Informatics in Medicine Unlocked 20 (2020) 100374

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

Data analytics for novel coronavirus disease

M. Rubaiyat Hossain Mondal [a,*], Subrato Bharati [a], Prajoy Podder [a], Priya Podder [b]

[a] Institute of Information and Communication Technology, Bangladesh University of Engineering and Technology, Dhaka, 1205, Bangladesh
[b] Dhaka National Medical College, Dhaka, 1100, Bangladesh

ARTICLE INFO

Keywords:
Coronavirus
COVID-19
Classification
Machine learning
Regression
SARS-CoV-2

ABSTRACT

This paper describes different aspects of novel coronavirus disease (COVID-19), presents visualization of the spread of the infection, and discusses the potential applications of data analytics on this viral infection. Firstly, a literature survey is done on COVID-19 highlighting a number of factors including its origin, its similarity with previous coronaviruses, its transmission capacity, its symptoms, etc. Secondly, data analytics is applied on a dataset of Johns Hopkins University to find out the spread of the viral infection. It is shown here that although the disease started in China in December 2019, the highest number of confirmed cases up to June 04, 2020 is in the USA. Thirdly, the worldwide increase in the number of confirmed cases over time is modelled here using a polynomial regression algorithm with degree 2. Fourthly, classification algorithms are applied on a dataset of 5644 samples provided by Hospital Israelita Albert Einstein of Brazil in order to diagnose COVID-19. It is shown here that multilayer perceptron (MLP), XGBoost and logistic regression can classify COVID-19 patients at an accuracy above 91%. Finally, a discussion is presented on the potential applications of data analytics in several important factors of COVID-19.

1. Introduction

A life threatening infectious coronavirus [1] started from Wuhan of China in December 2019. As of June 04, 2020, the disease has officially spread to 213 countries and territories infecting a total of 6,632,985 people causing a death of 391,136. The patients affected by this coronavirus show symptoms of pneumonia. The World Health Organization (WHO) has termed this particular virus as *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) and the associated disease as COVID-19 [2–6]. Recently, the WHO has declared COVID-19 as a pandemic. One of the similar coronaviruses reported in the last two decades is the outbreak of Severe Acute Respiratory Syndrome (SARS) in 2003 causing thousands of deaths. Another disease Middle East Respiratory Syndrome (MERS) emerged in 2012. A third virus known as Swine Acute Diarrhea Syndrome (SADS) was noted in the swine industry in 2017. These diseases originated from bats and they are pathogenic to humans or livestock [7–10]. Some of the main symptoms of this disease are the development of fever, cough, and respiratory problems including shortness of breath. According to WHO, the incubation period of this virus in most cases vary from 2 to 10 days. Unfortunately, there is no known cure for this disease, for example there is no vaccine for prevention of this disease. There is no specific antiviral treatment as well. The doctors help the patients to manage the symptoms. In many cases, the patients may develop pneumonia and experience failure of multiple organs leading to possible death. The COVID-19 is spread by respiratory droplets from the patients. In other words, the spread is severe when the infected patients cough or sneeze [10]. However, the spread can be minimized to some extent by taking hygiene measure including careful handwashing. Moreover, early detection of this disease can help in the containment of the virus. Because of the life threatening nature and no cure of COVID-19, significant research interest has been seen in this field since early January 2020. It is expected that data analytics can play an important role in investigating the salient features of COVID-19 and eventually find a vaccine for it. Therefore, this paper focuses on the application of data analytics on COVID-19. The contributions of this paper are summarized below:

1) The literature has been reviewed to find the important aspects of COVID-19.
2) Data analytics has been applied on a currently available dataset to visualize the spread of COVID-19 as of June 4, 2020.
3) A description is provided on the feature selection, regression and classification algorithms suitable for this disease. A polynomial regression model has been used to model the number of confirmed

* Corresponding author.
E-mail address: rubaiyat97@iict.buet.ac.bd (M.R.H. Mondal).

https://doi.org/10.1016/j.imu.2020.100374
Received 20 March 2020; Received in revised form 10 June 2020; Accepted 10 June 2020
Available online 15 June 2020

cases in the world. Moreover, feature selection and classification algorithms are applied on a dataset for data-driven diagnosis of COVID-19.

4) The possible applications of data analytics on the important aspects of the disease are presented.

The rest of the paper is organized as follows. Section 2 provides a literature survey on this virus. Section 3 uses data analytics to discuss the current situation of the spread of the virus across the globe. The application of regressions to model the confirmed cases and the application of classifiers to predict COVID-19 patients are discussed in Section 4. Section 5 discusses about the possible areas where data analytics can be used to get an understanding of the risks of COVID-19. Finally, Section 6 provides concluding remarks.

2. Survey on the research on COVID-19

Recently there has been tremendous research interest in COVID-19 focusing on its spread [1–8], origin and genomic structure [9–15], clinical characteristics [16–18], and drug discovery [13,19–21]. Particularly the interest is significant after the start of COVID-19 in late December 2019. Some of the important research results are described in the following.

2.1. Transmission of COVID-19

This section focuses on the transmission dynamics and reproductive number of COVID-19. There is a relation among the species of coronaviruses, geographical distributions of coronaviruses, bat species and reservoir hosts [1]. In the early outbreak of COVID-19 in China, work-related transmission has been an important factor, and the infection may have spread greatly among health care workers, transport workers, services and sales workers [2]. The transmission of COVID-19 is estimated by the collection and analyzation of spatiotemporal data [3]. In this regard, the basic reproductive number of COVID-19 is estimated using the individual case reports of 140 infected persons. With the consideration of case reports, the estimates of the epidemiology parameters, human travel data and infection data, the basic reproductive number is found to be in between 4.7 and 6.6 [3]. There may be underreporting cases of coronavirus during the period of January 1-15, 2020 [4]. The cumulative incidence of 5502 cases in China is estimated with a confidence interval of 95% [5]. The epidemic of COVID-19 may spread due to some untraced exposures other than the exposure in seafood marker in China [5]. Some patients do not show symptoms, while others have different types of symptoms [6]. Some patients have different symptoms such as pneumonia with problem identified in CT scan, acute respiratory distress syndrome, RNAaemia, etc. [6]. There may be undetected internationally imported cases too [7]. A transmission model in Ref. [8] estimated the basic reproductive number to be 3.11.

2.2. Origin of COVID-19

This section focuses on the time origin and genetic diversity of COVID-19. The increasing genetic diversity of this disease is indicated by phylogenetic, transmission network, and likelihood mapping analyses of the genome sequences [9]. The SARS and MERS diseases are caused by the SARS-CoV and MERS-CoV pathogens, respectively. SARS outbreak in 2002–2003 caused more than 8000 cases with 774 deaths in 37 countries. On the other hand, MERS outbreak in 2012 caused 2494 cases with 858 deaths in 27 countries over the world [9]. Both SARS-CoV and MERS-CoV are infectious disease zoonotic in origin that means these are spread from animal to human. Bats are regarded as the animal host source, while palm civets and camels are the intermediate means between bats and animals for SARS-CoV and MERS-CoV, respectively. Similar to SARS and MERS, COVID-19 may be originated from bats, but the intermediate carrier between bats to humans is yet to be cleared [9]. The concepts of the pathology and pathogenesis used for SARS-CoV and MERS-CoV are now applied to find the characteristics of COVID-19 [10]. COVID-19 is a class of β-coronavirus genus, having a 79.0% and a 51.8% nucleotide identity to SARS-CoV and MERS-CoV, respectively [10]. This virus has 96% similarity when compared with a bat coronavirus in terms of genome [9,10]. COVID-19 is a recombinant virus which may be a homologous recombination of a bat coronavirus and another coronavirus whose origin is not known [11]. Moreover, COVID-19 has some genetic similarity with bat coronavirus and codon usage bias with snake or snake's translation machinery [11]. Four structural proteins are essential for the formation of coronavirus during virion assembly [12].

