COVID-19 Data Analysis and Innovative Approach in Prediction of Cases

Abhijeet Kushwaha, Pallavi Vijay Chavan, and Vivek Kumar Singh

Abstract The world went into a standstill when the news of a global pandemic (Coronavirus Disease-19) caused by the Severe Acute Respiratory Syndrome Coronavirus 2 Virus was announced by the World Health Organization. In this chapter, we present a detailed investigation of the spread of COVID-19 in India, its neighboring countries, and other global hotspots. We also analyze the worst-hit states of India with respect to the rise in Confirmed cases and Death cases due to COVID-19 and investigate the growth factor of the same. We then build a Support Vector Machine for Regression analysis of confirmed cases in India and predict the future confirmed cases by analyzing the current growth curve. Our work finally extends to discuss the need for future research to find out the consequences of the COVID-19 crisis in areas of Economy, Agriculture, Health, Education, Secondary sector, and Service Sector.

Keywords COVID-19 · Artificial intelligence · Deep learning · Support vector machine · Regression · Data analysis

1 Introduction

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the virus which causes Coronavirus Disease (COVID-19) in humans. The first human cases of COVID-19 were identified in Wuhan City, China in December 2019. Just like the SARS outbreak in 2003, it is thought that SARS-CoV-2 entered the species barrier and initially affected human bodies, but more likely through an intermediate host,
that is another animal species more likely to be handled by humans—this could be a
domestic animal, a wild animal, or a domesticated wild animal and, as of yet, has not
been identified [1]. The spread of COVID-19 has brought all global superpowers and
developing nations to a standstill, as the infection has shown an exponential rate of
spread. To make matters worse, we are still in the process of understanding the virus
and finding vaccines to stop the spread. Until we do find a vaccine, social distancing
and home quarantine is the only way to stop the spread. This has led to a complete
nationwide lockdown in many countries including developing countries like India,
which has had very adverse effects on the country’s economy, healthcare facility and
administration. Thus, COVID-19 has presented itself as an unconventional situation
for the world to deal with, and therefore, it becomes all the more important to do a
comprehensive analysis of the data for COVID-19 cases. Before we begin our anal-
ysis and prediction on the current COVID-19 crisis in the world, it is important to
understand the following key terms: Coronavirus, COVID-19 and SARS-CoV-2. It
is also important to understand why COVID-19 escalated exponentially across the
world, and why it has been declared as a Pandemic by the World Health Organization.

Data about COVID-19 and its impact is released every day. However, to make
sense of the data, to clean the data and to develop prediction models of a high
accuracy is the big challenge. These challenges can be overcome by fields of Big
Data, Deep Learning and Artificial Intelligence. We take these factors as a motivation
to contribute this chapter and to bring out the novel approach of data analysis of
COVID-19. This chapter contributes the analysis through different factors like state
wise analysis in India, analysis of confirmed cases across India and abroad and the
prediction model to predict the possible future cases of COVID-19.

The organization of the chapter is as follows: The chapter begins with the intro-
duction in first section. Second section describes the state- of- art information of
COVID-19. Section 3 presents an intelligent data Analysis of COVID-19. This
section includes identifying and plotting global hotspots, analysis of spread in India
and the comparison with neighboring countries. Section 4 describes the Prediction of
future COVID-19 cases in India. Section 5 concludes the chapter with future scope
of research with COVID-19 in healthcare. The chapter at the end includes sufficient
references for the literature of the domain of healthcare and COVIDE0-19.

2 State-of-Art

The rapid spread of Corona virus, globally, has brought a lot of data that can be
analyzed to know the nature of its advancement. Every day, every moment data is
changing thus it is very challenging task for the researchers to deal with such big data.
This is important for health care industries to analyze this data with the help of some
big data algorithms, so that they come up with a solution. Scientist and researches
are trying their best to utilize machine learning algorithms to track current progress
of corona virus.
According to Ramgopal [2], it is found that in health sector vast amount of data is produced every day e.g. web of things. These stores require undertakings at many layers for better analysis of data stream. Author has presented a critical analysis of health care data. It is proposed to use high statistical analysis of big data to extract some hidden important information. Li et al. [3] picked official data and analyses transmission process of COVID-19. Authors have used both forward prediction and backward interference to analyze the data for better preparation by the affected countries. As work reported in [4], authors have used Susceptible-Infected-Removed (SIR) model. The data was collected on day basis from Jan 2020 for most affected countries by COVID-19. Simon et al. [5], presented a case study of novel Corona virus epidemic using Composite Monte-Carlo. The performance is further enhanced by GROOMS model. The results show a better prediction model for this epidemic forecast.

