ISSN 2350-109X
www.indianscholar.co.in

Indian Scholar

An International Multidisciplinary Research e-Journal

USING DATA ANALYTICS TO PREDICT SARS-COV-2 (SEVERE ACUTE RESPIRATORY SYNDROME CORONAVIRUS) PANDEMIC

Aditi Mallya
Euro school, Airoli
aditi.mallya@hotmail.com

Abstract

SARS-coV-2 has become a huge threat to humankind throughout the world. Machine learning (ML) techniques are used to analyze and interpret massive datasets and predict their output. ML techniques can be used to predict if a patient has been infected by SARS-coV-2 based on the symptoms stated by the WHO and CDC. First, the information on this virus is obtained which includes factors like its origin, its transmission capacity, its symptoms, its mutation rate, etc. Second, data analytics are applied to the existing data present. In this review, the methods for predicting future cases based on the existing data and as well as the algorithms used are discussed.

Keywords:*SARS-coV-2; Machine learning; Datasets; Data analytics*

INTRODUCTION

In this era of technology, data science and machine learning play an important role in many types of industries especially the health care industry. They make it convenient for medical professionals to manage their tasks.ML helps hospitals to maintain administrative processes and treat infectious disease [4][5]. Scientists are working and experimenting with machine learning (ML) to develop viable and acute solutions to diagnose and treat diseases. ML can identify diseases and virus infections more accurately so that patients' disease can be diagnosed at an early stage, the critically threatening stages of diseases can be avoided, and there can be fewer cases where the disease reaches an advanced stage. In the same manner, ML can be used to automate the task of predicting COVID-19 infection and help predict future COVID-19 infection counts [1].

The current pandemic, which has taken the world by storm, is caused by a virus named SARS-coV-2. It originated in Wuhan City, China in Dec. 2019. Coronaviruses are a type of virus that mainly causes respiratory diseases in humans whose severity can be very high (SARS - Severe Acute Respiratory Syndrome or MERS - Middle East Respiratory Syndrome) or very low (common cold or cough) [2]. First, we need to check the similarity of this virus with past outbreaks to get a clearer understanding of this pathogen.SARS-CoV-2 has a 79.5% similarity to SARS-CoV and 96% similarity to the bat coronavirus [7][8]. Coronavirusescomes under the Coronaviridae family. They are segregated into four genera: beta-, alpha-, gamma- and delta-coV. The viruses currently responsible for causing the disease in humans belong to alpha- or the beta-coV [2]. There are crown-like spikes present on the outer surface of the virus hence its

ISSN 2350-109X
www.indianscholar.co.in
Indian Scholar
An International Multidisciplinary Research e-Journal
MISA MEMBERS OF INTERNATIONAL SCHOOLS' ASSOCIATION

name, coronavirus. Coronaviruses are 65–125 nm in diameterand contain a single-stranded RNA as a nucleic material. (fig. 1) [3]

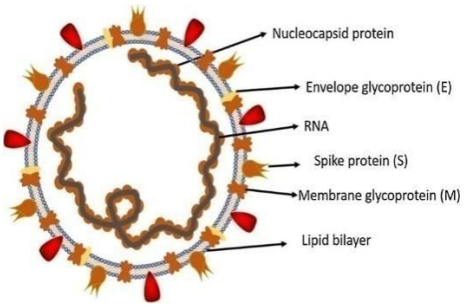

Fig. 1 [3]

The first cases which were reported from Wuhan proved to be a member of the beta-coV group [3]. Initially the coronavirus has an incubation period on 2-14 days in the human body, but recent data shows that the incubation period has increased from 14 to 20 or 28 days because of the mutation of the virus. it is transmitted by via an infected person's sneeze or cough [1]. The WHO declared COVID-19 as a global health emergency on the 30th of January 2020 and a global pandemic on the 11th of march, 2020 [6].

Theory

There are numerous methods or ML techniques which can be used to predict and forecast future pandemics. Some of them are linear regression (uses one independent variable to explain or predict the outcome of the dependent variable Y), logistic regression (a statistical analysis method used to predict a data value based on prior observations of a data set), using different types of models like predictive modelling, data visualizations and data sources which consists of data extracted from verified sources like John Hopkins University, WHO andDingXiangYuan (a website authorized by the Chinese government) [9] [1].

The dataset which is used and analysed here is provided the John Hopkins University available in Kaggle repository. [10] The dataset consists of 35775 records up till June 4, 2020, which includes state, country, longitude, latitude, date, confirmed cases, deaths and recovered cases. This dataset has records of cases in 213 countries and according to these records, these 213 countries have 6,632,985 confirmed cases, 391,136 death cases, 3,371,886 active cases and 2,869,963 of recovered cases. Table 1 shows the global spread of COVID-19 in 20 countries with the highest number of confirmed cases along with the other attributes like death and recovery rate. According to the table, USA has the highest number of confirmed cases.

