Methodical Analysis and Prediction of COVID-19 Cases of China and SAARC Countries

Sarika Agarwal and Himani Bansal

Abstract COVID-19 pandemic has become a major challenge for all the countries of the world. No medicine has been developed till now to cure it. Coronavirus (COVID-19) is the family of viruses that causes illness and has symptoms like the common cold, influenza, and severe acute respiratory syndrome (SARS) that spread via breathing droplets. Proper analysis and prediction of the COVID-19 patients and its increasing rate of spread will help the government and people to mitigate its effect. This gives a reason to analyze, compare, and predict the cases in India, China, and SAARC countries to make early decision for taking preventive measures to combat its effects in a timely manner. In this paper, we have analyzed COVID-19 cases from January 21, 2020 to June 25, 2020 and have predicted the cases of COVID-19 for the period of next two weeks using multiple linear regression and polynomial regression models of machine learning.

Keywords COVID-19 · Linear regression · Polynomial regression · Coronavirus · Prediction · Machine learning

1 Introduction

Corona means crown. Coronavirus has a crown-like structure which is known to be initiated from the animal and transmitted to a human. This virus is new to the human immune system to fight. This virus can stick to almost any substance and is one-nine hundredth of a width of a hair in size. As of today, coronavirus disease (COVID-19) has spread in almost all the countries. More than 215 countries and 5,607,791 people are affected by coronavirus disease as on May 26, 2020 [1].

S. Agarwal (✉) · H. Bansal
Department of CSE/IT, Jaypee Institute of Information Technology, Noida, India
e-mail: sarikagarwal.it@gmail.com

H. Bansal
e-mail: singal.himani@gmail.com

P. K. Singh et al. (eds.), *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*, Lecture Notes in Networks and Systems 203,
https://doi.org/10.1007/978-981-16-0733-2_41

Pneumonia without any cause was first identified in Wuhan, China, and the same was reported to the World Health Organization (WHO) office in China on December 31, 2019. This disease has spread to every province of mainland China and other countries also. In January 2020, the UK and Russia found two corona positive cases. In January, only one case was found in Sweden and Spain. Canada reported 4 cases. After analyzing data about symptoms of patients, it was declared as a health emergency. Then, WHO named this disease as coronavirus disease (COVID-19). COVID-19 is an abbreviated term, where CO denotes Corona, VI denotes virus, D denotes disease, and 19 refer to the year it was discovered. The symptoms of COVID-19 disease are the same as normal viral diseases like fever, tiredness, loss of smell, loss of taste, and dryness. Less commonly seen symptoms are diarrhea, conjunctivitis, a rash on skin, shortness of breath, and loss of speech. The symptoms differ in age group. It can cause more severe health issues to those whose age is above 60 [2]. If the patients are already suffering from diabetes, lung disease, and heart disease then it may lead to death. The transmission rate of this virus is indicated by the reproductive number (R_0—R-nought). R_0 tells how many people are infected from one person with that disease. WHO estimated R_0 for COVID cases between 1.4 and 2.5 on January 23, 2020 [3]. It means on an average, one person transmits COVID to 1–3 people. The total number of confirmed cases in January was 8096 worldwide. The confirmed infected cases increased exponentially up to 7,102,957 and deaths were 406,343 on June 8, 2020. Total persons recovered were 3,466,581. Figure 1 shows the exponential growth of world's confirmed, death, and recovered COVID cases data from January 22, 2020 to June 20, 2020. Recovery cases are almost 40 lakhs out of 80 lakhs cases on June 20, 2020 which shows that recovery rate is more than 50% throughout the world.

The person who gets recovered from corona develops antibodies in their body that fights against the coronavirus. Antibodies may help people to move to work without being scared and can protect that person from the re-infection of COVID-19. But no evidence has been found till now. Scientists are not sure as how long these antibodies will protect against coronavirus. Immunity against corona disease is

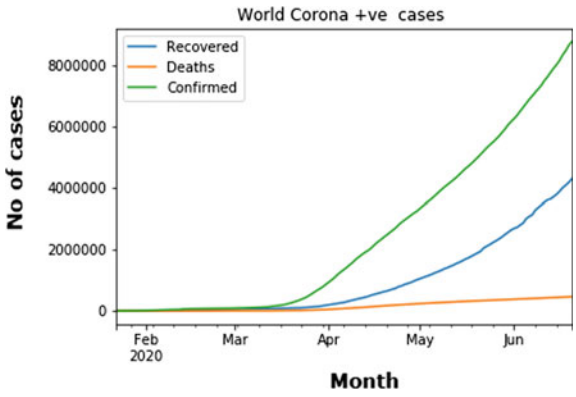


Fig. 1 World data statistics on COVID-19 from January 22, 2020 to June 20, 2020

developed in one or two weeks. The body starts fighting against viral without delay which is very natural. Three cells, namely macrophages, neutrophils, and dendritic, help to diminish the progress of the virus and prevent it from causing symptoms [4]. Analysis and prediction of such disease are necessary to combat its effect in a timely manner.