The genome of SARS-CoV-2 encodes non-structural proteins as well as structural proteins. The non-structural proteins are 3-chymotrypsin-like protease, papain-like protease, helicase, and RNA-dependent RNA polymerase (RdRp). The structural protein is spike glycoprotein. There are also accessory proteins [13]. In order to prevent any future outbreak of COVID-19, it is important to find the exact origin and the intermediate hosts of SARS-COV-2 [14]. It is found that the codon usage pattern of SARS-COV-2 is unique. When the codon usage of different viruses including SARS-CoV are taken into consideration, the codon usage of SARS-CoV-2 is very much different from that of humans. For the case of SARS-CoV-2, the codons are replaced by ones with lower human-preference. Because of this unique codon usage pattern, it is believed that SARS-CoV-2 was probably formed by evolving in an uncommon intermediate host for a long time. The codon biases are generated due to the evolution of genome composition [14]. When whole genome level is taken into consideration, SARS-CoV-2 has 79.5% similarity to SARS-CoV and 96% similarity to the bat coronavirus (RaTG13) [15]. This virus is still new in its evolutionary journey in human body hence it will be evolved in the future by changing codons, structures, etc. [14].

2.3. Clinical characteristics of COVID-19

The clinical symptoms, features, and parameters of COVID-19 are being investigated in a number of experiments and studies [16–18]. The study in Ref. [16] evaluates the clinical characteristics of COVID-19 for the case of nine pregnant women and possibility of intrauterine vertical transmission of this virus. The data related to clinical records, lab tests, and chest CT scans of those women are studied. According to the study [16], as of February 4, 2020, seven patients had fever, four had cough, three had myalgia, two had sore throat, two had malaise, two had fetal distress, five had lymphopenia, three had increased aminotransferase concentrations, and no one died. Furthermore, the study reports that no neonatal asphyxia was observed in 9 newborn babies. Results of these nine patients indicate that COVID-19 may not cause intrauterine fetal infections at pregnancy. However, the small sample size of nine patients is not good enough to come to a definite conclusion. The interaction between the human innate immune system and COVID-19 is studied in Ref. [17] to understand the virus induced inflammation of lung tissues. The results show that the infection can be controlled by humoral immunity, targeted immunotherapy can be useful [17]. The genome structure, entry of the virus into target cells, etc. information is reported in Ref. [18].

2.4. Drugs of COVID-19

Recently there has been significant research [19–21] in the development of drugs and vaccines for COVID-19. Some of the promising drugs target nonstructural proteins, while other drugs target viral entry and immune regulation pathways [19,20]. As of April 2020, some of the main potential post-infection therapies known as favipiravir (antiviral against influenza), remdesivir (antiviral), lopinavir/ritonavir (antiviral) and hydroxychloroquine/chloroquine (antiparasitic and antirheumatic)

– are in the final stage of human testing [13]. The development of vaccine against COVID-19 started just after the publication of genetic sequence of the causative virus SARS-CoV-2 on January 11, 2020. The first vaccine candidate started clinical trial on human on March 16, 2020. As of April 8, 2020, there are 78 confirmed active vaccine candidates of which 5 are into clinical development stage. Three of these are mRNA-1273, Ad5-nCoV and INO-4800 developed by Moderna, CanSino Biologicals, and Inovio, respectively. The remaining two LV-SMENP-DC and pathogen-specific aAPC are developed by Shenzhen Geno-Immune Medical Institute [19]. Very recently 5 more vaccines are also in the clinical trial stage. These are Covid-19/aAPC (Phase I), LV-SMENP-DC (modified DCs) (Phase I/II), bacTRL-Spike (Phase I), mRNA-1273 (Phase II), and Ad5-nCoV (Phase II) [21].

3. Data analytics on the status of the disease

The situation of this COVID-19 is changing very frequently. The work in Ref. [22] discusses about a number of datasets related to this virus. The authors in Ref. [22] also provide some visualization of this epidemic up to February 16, 2020. In this work, we analyze the dataset of June 04, 2020 provided by the Johns Hopkins University available in Kaggle repository [23]. This dataset has 35775 records as of June 04, 2020. The attributes of the dataset are Province/State, Country/Region, Latitude, Longitude, Date, Confirmed Cases, Deaths, and Recovered cases. We present data visualizations in terms of tables, bar charts, pie charts and other graphs. According to this dataset, 213 countries have confirmed cases of 6,632,985, death cases of 391,136, active cases of 3,371,886 and recovered cases of 2,869,963 by COVID-19 as of June 04, 2020. The death rate can be obtained in two ways. When the number of deaths is divided by the number of confirmed cases then death rate is 5.90%, however when number of deaths is divided by the number of closed cases (death + recovered) then death rate is 11.99%. For the rest of this paper, death rate will only be considered in terms of number of deaths over number of confirmed cases.

Table 1 shows 20 countries that have the highest confirmed cases. Apart from the number of confirmed infection cases, Table 1 also presents the number of recovered patients, the number of active cases, the number of deaths, death rate and recovery rate. It can be seen that most of the confirmed cases are in the USA. After the USA, the countries with most confirmed cases are Brazil, Russia, UK, Spain, Italy, India, and France. It can be seen from Ref. [22] that on February 16, 2020, China, Singapore and Japan have the first, second and third position in terms of the most number of patients having COVID-19. So, since February, the infection has spread very rapidly in other countries, while China, Singapore and Japan have managed the spread of the disease to some extent.

Next, COVID-19, a pandemic, is compared with other recent epidemics using four datasets available in Kaggle [24–27]. Table 2 provides a comparison of COVID-19 with EBOLA, SARS, H1N1 and MERS diseases. It can be seen that compared to EBOLA, SARS and MERS, the number of affected people by COVID-19 is much higher. However, H1N1 disease having a low mortality rate still has affected much more people than any of the epidemics mentioned in Table 2.

Next, using the dataset in Ref. [23], a number of illustrations are shown. Fig. 1 shows the bar chart of confirmed cases and death cases in different countries. It indicates that the number of confirmed cases and death cases of the disease are significantly higher in the USA than any other countries. Spain has the second and Italy has the third highest confirmed cases, while Italy has the second and the UK has the third highest death cases. An important point to note here that although China experienced the most number of confirmed cases up to February 16, 2020 [22], it is now in the 18th position in the world in terms of the number of confirmed cases and death cases. Fig. 2 shows the bar chart of death rate across the globe. In this case, death rate is represented as the ratio of number of deaths to number of confirmed cases. The largest death rate is in Yemen being 22.74%. The Belgium, France and Italy have the 2nd, 3rd and 4th highest death rates being 16.25%, 15.33% and 14.40%, respectively. Fig. 3 shows the plots of death rate and recover rate across the globe for a time span ranging from January 22, 2020 to June 04, 2020. In this case, recovery/death rate is defined as the ratio of recovery/death to confirmed cases. It can be seen that the recovery rate has significantly increased from January to first week of March, however, it decreased from second week of March to first week of April 2020. The recovery rate has increased since early April 2020 which indicates that the number of active cases has reduced as many active patients have recovered. On the other hand, the death rate has increased from below 3% in Mid-February 2020 to around 6.0% in June 2020. Fig. 4 presents

Table 2
Comparison of COVID-19 with other epidemics.

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

Table 1
Spread of COVID-19 across the globe.

