Internet of Things 11 (2020) 100222

Contents lists available at ScienceDirect

Internet of Things

journal homepage: www.elsevier.com/locate/iot

Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing☆

Shreshth Tuli[a,*], Shikhar Tuli[b], Rakesh Tuli[c], Sukhpal Singh Gill[d]

[a] Department of Computer Science and Engineering, Indian Institute of Technology Delhi, India
[b] Department of Electrical Engineering, Indian Institute of Technology Delhi, India
[c] Department of Biotechnology Engineering, University Institute of Engineering and Technology, Panjab University, Chandigarh, India
[d] School of Electronic Engineering and Computer Science (EECS), Queen Mary University of London, UK

ARTICLE INFO

Article history:
Received 6 May 2020
Revised 7 May 2020
Accepted 8 May 2020
Available online 12 May 2020

Keywords:
COVID-19
SARS-CoV-2
Coronavirus
Machine learning
Prediction
Cloud computing

ABSTRACT

The outbreak of COVID-19 Coronavirus, namely SARS-CoV-2, has created a calamitous situation throughout the world. The cumulative incidence of COVID-19 is rapidly increasing day by day. Machine Learning (ML) and Cloud Computing can be deployed very effectively to track the disease, predict growth of the epidemic and design strategies and policies to manage its spread. This study applies an improved mathematical model to analyse and predict the growth of the epidemic. An ML-based improved model has been applied to predict the potential threat of COVID-19 in countries worldwide. We show that using iterative weighting for fitting Generalized Inverse Weibull distribution, a better fit can be obtained to develop a prediction framework. This has been deployed on a cloud computing platform for more accurate and real-time prediction of the growth behavior of the epidemic. A data driven approach with higher accuracy as here can be very useful for a proactive response from the government and citizens. Finally, we propose a set of research opportunities and setup grounds for further practical applications.

1. Introduction

The novel Coronavirus disease (COVID-19) was first reported on 31 December 2019 in the Wuhan, Hubei Province, China. It started spreading rapidly across the world [1]. The cumulative incidence of the causitive virus (SARS-CoV-2) is rapidly increasing and has affected 196 countries and territories with USA, Spain, Italy, U.K. and France being the most affected [2]. World Health Organization (WHO) has declared the coronavirus outbreak a pandemic, while the virus continues to spread [3]. As on 4 May 2020, a total of 3,581,884 confirmed positive cases have been reported leading to 248,558 deaths [2]. The major difference between the pandemic caused by CoV-2 and related viruses, like Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS), is the ability of CoV-2 to spread rapidly through human contact and leave nearly 20% infected subjects as symptom-less carriers [4]. Moreover, various studies reported that the disease caused by CoV-2 is more dangerous for people with weak immune system. The elderly people and patients with life threatening diseases like cancer, diabetes, neurological conditions, coronary heart disease and HIV/AIDS are more vulnerable

☆ **Abbreviations:** ML, Machine Learning; SARS-CoV-2, Severe Acute Respiratory Syndrome Coronavirus 2, COVID-19, Coronavirus disease

* Corresponding author.
E-mail addresses: shreshthtuli@gmail.com (Shreshth Tuli), shikhartuli98@gmail.com (S. Tuli), rakeshtuli@hotmail.com (R. Tuli), s.s.gill@qmul.ac.uk (S.S. Gill).

https://doi.org/10.1016/j.iot.2020.100222
2542-6605/

to severe effects of COVID-19 [5]. In the absence of any curative drug, the only solution is to slow down the spread by exercising "social distancing" to block the chain of spread of the virus. This behavior of CoV-2 requires developing robust mathematical basis for tracking its spread and automation of the tracking tools for on line dynamic decision making.

There is a need for innovative solutions to develop, manage and analyse big data on the growing network of infected subjects, patient details, their community movements, and integrate with clinical trials and, pharmaceutical, genomic and public health data [6]. Multiple sources of data including, text messages, online communications, social media and web articles can be very helpful in analyzing the growth of infection with community behaviour. Wrapping this data with Machine Learning (ML) and Artificial Intelligence (AI), researchers can forecast where and when, the disease is likely to spread, and notify those regions to match the required arrangements. Travel history of infected subjects can be tracked automatically, to study epidemiological correlations with the spread of the disease. Some community transmission based effects have been studied in other works[1]. Infrastructure for the storage and analytics of such huge data for further processing needs to be developed in an efficient and cost-effective manner. This needs to be organized through utilization of cloud computing and AI solutions [7]. Alibaba developed cloud and AI solutions to help China, fight against coronavirus, predict the peak, size and duration of the outbreak, which is claimed to have been implemented with 98% accuracy in real world tests in various regions of China [8]. Different types of pneumonia can be resolved using ML-based CT Image Analytics Solution, which can be helpful to monitor the patients with COVID-19 [9]. Details can be seen in [10]. The development of vaccine for COVID-19 can also be accelerated by analysing the genome sequences and molecular docking, deploying various ML and AI techniques [11].

1.1. Motivation and our contributions

ML [12] can be utilized to handle large data and intelligently predict the spread of the disease. Cloud computing [13] can be used to rapidly enhance the prediction process using high-speed computations [7]. Novel energy-efficient edge systems can be used to procure data, in order to bring down power consumption. In this paper, we present a prediction model deployed using FogBus framework [14] for accurate prediction of the number of COVID-19 cases, the rise and the fall of the number of cases in near future and the date when various countries may expect the pandemic to end. We also provide a detailed comparison with a baseline model and show how catastrophic the effects can be if poorly fitting models are used. We present a prediction scheme based on the ML model, which can be used in remote cloud nodes for real-time prediction allowing governments and citizens to respond proactively. Finally, we summarize this work and present various research directions.

1.2. Article structure

The rest of the paper is organized as follows: Section 2 presents the prediction model and performance comparison. Section 3 provides discussions on the results, biases, implementation and possible deviations in future. Section 4 provides research opportunities and emerging trends. Finally, Section 5, concludes the work and describes the future research opportunities.

2. Prediction model and performance comparison

Machine Learning (ML) and Data Science community are striving hard to improve the forecasts of epidemiological models and analyze the information flowing over Twitter for the development of management strategies, and the assessment of impact of policies to curb its spread. Various datasets in this regard have been openly released to the public. Yet, there is a need to capture, develop and analyse more data as the COVID-19 grows worldwide [15,16].

The novel coronavirus is leaving a deep socio-economic impact globally. Due to the ease of virus transmission, primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes, countries which are densely populated need to be on a higher alert [17]. To gain more insight on how COVID-19 is impacting the world population and to predict the number of COVID-19 cases and dates when the pandemic may be expected to end in various countries, we propose a Machine Learning model that can be run continuously on Cloud Data Centers (CDCs) for accurate spread prediction and proactive development of strategic response by the government and citizens.

2.1. Dataset

The dataset used in this case study is the Our World in Data by Hannah Ritchie[2]. The dataset is updated daily from the World Health Organization (WHO) situation reports[3]. More details about the dataset are available at: https://ourworldindata.org/coronavirus-source-data.

[1] CDC transmission of CoV-2 https://www.cdc.gov/mmwr/volumes/69/wr/mm6915e1.htm
[2] Our World In Data: COVID-19 Dataset; source: https://github.com/owid/covid-19-data/tree/master/public/data/
[3] Situation Reports-WHO; source: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

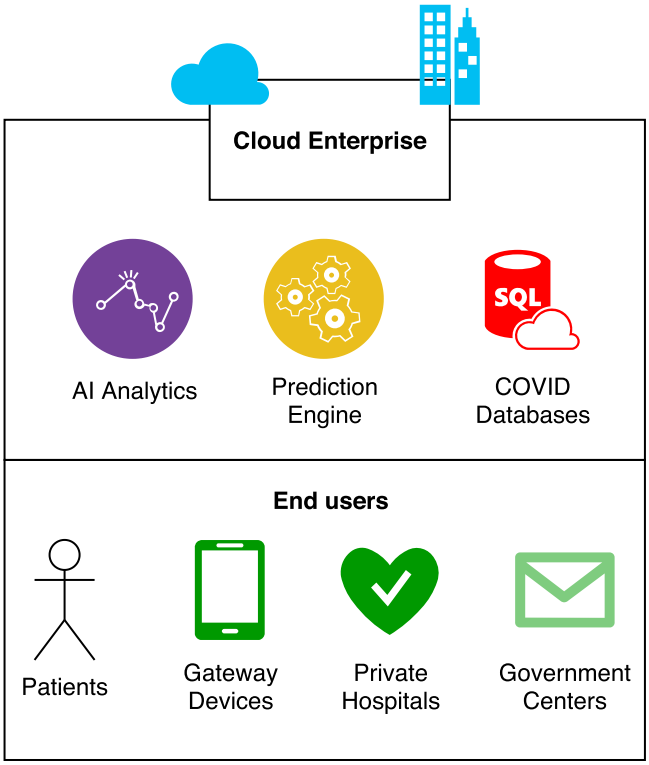


Fig. 1. Proposed Cloud based AI framework for COVID-19 related analytics.

