TIME SERIES FORECASTING OF CONFIRMED CASES AND DEATHS OF COVID-19 OUTBREAK USING MACHINE LEARNING AND DEEP LEARNING APPROACH

[1]Saqib Gulzar Bhat, [2]Sumaira Farooq, [3]Musadiq Amin

[1]Student, [2]Student, [3]Student
[1]Directorate of Information Technology and Support System,
[1]University of Kashmir, Srinagar, India

Abstract: Covid-19 has been responsible for the deaths of people in lakhs and millions of people have been affected worldwide. To avoid future deaths, it is of utmost importance to identify the future cases and virus spread rate in advance. It is an analytical and challenging real-world task to forecast accurately the spread of this virus. Therefore, we use day level information of COVID-19 spread for cumulative cases from whole world. The dataset used in this research is from Jhon Hopkin University which contains the spread of the virus from January 22, 2020 to till date. It is a daily updating dataset. We model the evolution of the COVID-19 outbreak, and perform prediction using Machine learning and Deep learning-based time series forecasting models for next 20 days. Effectiveness of the models are evaluated based on the mean absolute error, and mean square error. Our analysis can help in understanding the trends of the disease outbreak, and provide epidemiological stage information of adopted countries. Our investigations show that Deep learning approach is best for time series-based problems and more effective for forecasting COVID-19 prevalence. The forecasting results have potential to assist governments to plan policies to contain the spread of the virus.

Index Terms **- covid-19, Machine Learning, Deep Learning, Forecasting, Time Series, Pandemic, Jhon Hopkin University's dataset.**

1. INTRODUCTION

For More than 56 corers of confirmed cases and 6.3 million deaths have been reported across the world since the beginning of the COVID-19 outbreak. The virus has been responsible for affecting severely the physical and mental health status and economic conditions of the people all across the globe. The virus is called 'corona' virus because it shows the presence of a 'solar corona'-like image when it is observed under the electron microscope. In the past, these viruses had triggered many outbreaks, such as the Extreme Acute Respiratory Syndrome Coronavirus (SARS-CoV) in China and the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) in the Middle Eastern countries. COVID-19 is seen to be an infectious disease which is caused by Severe Acute Respiratory Coronavirus 2 (SARS-CoV-2) [1], [2]. Initially, the COVID-19 virus was identified in China in December 2019, after which it spread across all the countries in the world when people started coming in contact with the infected people and then travelled to different regions. Organs of Human respiratory system especially the lungs are severely affected by this virus.

Keeping the severity of this virus the WHO declared this disease as a pandemic during the initial phases of its transmission, indicating that it is a very severe and deadly disease [3]. It has been noticed that this virus causes death, either directly or through exacerbating pre-existing health problems. It also affects the physical and mental health of the people significantly. It is of utmost Importance to forecast the spread and number of potential covid-19 affected cases that may occur in advance using pre-compiled data as a large proportion of people are getting affected by the COVID-19 pandemic throughout the world. Many researchers, including data scientists, have been working intensely to determine ways to eradicate this disease completely from earth. Data scientists can effectively contribute to the research by designing prediction models that highlight the probable activities of this virus, which can further help in accurately predicting the spread of this virus. Hence, Machine and deep learning (DL) models are regarded as accurate tools which can help in developing prediction models. Though many neural networks (NNs) have been described in the past, the long short-term memory (LSTM) has been investigated in this work as they can use temporal data [4].

In this research work, Machine learning (linear regression) and LSTM deep-learning networks have been used. These algorithms were selected as they could analyze the time series data and accurately predict future trends [4]. These models showed considerable success in forecasting temporal data among other traditional methods.

2. LITERATURE SURVEY

Many Intensive research work is going on to evaluate and contain the worldwide disaster of COVID-19 on the human race. Research studies include predictions about the future cases [5], and analysis of the variables responsible for spread of the coronavirus [6].

In the literature, time series forecasting problems have been studied widely in which COVID-19 forecasting is an emerging problem. Forecasting models can be used to forecast the impact of the disease on the community which can help to control the epidemic.