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
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

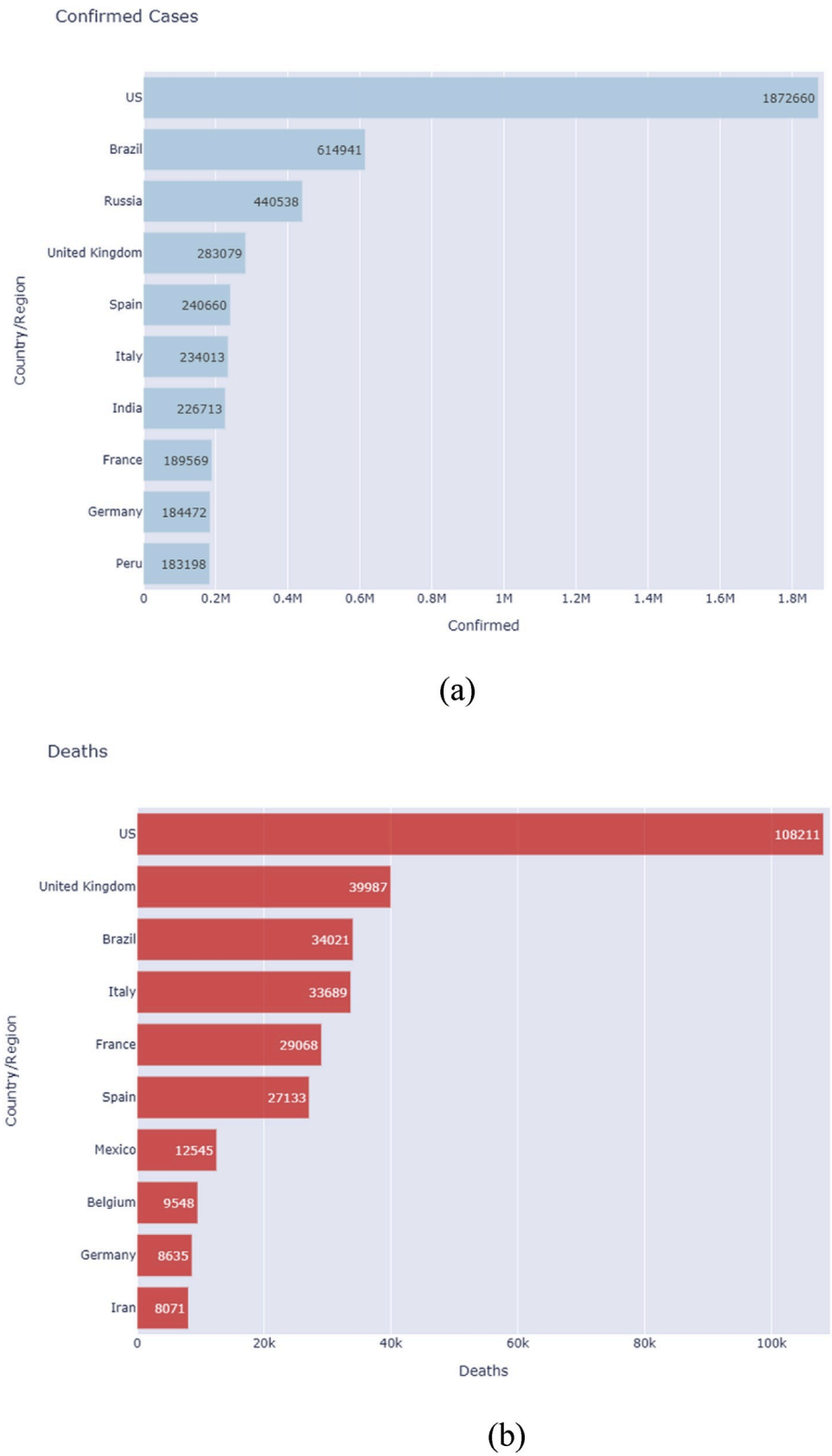


Fig. 1. Horizontal bar chart of cases across countries, (a) for confirmed cases (b) death cases.

graphs showing the death cases, recovered and active cases over time in the world. It can be seen from Fig. 4(a) that the death case has a significantly sharp increment since the first week of March 2020. From Fig. 4(b) it can be seen that the recovery cases have been steady with linear increment over time, but the active cases have experienced sharp increase after first week of March 2020. Fig. 5 illustrates the number of currently active cases of COVID-19. Fig. 5(a) is a pie chart indicating that the percentage of active cases in the world is currently 42.20% of the total confirmed cases. At the same time the recovered cases is 51.90% of the total confirmed cases. Fig. 5(b) is a bubble graph showing the number of active cases in the world in different countries. In the graph, the larger the bubbles, the higher the number of active cases. It

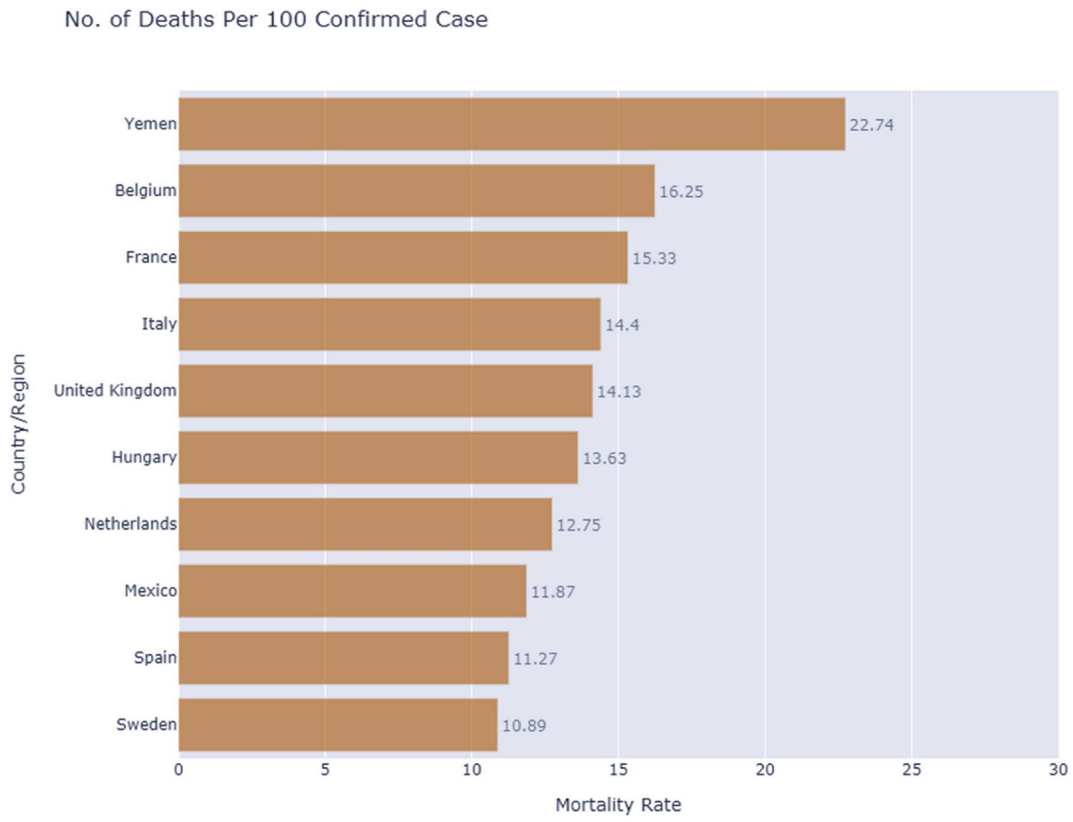


Fig. 2. Death rate in several countries.