This paper is divided into six sections. Section 2 gives the motivation behind taking up this study and our contribution in this study. In Sect. 3, comparative analytical study of COVID-19 in China, India, and other SAARC countries is done. Section 4 has results of methodical prediction of COVID confirmed, recovered and death cases. Section 5 details the comparative study and Sect. 6 marks the conclusion of the paper.

2 Motivation and Contribution

Novel coronavirus was first found in China and spread in almost all the countries. A number of COVID-19 cases are increasing day by day and we have limited sources which has become a problem for the government to provide the medical aid to all the infected persons. Early prediction of COVID-19 cases might be helpful for making necessary arrangements. This generated the need to analyze, compare, and predict the coronavirus cases in India, China, and SAARC countries using the multiple linear regression and polynomial regression models of machine learning.

With this aim, the authors have used a dataset from Kaggle [5] that contains datewise count of confirmed, death, and recovered cases of COVID-19 from different countries. The dataset contains datewise number of cases found, country of origin of the patient, exact count of confirmed, death, and recovered cases from January 22, 2020 to June 20, 2020. With this dataset, we got hold on the current situation of COVID-19 in the world and did comparative study of cases in China and SAARC countries. We then predicted COVID-19 cases in India, China, and SAARC countries from June 26, 2020 to July 10, 2020. South Asian Association for Regional Cooperation (SAARC) has Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka as its member countries.

3 Comparison of COVID-19 Among India, China, and Other SAARC Countries

China was the first country to report the case of COVID-19 from Wuhan seafood wholesale market. It was reported that COVID-19 was transmitted from animals [6]. Though it was transmitted from animals, it was concluded that it can spread from human to human. The first COVID case in China was reported on November 20, 2019. The Chinese government did not support the doctors. They even criticized who were willing to intimate others the seriousness of the new SARS Virus in the city

of Wuhan. After one month, in December 2019, 60 confirmed cases of SARS disease were found in Wuhan. The cases kept multiplying. On December 30, 2019, Health Commission in Wuhan asked local hospitals in Wuhan to report all the information about cases of pneumonia of unclear cause in the past week [7]. Dr. Li Wenliang also warned fellow doctors about the seriousness of the disease and advise them to wear protective clothing to avoid infection. Figure 2 shows the confirmed, recovered, and death cases of COVID-19 in China (till June 20, 2020).

We can easily predict from the graph that China has controlled the COVID-19 cases. Initially, the COVID-19 cases were rising simultaneously, the recovery rate was also increasing. By February 2020, China was able to control the increasing rate of COVID-19 and reduced the new cases by more than 90% [8]. Since, China did not stop the people to enter or move from the country; on January 13, 2020, the first case was confirmed outside China in Thailand. India also reported its first case on January 30, 2020. The patients had the travel history of China [9]. By June 11, 2020, India had total 286,579 confirmed cases. Out of them, 141,029 patients recovered (including 1 migration) while 8102 died. India currently has the largest number of confirmed cases in Asia. Figure 3 shows confirmed, death, and recovered cases in India till June 20, 2020. The graph shows that initially transmission rate was slow, but from the middle of April 2020, it is multiplying very fast.

All the SAARC countries (Afghanistan, Nepal, Bhutan, Sri Lanka, Bangladesh, India, Maldives, and Pakistan) have COVID-19 patients. Figure 4 shows India has the largest number of confirmed, recovered, and death cases in all the SAARC countries followed by Pakistan and Bangladesh. India has the highest population among all the SAARC countries, that may be one of the reasons for the fast transmission of COVID-19. Recovery cases are also fast in India followed by Pakistan and Bangladesh.

The death rate is 2.5% which means that patients are recovering from the disease. The effect of COVID-19 is mild in SAARC countries as compared with China. Figure 5 represents share of active, recovered, and dead patients in percentage across SAARC countries.