2.2. Cloud framework

The ML models are built to make a good advanced prediction of the number of new cases and the dates when the pandemic might end. To provide fail-safe computation and quick data analysis, we propose a framework to deploy these models on cloud datacenters, as shown in Fig. 1. In a cloud based environment, the government hospitals and private health-centers continuously send their positive patient count. Population density, average and median age, weather conditions, health facilities etc. are also to be integrated for enhancing the accuracy of the predictions. For this case study, we used three instances of single core *Azure B1s* virtual machines with 1-GiB RAM, SSD Storage and 64-bit Microsoft Windows Server 2016[4]. We used the HealthFog [12] framework leveraging the FogBus [14] for deploying multiple analysis tasks in an ensemble learning fashion to predict various metrics, like the number of anticipated facilities to manage patients and the hospitals. We analyzed that the cost of tracking patients on a daily basis, amortized CPU consumption and Cloud execution is 37% and only 1.2 USD per day. As the dataset size increases, computationally more powerful resources would be needed.

2.3. Machine learning model

Many recent works have suggested that the COVID-19 spread follows exponential distribution [18–20]. As per empirical evaluations and previous datasets on SARS-CoV-2 virus pandemic, many sources have shown that data corresponding to new cases with time has large number of outliers and may or may not follow a standard distribution like Gaussian or Exponential [21–24]. In recent study by Data-Driven Innovation Laboratory, Singapore University of Technology and Design (SUTD)[5], the regression curves were drawn using the Susceptible-Infected-Recovered model [25] and Gaussian distribution was deployed to estimate the number of cases with time. However, in the previously reported studies on the earlier version of the virus, namely SARA-CoV-1, the data was reported to follow Generalized Inverse Weibull (GIW) Distribution [26] better than Gaussian as shown in Fig. 2 (details of Robust Weibull fitting follow in the next section). Detailed comparison for SARS-

[4] Azure Cloud VMs: https://azure.microsoft.com/en-au/pricing/calculator/
[5] When Will COVID-19 End, DDI Lab, SUTD: https://ddi.sutd.edu.sg/when-will-covid-19-end

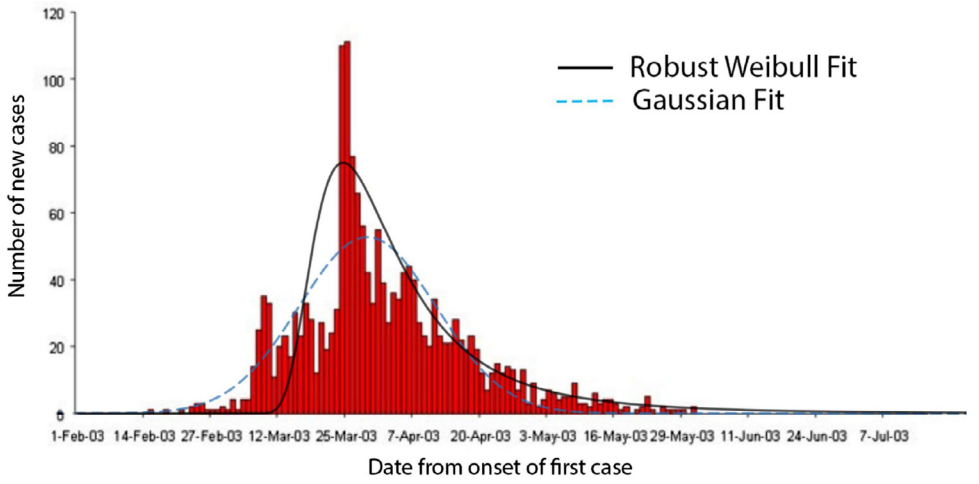


Fig. 2. Fit curves for SARS-CoV-1 pandemic for Hong Kong (SAR), China. Data source: WHO epidemic curves (https://www.who.int/csr/sars/epicurve/epiindex/en/index4.html).

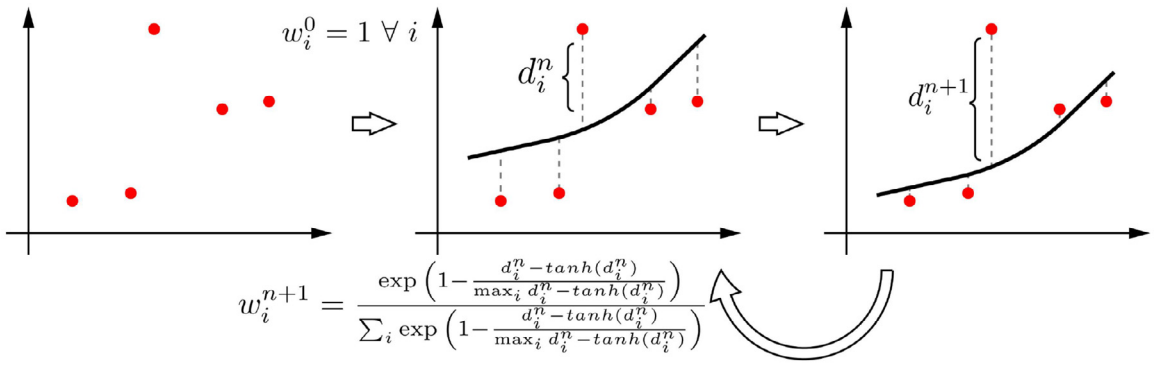


Fig. 3. Iterative weighting technique for robust curve fitting.

CoV-2 has been described in the next section. This fits the following function to the data:

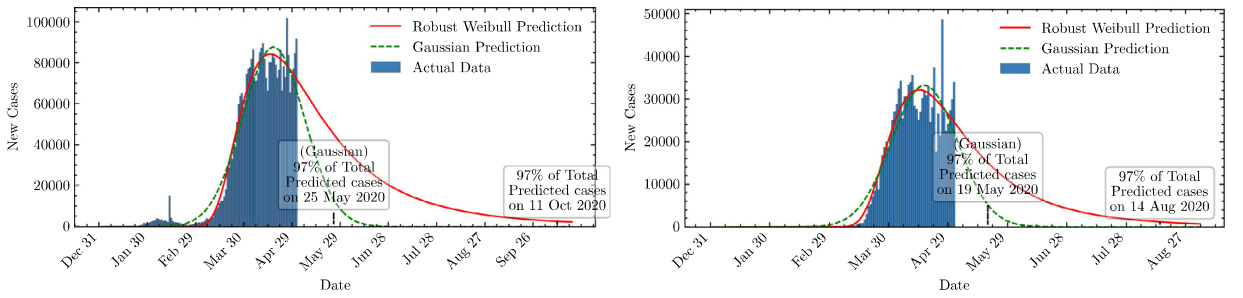
$$f(x) = k \cdot \gamma \cdot \beta \cdot \alpha^{\beta} \cdot x^{-1-\beta} \cdot \exp\left(-\gamma(\frac{\alpha}{x})^{\beta}\right). \tag{1}$$

Here, $f(x)$ denotes the number of cases with x, where $x > 0$ is the time in number of days from the first case, and $\alpha, \beta, \gamma > 0, \in \mathbb{R}$ are parameters of the model. Now, we can find the appropriate values of the parameters α, β and γ to minimize the error between the predicted cases ($y = f(x)$) and the actual cases ($\hat{y}$). This can be done using the popular Machine Learning technique of Levenberg-Marquardt (LM) for curve fitting [27]. However, as various sources have suggested, in initial stages of COVID-19 the data has many outliers and noise. This makes it hard to accurately predict the number of cases. Thus, we propose an iterative weighting strategy and call our fitting technique "Robust Fitting". A diagrammatic representation of the iterative weighting process is shown in Fig. 3.

