Time Series Analysis and Forecasting of COVID-19 using Machine Learning

A.K.M. Yasar, Md. Iftekhar Uddin Chy and Dr. Shahidul Islam Khan

Abstract. According to the World Health Organization (WHO), the COVID-19 pandemic is a serious global health warning that has been the most harmful respiratory virus since the 1918 H1N1 influenza pandemic. COVID-19 situation report till September 23, 2020, sum of 31 million substantiated cases and 1 million expired of life have been reported rapidly with 216 countries. We Collected the dataset from several organizations, such as WHO (World Health Organization), Directorate General of Health Services (DGHS) and Kaggle. The dataset contains information about confirmed cases, death cases and recovery cases. In this work, we have tried to make four models which are Linear Regression, SVR, Holt's Linear and Holt Winters Model to make a comparison of which model is working good for globally as well as for Bangladesh. The research shows that according to the Holt Winter model, the number of infected people will definitely increase on the date of October 3rd to 3,77, 592 total confirmed cases in Bangladesh and globally it will increase to 32,977,149 numbers on September 28th.

1 Introduction

1.1 Introduction

According to the World Health Organization (WHO), the COVID-19 (SARS-CoV-2) pandemic is a crucial worldwide health warning that has been the most detrimental respiratory virus since the 1918 H1N1 influenza pandemic. COVID-19 situation report till September 23, 2020, a whole of 31 million active cases and 1 million life expired have been reported rapidly with 216 countries.

In Bangladesh, the infected number is 3,53,844, and the total number of death is 5044 till September 23, 2020. SARS-CoV-2 (severe acute respiratory syndrome) is the

A.K.M. Yasar, Md. Iftekhar Uddin Chy and Dr. Shahidul Islam Khan
Department of CSE, International Islamic University Chittagong (IIUC)**.**
Chittagong, Bangladesh**.**
Email: akmyasar@gmail.com. , iftekhar.fahim97@gmail.com and
nayeemkh@gmail.com

reason for coronavirus disease (COVID-19). The communication of this virus moves person to person with the respiratory droplets. Experts, welfare workers, and other people who provide basic assistance types must be ensured consistent with the recommended clinical standards.

1.2 Background and Present Statement of the Problem

Though Covid-19 started back in 2019 from China it spread rapidly from the starting of 2020 worldwide. During the month of March 2020, the first infected patient (COVID-19) was found in Bangladesh as well. Since then this virus remains unstoppable to stop in Bangladesh as well as Worldwide.

Many works have been done worldwide based on Covid-19 so far globally. Scientists, researchers, everyone have been working to find the solution to this problem or supporting it by providing the perfect analysis, predictions, and suggestions that help to find a way to require steps, preparations, and protections. Based on COVID-19 research there are plethora of work globally but there is less work has been done in Bangladesh. It may be done privately but publicly there is some work that can help the people, health sectors, or Government to take measures steps before the situation gets worse.

In this situation, we need prediction tools and some time series analysis that can help project different scenarios, such as.

1. Number of possible conformations for new cases.
2. Number of possible death cases.
3. Number of recovered cases

As a result, prediction tools are useful for several different purposes. For better understanding, prediction models are important to possess a far better estimation of the disease and its possible threats. To be accurate, consistent with the Centers for Disease Control and Prevention (CDC), prediction models help answer the pandemic by informing decisions about planning, resource allocation and wish social distancing. Moreover, a proper time series analysis can help us to get an idea about the upcoming situations based on our past situations

2 Related Works

Dynamical Mathematical Modeling used to make a spreading analysis of the COVID-19 epidemic in Bangladesh [1]. In the study, a developed SIR (Susceptible-Infected-Recovered) model is used and a forecast is initiated to assume the COVID-19 cases in Bangladesh. The work analyzed the impact of before and after the situation of lockdown. The work is based on the situation and data between March, April, and May of 2020.

