research. Referring to the model development, it employs a combination of epidemiological model (SIR) and mathematical curve fitting methods to forecast the impact of the COVID-19 in India. The Covid-19 outbreak in India has been analyzed by comparing the dynamics of the pandemic in Madhya Pradesh, Odisha, and Chhattisgarh using the SIR models. It has been observed that the states experience a different pattern of disease transmission. The SIR model predicted the maximum number of infected individuals in Madhya Pradesh, Odisha, and Chhattisgarh as 27800, 25390, and 852. The complete control of the disease spread has been predicted in December 2020. From the SIR model, it is concluded that the infection growth rate in Madhya Pradesh is 5% higher than in Odisha and Chhattisgarh.

The best way to tackle this disease is to stay at home, use masks, restrain from traveling, avoid social gatherings, and frequently sanitizing and washing hands is to Flattening the curve in our country. Many countries across the globe have even imposed lockdowns to prevent the disease from spreading. Hence to eliminate community transmission of COVID, Govt. of INDIA can estimate additional lockdowns in the future through this model. With the pandemic spreading further and continuing even after June 2020, we expect our model to reflect the peak cases either in August or September 2020.

The epidemic response is a dynamic process and strategies may also need to change depending on the dynamics of the outbreak. The general public needs to understand this reality and have a right to know the facts. They should develop trust in the government's country-specific approaches and maximally support such efforts as all the interventions are executed with a valid reason and purpose. In India, an already existing public health system consisting of a battalion of field health staff is geared for needy interventions which are seen as a feasible task and an approach immensely helping to flatten the curve. We can extend this work by employing, optimization, Statistics modeling, ANN, ANFIS(Adaptive Neuro -Fuzzy Inference Systems), Sentimental analysis, etc to predict the future and take corrective measures.

7. REFERENCES

[1] Bai, H.X., Hsieh, B., Xiong, Z., Halsey, K., Choi, J.W., Tran, T.M.L., Pan, I., Shi, L.B., Wang, D.C., Mei, J. and Jiang, X.L., 2020. Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. Radiology, 296(2), pp.E46-E54.

[2] Punn, N.S., Sonbhadra, S.K. and Agarwal, S., 2020. COVID-19 epidemic analysis using machine learning and deep learning algorithms. MedRxiv.

[3] Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J.M., Yan, P. and Chowell, G., 2020. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infectious Disease Modelling, 5, pp.256-263.

[4] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[5] Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T.L., Duan, W., Tsoi, K.K.F. and Wang, F.Y., 2020. Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. IEEE Transactions on Computational Social Systems, 7(2), pp.556-562.

[6] Zhang, S., Diao, M., Yu, W., Pei, L., Lin, Z. and Chen, D., 2020. Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. International journal of infectious diseases, 93, pp.201-204.

[7] Hamzah, F.B., Lau, C., Nazri, H., Ligot, D.V., Lee, G., Tan, C.L., Shaib, M.K.M., Zaidon, U.H., Abdullah, A. and Chung, M.H., 2020. CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction. Bull World Health Organ, 1(32).

[8] Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., Yu, L., Ni, Q., Chen, Y., Su, J. and Lang, G., 2020. A deep learning system to screen novel coronavirus disease 2019 pneumonia. Engineering, 6(10), pp.1122-1129.

[9] Kanne, J.P., 2020. Chest CT findings in 2019 novel coronavirus (2019-nCoV) infections from Wuhan, China: key points for the radiologist.

[10] Chung, M., Bernheim, A., Mei, X., Zhang, N., Huang, M., Zeng, X., Cui, J., Xu, W., Yang, Y., Fayad, Z.A. and Jacobi, A., 2020. CT imaging features of 2019 novel coronavirus (2019-nCoV). Radiology, 295(1), pp.202-207.

[11] Ng, T.W., Turinici, G. and Danchin, A., 2003. A double epidemic model for the SARS propagation. BMC Infectious Diseases, 3(1), pp.1-16.

[12] Mackolil, J. and Mahanthesh, B., 2020. Mathematical Modelling of Coronavirus disease (COVID-19) Outbreak in India using Logistic Growth and SIR Models.

[13] Moremedi, G.M., Kaondera-Shava, R., Lubuma, J.M., Morris, N. and Tsanou, B., 2015. A Simple Mathematical Model for Ebola in Africa. Biomath Communications, 2(1).

[14] Du, S., Wang, J., Zhang, H., Cui, W., Kang, Z., Yang, T., Lou, B., Chi, Y., Long, H., Ma, M. and Yuan, Q., 2020. Predicting COVID-19 using hybrid AI model.

[15] Babbar, S., 2020. Battle with COVID-19 Under Partial to Zero Lockdowns in India. medRxiv.

[16] Beck, B.R., Shin, B., Choi, Y., Park, S. and Kang, K., 2020. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Computational and structural biotechnology journal, 18, pp.784-790.

[17] Mackolil, J. and Mahanthesh, B., 2020. Mathematical Modelling of Coronavirus disease (COVID-19) Outbreak in India using Logistic Growth and SIR Models.

[18] Stübinger, J. and Schneider, L., 2020, June. Epidemiology of coronavirus covid-19: Forecasting the future incidence in different countries. In Healthcare (Vol. 8, No. 2, p. 99). Multidisciplinary Digital Publishing Institute.

[19] Sujath, R., Chatterjee, J.M. and Hassanien, A.E., 2020. A machine learning forecasting model for COVID-19 pandemic in India. Stochastic Environmental Research and Risk Assessment, 34, pp.959-972.

[20] Greenfield, J., Sears, M., Nagrani, R., Mazzaferro, G., Widyastuti, A. and Austin, C.C., the RDA-COVID19-WG.(2020). Common Data Models and Full Spectrum Epidemiology: Epi-STACK architecture for COVID-19 epidemiology datasets. Data Sharing in Epidemiology, p.65.

[21] Official website for data collection https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide and https://www.covid19india.org/

[22] SIR model https://www.lewuathe.com/covid-19-dynamics-with-sir-model.html

[23] https://en.wikipedia.org/wiki/Curve_fitting