In [7], authors have performed forecasting evaluation study of the models using COVID-19-day level cases from 10 mostly affected states from Brazil. According to the authors, the stacking ensemble and SVR performed better as compared to ARIMA, CUBIST, RIDGE, and RF models for the adopted criteria.

In [8], the author has developed ARIMA (p, d, q) model and studied the COVID-19 epidemiological trend in the three most affected countries; Spain, Italy, and France of Europe continent using the data between 21 Feb to 15 April 2020. The author studied the various orders (p, d, q) of the model, and selected best performing order based on lowest values of MAPE for the three countries. He has suggested that ARIMA models are suitable for forecasting the COVID-19 prevalence for the upcoming days.

Chintalapudi et al. [9], adopted seasonal ARIMA model for forecasting of COVID-19 cases in Italy using the data till 31st March 2020. They have analysed the impact of two months lockdown in Italy, and observed decrement in the confirmed cases and growth in the recovered cases due to lockdown.

Alabi et al. [10] have adopted the Facebook Prophet model to forecast spread of COVID-19. They have performed prediction for confirmed and death cases. Their forecasting accuracy of Prophet was 79.6% for the data from WHO between 7th April to 3rd May 2020.

Parikshit et al. [11] have presented medical perspective of COVID-19, and prediction using Prophet model. They have recommended Prophet for prediction due to open-source algorithm, accuracy, and faster data fitting. Using the Prophet model, they have predicted infected cases worldwide as 1.6 million by the end of May 2020, and 2.3 million by the end of June 2020.

3. Data and Sources of Data

The data we used in this research work is compiled by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) from various sources including the World Health Organization (WHO), DXY.cn, BNO News, National Health Commission of the People's Republic of China (NHC), China CDC (CCDC), Hong Kong Department of Health, Macau Government, Taiwan CDC, US CDC, Government of Canada, Australia Government Department of Health, European Centre for Disease Prevention and Control (ECDC), Ministry of Health Singapore (MOH), and others. JHU CCSE maintains the data on the 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository on Github. This is a daily updating version of COVID-19 Data Repository. Fields available in the data include Province/State, Country/Region, Last Update, Confirmed, Suspected, Recovered, Deaths. The data consists of more than three different datasets named time_series_covid19_confirmed_global, time_series_covid19_recovered_global, time_series_covid19_deaths_global and csse_covid_19_daily_reports/04-25-2020. The time_series_covid19_confirmed_global contain total of confirmed cases till date, time_series_covid19_deaths_global contains the total number of deaths till date, time_series_covid19_recovered_global contains total recovered cases till date and csse_covid_19_daily_reports/04-25-2020 contains the country wise deaths, confirmed cases, active cases and recoveries. We used the time_series_covid19_confirmed_global, time_series_covid19_deaths_global and csse_covid_19_daily_reports/04-25-2020 datasets for our research.

4. Data Analysis through Visualization

All the datasets contains both numerical and Character type data. The Confirmed cases, and deaths dataset don not contains null or missing values except the daily report dataset, which has some missing values in Admin2, province state, lat, and long attributes. Confirmed cases dataset and Death dataset contains 285 rows and the columns are incremented by 1 each day. The Daily report dataset contains 12 columns and 3162 rows. The Confirmed cases dataset and Death dataset have first four columns (province/state, country/region, lat and long) which are irrelevant to our research, So, we first removed these columns from our datasets to make aur dataset ready for some calculations.

As of 4th of December 2022, there have been 645274750 confirmed cases and 6640913 Deaths with the covid-19 pandemic and the figures keep changing with each passing date.

	Total confirmed	total Deaths
0	645274750	6640913

Figure 1: total cases and deaths

Country with lowest mortality Ratio = New Zealand (0.001183)
Country with Highest mortality Ratio = MS Zaandam (0.222222) and Yamen (0.181381).
State with most cases = England (UK) with 19436864 cases, 171981 deaths, and 0.0088 as mortality rate
State with most Deaths = Suo Paulo (Brazil) with 172074 Deaths
State with highest mortality rate = Puebla (Mexico) with 186704 cases, 16510 deaths and 0.088429 as mortality rate
State with least Cases = Tibet (China) with only 1 case.