Work has been done for the Covid-19 data analysis and forecasting by applying some algorithm [2]. The work basically represents the concentration for the analysis of

India after including the world analysis. They tried linear regression and support vector machine regression in their work of forecasting for India.
A machine learning forecasting model for COVID-19 pandemic in India[3]. This system developed a model that could be useful to predict the spread of COVID-2019. This work will help in predicting and forecasting the near future by using the most frequent data of confirmed, recovered, and death cases across India They perform Linear regression, Multilayer perception, and Vector auto regression method for finding the pace of COVID-2019 cases in India.
Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions [4].This system developed a model to derive the epidemic curve and find confirmed cases across different provinces in China. They used the LSTM model, a type of RNN that has been used to process and predict various time series problems for predicting new infections over time. They also used Artificial Intelligence to train the SARS dataset for future prediction of the epidemics. They tried to show how these control measures impacted the containment of the epidemic during the lockdown period.
A Data-driven Approach to get the forecasting endpoint in Bangladesh[5]. In this work the researcher did the research with the data till June (During the lockdown). They implemented Time Series (TS) for prediction modeling approach and Recurrent neural network (RNN) which is used in temporal domains to learn sequential patterns. Recurrent LSTM networks can address the difficulties of traditional time series forecasting ways by taking nonlinearities of given COVID-19 dataset and can give output in a state of the art results on temporal data.
A time series model is made to forecast the confirmed and recovered cases of COVID-19 through a flexible family model[6]. In this work, they proposed the autoregressive time series model *TP–SMN–AR* model. It includes the symmetric Gaussian and asymmetric heavy-tailed non-Gaussian autoregressive time series models. Time series models have a statistical methodology that is functional to model time-indexed data and for forecasting. The work prefers to work with this family model to develop the time series of confirmed and recovered cases.

3 Proposed Techniques

3.1 Overview

We Collected the dataset from several organizations, such as WHO (World Health Organization), Directorate General of Health Services(DGHS) and Kaggle**.** We got different types of attributes from this data, attributes such as Date, S/No, State, Region, Latest Update, Confirmed, Recovered and Deaths. We've observed on the daily figures aggregated worldwide and Bangladesh of the three main update of interest which are confirmed cases, deaths and recoveries.

We have done prediction and time series analysis on our dataset. We have applied Linear Regression, Support Vector Regression, Holt's Linear Trend Model and Holt-

Winters' multiplicative method on the world dataset and Bangladesh. We will check the error values compare between them.

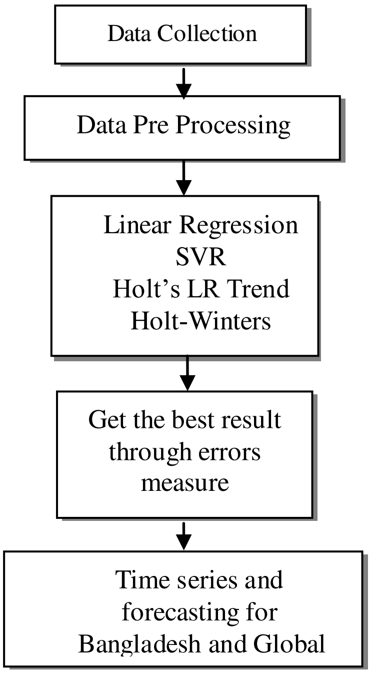

Data Collection
Data Pre Processing
Linear Regression
SVR
Holt's LR Trend
Holt-Winters
Get the best result through errors measure
Time series and forecasting for Bangladesh and Global

Fig 1. Overview of the work

3.2 Exploratory Data Analysis

It is an approach to understand and summarize the main characteristics of a given data. In our work, we have analyzed our datasets with different EDA methods and visualize those data to provide information regarding the outbreak of COVID-19 all over the world and Bangladesh.

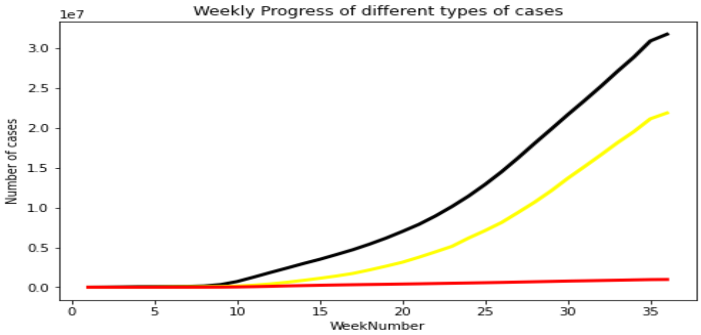


Fig. 2. Weekly Progress of different types of cases Globaly *(Black-Confirmed, Yellow-Recovered and Red-Deaths)*