Table 1. [7]

Country	Confirmed	Deaths	Recovered	Active	Death Rate	Recovered Rate
USA	1872660	108211	485002	1279447	5.78	25.90
Brazil	614941	34021	254963	325957	5.53	41.46
Russia	440538	5376	204197	230965	1.22	46.35

ISSN 2350-109X
www.indianscholar.co.in
Indian Scholar
An International Multidisciplinary Research e-Journal

Country	Confirmed	Deaths	Recovered	Active	Death Rate	Recovered Rate
UK	283079	39987	1219	241873	14.13	0.43
Spain	240660	27133	150376	63151	11.27	62.48
Italy	234013	33689	161895	38429	14.40	69.18
India	226713	6363	108450	111900	2.81	47.84
France	192330	29024	69573	93733	15.33	36.17
Germany	184472	8635	167909	7928	4.68	91.02
Peru	183198	5031	76228	101939	2.75	41.61
Turkey	167410	4630	131778	31002	2.77	78.72
Iran	164270	8071	127485	28714	4.91	77.61
Chile	118292	1356	21305	95631	1.15	18.01
Mexico	105680	12545	74758	18377	11.87	70.74
Canada	95269	7717	52184	35368	8.10	54.78
Saudi Arabia	93157	611	68965	23581	0.66	74.03
Pakistan	85264	1770	30128	53366	2.08	35.33
Mainland China	83027	4634	78328	65	5.58	94.34
Qatar	63741	45	39468	24228	0.07	61.92
Belgium	58767	9548	16048	33171	16.25	27.31

The COVID-19 pandemic is compared to four epidemics that has occurred in the past. This is done to get a clearer understanding of the mortality rate and the transmission capacity which is based on the number of people infected in total. The four other datasets [12] [13] [14] [15] are referred to in Ref. [7]. The four epidemics, which the COVID-19 pandemic is being compared to, are EBOLA, SARS, H1N1 AND MERS diseases. Table 2 shows the comparison of the pandemic to above named past epidemics. Based on the table, the number of infections in the current pandemic is higher than the SARS, EBOLA AND MERS epidemics but H1N1 disease still has the highest number of confirmed cases despite having a low mortality rate [7] [11] [12] [13] [14] [15].

ISSN 2350-109X
www.indianscholar.co.in

Indian Scholar

An International Multidisciplinary Research e-Journal

Table 2. [7]

Epidemic/pandemic	Start year	End year	confirmed	deaths	death rate
COVID-19	2019	–	6,632,985	391,136	5.90
SARS	2003	2004	8096	774	9.56
EBOLA	2014	2016	28646	11323	39.53
MERS	2012	2017	2494	858	34.40
H1N1	2009	2010	6724149	19654	0.29

Using the dataset in Ref. [10] which is from Ref. [7] several observations and trends can be drawn. The data shows the number of confirmed cases was substantially higher in the united states than any other country. Despite having the highest number of confirmed cases in February 2020, China in June 2020 is ranked 18th in terms of confirmed cases which shows that they had manged the spread. Fig. 2 shows the recovery rate and death rate across the world from January 22, 2020, to June 04, 2020. Death and recovery rate is defined by the ratio of recovery/death rate to confirmed cases. It shows that the recovery rate had increased from January to the first week of March The dataset also shows that there was a sharp increase regarding the death rate in the first week of March 2020. It also shows that the global percentage of active cases was 42.20% of all confirmed cases and the recovered cases

was 51.90% of all confirmed cases [7] [11].

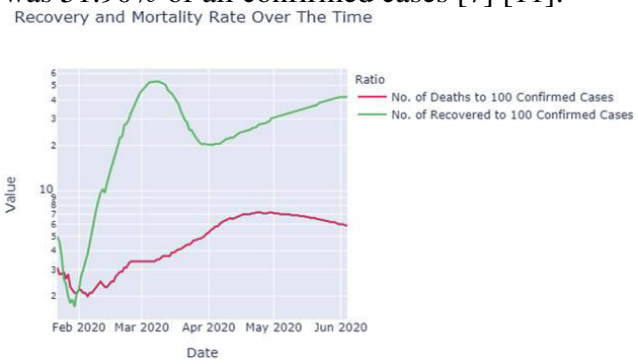

Fig. 2 [7] [10]

The application of data science on COVID-19 can be implemented by the use of various machine language classifiers [16]. The programming languages which can be used are Python, MATLAB etc. for prediction of an event or a problem, feature selection methods can be used to find the highest effect or factor affecting the problem [17] [18]. Then the classifiers can be used to conclude prediction results [17] [18]. Usage of the regression and feature selection process in algorithms is done and the regressions are expressed into formulas and the values are obtained which indicate the forecast of the COVID-19 pandemic.