As we know there were some NaN value in our daily report dataset. So, we first detect them and then removed them appropriately. To detect a NaN value, we make an empty list the search for NaN value in our dataset. On the presence of NaN values, we return the NaN index to the empty list, then remove the entire row with the help of indexes and then we delete the cases, deaths and recoveries corresponding to that index from total cases and total deaths.

Most affected countries:

USA has more than twice confirmed cases as compared to India, Brazil, and France. It also has thrice confirmed cases as compared Turkey, Russia, Kores, Italy and UK. USA also has more than 1/3 cases than rest of the countries labelled as others in the below figure.

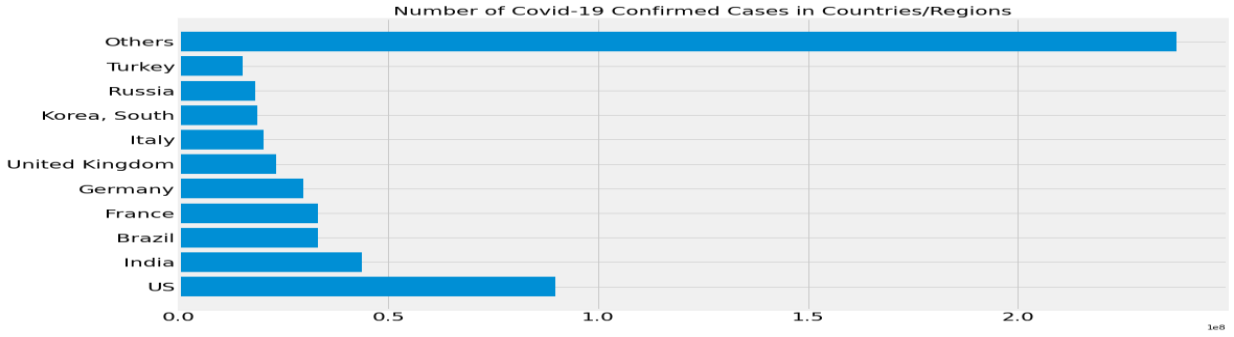

Figure 2: Top affected countries.

4.1 World cases and world deaths since beginning

World Deaths and world confirmed cases increased drastically. The graphs below show very less cases in 2022 as compared to cases in the year 2020 and 2021 and current trends show that covid-19 cases and covid-19 related deaths are decreasing. Every day there is decrease in the