The main idea is as follows. We maintain weights for all data points (i) in every iteration (n, starting from 0) as w_i^n. First, we fit a curve using the LM technique with weights of all data points as 1, thus $w_i^0 = 1 \ \forall \ i$. Second, we find the weight corresponding to every point for the next iteration (w_i^{n+1})) as:

$$w_i^{n+1} = \frac{\exp\left(1 - \frac{d_i^n - tanh(d_i^n)}{\max_i d_i^n - tanh(d_i^n)}\right)}{\sum_i \exp\left(1 - \frac{d_i^n - tanh(d_i^n)}{\max_i d_i^n - tanh(d_i^n)}\right)}. \tag{2}$$

Simply, in the above equation, we first take *tanhshrink* function defined as $tanhshrink(x) = x - \tanh(x)$ for the distances of all points along y axis from the curve (d_i). This is to have a higher value for points far from the curve and near 0 value for closer points. This, is then standardized by dividing with max value over all points and subtracted from 1 to get a weight



(i) World (ii) United States

Fig. 4. Comparison of predicted dates to reach 97% of the total expected cases by baseline Gaussian and proposed Robust Weibull models. The predicted end date of the pandemic in the baseline model are over-optimistic.

corresponding to each point. This weight is then standardized using *softmax* function so that sum of all weights is 1. The curve is fit again using LM method, now with the new weights w_i^{n+1}. The algorithm converges when the sum total deviation of all weights becomes lower than a threshold value.

2.4. Distribution model selection

To find the best fitting distribution model for the data corresponding to COVID-19, we studied the data on daily new confirmed COVID cases. Five sets of global data on daily new COVID-19 cases were used to fit parameters of different types of distributions. Finally, we identified the best performing 5 distributions. The results are shown in Table 1. We observe that using the iteratively weighted approach, the Inverse Weibull function fits the best to the COVID-19 dataset, as compared to the iterative versions of Gaussian, Beta (4-parameter), Fisher-Tippet (Extreme Value distribution), and Log Normal functions. When applied to the same dataset, Iterative Weibull showed an average MAPE of 12% lower than non-iteratively weighted Weibull. A step-by-step algorithm for iteratively weighted curve fitting using the GIW distribution (called "Robust Weibull") is given in Algorithm 1.

Algorithm 1 Robust Curve Fitting using Iterative weighting.

Require:
x : Input sequence of days from first case
y : Number of cases for each day in x
ϵ : Threshold parameter
procedure ROBUST CURVE FITTING
$w^0 \leftarrow$ Unit vector $[1] \times size(x)$
for iteration n from 0, step 1 **do**
$f \leftarrow$ LM(input = x, target = y, weights = w^n)
$d_i \leftarrow |f(x_i) - y_i| \ \forall \ i$
$$w_i^{n+1} \leftarrow \frac{\exp\left(1-\frac{d_i^n-\tanh(d_i^n)}{\max_i d_i^n-\tanh(d_i^n)}\right)}{\sum_i \exp\left(1-\frac{d_i^n-\tanh(d_i^n)}{\max_i d_i^n-\tanh(d_i^n)}\right)}$$
if $\sum_i |w_i^n - w_i^{n+1}| < \epsilon$ **then**
break
end for
end procedure

2.5. Analysis and interpretation

To compare the proposed "Robust Weibull fitting" model, we use the baseline proposed by Jianxi Luo from SUTD[3]. The comparison metrics include Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE) and Coefficient of determination (R^2). Table 2 shows the model predictions of the spread of the COVID-19 for every major country for which sufficient data was available and model fits had $R^2 > 0.5$ using the proposed model. As shown in the table, the proposed model performs significantly better than the baseline.

Table 1
Preliminary Evaluation of different models. We observe that iterative fitting of Inverse Weibull performs significantly better than iterative fitting of other distributions like Gaussian, Beta (4-parameter), Fisher-Tippet (Extreme Value distribution), and Log Normal. The lowest value of MSE/MAPE and highest values of R^2 among all distributions are shown in bold.

Country	MSE					R^2					MAPE				
	Weibull	Gaussian	Beta 4	Fisher-Tippet	Log Normal	Weibull	Gaussian	Beta 4	Fisher-Tippet	Log Normal	Weibull	Gaussian	Beta 4	Fisher-Tippet	Log Normal
World	**2.41E+07**	3.78E+07	2.99E+07	2.89E+07	2.99E+07	**0.98**	0.97	0.98	0.97	0.97	49.14	49.14	50.39	48.12	**46.19**
India	6.97E+03	7.09E+03	6.89E+03	**6.89E+03**	7.00E+03	0.97	0.97	**0.98**	0.97	0.97	**18.33**	18.33	18.49	21.69	20.69
United States	**8.37E+06**	1.11E+07	8.63E+05	9.47E+06	9.78E+06	**0.95**	0.93	0.94	0.93	0.94	**24.33**	24.33	40.23	71.64	111.63
United Kingdom	**2.00E+05**	2.22E+05	2.12E+05	2.02E+05	2.07E+05	**0.95**	0.95	0.95	0.95	0.95	21.46	21.46	20.43	21.52	**17.42**
Italy	**1.56E+05**	3.38E+05	2.10E+05	2.09E+05	2.35E+05	**0.96**	0.92	0.95	0.95	0.94	**14.98**	14.98	20.00	19.62	170.63

Table 2
Predictions and error comparisons. Country wise predictions using Robust Weibull model and error comparison between Robust Weibull and baseline Gaussian Model. We predict the total number of cases that will be reached, and the last case date i.e. when the model predicts new cases < 1. We also predict the date when the total number will reach 97% of the total expected cases. Such data is critical to prepare the healthcare services in advance. The fit comparison metrics (with proposed model as *W* and baseline model as *G*) show that Mean Square Error (MSE) and the Mean Absolute Percentage Error (MAPE) of the proposed model are lower than baseline for most cases. The coefficient of determination (R^2) is higher for the proposed model for most of the countries. The least MSE/MAPE and highest R^2 values among the two models are shown in bold. Data upto 4 May, 2020 was used to create these results.