We tried to visualize and analyze data between 22 January 2020 and 23 September 2020. However, a huge number of cases are reported in China compared to the rest of the world. We also showed Histogram for Top 15 countries that have caused the most number of death, Confirmed and Recovered cases. We also calculated the growth factor, mortality rates, average increasing number for all cases. We showed a plot for the recovery rate and mortality rate across the globe. We also calculated the total number of confirmed, recovered and death cases of Bangladesh. We find out the percentage of cases and show these values in the pie chart.

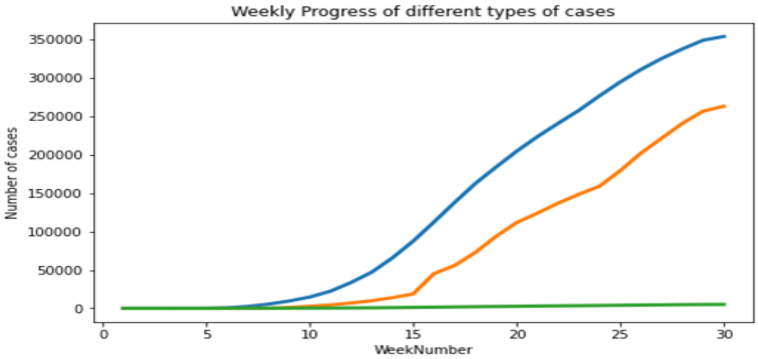


Fig. 3. Weekly Progress of different types of cases of Bangladesh
(Blue-Confirmed, Orange-Recovered and Green-Deaths)

3.3 Used Methods

Linear Regression: Linear regression models are wont to show or predict the connection between two variables. The factor that's being predicted is named the variable. The factors that are used to predict the price of the variable are called the independent variables. The main goal of the straightforward rectilinear regression is to think about the given data points and plot the simplest fit line to suit the model in the best way possible.

The two factors that are involved in simple rectilinear regression analysis are posted x and y. The equation that describes how y is said to x is understood because of the regression model.

The simple linear regression model is represented by:**.**

$$y = \beta0 + \beta1x1 \text{ ……………………………………..……………….(i)}$$

Support Vector Machine – Regression (SVR): Support Vector Regression is a regression algorithm. It supports linear and non-linear regressions. Support Vector Regression has some hyper parameters.

1. Kernel is a group of mathematical operations, which takes as input data and transforms its into the expected form. The most widely used kernels include Linear, Non-Linear, Polynomial and Sigmoid.
2. Epsilon, which is implemented to operate a line over data marks.

Holt's Linear Trend Method: It is Suitable for time series data with a trend component but without a seasonal component. It helps to forecast time series data that has a trend. Holt method adds the trend smoothing parameter β. The range of β is between 0 and 1. The easy exponential smoothing method doesn't compute for any trend or seasonal components, rather, it only uses the decreasing weights to forecast future outputs. This makes the tactic suitable just for statistics without trend and seasonality. When the trend is relatively constant then we use the Additive model. So then we can use Holt's Linear Trend Method for Time Series Data.

Forecast equation yt+h|t = lt+hbt ……………………………………….............*(ii)*

Level equation lt = αyt+(1−α)(lt−1+bt−1) ………………………………….......*(iii)*

Trend equation bt = β∗(lt−lt−1)+(1−β∗)bt−1 ……………………………..……….....*(iv)*

Where, lt denotes an estimate of the level of the series at time t, bt denotes an estimate of the trend (slope) of the series at time t, α is the smoothing parameter for the level, 0≤α≤1, and β∗ is the smoothing parameter for the trend, 0≤β∗≤1. As with simple exponential smoothing, the level equation here shows that lt is a weighted average of observation yt and the one-step-ahead training forecast for time t, here given by lt−1+bt−1. The trend equation shows that bt is a weighted average of the estimated trend at time t based on lt−lt−1 and bt−1, the previous estimate of the trend.