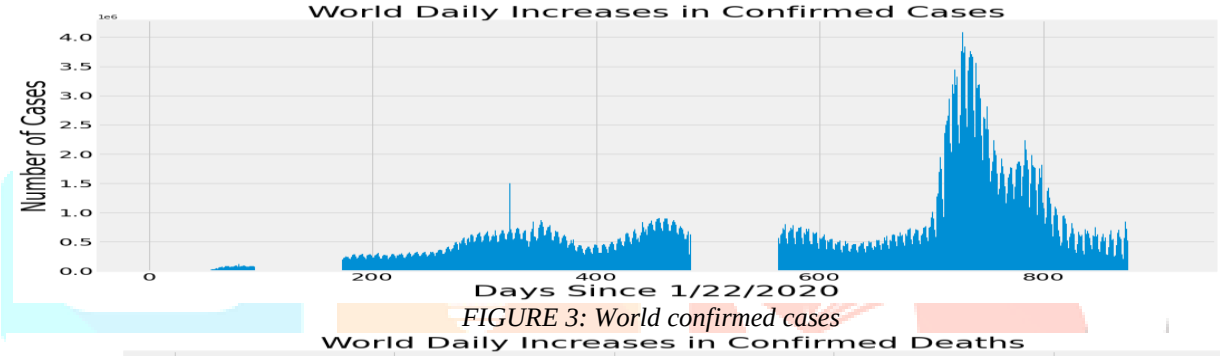

FIGURE 3: World confirmed cases

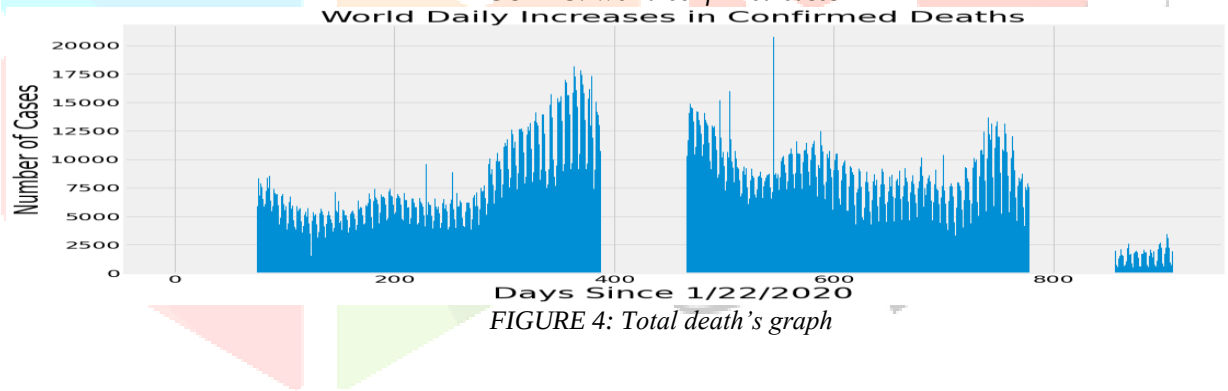

FIGURE 4: Total death's graph

4.2 Curve comparison for top affected countries:

1): Confirmed Cases

USA has more covid-19 cases as compared to other countries. India, Germany, France also has slight spike in covid cases. China and Spain have very less cases as compared to other countries. The curves curve reflects the fact.

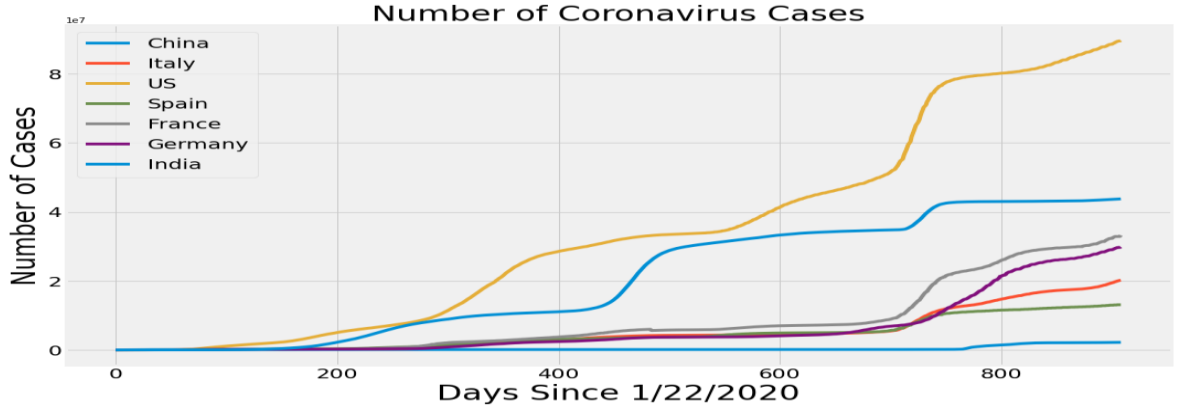

FIGURE 5: Countries cases comparison

2): Deaths

Deaths in USA and India were Very increasing rapidly. Every day there were deaths in lacks in both countries. China has very less Deaths due to covid-19.

