

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

1st Dr. Shahidul Islam Khan
Department of CSE,
International Islamic University
Chittagong, Bangladesh
nayeemkh@gmail.com

2nd A.K.M. Yasar
Department of CSE
International Islamic University
Chittagong, Bangladesh
akmyasar@gmail.com

3rd Md. Iftekhar Uddin Chy
Department of CSE
International Islamic University
Chittagong, Bangladesh
iftekhar.fahim97@gmail.com

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.

Index Terms—COVID-19, WHO, Confirmed case , 1918 H1N1, Error measure

I. INTRODUCTION AND BACKGROUND

A. 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 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.

B. Background and Present Statement

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.

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.

C. Motivations

In this situation, we need prediction models that can help us to create scenarios for upcoming days. Time series analysis and forecasting with the existing data with some proper method can make this work happen. Besides this, we can take help with comparing the root mean square error of every model through several methods. Holt's Model will work for future forecasting updates. There are fewer attributes of data available in Bangladesh so we tried out some techniques to make the analysis and forecasting with some different models than previously used.

II. LITERATURE REVIEW

Let us have a look into a review of the literature chapter where we will analyze the literature that relates to our work. There has been much work related to the time series analysis, forecasting, and prediction with several kinds of models which are good for the other countries which have a bigger population than Bangladesh. Bangladesh does not have that much bigger population than some other country. Our main goal is to see which models work most efficiently and we will see how the models are behaving globally as well.

A. Some related works have done previously

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. Where he found success using the Sigmoid Model. 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.

B. Some points to focus

[1] To make a spreading analysis of the COVID-19 epidemic in Bangladesh we saw that SIR model works well to make analysis because the data from the root level (thana, union) based data were on that time but after the time being it was unavailable for the government of Bangladesh to publish the root level data publicly. Only data were published divisionally for the people. So in that case simply a single model can't work to make the time series analysis as the SIR model itself needs various attributes of root level data so to work with this model it will be tough based on the data of 7 months ago. As India is a country with a big population so that linear regression or support vector regression might not work properly but as a Bangladesh is a population related to lower population so we would try to implement linear regression

and support vector regression. [2] Though the sigmoid can't be suitable for many countries as the analysing curve might not be in a 'S' shape all the time or not in any binary limitation. [3] As Vector auto regression model used to capture the relationship between multiple quantities as they change over time it will be tough for this model to capture. VAR models don't require the maximum amount knowledge about the forces influencing a variable as do structural models with equation. [4] SEIR model works with the root level of the data with many attributes. As we are using only confirmed cases, deaths and recovered cases in such cases SEIR doesn't work. This model is effective but not for all the countries which don't have availability of data in detail.[5]

III. METHODOLOGY

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.

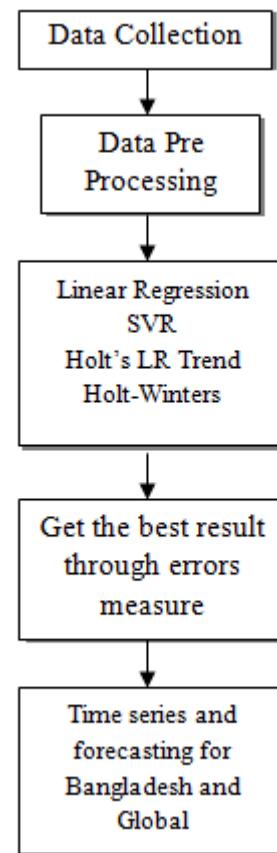


Fig. 1. Over view of the process

A. Description of the dataset

For our work, the dataset had been used for our analysis, visualizations, prediction and Time Series Analysis. Our dataset contains data from January-September 23, 2020. The dataset has 7 columns. It tracks the spread of the COVID-19 outbreak across the world. It consists of a number of confirmed, death, and recovered cases

B. 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. We have calculated Active and closed cases also visualized these data. We tried

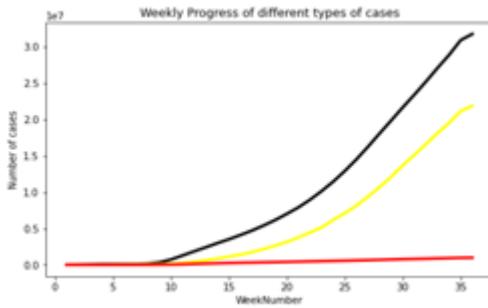


Fig. 2. Weekly Progress of different types of cases Globally (Black-Confirmed, Yellow-Recovered and Red-Deaths)

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.