Holt-Winters Multiplicative Method: Holt-Winters Seasonal Method taken for the forecasting time series that puts both trend and a seasonal combination. It comprises the forecast equation and three smoothing equations one for the level lt, one for the trend bt, and one for the seasonal component st, with corresponding smoothing parameters α, β∗ and γ. The formula for the multiplicative is:

$$
\begin{aligned}
\hat{y}_{t+h|t} &= (\ell_t + hb_t)s_{t+h-m(k+1)} \\
\ell_t &= \alpha\frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \\
s_t &= \gamma\frac{y_t}{(\ell_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m}
\end{aligned}
$$

3.4 Error Measure Techniques

Root Mean Square Error (RMSE): RMSE is quality deviation**.** Root mean square error is usually utilized in climatology, forecasting and multivariate analysis to verify experimental results

$$\text{RMS Errors} = \sqrt{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 / n}$$

. ……………………………………………..........*(v)*

Here, ŷi observed value for ith observation and yi predicted value and n number of observations.

Mean Absolute Error (MAE): Mean Absolute Error is calculated as the average of the forecast error values, where all of the forecast error values are forced to be positiv**e.** Based on absolute value of error

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |x_i - x|$$

....*(vi)*

Here, n is the number of errors**,** $\sum$ means add them all and $|\boldsymbol{x_i}.$ – x|means the absolute errors.

Mean Square Error (MSE): Mean Square Error (MSE) is defined as Mean or Average of the square of the difference between genuine and estimated values. It based on square of error. Lower the MSE, the closer is forecast to genuine.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

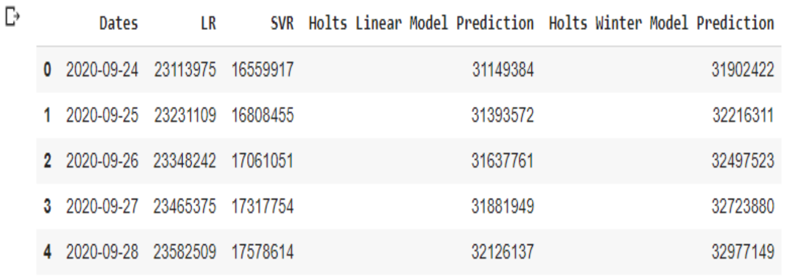
....*(vii)*

Here, n is data points on all variables, $y_i$is the vector of observed values of the variable being predicted and $\sim y_i$ is the predicted values.

4 Experimental Result

Our aim is to predict and Forecast Time Series Analysis. We use LR, SVR for Prediction and Holt's Linear Trend Model, Holt-Winters Multiplicative model for Time Series Analysis. We Calculate and compare RMSE, MAE and MSE values between these four models. We used these four models for the whole world and as well as Bangladesh. We mainly worked on the Prediction of Confirmed cases in the world with time series analysis.

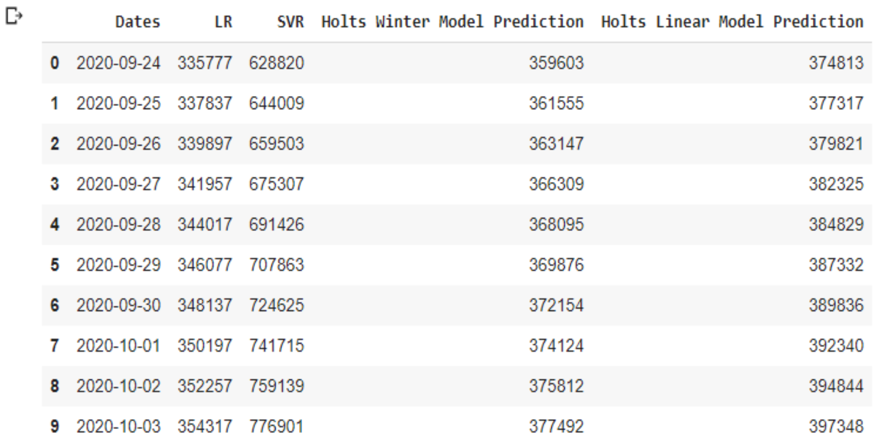
4.1 Time Series and Forecasting for Global