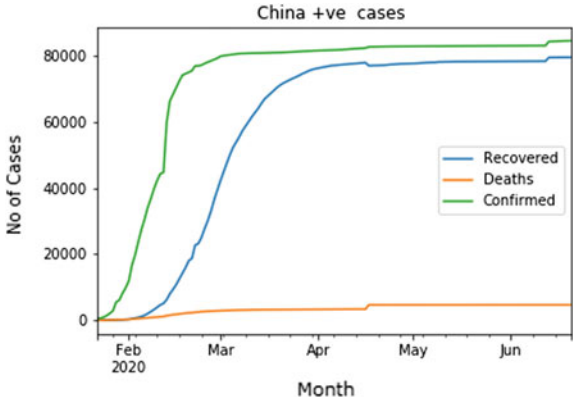


Fig. 2 Confirmed, recovered, and death cases of COVID-19 in China (till June 20, 2020)

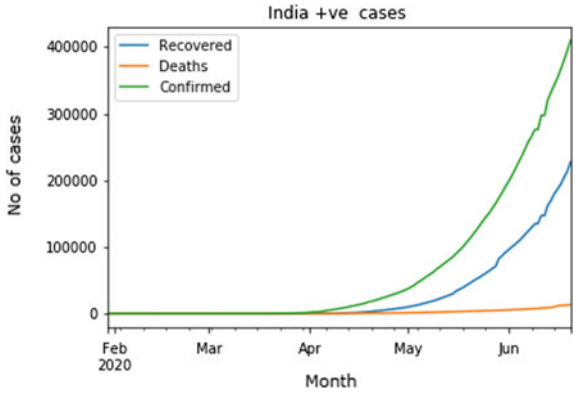


Fig. 3 Confirmed, recovered, and death cases of COVID-19 in India (till June 20, 2020)

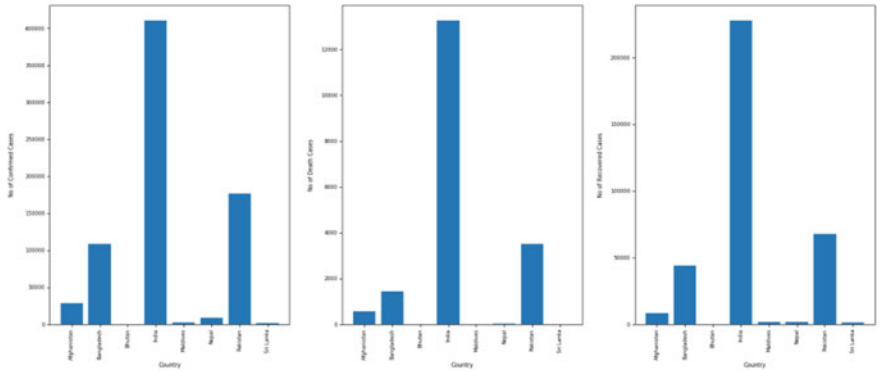


Fig. 4 COVID-19 confirmed, recovered, and death cases of SAARC countries till June 20, 2020

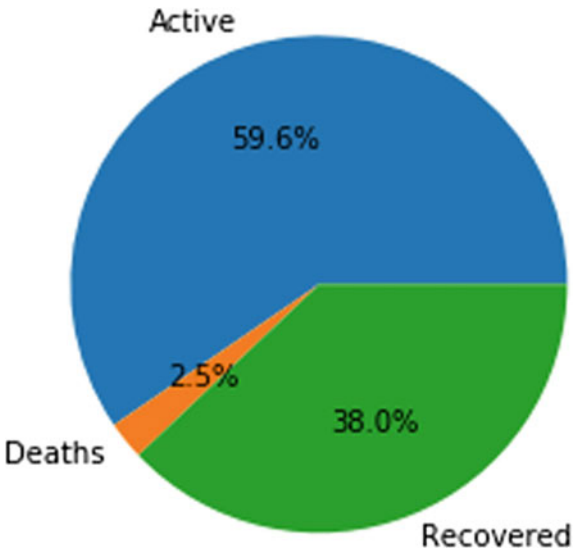
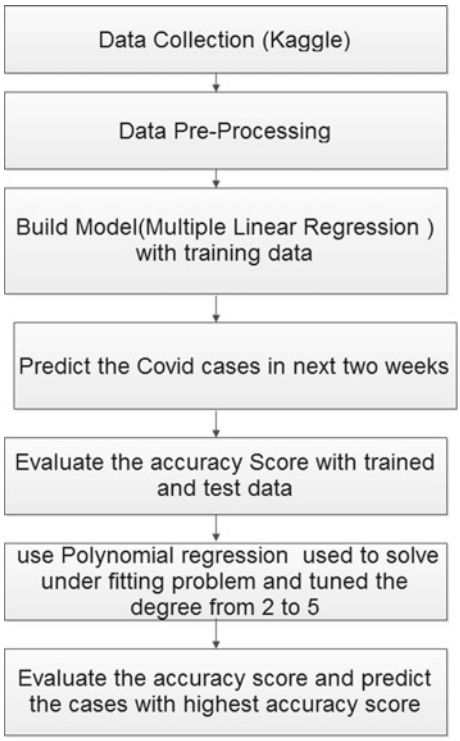