Country	Predictions of Robust Weibull Model			Fit comparison metrics					
	Total Cases	Date of last case	97% cases date	MSE (W)	MSE (G)	R^2 (W)	R^2 (G)	MAPE (W)	MAPE (G)
United States	1,937,724	11-Feb-22	14-Aug-20	**9.32E+06**	1.33E+07	**0.95**	0.92	**26.58**	1568.56
Russia	529,687	27-Nov-21	26-Sep-20	**5.50E+04**	5.92E+04	**0.99**	0.98	**24.53**	75.91
India	409,418	29-Oct-24	13-Aug-21	**8.40E+03**	9.11E+03	**0.97**	0.97	**22.38**	80.47
United Kingdom	331,124	31-Jul-21	18-Aug-20	**2.54E+05**	3.19E+05	**0.95**	0.93	**20.14**	211.72
Ukraine	254,087	10-Dec-41	18-Jan-31	**3.31E+04**	3.37E+04	**0.53**	0.52	**1842.80**	2079.70
Italy	253,022	7-Mar-21	27-Jun-20	**1.52E+05**	3.55E+05	**0.96**	0.91	**14.55**	1577.95
Spain	236,737	30-Sep-20	20-Apr-20	**4.67E+05**	6.59E+05	**0.93**	0.90	3682.04	**2917.90**
Turkey	234,218	22-Jun-23	30-Dec-20	2.27E+05	**1.49E+05**	0.92	**0.94**	**30.95**	555.87
Germany	181,369	17-Oct-20	2-May-20	**3.39E+05**	4.50E+05	**0.91**	0.88	1013.65	**582.89**
France	147,795	11-Oct-20	29-May-20	**3.93E+05**	4.14E+05	**0.84**	0.83	**32.36**	134.36
Qatar	143,779	1-Oct-22	18-Mar-21	3.71E+03	**3.46E+03**	**0.93**	0.93	99.14	**90.02**
Canada	139,331	7-Dec-21	31-Oct-20	**2.12E+04**	2.64E+04	**0.95**	0.94	**28.19**	210.56
Belarus	135,375	4-Jun-22	2-Feb-21	**1.28E+04**	1.28E+04	**0.83**	0.83	**1040.39**	1101.48
Iran	126,048	12-Mar-21	15-Jul-20	2.11E+05	**1.88E+05**	0.78	**0.80**	**1847.80**	2313.85
China	84,171	6-Jul-20	27-Mar-20	1.40E+06	**1.28E+06**	0.48	**0.53**	**114.04**	202.88
Sweden	68,671	25-Apr-22	15-Feb-21	**5.33E+03**	5.45E+03	**0.91**	0.91	**20.55**	151.04
Belgium	65,257	19-Nov-20	27-Jun-20	**3.88E+04**	4.10E+04	**0.88**	0.88	**18.76**	134.34
Bangladesh	53,127	19-Apr-22	22-Feb-21	**1.38E+03**	1.60E+03	**0.96**	0.96	**30.89**	118.80
Netherlands	53,057	28-Nov-20	2-Jul-20	**1.07E+04**	1.10E+04	**0.94**	0.94	**16.45**	140.98
United Arab Emirates	46,395	18-Jul-21	5-Nov-20	**3.30E+03**	3.49E+03	**0.91**	0.90	**840.72**	947.02
Portugal	37,302	7-Jun-21	12-Sep-20	**2.62E+04**	3.20E+04	**0.75**	0.70	**41.46**	222.40
Indonesia	35,581	19-Sep-21	20-Dec-20	1.24E+03	**1.22E+03**	**0.93**	0.93	**51.96**	124.06
Poland	35,113	22-Nov-22	1-Aug-21	**3.08E+03**	3.42E+03	**0.87**	0.86	**29.90**	110.45
Switzerland	31,407	26-Jul-20	13-May-20	**1.14E+04**	1.39E+04	**0.92**	0.90	**383.82**	476.28
Bahrain	30,258	21-Mar-23	12-Jan-22	1.05E+03	**1.04E+03**	**0.57**	0.57	**98.32**	102.20
Ireland	27,694	6-Sep-20	12-Jun-20	1.17E+04	**9.49E+03**	0.84	**0.87**	25.91	**21.83**
Singapore	24,088	19-Jul-20	28-May-20	**1.68E+04**	1.69E+04	0.82	**0.82**	**912.31**	1018.38
Dominican Republic	22,193	19-Jun-21	29-Apr-20	**1.79E+03**	1.85E+03	**0.81**	0.81	**304.96**	420.08
Romania	22,102	24-Dec-20	10-Aug-20	**1.98E+03**	2.29E+03	**0.91**	0.90	**16.83**	87.83
Algeria	19,188	6-Feb-22	16-May-21	**3.59E+02**	3.96E+02	**0.86**	0.85	**61.43**	147.61
Israel	18,167	3-Aug-20	26-May-20	**8.81E+03**	1.03E+04	**0.80**	0.77	**37.91**	137.87
Japan	17,614	27-Jul-20	29-May-20	1.12E+04	**1.08E+04**	0.74	**0.75**	**162.70**	202.00
Morocco	16,972	24-May-22	2-Aug-21	1.60E+03	**1.40E+03**	**0.69**	0.73	188.07	**171.78**
Serbia	16,426	24-Jan-21	27-Aug-20	2.49E+03	**2.36E+03**	**0.87**	0.87	**210.28**	229.10
Austria	15,781	9-Jun-20	30-Apr-20	**4.07E+03**	5.33E+03	**0.92**	0.89	**23.08**	34.08
Philippines	14,371	24-Nov-20	2-Apr-20	**5.08E+03**	5.49E+03	**0.65**	0.62	**543.57**	698.02
Denmark	13,282	26-Oct-20	17-Jul-20	1.94E+03	**1.81E+03**	0.81	**0.82**	**18.95**	104.47
Moldova	12,818	6-Feb-22	12-Jun-21	**8.78E+02**	9.57E+02	**0.75**	0.73	**36.51**	68.69
Hungary	11,077	19-Jul-22	22-Nov-21	5.85E+02	**5.66E+02**	0.64	**0.65**	**49.55**	71.00
South Korea	10,780	4-May-20	2-Apr-20	**3.35E+03**	3.88E+03	**0.87**	0.85	**55.81**	68.84
Finland	9158	21-Dec-20	5-Sep-20	**9.09E+02**	9.11E+02	**0.74**	0.74	**125.43**	188.74
Norway	8534	23-Jul-20	19-Apr-20	**1.73E+03**	1.79E+03	**0.80**	0.79	**187.65**	211.88
Czech Republic	8528	14-Jul-20	22-May-20	**1.34E+03**	1.56E+03	**0.85**	0.83	**20.11**	59.31
Malaysia	7080	6-Aug-20	7-Jun-20	**4.88E+02**	5.73E+02	**0.89**	0.87	**30.30**	112.20
Australia	6797	17-May-20	21-Apr-20	**2.54E+03**	2.78E+03	**0.81**	0.79	**31.85**	36.77
Oman	4871	4-Sep-20	23-Apr-20	6.01E+02	**5.98E+02**	**0.66**	0.66	**229.07**	232.19
Iraq	4113	20-Nov-20	20-Apr-20	**4.99E+02**	5.21E+02	**0.47**	0.45	**299.89**	354.05
Luxembourg	3887	29-May-20	2-May-20	**5.15E+02**	6.64E+02	**0.83**	0.79	**49.42**	127.81
Thailand	3044	30-May-20	2-Apr-20	**8.51E+02**	9.01E+02	**0.63**	0.61	**381.04**	399.02
Greece	2944	7-Jul-20	28-Apr-20	3.69E+02	**3.67E+02**	**0.66**	0.66	**137.87**	127.28
Croatia	2275	15-Jun-20	18-May-20	**8.44E+01**	1.04E+02	**0.88**	0.85	**20.64**	45.14
World	6,734,075	29-Jan-24	11-Oct-20	**2.91E+07**	4.92E+07	**0.98**	0.96	**47.53**	63.36

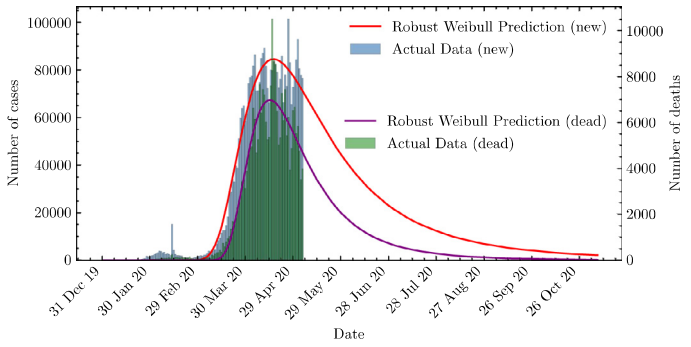


Fig. 5. Number of new cases and deaths for all countries.

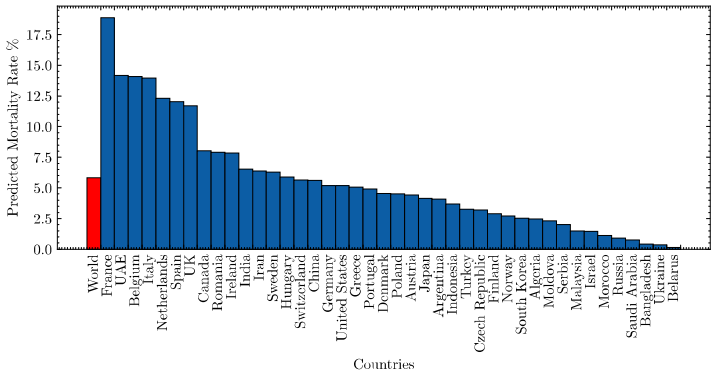


Fig. 6. Predicted Mortality Rate % for a few countries.

As shown in Fig. 4, the predictions of the baseline Gaussian model deployed by SUTD are overoptimistic. Following such models could lead to premature uplifting of the lockdown, causing adverse effect on management of the epidemic. Having better fit models, as proposed here, could help plan a better strategy, based on more accurate predictions and future scenarios.

Fig. 7 shows the total predicted number of cases for all countries across the globe. Here we have neglected those countries where the data is insufficient for making predictions, or the number of days for data is less than 30. As shown in Fig. 4 explained in model section, the fit curve can be used to predict the number of cases that will have to be dealt by the country, assuming the same trend continues. The figure illustrates that the maximum number of total cases will be in the North America region. The number of cases will also be high in the European continent, Russia and eastern Asia, including China, the original epicenter of the disease.

The model was also applied to the data corresponding to the number of deaths with time. Fig. 5 shows curves corresponding to both new cases and deaths across the world. Using the predicted total deaths, the expected mortality rate can be calculated as $100 \times \frac{\text{Predicted total deaths}}{\text{Predicted total cases}}$. The predicted mortality rates of the world and few countries are shown in Fig. 6.

3. Discussions

This section discuss about the biases in data, integration details with tracking systems and trend possibilities in the future.

3.1. Biases in data

The outbreak of SARS-CoV-2 and its corresponding diseases COVID-19 has received diverse responses from different countries. Countries like India, China and Australia have imposed partial to full nation-wide lock-downs leading to mixed repercussions [28–31]. Other countries like Sweden have imposed little to no restrictions. Such factors definitely affect the distribution of cases and hence the curve parameters.