	Dates	LR	SVR	Holts Linear Model Prediction	Holts Winter Model Prediction
0	2020-09-24	23113975	16559917	31149384	31902422
1	2020-09-25	23231109	16808455	31393572	32216311
2	2020-09-26	23348242	17061051	31637761	32497523
3	2020-09-27	23465375	17317754	31881949	32723880
4	2020-09-28	23582509	17578614	32126137	32977149

Table 1. Result of LR, SVR, Holt's Linear and Holt Winters for times series and forecasting (Global)

4.2 Times series and Forecasting for Bangladesh



	Dates	LR	SVR	Holts Winter Model Prediction	Holts Linear Model Prediction
0	2020-09-24	335777	628820	359603	374813
1	2020-09-25	337837	644009	361555	377317
2	2020-09-26	339897	659503	363147	379821
3	2020-09-27	341957	675307	366309	382325
4	2020-09-28	344017	691426	368095	384829
5	2020-09-29	346077	707863	369876	387332
6	2020-09-30	348137	724625	372154	389836
7	2020-10-01	350197	741715	374124	392340
8	2020-10-02	352257	759139	375812	394844
9	2020-10-03	354317	776901	377492	397348

Table 2. Result of LR, SVR, Holt's Linear and Holt Winters for times series and forecasting (Bangladesh)

4.3 Error measure for the used methods

After forecasting for both Bangladesh and Globally, now we will be focusing on 3error counting tools. We also compare the error and use values between all the models to find which gives better result. Here we focus onto the error measure techniques RMSE, MAE and MSE which will suggest us the lower error value will work better than others.

The Lower value of the error measure techniques will show us the method that work perfectly on this forecasting.

Comparison between Root Mean Square Error (RMSE)

Method	Global	Bangladesh
Linear Regression	2982454.405359	22198.15740944
SVR	6544407.20571	207887.17384
Holt'sLinear Trend	443268.712469	9279.9061379
Holt-Winters	121656.622203	2863.84309487

Table 3. RMSE value of methods for Global and Bangladesh

Comparison between Mean Absolute Error (MAE)

Method	Global	Bangladesh
Linear Regression	2608836.82085	22158.4814895
SVR	5211061.446367	205073.4631145
Holt'sLinear Trend	363524.412909	7220.5240359
Holt-Winters	93648.4776929	2790.34079234

Table 4. : MAE value of methods for Global and Bangladesh

Comparison between Mean Square Error (MSE)

Method	Global	Bangladesh
Linear Regression	8895034280050.623	492758192.3745023
SVR	42829265674252.33	43217077049.31404
Holt'sLinear Trend	196487151454.4432	86116657.9296865
Holt-Winters	14800333726.057678	8201597.27206297

Table 5. MSE value of methods for Global and Bangladesh

Among all of them the less error one model gets is the higher chance of work perfectly in the task. Here we see that the error for the Holt-Winters Multiplicative Model is lower than others in all error measure cases for both Globally and Bangladesh.

4.4 Data Visualization for Holt-Winters Multiplicative Model

Among the Four Models, Holt's Winter Model is performing better. Our data is random and it has seasonality and trending tendency that's why we have used Holt's

Winter Multiplicative Method. This model can handle complicated seasonal patterns. It uses exponential leveling to encode values from the past and forecast values for the present and future.