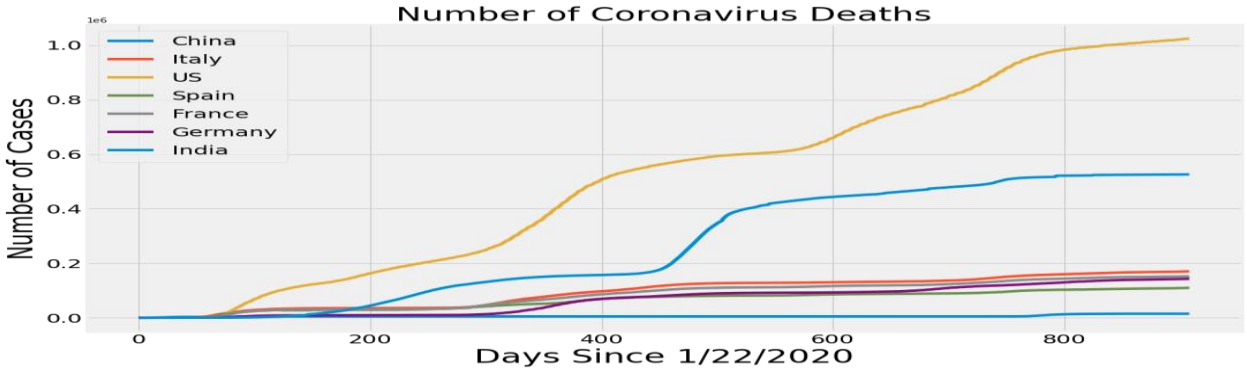

FIGURE 6: Countries deaths comparison

5. RESEARCH METHODOLOGY

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset. The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

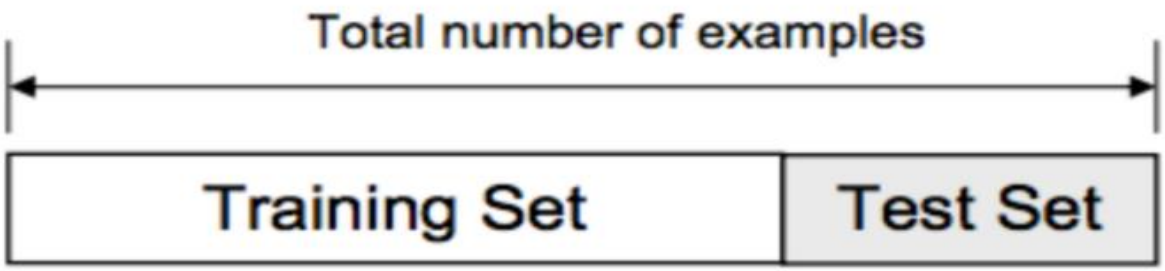

Figure 3: Train test split

We split our datasets into 75:25 ratios. Our training sets contains 75% of the entire datasets and test sets contains the 25% of the datasets.

The research we were working on is time series based and in time series-based projects prediction are done in correspondence to future time. We train our model on historical data and add future dates to our test dataset and model makes prediction for future dates according to the trends in historical data. We are making prediction about Covid-19 confirmed cases and deaths for next 20 days in future. In order to make this happen we added the dates of next 20 days to our dataset.

Before fitting a model and making predictions, we first scale our data using polynomial features and standard scaling. Often, the input features for a predictive modeling task interact in unexpected and often nonlinear ways. These interactions can be identified and modeled by a learning algorithm. Another approach is to engineer new features that expose these interactions and see if they improve model performance. Additionally, transforms like raising input variables to a power can help to better expose the important relationships between input variables and the target variable.

These features are called interaction and polynomial features and allow the use of simpler modeling algorithms as some of the complexity of interpreting the input variables and their relationships is pushed back to the data preparation stage. Sometimes these features can result in improved modeling performance, although at the cost of adding thousands or even millions of additional input variables. The formula for calculating the number of the polynomial features is **N (n, d) = C (n + d, d)** where **n** is the number of the features, **d** is the degree of the polynomial, **C** is binomial coefficient(combination).