Various researchers used machine learning models to predict the spread of COVID-19. Similar work is done on Brazilian stats by Mathew’s et al. [6]. The researchers used autoregressive integrated moving average, cubist, random forest (RF), Ridge regression and support vector regression for analyzing data.

Bhapkar et al. [7] proposed a graph based spread of COVID-19. Based on pandemic nature authors have divided graphs in four part. Relative analysis shows its hazardous nature. A very useful survey on forecasting methods used for COVID-19 pandemic is presented by Parikshit et al. [8]. Authors have presented analysis of various prediction models considering Day by Day data, data sources, prediction algorithms and all parameters. The day by day transforming nature of corona virus is still a challenge.

Authors in [9]. proposes an accurate forecasting model based on small dataset for COVID-19 cases. This model is able to forecast accurately the epidemic fate. The authors reported the limitations and unavailability of the data for accurate early forecasting. The most important challenge to the machine learning is discussed in this literature; the challenge is insufficient amount of data available for analysis. The model proposed in this literature is based on polynomial neural network and the corrective feedback. The predicted error is lowest with this model and the forecasting is at acceptable level.

3 COVID-19 Intelligent Data Analysis

In this section, we analyze COVID-19 data with respect to the neighboring countries viz. Pakistan, China, Afghanistan, Nepal, Bhutan, Sri Lanka and Bangladesh. In the next segment of analysis, COVID-19 spread in India with respect to major Global Hotspots is presented. The importance of this section is to clearly indicate the understanding about the trend of spread of COVID-19 on a global level. Based on this trend, an understanding of COVID-19 spread in India across the states is presented. This intelligent analysis helps to determine the states with highest confirmed cases, highest Mortality rate and highest Recovery rates.
3.1 Identifying and Plotting Global Hotspots

This section plots a choropleth map to visualize the global hotspots of confirmed cases of COVID-19 across the world. The data about confirmed COVID-19 cases throughout the world has been obtained from the data repository on GitHub by Johns Hopkins CSSE [10]. This dataset contains confirmed cases from the world till April 9, 2020. The choropleth map is shown in Fig. 1. It is obtained by using plotly library in Python in Jupyter Notebook. The choropleth map plots the confirmed COVID-19 cases across the world, where the darker hues of blue indicate a higher number of confirmed cases in a region in contrast to the lighter hues of blue that indicate lesser number of Confirmed cases respectively.

3.2 Analysis of COVID-19 Spread in India

Using the same dataset, analysis of COVID-19 spreading in different states of India is presented here. This analysis is plotted using pandas, numpy and matplotlib. The India’s data is extracted first from the combined data of multiple countries by setting ‘Country/Region’ in the dataset equal to ‘India’. In the next step, ‘Date’ column entries are converted to date time using pandas for an effective work using Date entries. The grouping all the data belonging to India date wise by aggregating the entries of ‘Confirmed’, ‘Recovered’ and ‘Death’ cases. Figure 1 is the plot of date wise Distribution of number of active cases in India. An increase or decrease in the
count of the active cases of COVID-19 will be an indication of whether the number of Death cases or Recovered cases are increasing or decreasing.

From Fig. 2 it is observed that the count of Active cases of COVID-19 in India is increasing. This is an indication that the number of closed cases (sum of Death cases and Recovered cases) is less with respect to the daily Confirmed cases. This graph indicates rapid spread of COVID-19 in India.

