A statistical and deep learning-based daily infected count prediction system for the coronavirus pandemic.

Vruddhi Shah Ankita Shelke Mamta Parab Jainam Shah
Ninad Mehendale

Received: date / Accepted: date

Abstract

We present new data analytics and simulation results that can help governments to plan their future actions and also help medical services to be prepared for the future in advance. We do not claim that we can predict with 100 % accuracy, but we have found during self-evaluation that accuracy with our predictions on new daily infected cases of Coronavirus, is within 6 % of the predicted values. As predicted by most researchers and simulations we did not expect the curve to be bell-shaped, but rather to be a combination of an exponential curve (phase 1) then a flat curve (phase 2), and finally linear decay (phase 3). Hence, we expect that the results mentioned in the manuscript help people in deciding their further strategies.

keywords

Covid-19 Simulations. Coronavirus Daily count

1 Introduction

The Chinese authorities alerted the world in December 2019, that a virus outspread had occurred in their community. Eventually, it spread to the other countries in the coming months, with cases multiplying within a span of a few days. This virus was Severe Acute Respiratory Syndrome-Related Coronavirus that causes Covid-19 or "Coronavirus". A virus is a small collection of genetic material bordered by a protein coat. It cannot replicate on its own. How long, a virus can remain on the surfaces is not certain. Coronavirus mostly spreads through these surfaces. It mainly spreads by the droplets, when people sneeze or cough or even when a person comes in contact with an infected one, and then touches his/her face by wiping his eyes or nose. Then it begins to spread very deep into the person's body. The targets of the virus are the lungs, the spleen, or the intestines.

The targets of the virus are the lungs, the spleen, or the intestines. out of which, the most noticeable effect of this virus can be seen in the lungs. A large number of epithelial cells are present at the border of the lungs. These epithelial cells also line the organs of our body and mucosa. Coronavirus injects its genetic material inside the epithelial cell by connecting to a particular receptor on the membrane of the victim's cell. The cell executes the new commands that are given by this genetic material, which are to replicate and reconstruct new coronaviruses. This adds up further by generating more and more duplicates of the original virus till a critical point is reached where the cell receives a final call which is to self-destruct.

And now the newly released coronaviruses starts attacking the other cells. An exponential increase in the number of infected cells is seen and within 10 days. A huge number of cells, about millions, are infected and the lungs would be housing more than a billion of these viruses. This virus has so far not caused any damage to the body, but will now show