Moreover, there is bias in data due to diverse travel histories and contact demographic histories of people from Wuhan [32]. Reports from health systems in Wuhan are overwhelmed and the only possible way of quantifying spread

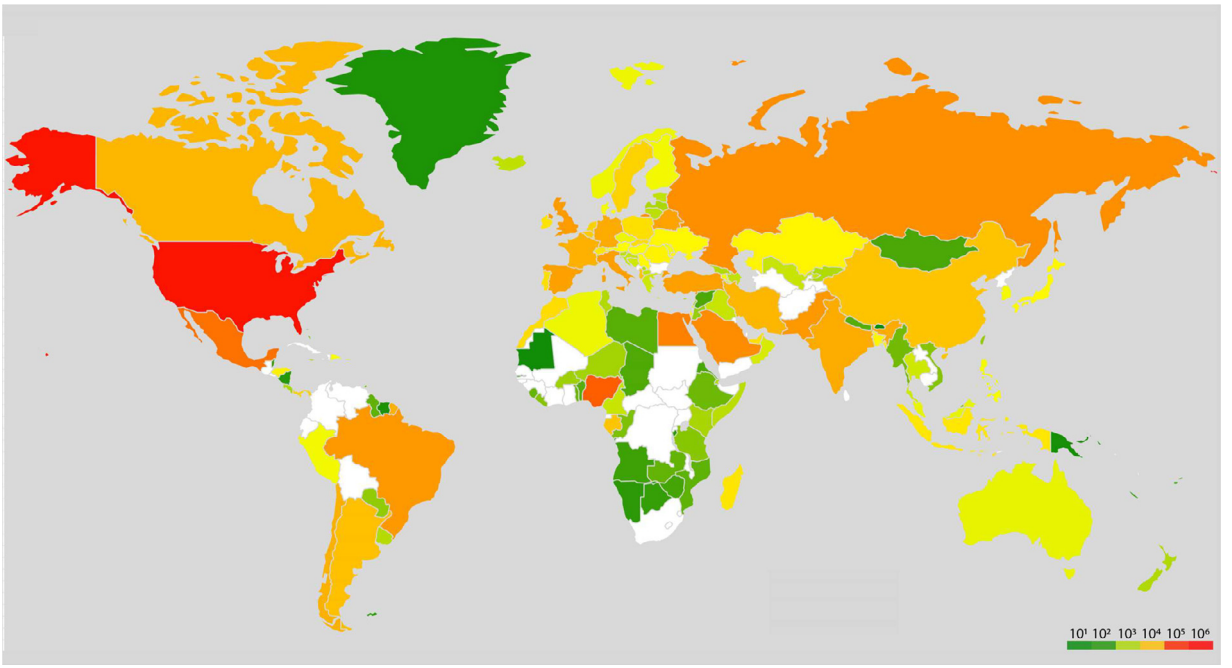


Fig. 7. Global heat-map for total predicted cases for different countries as on May 4, 2020 (countries with insufficient data for prediction are shown in white).

of coronavirus is through cumulative cases in each country [33]. The proposed GIW model is applied separately to each country to fit the model parameters to the distribution of new cases with time. The parameters themselves incorporate the biases from travel histories of citizens and migrants, lock-downs and social distancing measures taken specifically by each country. Having a holistic models that can take these indicators as quantified inputs to generate curve without having any training data would require development and collection of large datasets. Such models can be explored in future.

3.2. Leveraging tracking systems for near-real time predictions

With efficient and up-to-date tracking mechanisms, the spread of the disease can be traced. Once authorities have information on the spread of the virus, relevant decisions can be made including locking down target areas and increasing testing measures in adjacent areas. Only with systematic and planned testing can we mitigate the negative effects of the spread of this disease [34]. Government institutions can utilize cloud services to deploy such frameworks, feeding data from such tracking sensors and predict in near-real time the number of cases in the near future [35]. Further, if we frequently update the dataset and utilize other demographic indicators like population density, temperatures and age distribution in the proposed model, we can make more reliable and accurate predictions for the last expected case. This enables the authorities to lift the lock-down in a phased manner, thus keeping a check on the post-lockdown rise in cases.

3.3. Beyond the lock-downs

Currently, travel and group activities have been restricted world-over. As lock-downs are lifted, the number of new cases and deaths might change significantly from the proposed predicted trends. Other factors like virus mutations [36] would also affect the distribution in future. Hence, continuous work is required to ensure accurate predictions are made and correct measures can be taken.

4. Country-wise predictions

All data uptil 4 May 2020 has been used to generate the prediction results shown in Fig. 8 below:

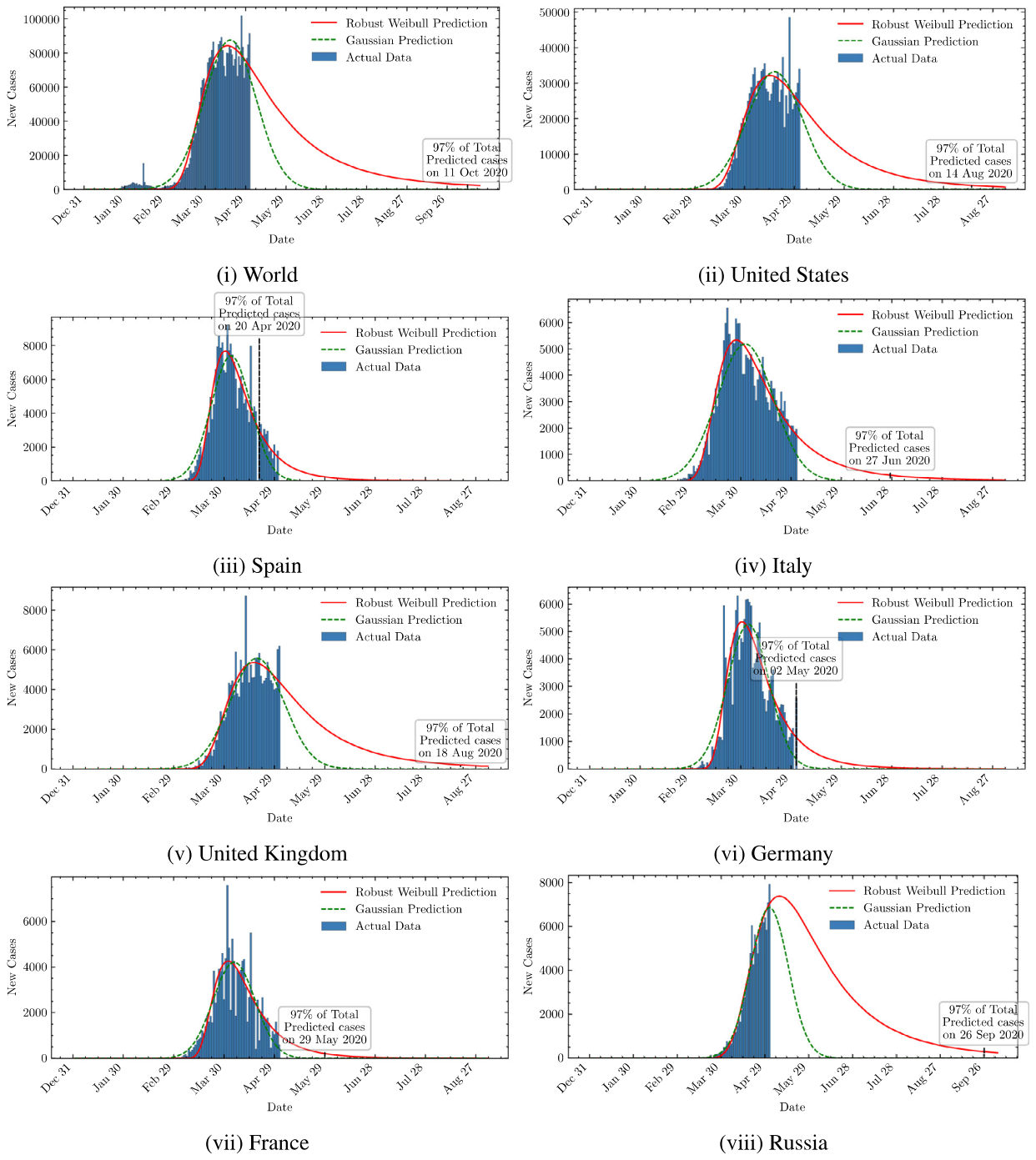
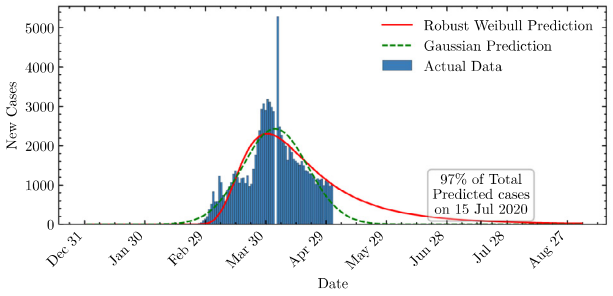
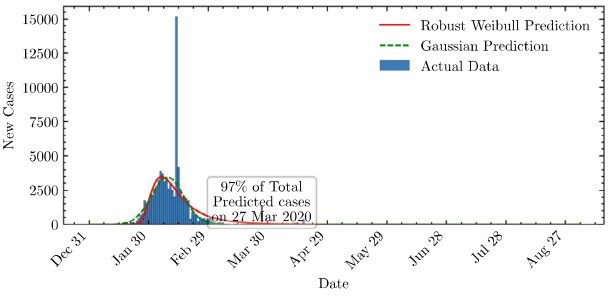


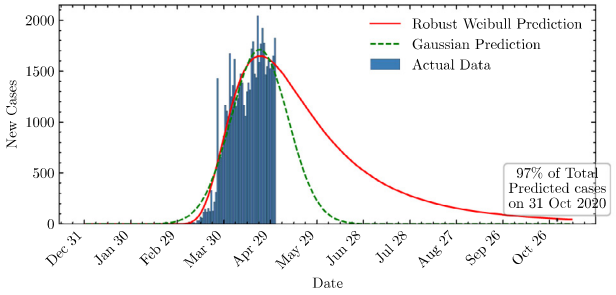
Fig. 8. New cases for different countries (continued).