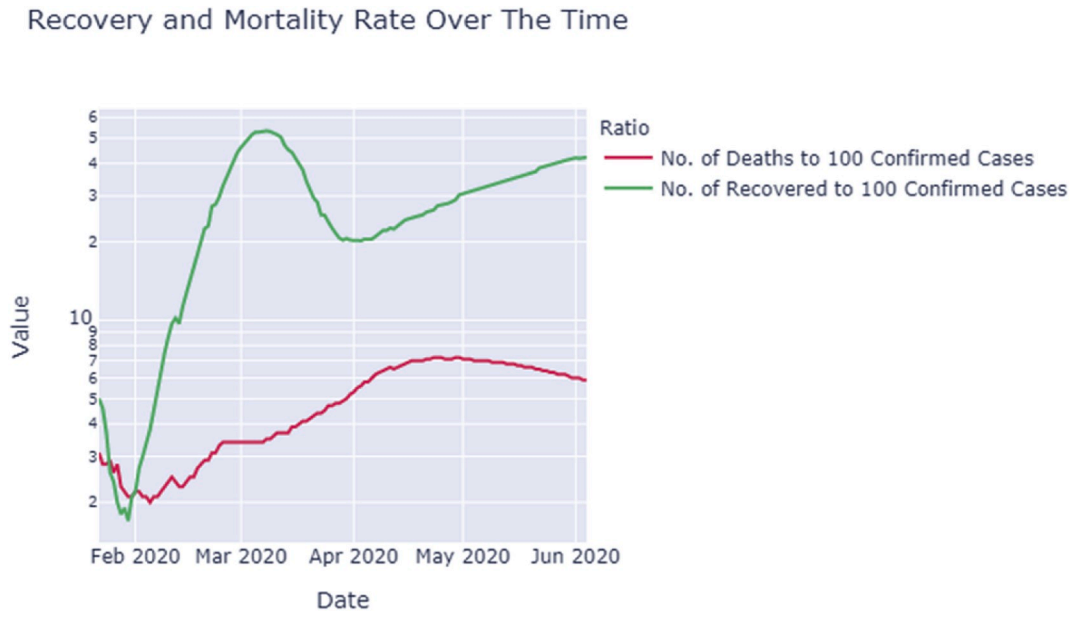
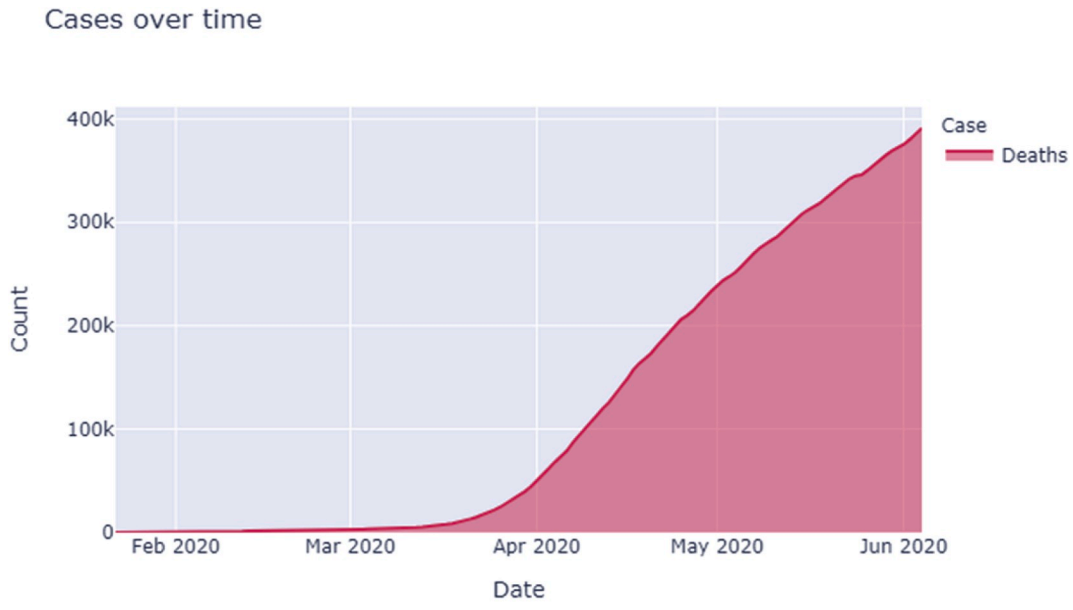


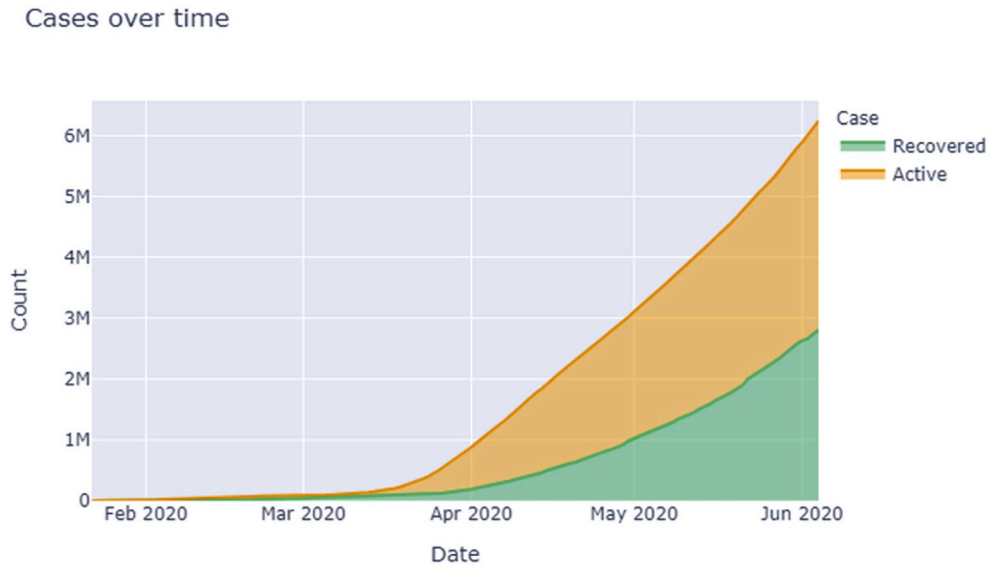
Fig. 3. Death rate and recovery rate worldwide.

can be seen that currently the USA has far more active cases than any other country.

Fig. 6 presents the cumulative number of confirmed, recovered and death cases from 22 January to June 4, 2020, Fig. 6(a) for the case of China and Fig. 6(b) for rest of the world. From Fig. 6(a), it is observed that since early March 2020, the number of confirmed cases and number of deaths in China have become somewhat stable, while the number of recovered patients has increased greatly. This indicates the success of China in overcoming the spread of the virus. On the other hand, Fig. 6(b) reflects that the number of confirmed cases, deaths and recovered cases are increasing outside China since early March 2020. Fig. 7 presents a horizontal bar chart of confirmed/recovered/death cases in Hubei province of China, other provinces of China and the rest of the world. Most of the cases in China are in the Hubei province. The bar chart shows that the number of infections and the number of deaths in Hubei are 68135 and 4512, respectively. On the other hand, number of infections

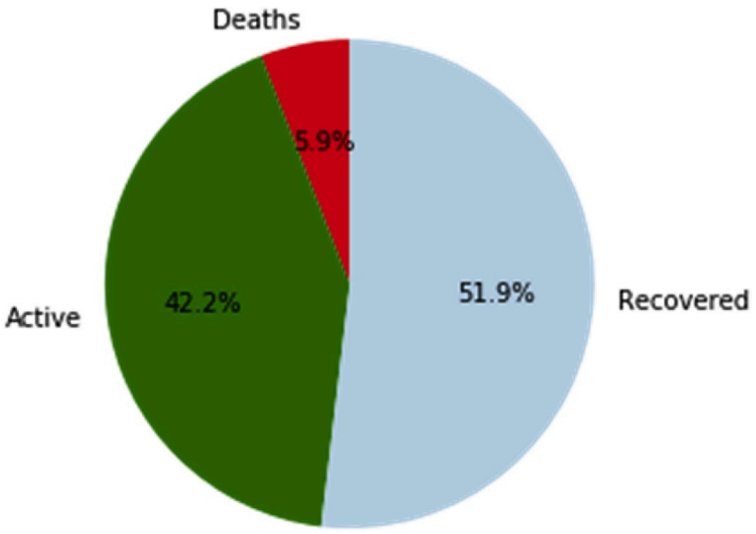


(a)