Now, we plot a graph of the overall count of Confirmed cases, Death cases and Recovered cases of COVID-19 in India, and analyze the reason of a rise in the number of Confirmed cases and number of Closed cases. Figure 3 indicates a sudden increase in the number of Confirmed cases after being flat and steady for a long time earlier. Also, the number of Recovered cases initially is represented as a flat curve, but they eventually start rising which is a positive sign. The number of Death cases isn’t rising as fast as the number of Confirmed cases and is mostly steady, which indicates that there have been less deaths per number of Confirmed cases. Also, as the number of Confirmed cases rises faster than both the number of Recovered cases and the number of Deaths combined, this validates the fact that the number of Active cases are rising. From Fig. 3 it is concluded that the number of closed cases is rising because of an increase in the number of Recovered cases, and not because of an increase in Death cases.

The density plot of Recovered cases and Death cases is shown in Fig. 4. From this figure it is concluded that the density of Recovered cases over a section is greater than the density of Death cases.

The recovery rate and mortality rate of the total Confirmed cases of COVID-19 cases in India is analyzed using following:

\[
\text{Recovery rate} = \left(\frac{\text{Number of Recovered cases}}{\text{Number of Confirmed cases}}\right) \times 100
\]
Mortality rate = \( \frac{\text{Number of Death cases}}{\text{Number of Confirmed cases}} \times 100 \)

The left graph and right graph in Fig. 5 depicts Recovery rate and Mortality rate of COVID-19 cases observed in India respectively. We can see that initially, the Recovery rate started rising while the mortality rate was still constant for Confirmed COVID-19 cases in India. However, with time, the recovery rate has significantly dropped while the mortality rate has significantly risen, which becomes a matter of concern for India.

The growth factor for the Confirmed cases, Recovered cases and Death cases is given below:

Growth factor of Confirmed cases = \( \frac{\text{New Confirmed cases of a day}}{\text{New Confirmed cases of Previous day}} \).

Growth factor of Recovered cases = \( \frac{\text{New Recovered cases of a day}}{\text{New Recovered cases of Previous day}} \).

Growth factor of Death cases = \( \frac{\text{New Death cases of a day}}{\text{New Death cases of Previous day}} \).

A growth factor above 1 indicates an increase in corresponding cases, while a growth factor above 1 is a positive sign if it is trending downwards. A growth factor constantly above 1 suggests exponential growth, while a constant growth factor of 1 suggests no change of any kind. Figure 6 represents a growth factor.
Fig. 4  Density plot for recovered cases and death cases of COVID-19 in India

Fig. 5  Recovery rate and mortality rate due to COVID-19 in India
3.3 Comparison of Spread of COVID-19 in India with Neighboring Countries

Having analyzed the general spread of COVID-19 in India, in this section the comparison of spread with respect to India’s neighboring countries is shown.

In this analysis, the number of Confirmed cases and increase in the number of Confirmed cases in India and its comparison with its neighboring countries is presented. The logarithmic scale is chosen for an effective visualization of data and shown in Fig. 7. From this figure, it is observed that China is the most affected neighbor of India, as the spread of COVID-19 began from China. After China, cases in India rose, then stayed constant, and then rose again very fast. We also see that initially there were no cases Confirmed in Pakistan for a long time, until the last week of February when confirmed cases in Pakistan started rising. Rise of Confirmed cases in Nepal started very late and is increasing. It is also infer that after a sudden shoot in Confirmed cases in China, the curve is now constantly flat, which indicates that only China in the neighboring countries has managed to contain the spread of COVID-19. Thus, a flat and constant curve is what the countries would have to aim for to conclude that the spread of COVID-19 in their regions has been contained.

Let us now analyze the trend in the number of Recovered cases of India and its neighboring countries. It is evident from Fig. 8 as China started with the recovery very early, and the rising curve suggests that China has seen a significant amount of Recovery of its Confirmed COVID-19 citizens. After China, India and Pakistan have seen a significant rise in the amount of Recovered cases, followed by Bangladesh, Sri Lanka and Bhutan.
Following Fig. 9 shows the analysis of daily increase in Recovered cases of COVID-19 in neighboring countries. Figure 10 is an analysis of the trend in the number of Death cases of India and neighboring countries.