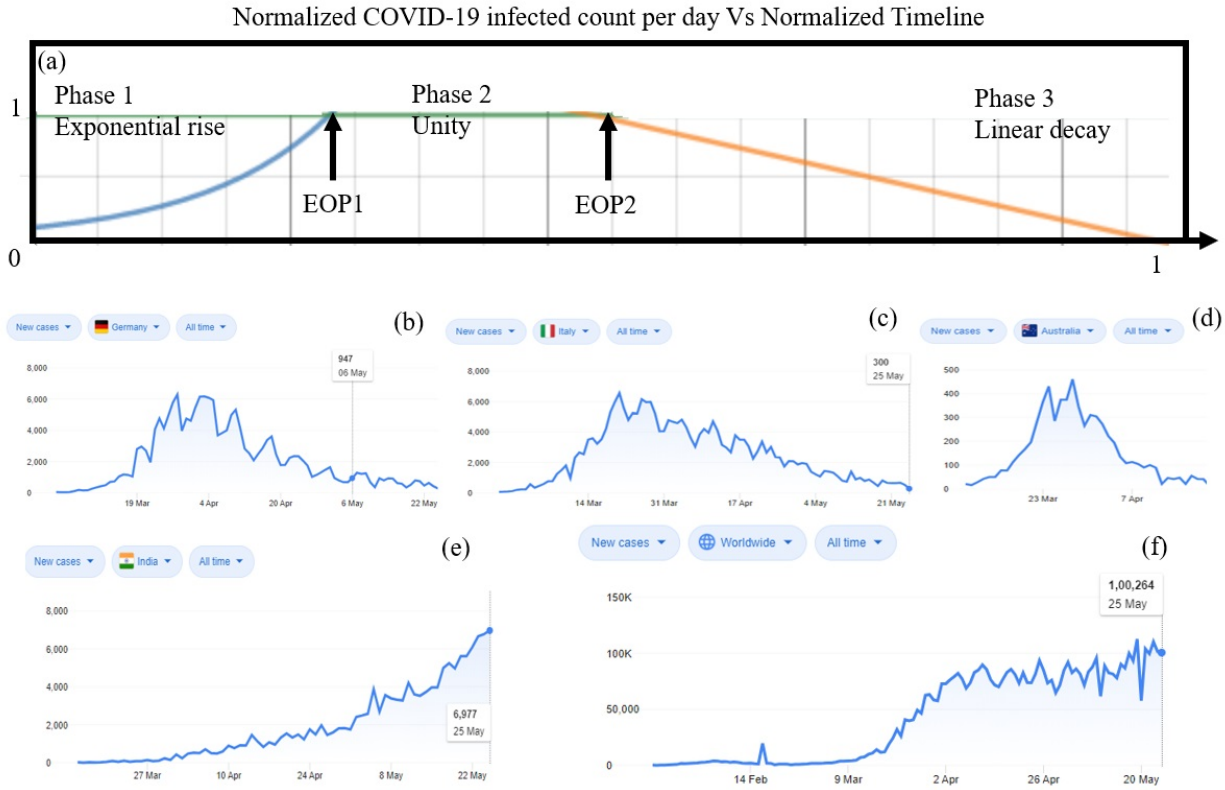

Figure 1: (a) The normalized curve of COVID-19 infected new patient's daily count versus normalized timeline. Three major counties i.e. Germany (b), Italy (c), and Australia's (d) data reported as countries already reached phase 3. India's (e) data reported as phase 1 country. The whole world (f) has reached phase 2.

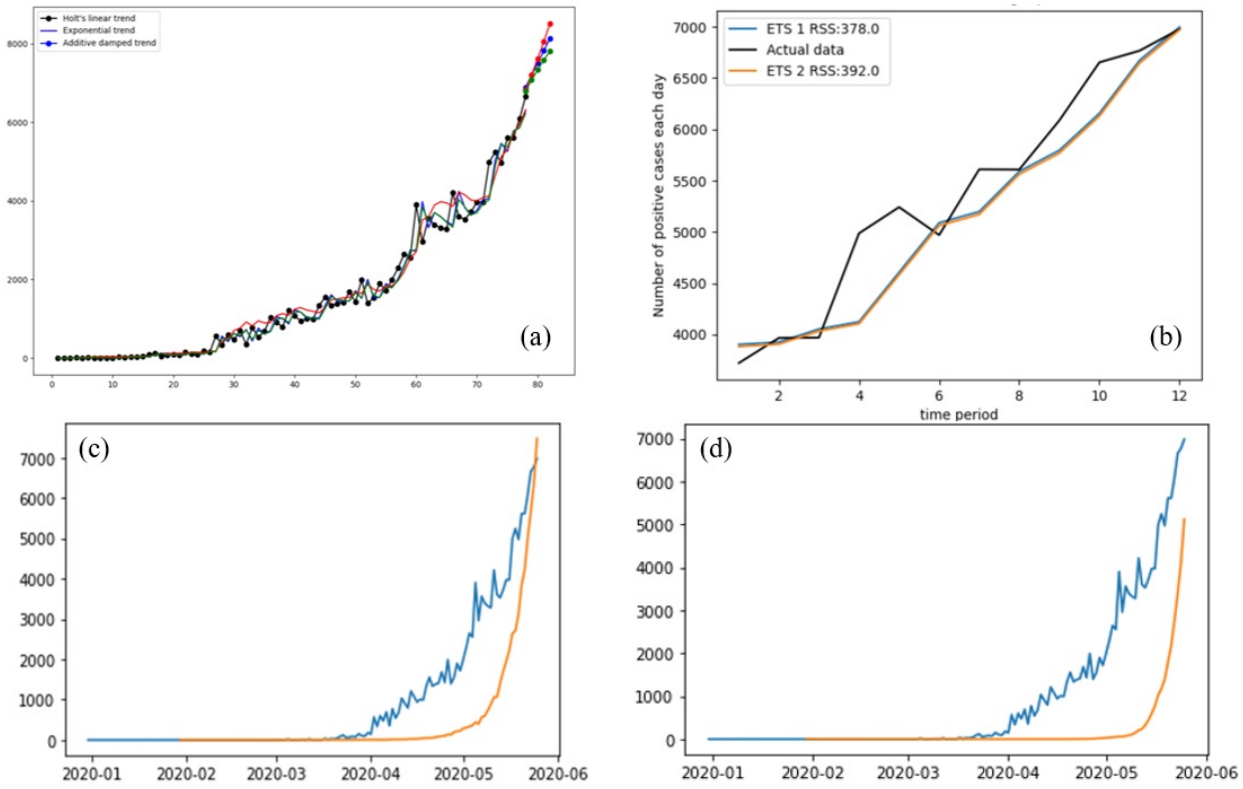

Figure 2: (a) The output of the Holt-Winter algorithm. Line in black shows actual data, line in red shows exponential trend, line in blue shows holt's linear trend, line in green shows additive damped (b) output of ETS algorithm. Line in black shows actual data, line in red shows exponential trend and line in green additive damped. The output of the Residual Sum of Squares (RSS) is 378 people (c) the output of AR model. Line in blue shows actual data and line in orange shows the predicted results. The RMSE value for the model is 1345 (d) the output of MA model. Line in blue shows actual data and line in orange shows the predicted results. The RMSE value for the model is 1875