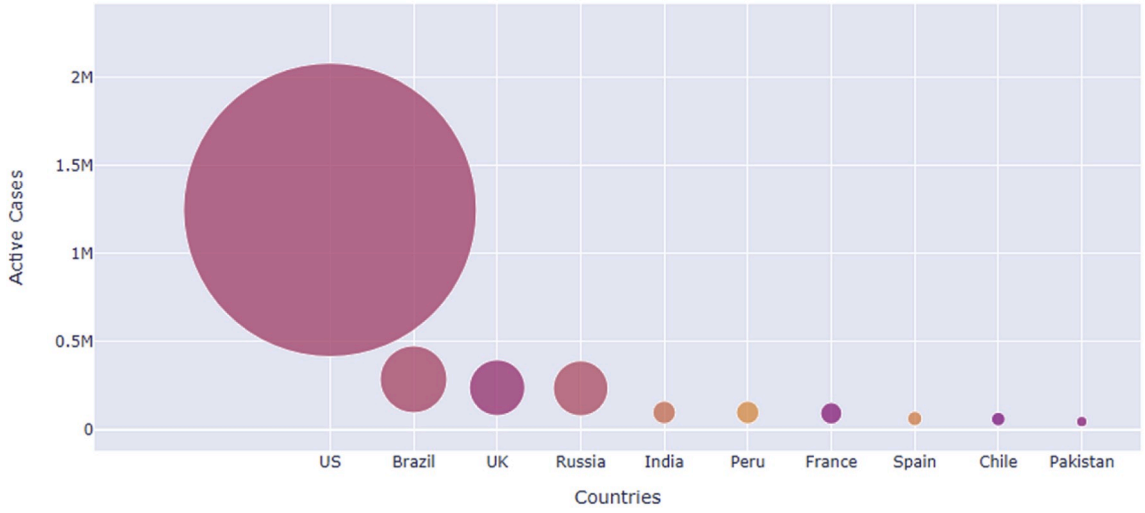


(b)

Fig. 4. Graph showing the cases over time for (a) death cases (b) recovered and active cases.



(a)



(b)

Fig. 5. Illustration of the (a) percentage of active cases in the world via pie chart (b) number of active cases in different countries via bubble graph.

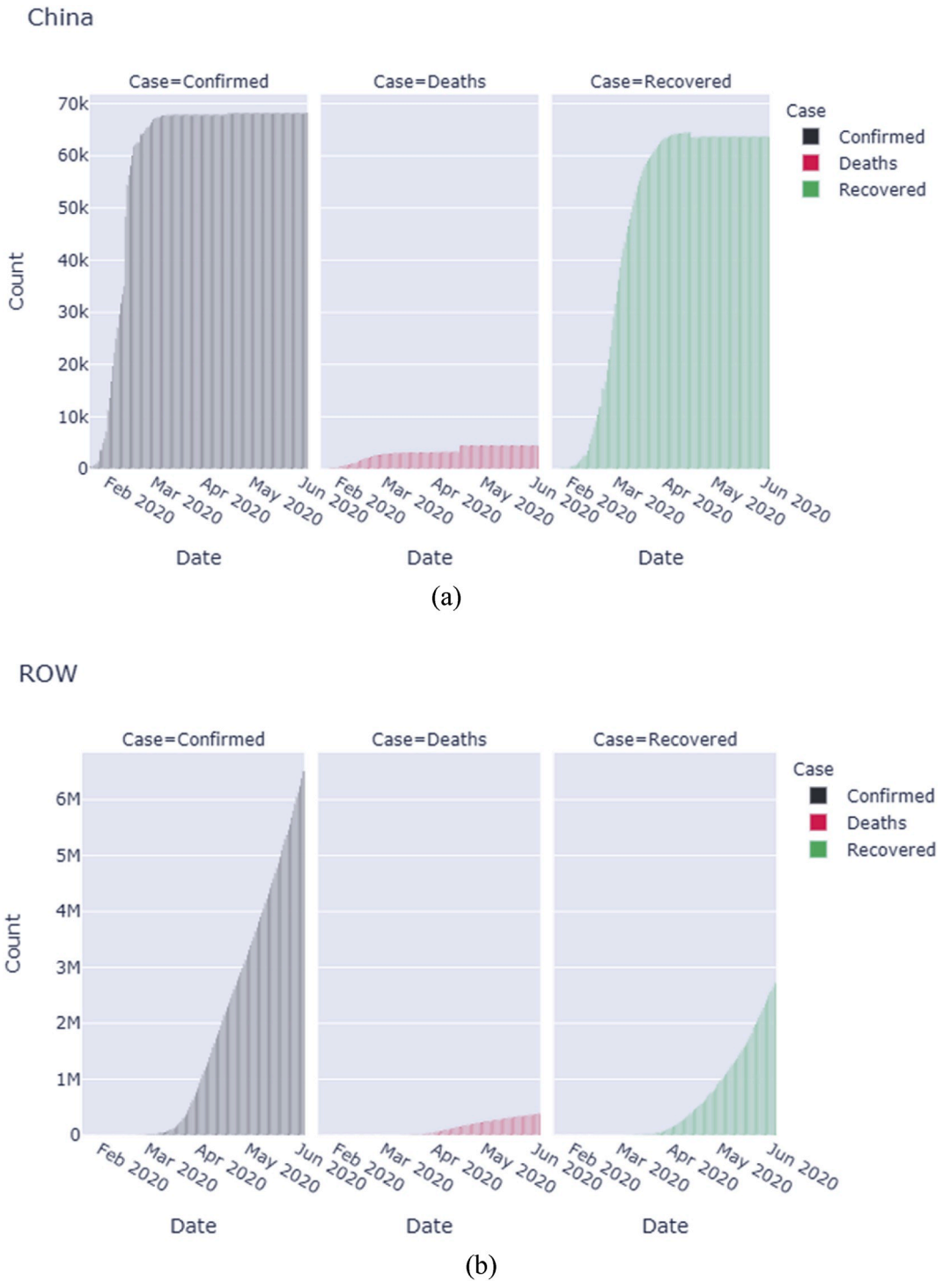


Fig. 6. Graphs showing the confirmed, recovered and death cases over time, (a) for China (b) for rest of the world (ROW).

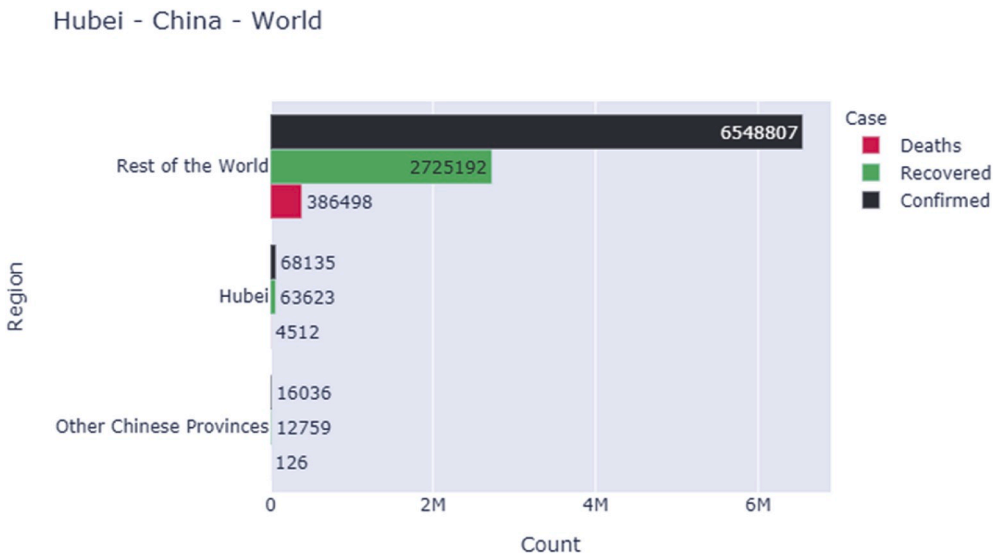


Fig. 7. Horizontal bar chart showing the confirmed, recovered and death cases for Hubei, other Chinese provinces and for the rest of the world.