Fig. 5 Share of active, recovered, and dead patients in SAARC countries

4 Prediction of COVID-19 Cases in India, China and SAARC Countries

The spread of COVID-19 is unstoppable and has been declared as a pandemic. It has infected more than 8,007,804 people in the world by June 15, 2020 and more than 50% were recovered throughout the world. Prediction of confirmed cases of Corona virus disease were also done by many techniques like ARIMA [15]. We have tried to predict the cases in SAARC countries till July 20 through multiple linear regression. We have preprocessed the data by dividing the date column in day, month, year and eliminating the extra information like gender and symptoms. Date, confirmed, recovered and death cases of India, China, and other SAARC countries were used to train the model. The period of data considered is from January 22, 2020 to June 12, 2020. We have focused only on SAARC countries and China COVID-19 cases. We have used multiple linear regression to train and predict the COVID cases. Polynomial regression is also used to overcome the problem of underfitting. Figure 5 shows the flow of our implementation of prediction of COVID-19 cases.



Multiple linear regression is a machine learning algorithm used to predict the output and take more than one feature. The calculation for multiple linear regression is shown in Eq. 1. β_0, β_1, β_n are coefficient, x_1, x_2, x_n are independent variable and $\in$ is an intercept. Multiple linear regression is used to find the relationship between

the dependent and independent variables.

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \in \quad (1)$$

Polynomial regression is also a kind of regression that is used to overcome the problem of underfitting. Underfitting is a situation when our model does not learn enough from training data resulting in unreliable prediction. To overcome underfitting, we increased the features of the model. The model in Eq. 1 is transformed into Eq. 2.

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_3^3 + \ldots + \beta_n x_n^n \quad (2)$$

Polynomial regression with degree 1 is same as linear regression. To increase the features of our model, we have to tune the degree parameter to the extent that gives maximum accuracy and minimum variance between the training and testing model. Selecting a degree is a challenging task [10]. If the degree of the polynomial is less, it will lead to the problem of underfitting, i.e., not be able to fit the model properly. If the value of the degree of the polynomial is greater than actual, it will lead to the problem of overfitting. Overfitting problem can be solved through regularization.

We have used multiple linear regression for predicting the confirmed, recovered, and death cases of COVID-19 in SAARC countries and China. The accuracy score for the trained model came to be 0.11623268358995764 and the accuracy score of the test model came as 0.0938602714826949. This shows that accuracy is not as desired. Since the model trained by us was showing the underfitting problem, polynomial regression was also used. Table 1 gives the predicted datewise (daily) confirmed cases of SAARC countries and China from June 26, 2020 to July 10, 2020 obtained from polynomial regression at degree 2.

Polynomial regression of degree 2, 3, and 4 was applied for the prediction made. Table 2 shows the training and testing accuracy score at different degrees of polynomial regression.

As degree is increasing, the variance is also increasing between training and testing models. By analyzing Table 2, polynomial regression at degree 2 and 3 is good to predict the confirmed cases.

5 Results

The COVID-19 virus originated from China and spread almost all over the world. The first COVID case in SAARC countries was on January 23, 2020, in Nepal. After that, it was reported across all the SAARC countries.