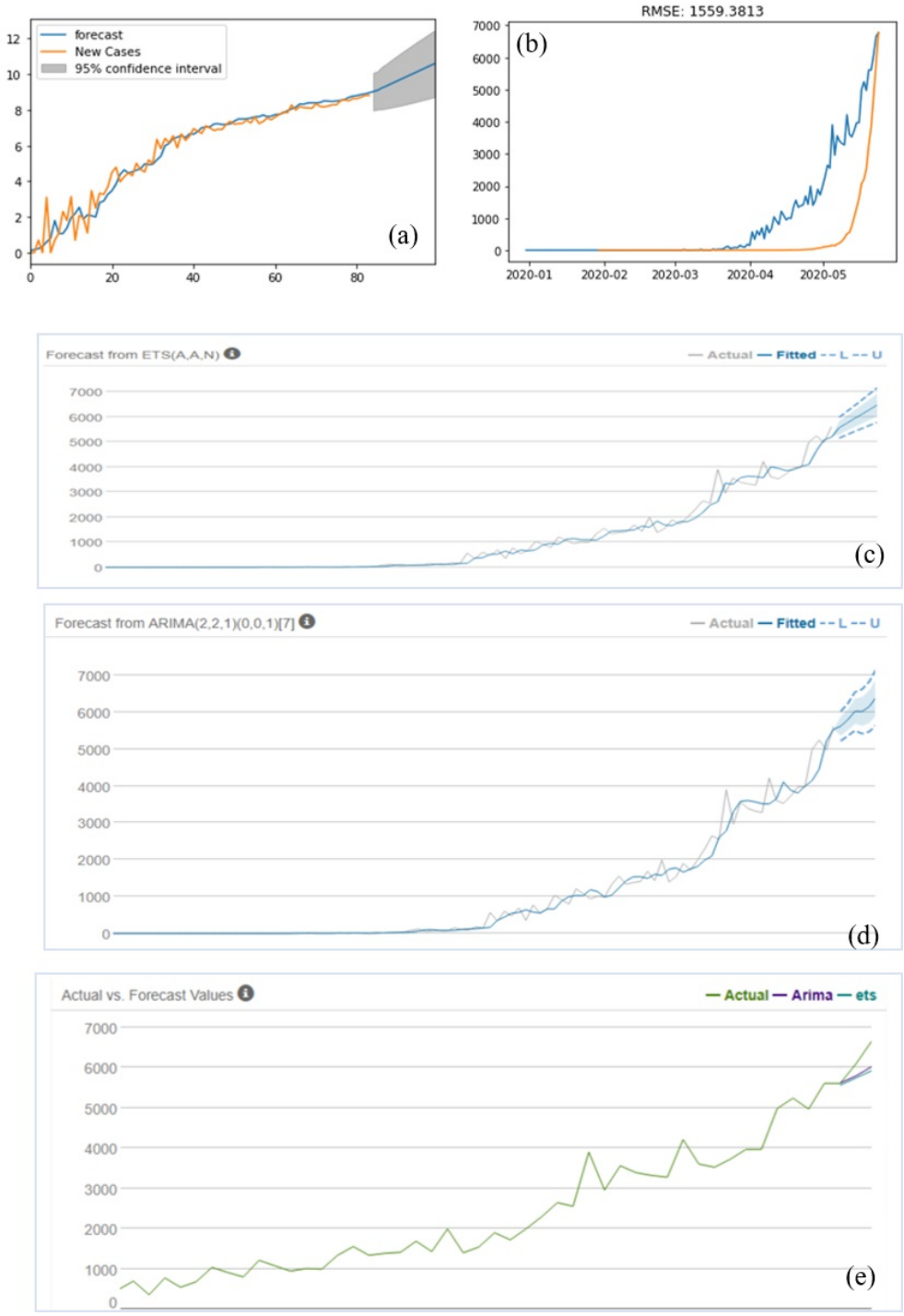

Figure 3: (a) Output for ARIMA log values. The blue line indicates the forecast, the orange line indicates the new cases and the grey part indicates a 95 % confidence level. (b) Output for ARIMA after taking exponents. The RMSE value is 1559. (c) Output for ETS forecasting. The RMSE is 207. The grey line indicates actual data, the blue line indicates the fitted data. L and U indicate lower bound and upper bound. (d)Output for Arima Forecast. The RMSE is 199. The grey line indicates actual data, the blue line indicates the fitted data. L and U indicate lower bound and upper bound. (e) Output for Actual vs. forecast values. The green line indicates actual values, the purple line indicates Arima values and the blue line indicates ETS.