From Fig. 10, it is evident that China has the highest number of Death cases of all India’s neighboring countries, however the curve has become constant suggesting that China has contained the spread of COVID-19. India, Pakistan, Bangladesh, Afghanistan and Sri Lanka have seen a rise in the number of Death cases due to COVID-19. However, until April 15, 2020 when the above data was graphed, Bhutan and Nepal have seen no death cases for COVID-19.
3.4 Comparison of COVID-19 Spread in India with Global Hotspots

This section investigates the trends of Confirmed cases, Recovered cases and Death cases due to COVID-19 in India and other global hotspots like Spain, Italy, USA,
Fig. 11  Comparison of confirmed cases in India and other global hotspots (logarithmic scale)

Iran, France and the United Kingdom’s. Figure 11 is the investigation of the number of Confirmed cases in India and its comparison with the global hotspots.

From Fig. 11 it is concluded that India has lowest number of Confirmed cases with respect to other global hotspots. In Fig. 12, the number of Confirmed cases in other countries had increased and are now decreasing, but the daily increase in number of Confirmed cases in USA is still rising and is the highest.

Fig. 12  Comparison of daily increase in number of confirmed cases in India and global hotspots due to COVID-19
At this point, we plot the number of Recovered cases in India and in global hotspots, and the increase in daily number of Recovered cases in India and global hotspots. From Fig. 13, it is concluded that the number of Recovered cases in India and other global hotspots have been rising. We also observe in Fig. 14 that the daily increase in the number of Recovered cases was constant at first, and then started increasing.

**Fig. 13** Comparison of number of recovered cases in India and global hotspots due to COVID-19

**Fig. 14** Comparison of daily increase in number of recovered cases
In the next segment of analysis, we plot the number of Death cases in India and in global hotspots, and the increase in daily number of Death cases in India and global hotspots.

From Fig. 15 it is observed that the number of Deaths in all the major hotspots started rising very fast, while in India, the number of death cases started rising comparatively late. However, we can start to notice that the curves of all the plotted countries are flattening, which indicates a slow rise in Death cases.

From Fig. 16 it is concluded that the increase in daily Death cases in India has been nearly constant when compared to other global hotspots, whose graph is highly rising.

### 3.5 Comparison of COVID-19 Spread in the States of India

This section analyses and evaluates the spread in the states of India. The dataset used for plotting all graphs in this section is obtained from the following two sources: The Ministry of Health and Family Welfare, Government of India [11] and the worldwide web, covid19india [12]. The tools used are pandas for Data Analysis, numpy for Computing and matplotlib and seaborn for plotting the results.

The state’s data is extracted from the combined data of multiple states by setting “State/Union Territory” in the dataset equal to that corresponding state or Union Territory. We then convert the ‘Date’ column entries to `datetime` using pandas for an effective work using Date entries. Finally, we group all the data belonging to states, date wise by aggregating the entries of ‘Confirmed’, ‘Recovered’ and ‘Death’ cases.

Following States are considered for the analysis: Assam, Arunachal Pradesh, Andhra Pradesh, Rajasthan, Bihar, Kerala, Chhattisgarh, Madhya Pradesh, Gujarat,
Haryana, Telangana, Himachal Pradesh, Jharkhand, Karnataka, Nagaland, Manipur, Meghalaya, Mizoram, Odisha, West Bengal, Punjab, Sikkim, Uttar Pradesh, Tamil Nadu, Maharashtra, Tripura, Uttarakhand, Goa. Following Union Territories are considered for the analysis: The Government of NCT of Delhi, Daman & Diu, Dadra and Nagar Haveli, Puducherry, Andaman and Nicobar Islands, Chandigarh, Jammu and Kashmir, Ladakh and Lakshadweep.

The analysis of the number of Confirmed cases of COVID-19 in all states of India is presented in Fig. 17. In this figure, it is observed that the five regions (State/UT) with the highest number of COVID-19 cases are Maharashtra, Delhi, Tamil Nadu, Rajasthan and Madhya Pradesh respectively in descending order. Therefore, we compare and contrast these five regions in further analysis of the number of Confirmed cases in India.