Standard Scaler standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation. Standard Scaler does not meet the strict definition of scale I introduced earlier. Standard Scaler results in a distribution with a standard deviation equal to 1. The variance is equal to 1 also, because variance = standard

deviation squared. And 1 squared = 1. Standard Scaler makes the mean of the distribution approximately 0. Below is the formula of standard scaler. The formula scales the entire data with zero mean and unit variance altogether

$$z = \frac{x - \mu}{\sigma}$$

Equation 1: standard scaler formula

We use the linear regression by setting it's Normalize parameter to True and Fit Intercept parameter to False. For LSTM we used 2 Dense layers. We also used the Dropout Value of 0.6 and 0.6 after each layer to avoid overfitting. Uniform weight distribution is used as kernel initializer and Relu is used as activation function. 100 epochs with 0.001 as call backing with and 20 as patience. Input Batch size for a layer is 32.

6. Results

Both Algorithms (Linear regression and LSTM) worked while forecasting confirmed cases. Linear regression performed most well while forecasting covid confirmed cases than LSTM. There is little difference between the confirmed cases forecasting by Linear regression and LSTM. In fact, Linear regression forecasted almost same total confirmed cases as of mentioned in the dataset from the beginning. In case of total deaths prediction, performance of both the implemented algorithms are not much promising. Still LSTM prediction is more accurate while predicting the deaths as compared to Linear regression prediction. So, we can say that Linear regression has shown better results while forecasting the total confirmed cases and LSTM has shown better results while forecasting the total deaths till date. below are the graphs of two implemented algorithms–Polynomial regression and LSTM on total Confirmed Cases data and total Covid Deaths Data. Graphs show the same results mentioned by us above. The performance of the Linear Regression was better than LSTM. Linear regression and LSTM predicts increase in covid-19 cases and there will be significant decrease in deaths for next 20 days. Overall, the performance of the of Linear regression and LSTM is better in forecasting the total deaths as the Actual data curve and prediction curve are close to each other.

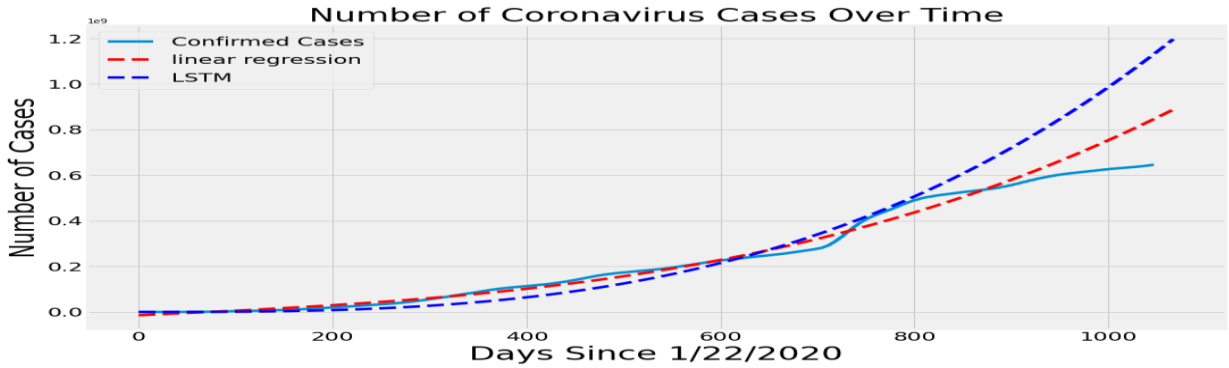

Figure 4: Total Cases prediction by both algorithms.

