Predictive Analysis of COVID-19 Virus Pandemic Spread with the Support of Efficient Machine Learning Techniques

Dr.M.Pompapathi[1], K Shyam Sunder Reddy[2]

[1]*Associate Professor,* [2]*Assistant Professor*

[1,2]*Department of IT*

[1]*RVR & JC College of Engineering, Guntur, Andhra Pradesh*

[2]*Vasavi College of Engineering, Hyderabad*

[1]*manasani.pompapathi@gmail.com,* [2]*shyamd4@staff.vce.ac.in*

Abstract

COVID-19, the pandemic that has got the whole world worried, also called 2019-nCoV or 2019 novel coronavirus, is caused due to the virus called SARS CoV-2 which happens to be one of the kind of coronaviruses. The medical fraternity has not been able to find a cure to this dreadful disease till day. However, rigorous work is being carried out towards the same. During such times, it becomes difficult for the governments of various countries across the globe to maintain integrity and peace among the minds of the people. Therefore, for the government to take thoughtful decisions in the interest of the people of their country, they require a model or a way in which they can predict and analyze the way the disease is spreading across their territory. This paper puts forward a statistical machine learning model which uses the regularly updated data set consisting number of confirmed cases, active cases, recovered cases and deaths caused due to COVID-19 to predict and visualize the spread of the disease in different parts of the world. Thus, providing support to the decisions taken by the government officials.

Keywords: *COVID-19, Machine Learning, SARS CoV-2, Geospatial Analysis, ARIMA Model, 2019-nCoV*

1. Introduction

The whole world is currently on high alert due to the prevailing pandemic called the COVID-19, more specifically known as the ‘SARS CoV-2’ virus epidemic. There are numerous corona viruses that have been known to the human race till day. Yet the nature surprises us with something new and challenging every time, and this time it’s SARS CoV-2.

Firstly, it’s important to get familiar with its name’s nomenclature. COVID-19 is an acronym for “Corona Virus Disease-2019”. Similarly, SARS is an acronym for “Severe Acute Respiratory Syndrome”.

Bats are a natural reservoir for many viruses, that is to say, the virus can replicate inside the mammal without taking a toll on it. The bats then pass it on to other animals and finally the virus is passed down to humans. This is what happened in Wuhan, China. After the first person was infected, the virus spread from person to person, akin to the escalating wild fire in the forests, by people coughing and sneezing. Just a single cough can spread 3000 droplets all over the place. In case, the virus lands in air, it can stay on a particle and be suspended in the atmosphere for 3 hours. If it lands on a cardboard box, it can last there for about 24 hours and if it goes on to a plastic, it can stay there for to 2-3 days. So basically, the world is up against an enemy, which can neither been see nor know where it is in our surroundings.

Now it’s time to get into details of the virus, it’s structure and various other factors. Corona virus is a family of viruses that have little spikes on the surface that look like a crown, which is how it got its name. There are many types of corona viruses as stated previously and they don't have any special names and usually they cause common cold. There is SARS corona virus and its outbreak happened in the year 2002. Then there is MERS corona virus and its outbreak happened in the year 2012. Here, MERS is an acronym for "Middle East Respiratory Syndrome".When the virus enters an individual's body, it uses the proteins (crown-shaped structure) on its surface as shown in figure 1, to get into one's cells. These proteins are like a key that unlocks receptors on the outer surface of the cell. It is believed that specifically the ACE-2 receptors on human lungs are being used by the corona virus to get inside of our cells. Once the virus is inside, it releases the RNA, it's genetic material, and tricks our body cells to make copies of the virus. Therefore, our body cells become a factory that produces the virus, counting up to some thousands without even realizing. In the initial stages, when the virus is replicating one doesn't actually show any kind of illness symptoms. This is called the "Incubation Period". This period of incubation varies from virus to virus. But, in the case of SARS CoV-2 it is believed that the incubation period is 4 days.

As far as the symptoms are concerned, few common and easily identifiable symptoms like fever, sore throat, cough and difficulty in breathing were recognised in the patients infected by the virus. Less common symptoms are nausea, vomiting, diarrhoea which was observed in 3-5% of the infected people. Most of the people getting infected by the virus can recover, there's no need to worry or panic about it. But few people develop, what can be called as, Acute Respiratory Distress Syndrome (ARDS) and this is the condition where the lungs develop inflammation. The ability of acquiring oxygen out of air is damaged and eventually there is a shortage of oxygen supply for the red blood cells (RBC). There are many factors that cause ARDS, like sepsis, pneumonia and even pancreatitis.

So, the overall motive of this paper is to control the spread of this disease by predicting and analysing the situation in various countries and coming up with some numbers based on which the government and health facilities of a country can come with measures to stop the spread of the virus.