The number of Recovered cases of COVID-19 in all states of India is shown in Fig. 18. Analyzing Fig. 18, we can conclude that the five regions (State/UT) with the highest number of Recovered COVID-19 cases are Maharashtra, Kerala, Rajasthan, Telangana and Tamil Nadu respectively in descending order. Therefore, we compare and contrast these five regions during analysis of the number of Recovered cases in India.

The analysis of the number of Death cases of COVID-19 in all states of India is shown in Fig. 19. Analyzing this figure, we can conclude that the five regions (State/UT) with the highest number of Death in COVID-19 cases are Maharashtra, Madhya Pradesh, Gujarat, Delhi and Telangana respectively in descending order. Therefore, we will compare and contrast these five regions during analysis of the number of Death cases in India.
Fig. 17  Confirmed cases of Indian States and UTs as of April 15

Fig. 18  Recovered cases of Indian States and UTs as of April 15
The trends in the number of Confirmed cases, Cured cases and Deaths cases of all the States and Union Territories of India are plotted ahead. Following is the trend in the number of Confirmed cases of COVID-19 in all States and Union Territories.

Investigating Figs. 20 and 21, we see that the trend in Confirmed cases in all States and Union Territories seem to be fuzzy at first, however after a point in the graph, they start repeating a common shape of the curve.

Following is the trend in the number of Cured Cases of COVID-19 in all States and Union Territories.

Analyzing Figs. 22 and 23, we see that in all States and Union territories in India, the number of Recovered cases initially keeps increasing and decreasing, however after a point in the graph, the curve starts repeating a common pattern.

The trend in the number of Death Cases of COVID-19 in all States and Union Territories is plotted.

From Figs. 24 and 25 it is concluded that after a quick increase in Deaths in majority states and UTs, a large number of which now see constant rise in Death cases. We will now analyze the five highest states in terms of Cured cases, Confirmed cases and Death Cases of COVID-19.

There are five states in India where in the CIVOD-19 spread is rapidly increasing every day. At this point, we plot the number of Confirmed cases of the highest five states, viz. Maharashtra, Delhi, Tamil Nadu, Rajasthan and Madhya Pradesh. Figure 26 is the representation of confirmed cases in these five regions. Figure 27 represents the number of Recovered cases in highest five regions and Fig. 28
Fig. 20  Confirmed cases of all Indian states as of April 15 (Logarithmic scale)

Fig. 21  Confirmed cases of all Indian UTs as of April 15 (Logarithmic scale)
Fig. 22  Cured cases of all Indian states as of April 15 (Logarithmic scale)

represents the number of death cases in top five regions.

Finally, the Mortality and Recovery rates of the top 5 regions of India are analyzed and plotted. Figure 29 indicates a plot of the Mortality rate in the top five regions of India. Figure 30 represents the plot of recovery rate in top five regions.
Fig. 24  Death cases of all Indian states as of April 15 (Logarithmic scale)

Fig. 25  Death cases of all Indian UTs as of April 15 (Logarithmic scale)


4 Prediction of Future COVID-19 Cases in India

The art of prediction depends on how well we understand the current scenario, whose future we want to predict. We devoted the complete Sect. 3 to analyze the different trends in data visualization by looking at as many perspectives as possible, and derive accurate results. In this chapter, Support Vector Machine (SVM) Regression model is used for prediction. This model predicts the future possible number of COVID-19 cases.

SVM is a popular machine learning tool for classification and regression. SVM regressor of polynomial kernel is created with a degree of 10. The same dataset which was used for analysis is used for training purpose and to fit the model. The model gives Root Mean Square Error for SVR as 431.6616564645887. Figure 31 is the plot of data used for training and the fitted curve using SVM regression. The R-squared score obtained for the model is 0.942. Below are the prediction values for the next 10 days (Table 1).