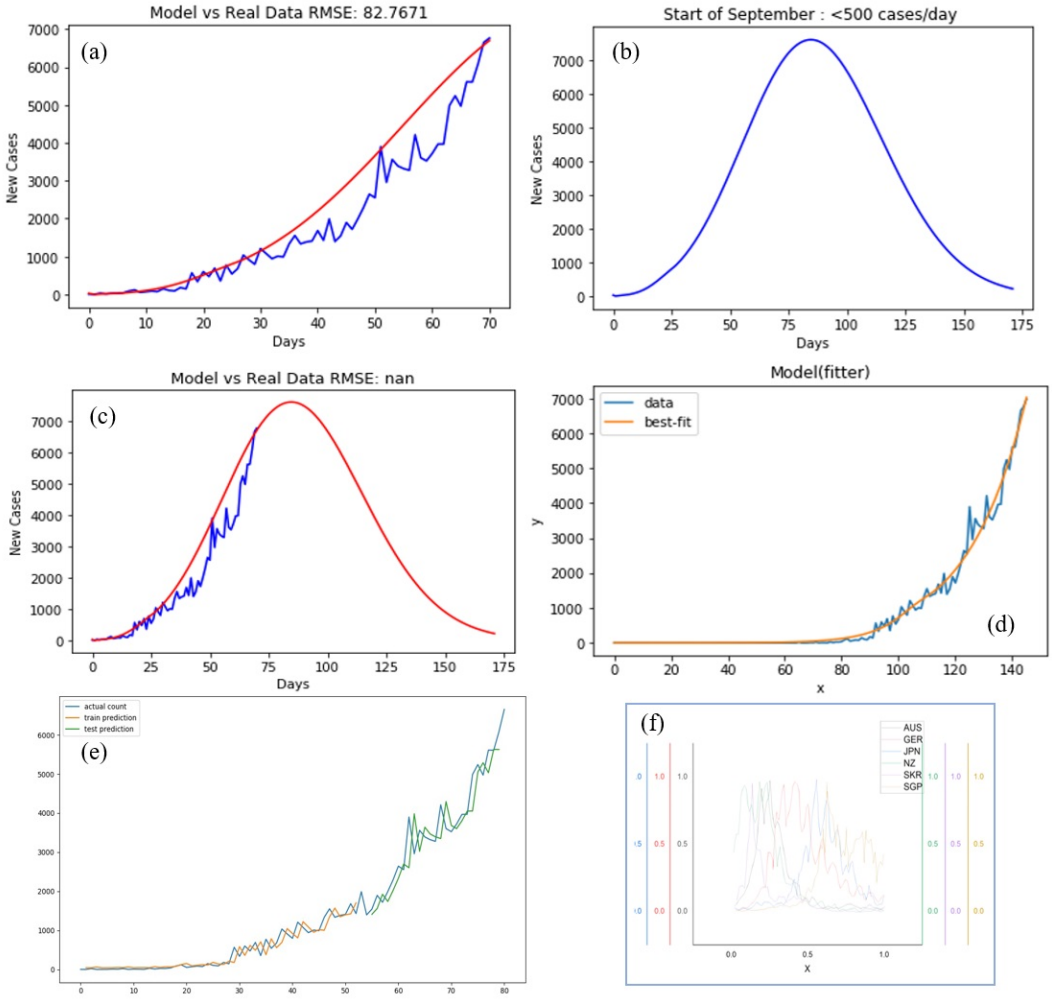

Figure 4: (a) Indicates actual v/s fitted values. The line in blue indicates the predicted new cases every day and line in orange indicates actual numbers of cases every day. (b) Indicates the prediction of Covid-19 patients count. Its bell-shaped curve traditionally assumed by all researchers. (c) Indicates the prediction v/s fitted graph of Covid-19 patients. (d) Graph showing the output of Extended SIR model 2 (e) Indicates the results of the RNN algorithm. Lines in orange and green indicate train and test prediction and line in blue indicates the actual count of patients (f) Indicates Normalized Graphs of Countries at Decay Phase.