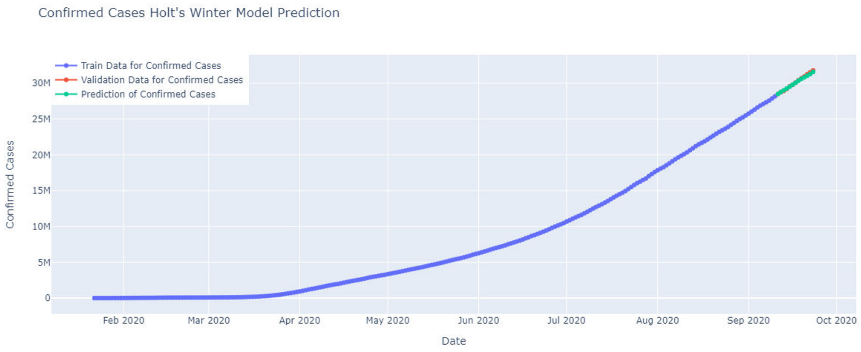


Fig. 4.Visualization of Forecast for Globally Confirmed cases by Holt Winters

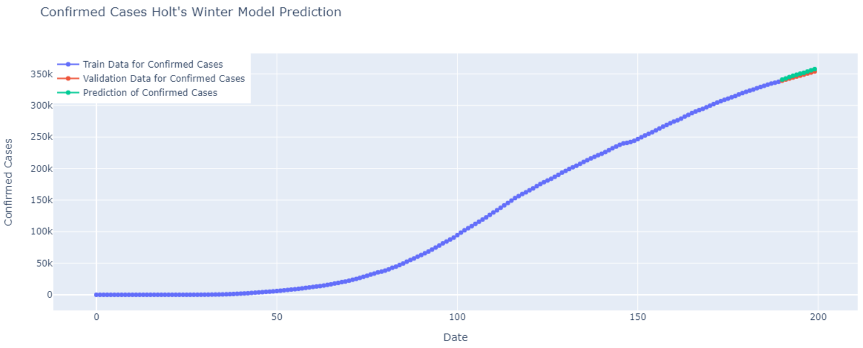


Fig. 5. Visualization of Forecast for Globally Confirmed cases by Holt Winters

When we expect regular fluctuations in the data, Holt- Winters model attempts to map the seasonal behavior. That's why Holt-Winters Model fits the data best and performing better.

5 Conclusion

In this study, some AI models predict the worldwide adjust- ment of COVID-19 mortality as well as for the Bangladesh perspective. We have investigated this information and found that the number of infected numbers continued to increase from April 2020. We have found that the Holt-Winters model is working better than the other three models we have used in our research.

In addition, according to the Holt Winter model, the number of infected people will definitely increase on the date of Oc- tober 3rd to 3,77,592 total confirmed cases in Bangladesh and globally it will increase to 32,977,149 numbers on September 28th.

Experts, welfare workers, and other people who provide basic assistance types must be ensured consistent with the recommended clinical standards. Due to people's frivolous conduct, the disease spread later, just as infected person can multiply the number of cases.

Future Work: In the future there can be more attributes in detail can be used for getting more accurate value. Terms of many different methods that require values in detail can be used to get the more proper result in the end. It is also can be looked over the gender, weather, age, etc to find if there any possibility for them to affect the numbers. In those terms, some more modern techniques can be applied in future research.

References

1. A. Arifutzzaman, A. Fargana, and A. A. Rakhimov, "*Spreading Analysis of COVID-19 Epidemic in Bangladesh by Dynamical Mathematical Modelling*," 2020, doi: 10.1101/2020.06.12.20130047.

2. S. Sengupta and S. Mugde, "*Covid-19 Pandemic Data Analysis and Forecasting using Machine Learning Algorithms*," *medRxiv*, p. 2020.06.25.20140004, 2020, [Online].

.

3. R. Sujath, J. M. Chatterjee, and A. E. Hassanien, "*A machine learning forecasting model for COVID-19 pandemic in India*," *Stoch. Environ. Res. Risk Assess.*, vol. 34, no. 7, pp. 959–972, 2020, doi: 10.1007/s00477-020-01827-8.

4. Z. Yang *et al.*, "*Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions*," *J. Thorac. Dis.*, vol. 12, no. 3, 2020, [Online]. Available: http://jtd.amegroups.com/article/view/36385.

5. A.-E. E. Hridoy *et al.*, "*Forecasting COVID-19 Dynamics and Endpoint in Bangladesh: A Data-driven Approach*," pp. 0–3, 2020, doi: 10.1101/2020.06.26.20140905.

6. M. Maleki, M. R. Mahmoudi, D. Wraith, and K.-H. Pho, "*Time series modelling to forecast the confirmed and recovered cases of COVID-19*," *Travel Med. Infect. Dis.*, vol. 37, p. 101742, 2020, doi: https://doi.org/10.1016/j.tmaid.2020.101742.

7. R. Gupta and S. K. Pal, "*Trend Analysis and Forecasting of COVID-19 outbreak in India*," *medRxiv*, 2020, doi: 10.1101/2020.03.26.20044511.

8. T. Chakraborty and I. Ghosh, "*Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis*," *Chaos, Solitons & Fractals*, vol. 135, p. 109850, 2020, doi: https://doi.org/10.1016/j.chaos.2020.109850.