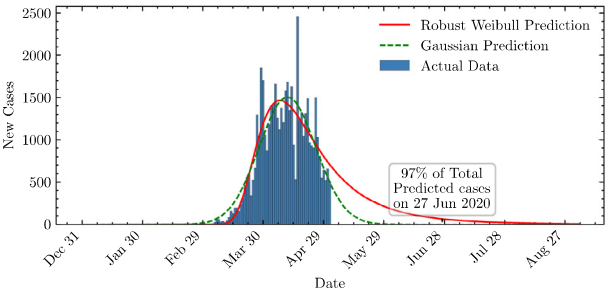
(ix) Iran



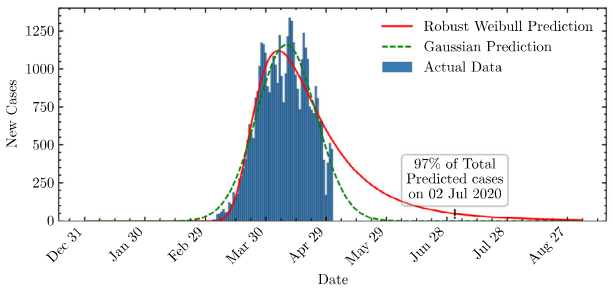
(x) China



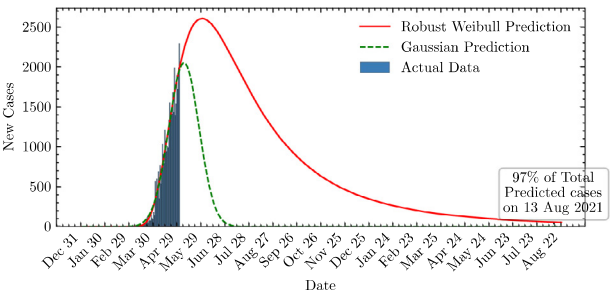
(xi) Canada



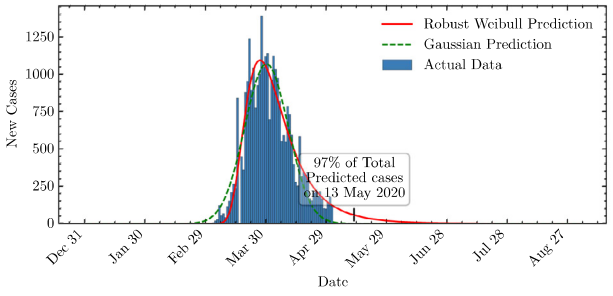
(xii) Belgium



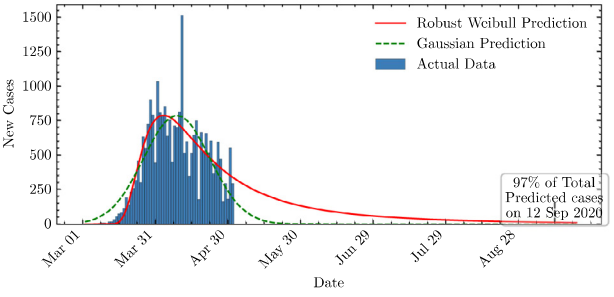
(xiii) Netherlands



(xiv) India

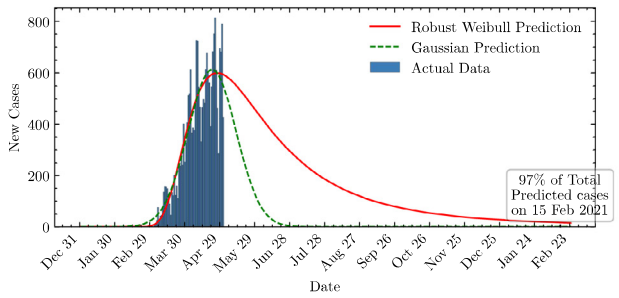


(xv) Switzerland

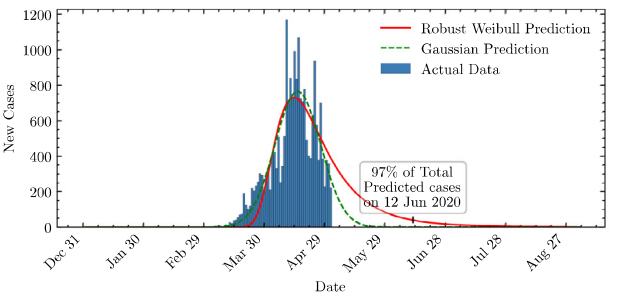


(xvi) Portugal

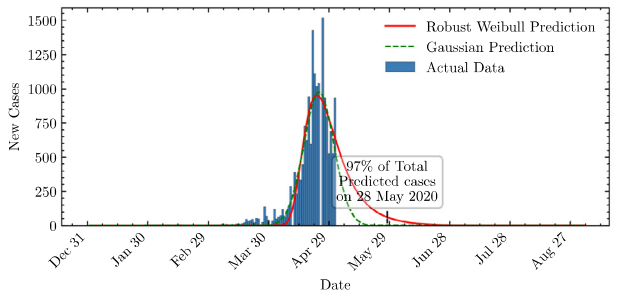
Fig. 8. Continued



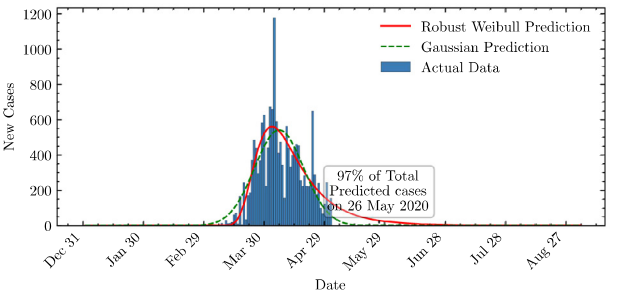
(xvii) Sweden



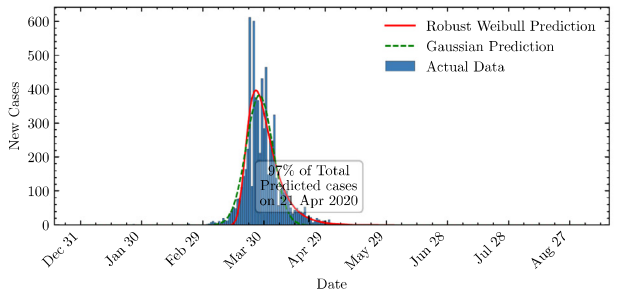
(xviii) Ireland



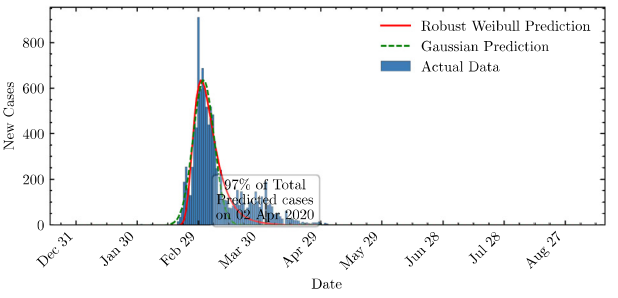
(xix) Singapore



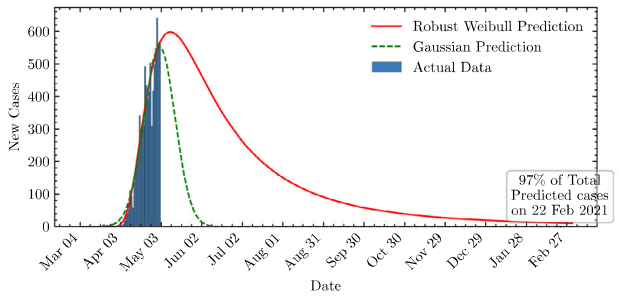
(xx) Israel



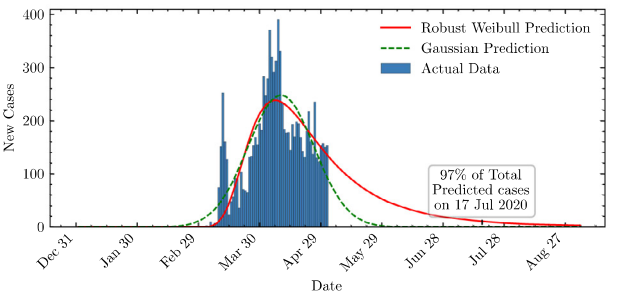
(xxi) Australia



(xxii) South Korea

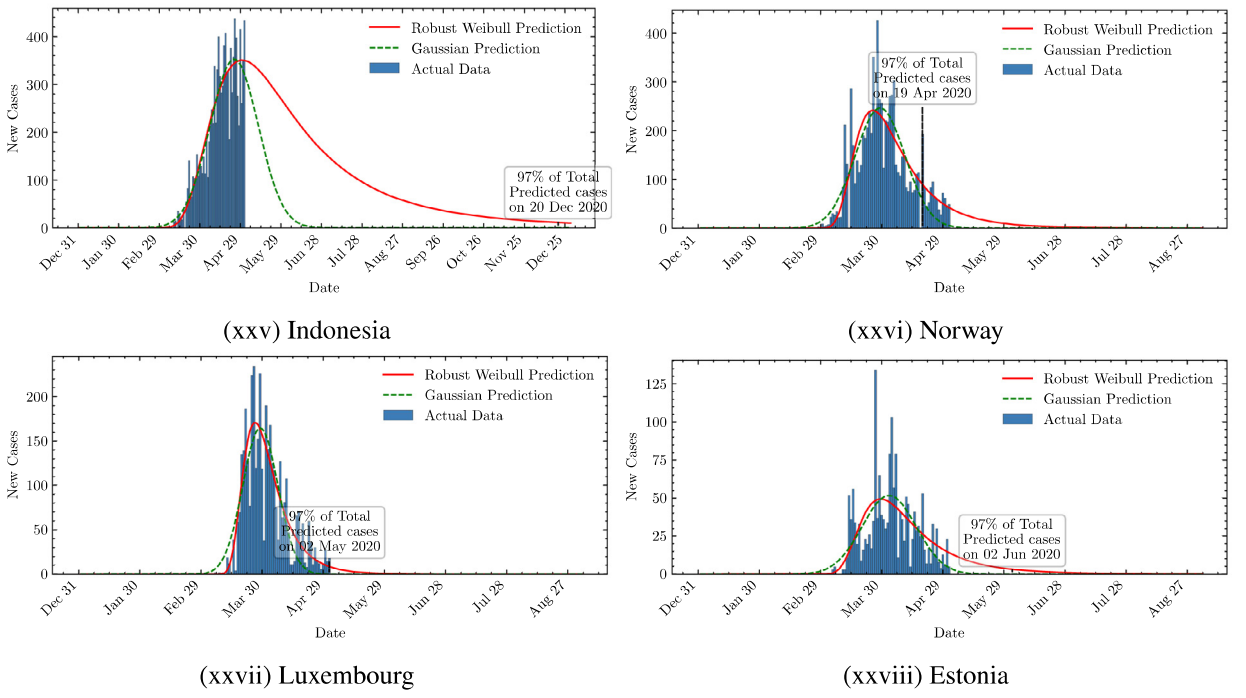


(xxiii) Bangladesh



(xxiv) Denmark

Fig. 8. Continued



(xxv) Indonesia (xxvi) Norway

(xxvii) Luxembourg (xxviii) Estonia

Fig. 8. Continued

5. Research opportunities and emerging trends

The COVID-19 pandemic has opened several new directions of research for the current and future pandemics. The prominent research opportunities are described as follows.

1. **Incorporating other indicators:** Important parameters like population density, distribution of age, individual and community movements, level of healthcare facilities available, strain type and virulence of the virus etc., need to be included in the regression model to further enhance the prediction accuracy.
2. **Integrating with other time series models:** Models like ARIMA [37] can be integrated with Weibull function for further time series analysis and predictions.
3. **Predicting protein structure of CoV-2:** AI can be utilized to predict the structure and function of various proteins associated with CoV-2 and their interaction with the host human proteins and cellular environment. The contribution of various socio-economic variables that determine the vulnerability, spread and progression of the epidemic can be predicted by developing suitable algorithms. This can help efficiently decide resource allocation in large countries with limited healthcare resources.
4. **Analyzing social media data using AI:** We can also explore and analyze social media data for real time collection of epidemiological data related to COVID-19 [15].
5. **Contact-less treatment and drug delivery using Robotics:** AI based Robots can be used to perform contact-less delivery and treat patients remotely to reduce involvement of medical staff with infected people. Further, there have been considerable improvements in air quality across the globe due to COVID-19 enforced lock-downs.