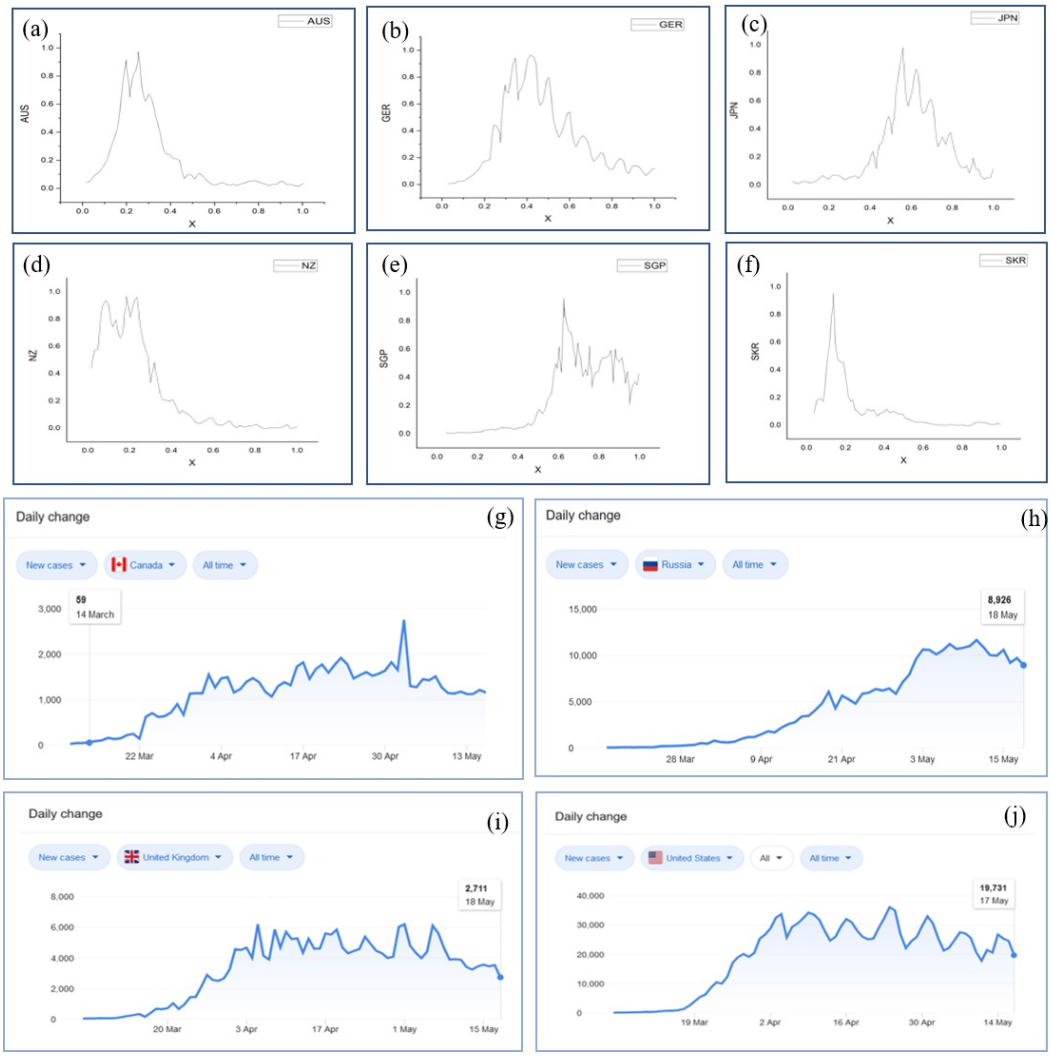

Figure 5: Graph indicating (a) Australia, (b) Germany, (c) Japan (d) New Zealand (e) Singapore (f)South Korea, decay phase. (g) Canada indicating a flat phase in the curve. As of 20 May, the number of confirmed cases is 78072, and 39228 people had recovered. 5842 deaths were reported. (h) Russia indicating a flat phase in the curve. As of 20 May, the number of confirmed cases is 291000 and 70209 people had recovered. 2722 deaths were reported. (i) The United Kingdom indicating a flat phase in the curve. As of 20 May, the number of confirmed cases is 246000, and 34796 deaths were reported. (j) The United States indicating a flat phase in the curve. As of 20 May, the number of confirmed cases is 1540000 and 288000 people had recovered. 90608 deaths were reported.