5 Conclusions

With this study, we aimed to investigate the COVID-19 crisis using the official datasets available in the public domain, and the very effective Data Visualization
and Machine learning tools. We investigated the spread of COVID-19 using a choropleth map, then went on to investigate the spread throughout the world, in India, in the neighboring countries of India, and in the world’s major hotspots of COVID-19 spread. With the help of this study, we were able to visualize and draw conclusions towards the patterns of spread of COVID-19, along with visualizing the rise in mortality rates, recovery rates, growth factors, density plots and predict the number of Future cases with fair accuracy. Data Analytics tools and Intelligent models have made it possible for us to model and visualize the spread of COVID-19. With the help of the predicted values of Confirmed cases, the government and all officials can work towards equipping themselves with all the healthcare requirements like drugs, equipment, and quarantine spaces beforehand. The world going into a complete lockdown has impacted the economy, and therefore all the interlinked services like Agriculture, Service sector, and Industrial Production. More research can be conducted on
these individual matters to realize the impact of the COVID-19 on the world after the COVID-19 crisis.
Fig. 29  Highest mortality rates of the top five regions in India

Fig. 30  Highest recovery rates of the top five regions in India
Fig. 31  Predicted confirmed cases from SVM regression model
| Sr. No | Date       | SVM predictions |
|-------|------------|-----------------|
| 1     | 2020-04-16 | 14,848.68537    |
| 2     | 2020-04-17 | 16,690.76385    |
| 3     | 2020-04-18 | 18,736.04939    |
| 4     | 2020-04-19 | 21,004.22921    |
| 5     | 2020-04-20 | 23,516.63986    |
| 6     | 2020-04-21 | 26,296.38409    |
| 7     | 2020-04-22 | 29,368.45403    |
| 8     | 2020-04-23 | 32,759.86395    |
| 9     | 2020-04-24 | 36,499.78771    |
| 10    | 2020-04-25 | 40,619.70376    |

References

1. who.int. 2020. [online] Available at: https://www.who.int/news-room/q-a-detail/q-a-coronavir uses. Accessed 19 Apr 2020
2. Kashyap, R.: Big data analytics challenges and solutions, big data analytics for intelligent healthcare management, a volume in advances in ubiquitous sensing applications for healthcare, Elsevier, pp 19–41 (2019)
3. Bora, D.J.: Big data analytics in healthcare: a critical analysis, big data analytics for intelligent healthcare management, a volume in advances in ubiquitous sensing applications for healthcare, pp. 43–57. Elsevier (2019)
4. Li, L., Yang, Z., Dang, Z., Meng, C., Huang, J., Meng, H., Wang, D., Chen, G., Zhang, J., Peng, H., Shao, Y.: Propagation analysis and prediction of the COVID-19. Infect. Dis. Model. 5, 282–292 (2020)
5. Ahmetolan, S., Bilge, A.H., Demirci, A., Peker-Dobie, A., Ergonul, O.: What can we estimate from fatality and infectious case data? A case study of Covid-19 pandemic (2020). arXiv preprint arXiv:2004.13178
6. Fong, S.J., Li, G., Dey, N., Crespo, R.G., Herrera-Viedma, E.: Composite monte carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. Appl. Soft Comput. 106282 (2020)
7. Ribeiro, M.H.D.M., da Silva, R.G., Mariani, V.C., dos Santos Coelho, L.: Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil, Chaos, Solitons & Fractals, 109853 (2020)
8. Bhapkar, H.R., Mahalle, P., Dhotre, P.S.: Virus graph and COVID-19 pandemic: a graph theory approach, Preprints, 2020040507 (2020)
9. Fong, S., Li, G., Dey, N., Crespo, R.G., Herrera-Viedma, E.: Finding an accurate early forecasting model from small dataset: a case of 2019-nCoV novel coronavirus outbreak. Int. J. Interact. Multimedia Artif. Intell. 6, 132–140 (2020)
10. GitHub (2020) Cssegisanddata/COVID-19 [online] Available at: https://github.com/CSSEGISandData/COVID-19/blob/master/README.md
11. mohfw.gov.in (2020) [online] Available at: https://www.mohfw.gov.in/
12. covid19india.org. 2020. COVID-19 Tracker| India. [online] Available at: https://www.covid19india.org/