Table 3
Values of the metric and parameters for polynomial regression of degree 2.

Metric/Parameter	20% Test Size	10% Test Size
R^2	0.8228	0.6718
adjusted R^2	0.986	0.992
a	639.55 (t-stat: 40.006)	603.0274 (t-stat: 50.056)
b	−3.406e+04 (t-stat: 19.266)	−3.073e+04 (t-stat: 20.574)
c	3.21e+05 (t-stat: 7.844)	2.744e+05 (t-stat: 7.073)

and deaths in other Chinese provinces are 16036 and 126, respectively. This indicates that China has successfully managed to stop the spread of COVID-19 from Hubei to other provinces. It can also be seen that number of infections and deaths in the rest of the world are 6,548,807 and 386,498, respectively.

4. Application of data analytics on COVID-19

This Section provides the methodology of data analytics on COVID-19. For this, different machine learning classifiers can be used. These operations can be done using different programming languages including Python, R-studio, MATLAB, etc. For classification, regression or prediction of a particular problem, feature selection methods can be used to find the features that have the highest impact on that problem [28,29]. Then different classifiers and regression models can be applied to obtain the classification or prediction results [28,29].

4.1. Feature selection and classification and regression algorithms

Using the concept of feature selection reported in Ref. [28], the feature selection process is shown in Algorithm 1.

Algorithm 1. Feature Selection Process

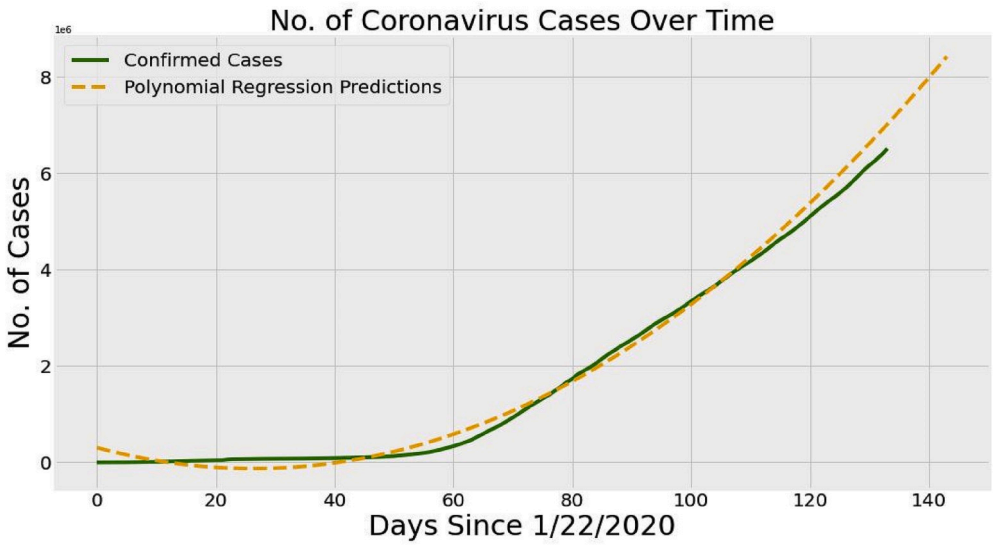


Fig. 8. Number of confirmed cases versus the number of days starting from January 22, 2020. The green solid line is for actual data and the dashed lines are for predicted ones. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Input: A dataset

Output: Selected features with higher ranking

Process:

1. Import several packages and classes
2. Load the data
3. Divide the data and assign the full dataset (except the target) to x
4. Assign the target to y
5. Obtain n number of features
6. Find the ranking or scores of features
7. Find the new variable x_new with only the n best features

Again, using the concept of classification reported in Ref. [28], the process of classification and regression are shown in Algorithm 2:

Algorithm 2. Classification/Regression Process

Input: A dataset

Output: Predicted value, accuracy value

Process:

1. Import several packages and classes for the particular classifier/regressor
2. Load the data
3. Delete any null columns
4. Label the target
5. Divide the data and assign the full dataset (except the target) to x. If feature selection is done, then x is converted to x_new using algorithm 1
6. Assign the target to y
7. Split the dataset into training and testing portions:

(i) for holdout method, randomly split the dataset into a testing group and a training group

(ii) for k fold cross validation method, split the dataset into k equal sized groups where one group is used for testing and the rest for training, and finally average the performance across k predictions

8. Create an instance of the model of a particular classifier/regressor
9. Fit the model instance with training data
10. Predict the target with testing data
11. Evaluate the classification predictions in terms of accuracy, and evaluate the regression predictions in terms of coefficient of determination.

4.2. Regression model for predicting confirmed cases

For the case of predicting a continuous quantity output regression is used, whereas classification is suitable for predicting a discrete class label output. Hence for modelling the number of confirmed cases over time and predicting the future evolution, regression is considered. There

are a number of regressors in machine learning, in this paper, two widely used algorithms named linear regression and polynomial algorithms are considered. Linear regression is popular for simple algorithm and for a having well-known features. Linear regression finds an equation that ensures the smallest difference among the observed sample values and the corresponding fitted sample values. Polynomial regression is a special form of regression where the relationship between independent and dependent variable are modelled as a polynomial of n-th degree in the independent variable. The regression prediction can be evaluated by a number of metrics, but we will consider one of the popular metrics known as coefficient of determination or R-squared value (R^2). This R^2 term is considered as a goodness of fit of a model and measures how good the predictions approximate or fit the actual data samples. In other words, it measures the scatter of the data points around the fitted regression line. A value of $R^2 = 1$ means that the predictions from the regression model perfectly fit to the actual data samples. In this case, dataset [23] is used for regression where the number of days is the independent variable and the confirmed cases is the dependent variable. The regression line is plotted using the steps described in Algorithm 2 and then applying the *plt()* function.

Recently there has been a number of research papers that consider regression for predicting the number of confirmed COVID-19 cases, for example the work in Ref. [30] develops a regression model for the case of India. In the following, regression models are used on the number of confirmed cases in the world available in the dataset [23]. The regression part version of Algorithm 2 reported in Section 4.1 is used to model the trend of the confirmed cases over time. In this paper, linear regression and polynomial regressions methods are taken into consideration. A linear regression can be expressed as

$$Y = aX + b \tag{1}$$

where X and Y are the input and output variables, while a and b are parameters of the regression analysis. A polynomial regression of degree two can be expressed as

$$Y = aX^2 + bX + c \tag{2}$$

where a, b and c are parameters of the regression analysis. Similarly, higher order polynomial regressions can be mathematically described as shown in Ref. [30]. The values of a, b and c are estimated by a type of linear least squares known as ordinary least squares method.

Different test and training portions are considered. Similar to the literature [30], both linear regression and polynomial regression of different orders are compared in terms of coefficient of determination or R-squared value (R^2) which is statistically measure closeness of the data points to the fitted regression line, and adjusted R^2 which adjusts the statistical measure based on the number of input variables. The difference between R^2 and adjusted R^2 is that the value of R^2 does not decrease with the increase of input variables that have low impact on output, while adjusted R^2 penalizes for the addition of input variables that are uncorrelated with the outcome variable. Hence, adjusted R^2 is a better metric than R^2 for practical use case scenarios and for fairly comparing different models. Experiments are performed in Python language to compare the different models in modelling the number of confirmed cases in the world. It is observed from the experiment that the highest value of R^2 and adjusted R^2 are obtained for polynomial regression of degree 2. For a test size of 20% and a training size of 80% of the data samples, polynomial regression of degree 2 exhibits R^2 of 0.8228 and adjusted R^2 of 0.986. An adjusted R^2 of 0.986 means that 98.60% of the variance in the confirmed cases can be explained from this regression model. In other words, this polynomial regression model has 98.60% accurately estimated the confirmed cases in the world up to June 04, 2020. The values of R^2 for polynomial regression of degree 3 to 5, and linear regression are lower than 0.50 and thus unacceptable. Table 3 shows the values of the metric R^2, adjusted R^2, and parameters a, b and c for polynomial regression of degree 2 for test sizes of 20% as well as 10%. It can be seen that the value of R^2 is significantly higher for a test size of 20%, while adjusted R^2 is slightly higher for a test size of 10%. Table 3 presents the values of a, b and c along with their values of t-statistics (denoted as t-stat) which is the ratio of a parameter divided by its standard error. Fig. 8 shows the actual and predicted confirmed cases for polynomial regression of order 2 for test size of 20%. In this figure, x-axis indicates the number of days 0 indicates January 22, 2020 and y-axis represents the number of confirmed cases. From the dashed line which represents the predicted values one can predict the confirmed infection cases for a future day. For example, if the curve is extended for 30 more days, we can estimate what will be number of cases after 1 month assuming the increase in number follows the current pattern. Table 4 presents the predicted confirmed cases for future days with respect to June 4, 2020, the last day of the dataset. The predicted values are calculated using the model for polynomial regression of degree 2 shown in equation (2), and the values of the parameters of the model (test size 20%) given in Table 3.