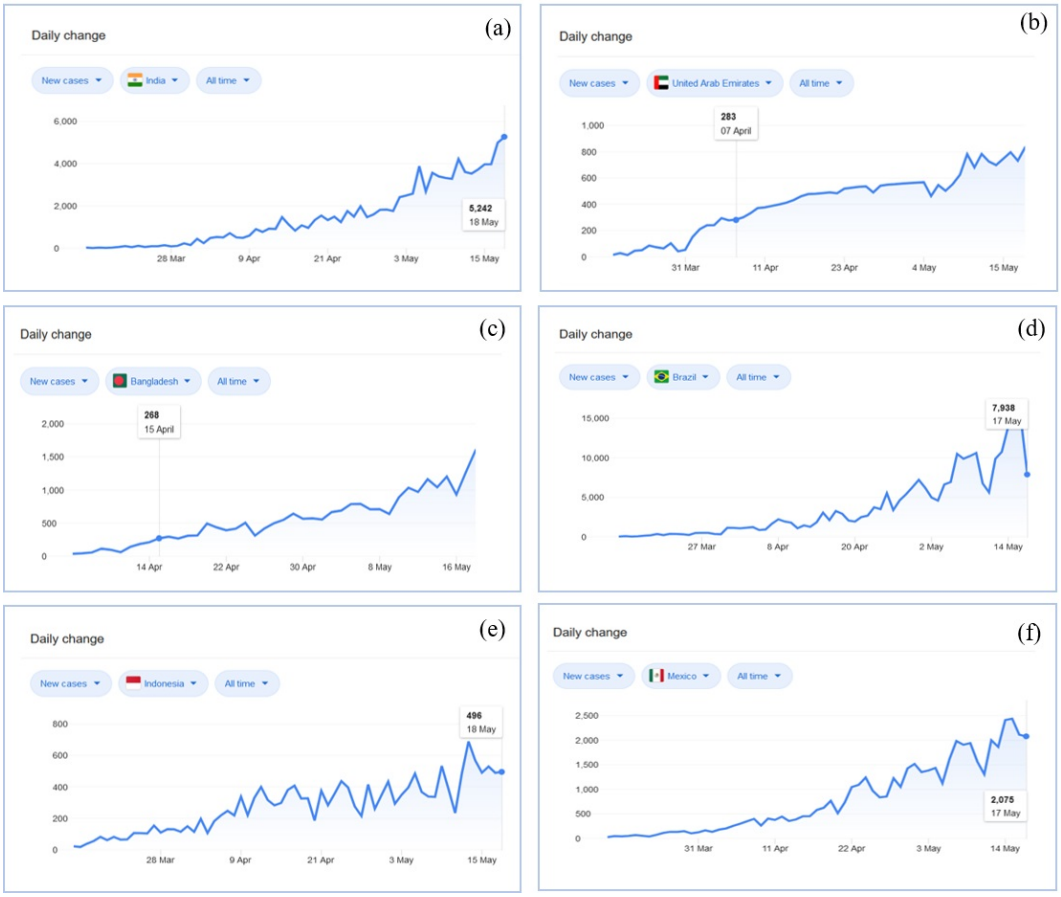

Figure 6: (a) India indicates a rise in the curve. As of 20 May, the number of confirmed cases is 78072, and 39228 people had recovered. 3029 deaths were reported. (b) The United Arab Emirates indicating a rise in the curve. As of 20 May, the number of confirmed cases is 24190, and 9557 people had recovered. 224 deaths were reported. (c) Bangladesh showing a rise in the curve. As of 20 May, the number of confirmed cases is 23870, and 4585 people had recovered. 349 deaths were reported. (d) Brazil showing a rise in the curve. As of 20 May, the number of confirmed cases is 254000 and 100000 people had recovered. 16792 deaths were reported. (e) Indonesia showing a rise in the curve. As of 20 May, the number of confirmed cases is 18010, and 4324 people had recovered. 1191 deaths were reported. (f)Mexico showing a rise in the curve. As of 20 May, the number of confirmed cases is 49219 and 33329 people had recovered. 5177 deaths were reported.

a tremendous impact on the immune system. We all know that the immune system is meant for our protection, but it requires dense regulation as immune cells can also be harmful. Corona infects some of the immune cells and creates a disturbance by confusing them when these cells are poured into the lungs to fight the virus. These immune cells communicate through tiny proteins carrying information called as the cytokines. These cytokines are responsible for carrying out important immune reactions. Coronavirus causes the infected immune cells to exaggerate; it puts the immune system into a state of confusion and sends more antibodies than it should and thereby wasting a lot of its resources and causes great damage. Two types of cells in general create the disturbance. The first type of cells are Neutrophils, which are capable of killing the outside material, are sometimes even capable of killing host body cells. They arrive in large numbers and kill as many friends as enemies. The other being, the killer T-cells that ask the infected cells to commit a controlled suicide. And in this situation of conflict, they tend to ask the healthy cells to destroy themselves too. And as huge number of these cells arrive, more damage is caused to the healthy lung tissues as well. This leads to lifelong disabilities and causes permanent irreversible damage. But in majority of the cases, the immune system slowly regains power. The recovery begins and they start killing the infected cells and also recognize the viruses that try to infect the other new cells. In most of the cases, the patients can be mildly symptomatic. But some cases tend to become critical, and in this case, millions of epithelial cells die, the lungs lining are also damage, and the alveoli gets infected with the virus damaging the respiration. Such patients face difficulties in breathing and need ventilators to survive. For weeks, the immune system has fought with the virus to its full capacity and made antiviral weapons to fight the coronavirus. But if the virus rapidly multiplies and enters the bloodstream, then death is more likely.