6. **Climate Change:** There have been considerable improvements in air quality across the globe due to COVID19 enforced lock-downs. However, there is a prevailing conjecture of the revenge pollution following these lock-downs [38]. More extensive studies considering age distributions and demographics with other characteristics van be studied as part of future work.
7. **Risk assessment:** The risk of severe disease related with COVID-19 for people with different age can be predicted using AI. Using such algorithms, proactive measures can be taken to prevent virus being spread to sensitive groups of the society.
8. **Real time sensors and visual imaging:** AI based proactive measures can be taken to prevent the spread of the virus to sensitive groups in the society. Real time sensors can be used, for example in traffic camera or surveillance, which track COVID-19 symptoms based on visual imaging and tracking apps, and inform respective hospitals and administrative authorities for punitive action [39]. Tracking needs to cover all stages from ports of entries to public places and hospitals [40].

The research directions and challenges are summarized in Fig. 9.

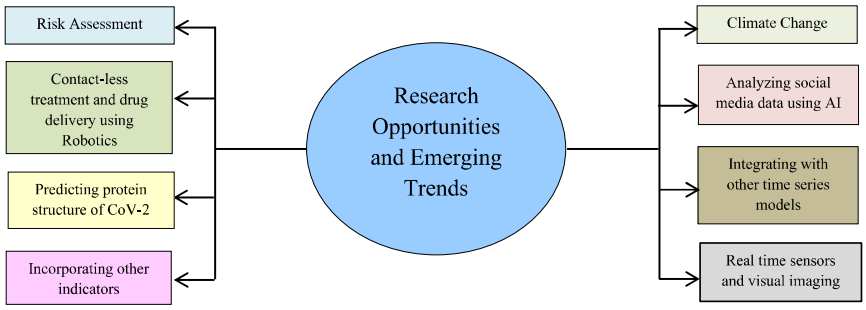


Fig. 9. Future Research Directions and Open Challenges.

6. Summary and conclusions

In this study, we have discussed how improved mathematical modelling, Machine Learning and cloud computing can help to predict the growth of the epidemic proactively. Further, a case study has been presented which shows the severity of the spread of CoV-2 in countries worldwide. Using the proposed Robust Weibull model based on iterative weighting, we show that our model is able to make statistically better predictions than the baseline. The baseline Gaussian model shows an over-optimistic picture of the COVID-19 scenario. A poorly fitting model could lead to a non optimal decision making, leading to worsening of public health situation.

Software Availability

Our prediction model is available online at https://github.com/shreshthtuli/covid-19-prediction. The dataset used for this work is the *Our World Dataset*, available at https://github.com/owid/covid-19-data/tree/master/public/data/. Few interactive graphs can be seen at https://collaboration.coraltele.com/covid/.

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Manmeet Singh (IITM, India) for his valuable comments, useful suggestions and discussion to improve the quality of the paper. We would like to thank the editor, area editor and anonymous reviewers for their valuable comments and suggestions to help and improve our research paper.