4.3. Classification for diagnosis of COVID-19 patients

In the following machine learning algorithms are used for automatic diagnosis of COVID-19. Currently there is a lack of acceptable datasets that could be used for prediction of COVID-19 in patients. One dataset is recently uploaded at [31] which contains 5644 samples with 111 attributes provided by Hospital Israelita Albert Einstein, Brazil. This dataset contains anonymized data samples collected by RT-PCR and additional laboratory tests during a visit to the hospital. In a recent preprint paper [32], some results are provided on the diagnosis of COVID-19 patients using a subset of the dataset in Ref. [31].

First of all, preprocessing is done to process and standardize null values and categorical data. The attributes that have more than 99.80% null values in positive samples are dropped as these attributes are unlikely to predict positive cases. The records with mostly null values are also removed. This way the processed dataset has 1091 records and 61 columns. The target feature is converted to make positive samples or virus infected patients as '1' and negative samples or normal persons as '0'. Next, feature selection is performed using the steps shown in Algorithm 1 of Section 4.1. This provides us the ranking of all the features. The most 10 influential features are shown in Table 5. It can be seen that 'Serum Glucose' has the best ranking (rank 1) and thus is the most influential attribute among all the attributes in the dataset. The second and third best attributes are 'Respiratory Syncytial Virus' and 'Influenza A', respectively.

Next, classification algorithms are applied on the processed dataset to classify the COVID-19 positive patients from all suspected patients in the dataset. There are many popular classification algorithms including support vector machine (SVM), k nearest neighbors (kNN), XGBoost, multilayer perceptron (MLP), logistic regression (LR) and decision tree (DT), random forest (RF), majority voting and other ensemble methods. The parameters of these classifiers can also be varied to find the best result for a particular scenario. In this case, XGBoost, MLP, KNN, LR and DT are considered as these provide more promising results than others in our experiments. These well-known classifiers are briefly described in the following [28,29]. XGBoost is a DT-based ensemble classifier having a gradient boosting framework. It is designed for optimal hardware usage. MLP is a form of feed forward artificial neural network and composed of multiple layers of artificial neurons termed as perceptrons. In MLP algorithm, a perceptron uses an activation function which maps the weighted input of each neuron. In KNN, classification is performed by a plurality vote of the k neighbors where the output class is the one most closest to the neighbors. kNN is a non-parametric simple and versatile algorithm. Its implementation is easy, but the functioning gets slower with increase in the number of predictor variables. LR functions according to the concept of probability. LR is a form of classification algorithm where the observations are assigned to discrete classes with a logistic sigmoid function. This algorithm can be of three different types:

Table 4
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 5
Ranking of the features of the dataset.

Name of the feature	Rank
Serum Glucose	1
Respiratory Syncytial Virus	2
Influenza A	3
Influenza B	4
CoronavirusNL63	5
Coronavirus HKU1	6
Parainfluenza 3	7
Adenovirus	8
Parainfluenza 4	9
Coronavirus229E	10

Table 6
Performance Comparison of the classifiers.

Classifier	Precision	Recall	F1 Score	AUC value	Testing Accuracy
MLP	93%	93%	93%	96.70%	93.13%
LR	92%	93%	92%	96.60%	92.12%
XGBoost	92%	92%	92%	96.30%	91.57%
KNN	89%	89%	89%	93.70%	88.91%
DT	87%	87%	87%	94.40%	86.71%

binomial, multinomial and ordinal. In DT classifier, data are continuously split into subsets, then split into even smaller subsets, and this process continues until the data within the subsets become near homogenous. The structure of a DT is similar to flow-chart where each non-leaf nodes represent a test on an attribute, a branch represents the outcome of a test and each terminal node represent a class label. Moreover, DT algorithm is easy to interpret and performs well with large datasets.

For our experiments, the steps in Algorithm 2 of Section 4.1 that are for classifiers and for cross validation are considered. In this case a 20 fold cross validation is performed to split the dataset into training and testing portions. These performance of the classifiers are measures in terms of precision, recall, F1 score, area under the receiver operating characteristic curve (ROC) termed as AUC value, testing accuracy. Table 6 shows the performance comparison of these algorithms. It can be seen that MLP has the highest testing accuracy value of 93.13%, while LR and XGBoost have the testing accuracy values of 92.12% and 91.57%, respectively. This means that the MLP can 93.13% accurately classify a negative (normal) case as normal and a positive (infected) case as infected. In other words, any classification test with MLP for this dataset has 93.13% probability to be correct. Table 6 also shows that MLP has the highest AUC value 96.70%. This means that the classifier can successfully separate the positive and negative classes of the COVID-19 suspected patients with 96.70% probability. In addition, MLP has the highest values of precision, recall and F1 score, all being 93%. Note that the 93% recall value of MLP means that the classifier identifies patients having the disease at 93% accuracy. In other words, if a patient is diagnosed as negative (normal) there is only 7% chance that the test is inaccurate. Both XGBoost and LR have precision, recall and F1 score values equal to or greater than 92%. Hence, among the classifiers considered, MLP, XGBoost and LR are good choices for classifying this specific dataset and thus predict COVID-19 patients reliably.

The performance of the classifiers in our work are better than that of the work in Ref. [32] in terms of recall, AUC value and testing accuracy. For example, the highest value of AUC value in Ref. [32] is 84.70% compared to our highest AUC value of 96.70%, and the highest F1 score in Refs. [32] is 78.10% compared to our highest F1 score of 93%. The difference in the results of [32] and our work are due the fact that the work in Ref. [32] considers only 245 suspected patients and 15 attributes. On the other hand, our work considers 1091 probable patients and 61 attributes after preprocessing. Moreover, the work in Ref. [32] uses the holdout method with 30% testing data and 70% training data. On the other hand, our work considers cross validation method for splitting the testing and training data samples.

5. Possible scope of data analytics

In future modeling the processes of COVID-19 can contribute to improving our understanding in this disease. This models and approaches will provide real time decisions to shape the research on this disease. Data analysis can be useful in fighting this issue. In future, data analytics can be applied in COVID-19 particularly in the following domains:

1. To predict how the virus spreads that is the transmission possibility. To provide information about possible zones that may have clusters of the virus. This will identify possible risky zones and people may avoid these areas.
2. To develop a system that will continuously observe the conditions of a suspected person, and to automatically predict whether he/she has COVID-19 or not. This kind of warning system can help to detect cases of COVID-19 and reduce the number of undetected cases. For this, the symptoms of previous patients have to be stored in a repository. For example, the images of CT scans of normal people and coronavirus affected patients can be stored for training the system, while any new patient's CT scan image can be compared with the training images to find the similarity score. The main features of these patients can be found out from the application of machine learning. With the increase in the number of patient data, the training dataset will increase making the system's prediction work better.
3. To find the basic reproductive number of this virus.
4. To find the effect on pregnancy of different stages on COVID-19 patients and their newborns.
5. To find the effect of this virus on patients having heart disease, lungs problem, kidney disease, liver disease, cancers and other diseases.
6. To find the effect of isolation or quarantine on the spread of this coronavirus. For example, we can analyze the data of China, where the spread of this disease has slowed down after applying strict quarantine measures.
7. To find the effect of travel ban and school closure on the spreading of the virus.
8. To predict the effect of different vaccinations and treatments on COVID-19 patients. This prediction can be done by analyzing datasets of other viral diseases that have vaccination/treatment attributes.