There are two possibilities for a pandemic like a corona: rapid and slow. A rapid pandemic costs many lives which would be devastating. The worst-case for a rapid pandemic begins with a very high rate of infection as there are no countermeasures taken to slow it down beforehand. In a rapid pandemic a large number of people would get sick and all at the same time. And if these numbers become too large, then the healthcare facilities would be unable to grip it. Not enough resources, like medical staff or medical equipments like ventilators, would be left to help those in need of them. People will die untreated in such a case. If the health care workers get sick themselves, the capacity of the health care systems to operate properly will fall drastically. And if so happens, then decisions will have to be made about who gets to live and who doesn't. The number of deaths will tend to rise significantly in such a scenario.

Our prediction model helps decision-makers like government and medical supervisors to do the right thing. We expect more and more lives are saved and though this curve is unavoidable we can reduce its peak and slopes. This entire manuscript just provides information about likely next day count and even people can figure out when the end of phase 1 (EOP1) and end of phase 1 (EOP2) can arrive and also in how many days the world or their country will be free from this pandemic crisis.

2 Methodology

We have used multiple time series prediction codes that are useful to the public and for people's benefit we are giving away all the codes with this manuscript. Algorithms used for prediction of next COVID-19 infected patient count daily rise value are

2.1 Holts forecasting model

The simple exponential smoothing to allow the forecasting of the data with no trends was extended by Holt [1]. A forecast equation and two smoothing equations are involved in Holt's method [2](one for the level and one for the trend) where the level smoothing parameter is $0 \leq \alpha \leq 1$ is, and the trend smoothing parameter is $0 \leq \beta \leq 1$. For the long-term forecast, the forecasting with Holt's method will increase or decrease indefinitely and can extend into the future. In that case, we use the Damped trend

method having a damping parameter $0 \leq \phi \leq 1$ to prevent the forecast from giving extremely large value.

Three different approaches were used here:

- Holt's linear trend
- Exponential trend
- Additive damped trend

2.2 Error trend seasonality (ETS) model

ETS stands for error trend seasonality [3], based on the nature of data, we can choose the model to forecast values. Other parameters such as trend, damped, seasonal, seasonal periods, box-cox transform, remove bias are adjusted and RMS values are calculated for all to get the top three best working models for our data.

2.3 AR, MA, and ARIMA Model

Auto-Regressive(AR) is used to describe certain time-varying processes and is a representation of a type of random process in statistics, econometrics, and signal processing [4]. The moving average(MA) model is used in time-series analysis and is a common approach for modeling univariate time series[4]. A class of models that explains a given time series based on its own past values is 'Auto-Regressive Integrated Moving Average'(ARIMA) [5]. It uses its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

2.4 Extend susceptible infected recovered((SIR) model

SIR model [6] is used to simulate disease and hence help us find trends as well as possible future predictions for the spread of disease. In the Extended model we can include more parameters like Exposed, that is, people who have been exposed but not yet showing symptoms. Further, we add cases that have been isolated and deaths, to get closer to the real world. Also, parameters like incubation period for the virus, mortality rate, number of people an infected infects per day, etc. are taken from real-world data as well as manual fitting of data.

3 Extended - SIR MODEL 2 (Curve Fitting)

Explanation: Similar to previous, but we use limit module for python for best-fit curve fitting and also add more parameters like age group, critical conditions, number of beds leading to triage, etc.

4 Forecasting with Neural networks

An artificial recurrent neural network used in deep learning is Long short-term memory (LSTM) having feedback connections. It processes single data points and also entire sequences of data. It overcomes the vanishing gradient problem and is a recurrent neural network trained in backpropagation through time [7].

5 Results

Table 1 indicates the Actual count and predicted count of COVID-19 patients for eight different time series models. The actual count patient was given as 6835 to model. Among all the models the nearest count predicted by the RNN model which was 6872. The Root Mean Square Error (RMSE) fount to be for this model was 35.7. The most inefficient model found among all models was ARIMA. ARIMA estimated a count of 6835 patients as 7551. The RMSE value found in this model was 194.6. Holts models 1, 2, and 3 respectively detected estimated count as 7271, 7481, 7239 with RMSE values of 34.88, 34.132, and 33.409. The models named ETS, ETS 1 and ETS 2 estimated count of patients 7140, 7261 and 7234 with RMSE value of 211.3, 378 and 392. According to the table, we found that the RNN model was nearly able to predict the count of COVID-19 patients correctly.