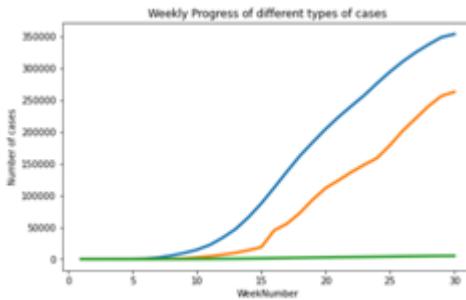


Fig. 3. Weekly Progress of different types of cases of Bangladesh (Blue-Confirmed, Orange-Recovered and Green-Deaths)

C. Proposed Techniques

Linear Regression: 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: $y = B_0 + B_1 x$

Support Vector Regression: It supports linear and non-linear regressions. Support Vector Regression has some hyper parameters. 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. 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 . The range of 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.

$$\text{Forecast equation } \hat{y}_{t+h|t} = l_t + h b_t$$

$$\text{Level equation } l_t = y_t + (1 - \alpha)(l_{t-1} + b_{t-1})$$

$$\text{Trend equation } b_t = (l_t - l_{t-1}) + (1 - \beta)b_{t-1}$$

Where, l_t denotes an estimate of the level of the series at time t , b_t denotes an estimate of the trend (slope) of the series at time t , α is the smoothing parameter for the level, 01, and β is the smoothing parameter for the trend, 01

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 l_t , one for the trend b_t , and one for the seasonal component s_t , with corresponding smoothing parameters α , β , and γ . From our dataset we can say that seasonality components are increasing, that's why we will use Holt-Winters multiplicative model. The formula for the multiplicative is

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

In this method seasonal variation changes in relation to the overall changes the data. So, if the data is trending upward, the seasonal differences grow proportionally as well. It is easy to identify the frequency of seasonality when one figures out which type of seasonality contain in the data.

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

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

Here, \hat{y}_i observed value for i th observation and y_i 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 positive. Based on absolute value of error

Here, n is the number of errors means add them all $|x_i - \hat{x}_i|$ means the absolute errors.

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

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.

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

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

IV. EXPERIMENTS RESULTS

After forecasting for both Bangladesh and Globally now we focuses onto the error measure techniques RMSE, MAE and MSE.

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

Table 1: RMSE value of methods for Global and Bangladesh

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

Table 2: MAE value of methods for Global and Bangladesh

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

Table 3: 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.

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.

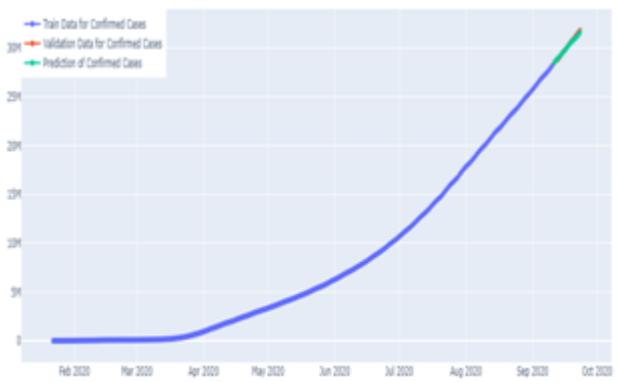


Fig. 4. Visualization of Forecast for Globally Confirmed cases by Holt Winters

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.

V. CONCLUSION

In this study, some AI models predict the worldwide adjustment 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 October 3rd to 3,77,592 total confirmed cases in Bangladesh and globally it will increase to 32,977,149 numbers on September 28th.

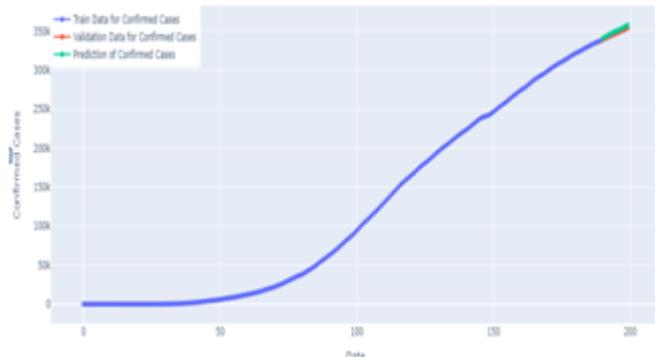


Fig. 5. Visualization of Forecast for Bangladesh's Confirmed cases by Holt Winters

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.

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. Arifuzzaman, A. Fargana, and A. A. Rakimov, "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]. Available: <http://medrxiv.org/content/early/2020/06/26/2020.06.25.20140004.abstract>
- [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>.