ISSN 2350-109X
www.indianscholar.co.in

Indian Scholar

An International Multidisciplinary Research e-Journal

RESULT

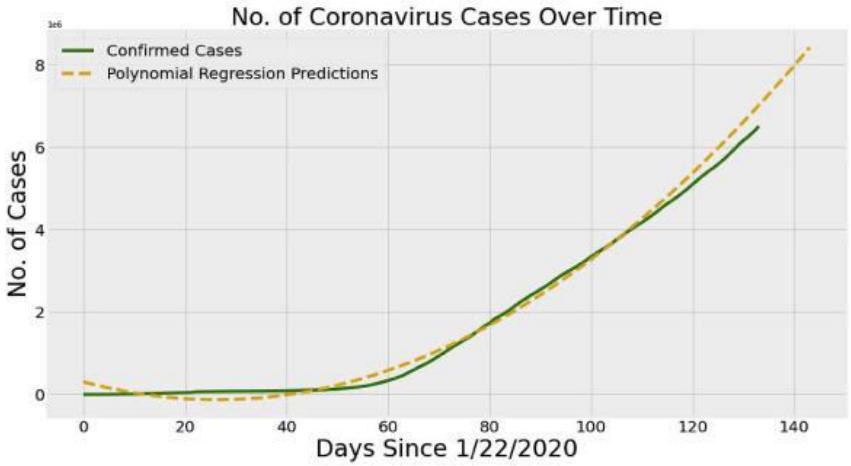

Fig. 3 [7][10]

Fig. 3 shows the number of confirmed cases plotted against the number of days starting from January 22, 2020. The green solid line is for actual data and the dashed lines are for predicted ones obtained from the created algorithms. Table 3. Shows the estimated values of the predicted confirmed cases.

Future dates	Predicted confirmed cases
June 5, 2020	7286283
June 6, 2020	7423297
June 7, 2020	7561566
June 8, 2020	7701090
June 9, 2020	7841868
June 10, 2020	7983901
June 11, 2020	8127188
June 12, 2020	8271730
June 13, 2020	8417526

Table 3. [7] [10] [17] [18]

CONCLUSION

In conclusion, the dataset which has been referred to here [10] can play a vital role in monitoring and predicting outbreaks such as COVID-19. The results concluded shows that a gradual increase in the confirmed cases was expected. With the help of ML and the availability of appropriate datasets, accurate predictions can be made which will help countries and especially their healthcare industry prepare for worst. The ongoing COVID-19 pandemic has deeply affected the status of many countries and has resulted in a worldwide emergency. The aim of this

study was to forecast the number of confirmed cases of SARS-coV-2 so it could help industries prepare well before an event of major crisis.

REFERENCES

1. Painuli, D., Mishra, D., Bhardwaj, S., & Aggarwal, M. (2021). Forecast and prediction of COVID-19 using machine learning. *Data Science for COVID-19*, 381–397. https://doi.org/10.1016/B978-0-12-824536-1.00027-7
2. WHO/2019-nCoV/FAQ/Virus_origin/2020.1
3. Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R. (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of advanced research*, *24*, 91–98. https://doi.org/10.1016/j.jare.2020.03.005
4. Shirsath, S.S. and Patil, S., 2018. Disease prediction using machine learning over big data. *International Journal of Innovative Research in Science, Engineering and Technology*, *7*(6), pp.6752-6757.
5. Sreeja, S., Bhavya, L., Swamynath, S. and Dhanuja, R., 2019. Chest x-ray pneumonia prediction using machine learning algorithms. *Int. J. Res. Appl. Sci. Eng. Technol*, *7*(04), pp.3227-3230.
6. David J Cennimo Discusses Coronavirus Disease 2019 (COVID 19). Available from: https://emedicine.medscape.com/article/2500114-overview.
7. M. Rubaiyat Hossain Mondal, Subrato Bharati, PrajoyPodder, Priya Podder, Data analytics for novel coronavirus disease ,Informatics in Medicine Unlocked ,Volume 20,2020100374,ISSN 2352-9148, https://doi.org/10.1016/j.imu.2020.100374.
8. Zhou, P., Yang, XL., Wang, XG. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579,** 270–273 (2020). https://doi.org/10.1038/s41586-020-2012-7
9. Binti Hamzah FA, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, et al. CoronaTracker: Worldwide COVID-19 Outbreak Data Analysis and Prediction. [Preprint]. Bull World Health Organ. E-pub: 19 March 2020. doi: http://dx.doi.org/10.2471/BLT.20.255695
10. COVID-19 Dataset | Kaggle
11. Dey, SK, Rahman, MM, Siddiqi, UR, Howlader, A. Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach. *J Med Virol*. 2020; 92: 632–638. https://doi.org/10.1002/jmv.25743
12. Ebola | 2014-2016 | Western Africa Ebola Outbreak | Kaggle
13. MERS Outbreaks data 2012-2019 | Kaggle
14. Ebola Cases, 2014 to 2016 | Kaggle
15. SARS 2003 Outbreak Dataset | Kaggle
16. Sumayh S. Aljameel, Irfan Ullah Khan, Nida Aslam, Malak Aljabri, Eman S. Alsulmi, "Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients", *Scientific Programming*, vol. 2021, Article ID 5587188, 10 pages, 2021. https://doi.org/10.1155/2021/5587188
17. Bharati S, Podder P, Mondal R, Mahmood A, Raihan-Al-Masud M. Comparative performance analysis of different classification algorithm for the purpose of prediction of lung cancer. InInternational Conference on Intelligent Systems Design and Applications 2018 Dec 6 (pp. 447-457). Springer, Cham.
18. Raihan-Al-Masud M, Mondal MR. Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. Plos one. 2020 Feb 6;15(2):e0228422.