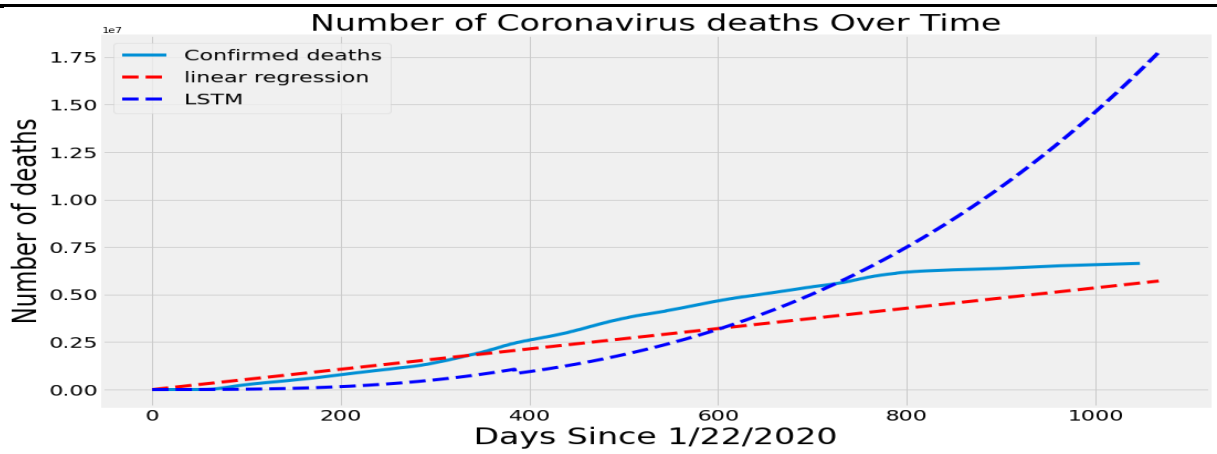

Figure 5: Total Deaths prediction by both algorithms

Initially both the implement algorithms on total confirmed cases data were predicting same numbers cases as were given in dataset, but with the advent of data the accuracy of both the algorithms starts decreasing drastically.

	Mean Absolute Error	**Mean Square Error**
LSTM	Deaths (4956354.600760456) Cases (145679538.3878327)	Deaths (31531631167095.17) Cases (3.4553418135601624e+16)
Linear Regression	Deaths (3560271.313475994) Cases (6811230.45184578)	Deaths (14962315469979.146) Cases (7602388440329945.0)

Linear Regression total cases Prediction for upcoming 20 days

	Date	Predicted number of Confirmed Cases Worldwide
0	12/05/2022	847827153.0
1	12/06/2022	849892594.0
2	12/07/2022	851961579.0
3	12/08/2022	854034111.0
4	12/09/2022	856110194.0
5	12/10/2022	858189832.0
6	12/11/2022	860273027.0
7	12/12/2022	862359783.0
8	12/13/2022	864450103.0
9	12/14/2022	866543992.0
10	12/15/2022	868641451.0
11	12/16/2022	870742485.0
12	12/17/2022	872847096.0
13	12/18/2022	874955289.0
14	12/19/2022	877067066.0
15	12/20/2022	879182431.0
16	12/21/2022	881301387.0
17	12/22/2022	883423938.0
18	12/23/2022	885550087.0
19	12/24/2022	887679837.0

Table 1: Total cases for future 20 days

LSTM total deaths prediction for upcoming 20 days

	Date	Predicted number of deaths Worldwide
0	12/05/2022	15880182.0
1	12/06/2022	15926456.0
2	12/07/2022	15972824.0
3	12/08/2022	16019277.0
4	12/09/2022	16065822.0
5	12/10/2022	16112456.0
6	12/11/2022	16159181.0
7	12/12/2022	16205994.0
8	12/13/2022	16252900.0
9	12/14/2022	16299894.0
10	12/15/2022	16346982.0
11	12/16/2022	16394158.0
12	12/17/2022	16441427.0
13	12/18/2022	16488783.0
14	12/19/2022	16536234.0
15	12/20/2022	16583774.0
16	12/21/2022	16631403.0
17	12/22/2022	16679125.0
18	12/23/2022	16726939.0
19	12/24/2022	16774842.0