Figure 6 shows exponential growth of COVID cases in all the SAARC countries till June 20, 2020. All the countries are trying hard to control COVID. Preventive measures like lockdown, social distancing, and regular hand wash have been taken. In spite of all these, our prediction shows that there will be an exponential growth

Table 1 Two weeks prediction of COVID-19 patients in SAARC countries and China

Date	Afghanistan	Nepal	Bhutan	Srilanka	Bangladesh	India	Maldives	Pakistan	China
26/6/20	17,516	20,009	22,501	24,994	27,486	29,978	32,471	34,963	3336
27/6/20	17,770	20,263	22,755	25,247	27,740	30,232	32,725	35,217	3346
28/6/20	18,024	20,517	23,009	25,501	27,994	30,486	32,979	35,471	3356
29/6/20	18,278	20,770	23,263	25,755	28,248	30,740	33,233	35,725	3365
30/6/20	18,532	21,024	23,517	26,009	28,502	30,994	33,486	35,979	3375
1/7/20	18,303	20,796	23,288	25,780	28,273	30,765	33,258	35,750	3472
2/7/20	18,557	21,049	23,542	26,034	28,527	31,019	33,512	36,004	3482
3/7/20	18,811	21,303	23,796	26,288	28,781	31,273	33,765	36,258	3492
4/7/20	19,065	21,557	24,050	26,542	29,035	31,527	34,019	36,512	3502
5/7/20	19,319	21,811	24,304	26,796	29,288	31,781	34,273	36,766	3512
6/7/20	19,573	22,065	24,557	27,050	29,542	32,035	34,527	37,020	3521
7/7/20	19,826	22,319	24,811	27,304	29,796	32,289	34,781	37,273	3531
8/7/20	20,080	22,573	25,065	27,558	30,050	32,542	35,035	37,527	3541
9/7/20	20,334	22,827	25,319	27,812	30,304	32,796	35,289	37,781	3551
10/7/20	20,588	23,081	25,573	28,065	30,558	33,050	35,543	38,035	3560

Table 2 Training and testing accuracy score at different degree of polynomial regression

Degree	Testing accuracy sore	Training accuracy score
2	0.12756589341174684	0.16024107460679723
3	0.17943516161710302	0.22900462331233495
4	0.21600801519253746	0.3014880334951651
5	0.32105688124553533	0.40943900115182696

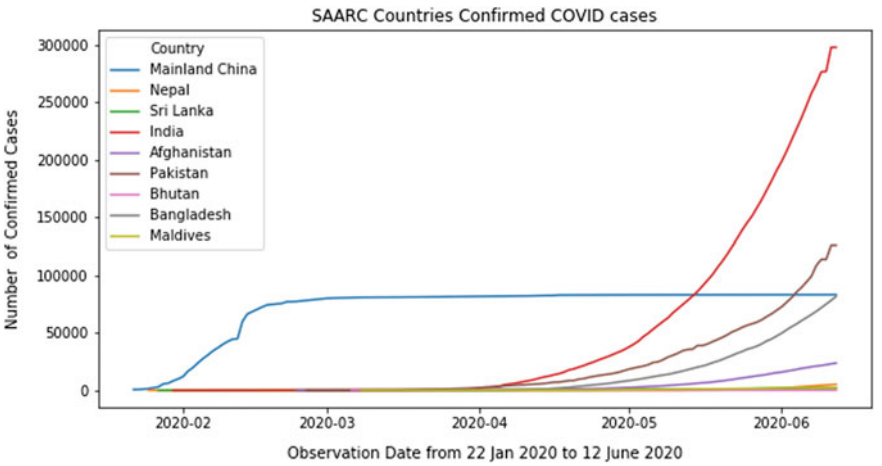


Fig. 6 Confirmed COVID cases in SAARC countries and China

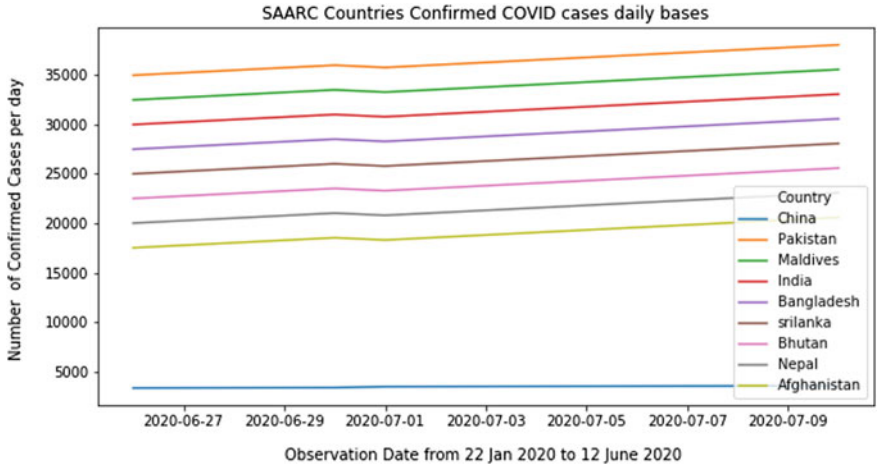


Fig. 7 Daily confirmed COVID cases prediction from June 26, 2020 to July 10, 2020

from June 26, 2020 to July 10, 2020. China have controlled the cases, but still cases will be found as per our prediction.