As shown in figure 1 (a), after analyzing Covid-19 new patients count per day for more than 100 coun-

Model Name	Actual count	Estimated count	RMSE values
Holts model 1	6835	7271	34.88
Holts model 2	6835	7481	34.132
Holts model 3	6835	7239	33.409
ETS model 1	6835	7261	378
ETS model 2	6835	7234	392
ETS	6835	7140	211.3
ARIMA	6835	7551	194.6
RNN	6835	6872	35.7

Table 1: Comparison of different COVID-19 count prediction methods

tries, we have found an interesting result as shown in figure 1 (a). All the predictions in the mentioned curve are with the assumption, that there will not be any vaccine or medicine developed during the normalized timeline period. In the case of a vaccine gets developed and distributed the nature of the curve will be affected. COVID new patients count rises exponentially during phase 1 till a point EOP1 (i.e. End Of Phase 1). This exponential rise in the curve is dependent on various factors like lockdown in the country, sanitization, social distancing, atmospheric temperature variations, weather conditions like rain, snow, etc. also infection spread rate and many more.

After EOP1 the number of cases increasing per day remains constant till EOP2 (i.e. End Of Phase 2), where the herd immunity begins to develop. This means that a saturation level for the spreading of the virus has been reached. This is the phase where the highest limit of COVID to spread has reached and thereafter cannot increase further. This limit is also dependent on the factors mentioned above for exponential rise. The period of saturation between EOP1 and EOP2 is typically 1/3 rd of the exponential rising curve timing, assuming the same conditions are maintained. After the saturation phase ends, i.e. EOP2 onwards, phase of linear decay starts i.e. phase 3. In Phase 3 reduction happens n count, because most of the people either remained asymptomatic and became immune or symptomatic patients are already been cured or isolated from the community. Also, antibodies would have started to produce due to intercellular signal transfer. There is a slow linear decay in phase 3. The important findings were derived from the curves seen below in figure 1 (b), (c), and (d). The population density of the countries whose graphs are shown is large. According to our study, India (fig. 1 (e)) has not reached EOP1 but the world (fig. 1 (f)) has already crossed EOP1 and is now in the transition phase 2 from EOP1 to EOP2. This means that when the number of cases in India would still be rising, a linear decay would be observed in other countries, and slowly the world will be proceeding towards EOP3. The graphs in Figure 1 (b), (c), and (d) also show three major countries impacted by COVID-19 i.e. Italy, Germany, and Australia. In the case of Australia, EOP1 was reached very fast compared to other places, and the time interval for transiting from EOP1 to EOP2 was also less. It reached its linear decay phase quickly too. Whereas in the case of Germany, the situation was opposite to that seen in Australia. Germany and Italy show similar normalized nature of the curves because they both lie in the southern hemisphere, though their population density is different. Their nature of curves is similar and not the number of cases. We have designed prediction software using multiple algorithms. Its sole purpose is to benefit the society in these tough times. Codes are also given as a part of this manuscript in the appendix.

6 Conclusions

We have found that the Coronavirus trend is not a regular bell-shaped curve but it is more like a 3 phase

curve. The majority of the countries worldwide are now in the transition phase from EOP1 to EOP2 but still, there are many countries like India which are in the rising phase. With this manuscript, we would like to alert the population about the trend that we have analyzed with multiple models and also would like to give away the codes we developed for future researchers to develop further.

Compliance with Ethical Standards:

Conflicts of interest

Authors N. Mehendale, V. Shah, J. Shah, M. Parab and A. Shelke declares that they have no conflict of interest.

Involvement of human participant and animals

This article does not contain any studies with humans and animals performed by any of the authors. All the necessary permissions were obtained from Institute Ethical committee and concerned authorities.

Information about informed consent

Informed consent was not required as there were no human participants.

Funding

There was no funding involved.

References

[1] P.S. Kalekar, Time series forecasting using holt-winters exponential smoothing, Kanwal Rekhi School of Information Technology **4329008**(13) (2004)

[2] J.W. Taylor, Exponential smoothing with a damped multiplicative trend, International journal of Forecasting **19**(4), 715 (2003)

[3] G. Jain, B. Mallick, A study of time series models arima and ets, Available at SSRN 2898968 (2017)

[4] M. Valipour, M.E. Banihabib, S.M.R. Behbahani, Comparison of the arma, arima, and the autoregressive artificial neural network models in forecasting the monthly inflow of dez dam reservoir, Journal of hydrology **476**, 433 (2013)

[5] A.C. Harvey, A unified view of statistical forecasting procedures, Journal of forecasting **3**(3), 245 (1984)

[6] D. Wang, X. Zhao, Empirical analysis and forecasting for sars epidemic situation, Beijing da xue xue bao. Yi xue ban= Journal of Peking University. Health sciences **35**, 72 (2003)

[7] J. Vermaak, E. Botha, Recurrent neural networks for short-term load forecasting, IEEE Transactions on Power Systems **13**(1), 126 (1998)