Table 2: Total deaths for future 20 days

7. Conclusion

The COVID-19 pandemic has severely affected the lives of people in every country across the globe. This condition was worse in certain areas. Currently, there is a cure for this disease and even the odds of predicting the severity of this pandemic are very much in number. Still, deep learning and machine learning models could be applied to make predictions about this disease. For this purpose, we used the time series datasets, collected by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) from various sources including the World Health Organization (WHO), DXY.cn, BNO News, National Health Commission of the People's Republic of China (NHC), China CDC (CCDC), Hong Kong Department of Health, Macau Government, Taiwan CDC, US CDC, Government of Canada, Australia Government Department of Health, European Centre for Disease Prevention and Control (ECDC), Ministry of Health Singapore (MOH), and others, for proposing two DL and ML-based prediction models. one ML-based prediction models, Linear Regression and one DL Based model, LSTM, have been evaluated using time series data from three datasets. We used the Python language for developing and implementing the models. In step 1, we Transformed our dataset using standard scaling and polynomial Feature to scale the data. Before transforming the data, we make the exploratory data analysis and reveal the most and least covid affected country. We also reveal the country with highest and lowest mortality rate. Also, we mine the top most 20 affected countries and states due to covid. Among the two implemented algorithms Linear Regression performs most well on total Cases dataset and LSTM performed slightly well on deaths data. After model implementation, we saw what will be the trend of covid related confirmed cases and deaths all over the world for next 20 days in future

8. References

1. P. M. Hosseiny, S. Kooraki, A. Gholamrezanezhad, S. Reddy, L. Myers Radiology perspective of coronavirus disease 2019 (COVID-19): lessons from severe acute respiratory syndrome and Middle East respiratory syndrome
Am. J. Roentgenol., 214 (5) (2020), pp. 1078-1082

2. A. Kumari, M. Sood Implementation of SimpleRNN and LSTMs based prediction model for coronavirus disease (Covid-19)
Editor (Ed.), Book Implementation of Simple RNN and LSTMs Based Prediction Model for Coronavirus Disease (Covid-19), IOP Publishing (2021)
pp. 012015

3. M. Jamal, M. Shah, S.H. Almarzooqi, H. Aber, S. Khawaja, R. El Abed, Z. Alkhatib, L.P. Samaranayake
Overview of transnational recommendations for COVID-19 transmission control in dental care settings
Oral Dis., 27 (S3) (2021), pp. 655-664

4. H. Apaydin, H. Feizi, M.T. Sattari, M.S. Colak, S. Shamshirband, K.-W. Chau
Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting
Water, 12 (5) (2020), p. 1500, 10.3390/w12051500

5. Xiaolei Zhang, Renjun Ma and Lin Wang, "Predicting turning point duration and attack rate of covid-19 outbreaks in major western countries", Chaos Solitons & Fractals, pp. 109829, 2020.

6. Barbara Oliveiros, Liliana Caramelo, Nuno C Ferreira and Francisco Caramelo, "Role of temperature and humidity in the modulation of the doubling time of covid-19 cases", medRxiv, 2020.

7. Matheus Henrique, Dal Molin Ribeiro, Ramon Gomes da Silva, Viviana Cocco Mariani and Leandro dos Santos Coelho, "Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for Brazil", Chaos Solitons & Fractals, pp. 109853, 2020.

8. Muccio Fanelli and Francesco Piazza, "Analysis and forecast of covid-19 spreading in China Italy and France", Chaos Solitons & Fractals, vol. 134, pp. 109761, 2020.

9. Nalini Chintalapudi, Gopi Battineni and Francesco Amenta, "Covid-19 disease outbreak forecasting of registered and recovered cases after sixty-day lockdown in Italy: A data driven model approach", Journal of Microbiology Immunology and Infection, 2020.

10. Barbara Oliveiros, Liliana Caramelo, Nuno C Ferreira and Francisco Caramelo, "Role of temperature and humidity in the modulation of the doubling time of covid-19 cases", medRxiv, 2020.

11. Parikshit N Mahalle, Nilesh P Sable, Namita P Mahalle and Gitanjali R Shinde, "Data analytics: Covid-19 prediction using multimodal data", Preprints, 2020. first pass the betas coefficients are computed by using regression.