In addition to the above issues, research is required on the possible effect of the coronavirus in developing countries that are highly populated but have poor health care systems. In these countries many people are not accustomed to maintaining personal hygiene appropriately. Lack of proper education and awareness is also a big problem. Lack of proper testing kits means many are not tested. Moreover, there is lack of good

lab facilities, skill personnel resulting in possible error in processing and analysis. So, many patients may remain undetected. Therefore, it is important to find an inexpensive and cheap artificial intelligent system for data driven diagnosis of COVID-19 patients in these countries.

6. Conclusion

COVID-19 has become a pandemic across the globe since its first official appearance in China in December 2019. As of June 04, 2020, the disease is now in 213 countries and territories, with the USA having the highest confirmed cases of the world. Since early March 2020, the situation in China has improved due to a number of steps taken including strict quarantine measures and travel bans. Currently there is no effective treatment for COVID-19, the existing treatments are only for the symptoms. Hence, investigation into the pathogenesis of this infection is important for finding its appropriate treatment. It is shown here that the current increase in the number of confirmed cases over time can be modelled by polynomial regression of degree 2. However, new models have to be developed to reliably predict the future number of confirmed cases considering continuation and relaxation of lockdowns and other aspects. It is also shown in this work, that MLP, XGBoost and LR can reliably classify COVID-19 patients found in a dataset of Hospital Israelita Albert Einstein, Brazil. However, the effectiveness of the classifiers should be validated on a more reliable and larger dataset. With the use of machine learning and with the availability of data on important features, automatic predictions can be made on the possibility of a suspected person to have COVID-19. In future it is believed that different types of coronavirus outbreaks will continue originating from animals for example bats. Hence continuous research is required in the investigation of current and any future coronaviruses.

Ethical statement

All authors have reported that they have no relationships relevant to the contents of this paper to disclose. All the ethical guidelines were followed during the research work.

Declaration of competing interest

The authors declare that they have no competing interests.

Acknowledgement

The authors would like to thank the Johns Hopkins University for open sourcing their dataset.

References

[1] Fan Y, Zhao K, Shi Z, Zhou P. Bat coronaviruses in China. Viruses 2019;11(210).
[2] Lan F-Y, Wei C-F, Hsu Y-T, Christiani DC, Kales SN. Work-related COVID-19 transmission in six Asian countries/areas: a follow-up study. PloS One 2020;15(5): e0233588. https://doi.org/10.1371/journal.pone.0233588.
[3] Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner NW, Ke R. The novel Coronavirus 2019-nCoV is highly contagious and more infectious than initially estimated. arXiv, https://doi.org/10.1101/2020.02.07.20021154; 2020.
[4] Zhao S. Estimating the unreported number of novel Coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data-driven modelling analysis of the early outbreak. J Clin Med 2020;9(2):388.
[5] Nishiura H. The extent of transmission of novel Coronavirus in Wuhan, China, 2020. J Clin Med 2020;9(2):330.
[6] Huang C. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395:497–506.
[7] De Salazar PM, Niehus R, Taylor A, Buckee C, Lipsitch M. Using predicted imports of 2019-nCoV cases to determine locations that may not be identifying all imported cases. medRxiv, https://doi.org/10.1101/2020.02.04.20020495; 2020.
[8] Read JM, Bridgen JRE, Cummings DAT, Ho A, Jewell CP. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. medRxiv, https://doi.org/10.1101/2020.01.23.20018549; 2020.
[9] Li X, Wang W, Zhao X, et al. Transmission dynamics and evolutionary history of 2019-nCoV. J Med Virol 2020;92:501–11.
[10] Liu J, Zheng X, Tong Q, et al. Overlapping and discrete aspects of the pathology and pathogenesis of the emerging human pathogenic coronaviruses SARS-CoV, MERS-CoV, and 2019-nCoV. J Med Virol 2020;92:491–4.
[11] Ji W, Wang W, Zhao X, Zai J, Li X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. J Med Virol 2020;92:433–40.
[12] Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. J Med Virol 2020;92:418–23.
[13] Li G, De Clercq E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). Nat Rev Drug Discov 2020;19(3):149–50. https://doi.org/10.1038/d41573-020-00016-0. PMID 32127666.
[14] Wang Xiaolong. The codon usage pattern of the novel coronavirus is drastically different from those of other pathogenic viruses. 2020. https://doi.org/10.21203/rs.3.rs-18966/v1.
[15] Zhou P. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579(7798):270–3.
[16] Chen H. Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: a retrospective review of medical records. Lancet 2020;395(10226):809–15.
[17] Li G, Fan Y, Lai Y, et al. Coronavirus infections and immune responses. J Med Virol 2020;92:424–32.
[18] Ashour HM, Elkhatib WF, Rahman MM, Elshabrawy HA. Insights into the recent 2019 novel coronavirus (SARS-CoV-2) in light of past human coronavirus outbreaks. Pathogens 2020;9(3):186.
[19] Thanh Le Tung, Andreadakis Zacharias, Kumar Arun, Román Gómez, Raúl, Tollefsen Stig, Saville Melanie, Mayhew Stephen. The COVID-19 vaccine development landscape. Nat Rev Drug Discov 9 April 2020. https://doi.org/10.1038/d41573-020-00073-5. ISSN 1474-1776.
[20] Sanders JM, Monogue ML, Jodlowski TZ, Cutrell JB. Pharmacologic treatments for coronavirus disease 2019 (COVID-19): a review. JAMA. Published online April 13, 2020. doi:10.1001/jama.2020.6019.
[21] Shih H-I, Wu C-J, Tu Y-F, Chi C-Y. Fighting COVID-19: a quick review of diagnoses, therapies, and vaccines. Biomed J 2020. https://doi.org/10.1016/j.bj.2020.05.021. ISSN 2319-4170.
[22] 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;92:632–8.
[23] https://www.kaggle.com/imdevskp/corona-virus-report. [Accessed 4 June 2020].
[24] https://www.kaggle.com/imdevskp/ebola-outbreak-20142016-complete-dataset. [Accessed 4 June 2020].
[25] https://www.kaggle.com/imdevskp/mers-outbreak-dataset-20122019. [Accessed 4 June 2020].
[26] https://www.kaggle.com/kingburrito666/ebola-cases. [Accessed 4 June 2020].
[27] https://www.kaggle.com/imdevskp/sars-outbreak-2003-complete-dataset. [Accessed 4 June 2020].
[28] Raihan-Al-Masud M, Mondal MRH. Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. PloS One 2020;15(2): e0228422.
[29] 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. In: Abraham A, Cherukuri A, Melin P, Gandhi N, editors. Intelligent systems design and applications. 2018, vol. 941. Advances in Intelligent Systems and Computing; 2020. p. 447–57.
[30] Yadav RS. Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India. In: Int. j. inf. tecnol. Springer; 2020. https://doi.org/10.1007/s41870-020-00484-y.
[31] https://www.kaggle.com/einsteindata4u/covid19. [Accessed 4 June 2020].
[32] de Moraes Batista Andre Filipe, Miraglia Joao Luiz, Rizzi Donato Thiago Henrique, Porto Chiavegatto Filho Alexandre Dias. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. medRxiv, https://doi.org/10.1101/2020.04.04.20052092; 2020. 04.04.20052092.