There is an exponential growth of COVID cases in India. Though Nepal was the first country that confirmed the COVID cases among all SAARC countries, Nepal COVID cases are not increasing exponentially. After India, Pakistan has shown exponential growth in COVID cases. The third country is Bangladesh where the COVID cases are increasing. The growth of COVID-19 cases is also controlled in China. Figure 7 shows the prediction of confirmed cases from June 26, 2020 to July 10, 2020 in SAARC countries and China.

6 Comparison with Other Schemes

Many models of machine learning have been used by the authors in literature to predict the cases of COVID-19. Some of the models used are traditional regression, multiple linear regression, polynomial regression and long short-term memory (LSTM). Jha et al. [11] predicted 7003 deceased cases by September 1, 2020 in Texas using Bayesian model. The author argues that prior distribution is set by the COVID experts and can be useful for small datasets. Tobias et al. [12] predicted confirmed cases of COVID-19 in Italy and Spain under lockdown using quasi-Poisson regression model. The quasi-Poisson model is a linear function of the mean. Gu et al. [13] applied cubic regression equations, which used the number of days as the input variable to predict the conformed COVID-19 cases in China and world. Pavlyshenko [14] used logistic curve to model COVID-19 spread. Different authors used different models but the efficiency of the model is based on the accuracy score of the training and testing data. No doubt error rate is also to be considered. We have used multiple

linear regression and have calculated the accuracy score of both trained and test data. We have used 80% data for training set and 20% data for testing. Linear regression model is not able to learn enough from the training set and gives the problem of underfitting. Then we tuned the model with polynomial regression at different degrees from 2 to 5 and again calculated accuracy score. Finally, we have predict the COVID-19 cases of next two weeks with polynomial regression having degree 3 showing the highest accuracy score.

7 Conclusion

We did not include population, age, and land size of the country. Population size may also affect the cases of COVID patients. Age of the population also affects the cases, as if the patients are younger, the recovery rate may increase and vice versa. In our prediction, we realized that the cases in all SAARC countries will increase exponentially. More precautions must be taken. No doubt mental stress is also increasing among the people. COVID-19 has badly affected business, employment, and people who are living below the poverty line. The government is also doing well in providing medical assistance to COVID patients. It is time to ramp up the preventive measures and the precautions to be taken.

References

1. World Health Organization. Rolling updates on coronavirus disease (COVID-19). Accessed from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen
2. How does COVID-19 affect different age groups. Accessed from https://www.nwhn.org/how-does-covid-19-affect-different-age-groups/
3. Worldometer. Accessed from https://www.worldometers.info/coronavirus/
4. World Health Organization. "Immunity Passports" in the context of COVID-19. Accessed from https://www.who.int/news-room/commentaries/detail/immunity-passports-in-the-context-of-covid-19
5. Kaggle. Accessed from https://www.kaggle.com/search?q=covid+19+dataset
6. Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Di Napoli R (2020) Features, evaluation and treatment coronavirus (COVID-19). In: Statpearls [internet]. StatPearls Publishing. The Lancet
7. BBC News. Coronavirus: what did China do about early outbreak? Accessed from https://www.bbc.com/news/world-52573137
8. Remuzzi A, Remuzzi G (2020) COVID-19 and Italy: what next? The Lancet
9. Wikipedia. COVID-19 pandemic in India. Accessed from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_India
10. Pandey G, Chaudhary P, Gupta R, Pal S (2020) SEIR and Regression Model based COVID-19 outbreak predictions in India. arXiv preprint arXiv:2004.00958
11. Jha PK, Cao L, Oden JT (2020) Bayesian-based predictions of COVID-19 evolution in Texas using multispecies mixture-theoretic continuum models. Comput Mech 1–14
12. Tobias A (2020) Evaluation of the lockdowns for the SARS CoV 2 epidemic in Italy and Spain after one month follow up. Sci Total Environ 725:

13. Gu C, Zhu J, Sun Y, Zhou K, Gu J (2020) The inflection point about COVID-19 may have passed. Sci Bull 5:98
14. Pavlyshenko BM (2020) Regression approach for modeling COVID-19 spread and its impact on stock market
15. Chakraborty T, Ghosh I (2020) Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis. Chaos Solitons Fractals 109850