References

[1] C. Wang, P.W. Horby, F.G. Hayden, G.F. Gao, A novel coronavirus outbreak of global health concern, The Lancet 395 (10223) (2020) 470–473.
[2] Coronavirus - worldometer, link: https://www.worldometers.info/coronavirus/, [online accessed]
[3] G. Li, E. De Clercq, Therapeutic options for the 2019 novel coronavirus (2019-ncov), 2020.
[4] S. Mallapaty, What the cruise-ship outbreaks reveal about COVID-19, Nature 580 (7801) (2020). 18–18
[5] K. Liu, Y. Chen, R. Lin, K. Han, Clinical features of COVID-19 in elderly patients: a comparison with young and middle-aged patients, J. Infect. (2020).
[6] S. Zhao, Q. Lin, J. Ran, S.S. Musa, G. Yang, W. Wang, Y. Lou, D. Gao, L. Yang, D. He, et al., Preliminary estimation of the basic reproduction number of novel coronavirus (2019-ncov) in china, from 2019 to 2020: adata-driven analysis in the early phase of the outbreak, Int. J. Infect. Dis. 92 (2020) 214–217.
[7] S. Tuli, S. Tuli, G. Wander, P. Wander, S.S. Gill, S. Dustdar, R. Sakellariou, O. Rana, Next generation technologies for smart healthcare: challenges, vision, model, trends and future directions, Internet Technol. Lett.e145.
[8] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, The Lancet 395 (10223) (2020) 497–506.
[9] A. Depeursinge, A.S. Chin, A.N. Leung, D. Terrone, M. Bristow, G. Rosen, D.L. Rubin, Automated classification of usual interstitial pneumonia using regional volumetric texture analysis in high-resolution CT, Invest. Radiol. 50 (4) (2015) 261.
[10] S. Jin, B. Wang, H. Xu, C. Luo, L. Wei, W. Zhao, X. Hou, W. Ma, Z. Xu, Z. Zheng, et al., AI-assisted CT imaging analysis for covid-19 screening: building and deploying a medical ai system in four weeks, medRxiv (2020).
[11] M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics, Nat. Rev. Genet. 16 (6) (2015) 321–332.
[12] S. Tuli, N. Basumatary, S.S. Gill, M. Kahani, R.C. Arya, G.S. Wander, R. Buyya, HealthFog: an ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated iot and fog computing environments, Future Generat. Comput. Syst. 104 (2020) 187–200.
[13] S.S. Gill, S. Tuli, M. Xu, I. Singh, K.V. Singh, D. Lindsay, S. Tuli, D. Smirnova, M. Singh, U. Jain, et al., Transformative effects of IoT, blockchain and artificial intelligence on cloud computing: evolution, vision, trends and open challenges, Internet Things 8 (2019) 100118.
[14] S. Tuli, R. Mahmud, S. Tuli, R. Buyya, Fogbus: a blockchain-based lightweight framework for edge and fog computing, J. Syst. Softw. (2019).
[15] E. Chen, K. Lerman, E. Ferrara, COVID-19: The First Public Coronavirus Twitter Dataset, arXiv preprint: 1603.07252(2020).
[16] J.P. Cohen, P. Morrison, L. Dao, COVID-19 image data collection, arXiv preprint: 2003.11597(2020).
[17] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, M. Wang, Presumed asymptomatic carrier transmission of COVID-19, JAMA (2020).

[18] B.F. Maier, D. Brockmann, Effective containment explains sub-exponential growth in confirmed cases of recent COVID-19 outbreak in mainland china, arXiv preprint: 2002.07572(2020).
[19] Y. Li, M. Liang, X. Yin, X. Liu, M. Hao, Z. Hu, Y. Wang, L. Jin, Covid-19 epidemic outside china: 34 founders and exponential growth, medRxiv (2020).
[20] M. Raygoza, Covid-19, exponential growth, and the power of showing up in social solidarity: The math behind the virus(2020).
[21] Y. Bai, Z. Jin, Prediction of SARS epidemic by BP neural networks with online prediction strategy, Chaos Soliton. Fractal. 26 (2) (2005) 559–569.
[22] Y.-H. Hsieh, J.-Y. Lee, H.-L. Chang, Sars epidemiology modeling, Emerging Infect. Dis. 10 (6) (2004) 1165.
[23] D. Lai, Monitoring the SARS epidemic in china: a time series analysis, J. Data Sci. 3 (3) (2005) 279–293.
[24] W. Wang, S. Ruan, Simulating the SARS outbreak in beijing with limited data, J. Theor. Biol. 227 (3) (2004) 369–379.
[25] D. Smith, L. Moore, The SIR model for spread of disease: the differential equation model, Loci.(originally Convergence.) Https://Www. Maa. Org/Press/Periodicals/Loci/Joma/The-Sir-Model-for-Spread-of-Disease-the-Differential-Equation-Model (2004).
[26] F.R.S. De Gusmao, E.M.M. Ortega, G.M. Cordeiro, The generalized inverse weibull distribution, Statist. Papers 52 (3) (2011) 591–619.
[27] J.J. Moré, The levenberg-marquardt algorithm: implementation and theory, in: Numerical analysis, Springer, 1978, pp. 105–116.
[28] P. Pulla, Covid-19: India imposes lockdown for 21 days and cases rise, 2020,
[29] O. Analytica, Japan'S partial COVID-19 lockdown may be insufficient, Emerald Expert Briefings(oxan-es).
[30] J. Thornton, Covid-19: A&e visits in england fall by 25% in week after lockdown, 2020.
[31] Y. Zhang, B. Jiang, J. Yuan, Y. Tao, The impact of social distancing and epicenter lockdown on the COVID-19 epidemic in mainland China: a data-driven SEIQR model study, medRxiv (2020).
[32] R. Niehus, P. Martinez de Salazar Munoz, A. Taylor, M. Lipsitch, Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depend on reported cases in international travelers (2020).
[33] J.T. Wu, K. Leung, G.M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study, The Lancet 395 (10225) (2020) 689–697.
[34] Covid-19 science: Why testing is so important | american heart association, (https://www.heart.org/en/news/2020/04/02/covid-19-science-why-testing-is-so-important). (Accessed on 05/07/2020).
[35] R. Mancini, S. Tuli, T. Cucinotta, R. Buyya, iGateLink: A Gateway library for linking IoT, edge, fog and cloud computing environments, Proc. Int. Conf. Intell. Cloud Comput. (2019).
[36] X. Tang, C. Wu, X. Li, Y. Song, X. Yao, X. Wu, Y. Duan, H. Zhang, Y. Wang, Z. Qian, et al., On the origin and continuing evolution of SARS-cov-2, Natl. Sci. Rev. (2020).
[37] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, M. Ciccozzi, Application of the ARIMA model on the COVID-2019 epidemic dataset, Data Brief (2020) 105340.
[38] There's an unlikely beneficiary of coronavirus: the planets, link: https://edition.cnn.com/2020/03/16/asia/china-pollution-coronavirus-hnk-intl/index.html, [online accessed].
[39] H.S. Maghdid, K.Z. Ghafoor, A.S. Sadiq, K. Curran, K. Rabie, A Novel AI-enabled Framework to Diagnose Coronavirus COVID 19 using Smartphone Embedded Sensors: Design Study, arXiv preprint: 2003.07434(2020).
[40] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, et al., A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19), medRxiv (2020).

Shreshth Tuli is an undergraduate student at the Department of Computer Science and Engineering at Indian Institute of Technology - Delhi, India. He is also a co-founder of Qubit Inc. company which works on providing next generation solutions for industrial problems. He is a national level Kishore Vaigyanic Protsahan Yojana (KVPY) scholarship holder for excellence in science and innovation. He has worked as a visiting researcher at the Cloud Computing and Distributed Systems (CLOUDS) Laboratory, Department of Computing and Information Systems, the University of Melbourne, Australia. His research interests include Internet of Things (IoT), Fog Computing, Network Design, and Artificial Intelligence.

Shikhar Tuli is an undergraduate student at the Department of Electrical Engineering at Indian Institute of Technology - Delhi, India. He is the founder and CEO of Qubit Inc. He has worked remotely with the Cloud Computing and Distributed Systems (CLOUDS) Laboratory, Department of Computing and Information Systems, the University of Melbourne, Australia in the realization of the FogBus framework. He has also worked at the Embedded Systems Laboratory, EPFL, Switzerland in the design of low-power and physics-optimized Edge devices made from emerging Non-Volatile Memories. His research interests include Internet of Things (IoT), In-memory and Neuromorphic computing architectures and Nanoelectronics. He specializes in designing novel hardware technologies that are valuable to both industry and academia.

Rakesh Tuli is Senior Research Advisor and J C Bose National Fellow, UIET, Panjab University, Chandigarh. Before this, he was Executive Director, National Agri-Food Biotech Institute, Mohali; and Director, National Botanical Research Institute, Lucknow. He has many publications in reputed journals and conferences including Nature Biotechnology. His research includes Genomics and Transgenic Approaches to Improving Plants for Agricultural and Health/Medicinal Applications.

Sukhpal Singh Gill is a Lecturer (Assistant Professor) in Cloud Computing at School of Electronic Engineering and Computer Science, Queen Mary University of London, UK. Prior to this, Dr. Gill has held positions as a Research Associate at the School of Computing and Communications, Lancaster University, UK and also as a Postdoctoral Research Fellow at CLOUDS Laboratory, The University of Melbourne, Australia. His research interests include Cloud Computing, Fog Computing, Software Engineering, Internet of Things and Healthcare. For further information, please visit http://www.ssgill.me.