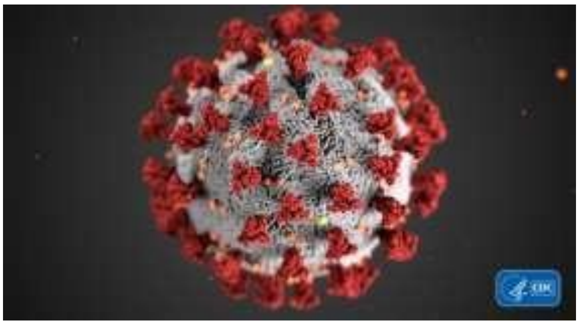


Figure 1 Structure of SARS CoV-2

2. Literature Survey

The system currently in use for the prediction and analysis of COVID-19 is able to give the details of the reported cases of people getting infected by the virus in various parts of the country.

Based on the reported cases, the government and health facilities of the country are taking measures to contain the virus. But this isn't the efficient way to stop the spread of a contagious disease like COVID-19. The government and health facilities must be provided with prior predictions on how the virus could spread, thus sparing government some time to take wise decisions, to make the situation better and reduce the burden on the health facilities of the country.

Authorities are finding it difficult to build an effective and efficient models for predicting the outbreak of such viruses as the infection continues to spread and the social media traffic around it accumulates, so the amount of noise that accumulates is huge, which has to be filtered through before meaningful trends can be derived out of it.

Hence, so far, most models employed for tracking and forecasting do not use AI methods or techniques. Instead, most forecasting models are established on epidemiological models, so-called SIR models, which is an acronym for the population of an area that is "Susceptible, Infected and Removed".

3. Implementation

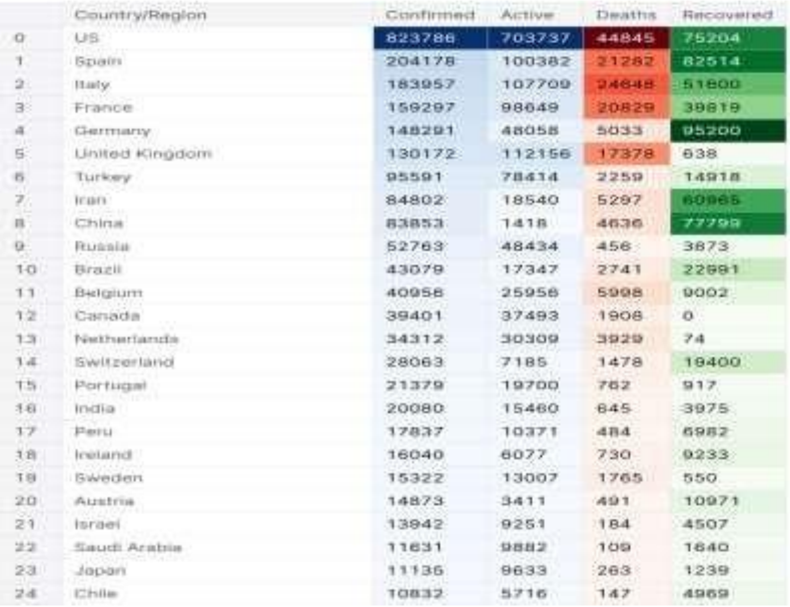
This paper puts forward the work carried out in predicting and analysing the current prevalent pandemic of SARS CoV-2 or COVID-19. This is done using the statistical technique of time series forecasting as the basic concept to train the machine learning model in predicting various parameters concerned with the outbreak of SARS CoV-2.

3.1 Exploratory Data Analysis

This section of the paper deals with implementing the tracking of COVID-19 pandemic by importing the required data sets as shown in figure 3.1, that are being updated on a regular basis and writing few lines of code in Python language, which also happens to the language in high-demand across the globe.

The data set has the following attributes:

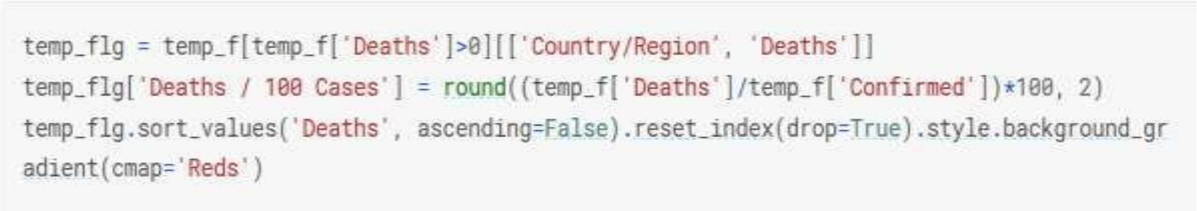
- Province/State	Country/Region
- Lat (Latitude)	Long (Longitude)
- Date	Confirmed Cases
- Number of Deaths	Number of Recoveries
- Total Number of Cases	Currently Active Cases



	Country/Region	Confirmed	Active	Deaths	Recovered
0	US	823786	703737	44845	75204
1	Spain	204178	100382	21282	82514
2	Italy	183957	107709	24648	51600
3	France	159297	98649	20829	39819
4	Germany	148291	48058	5033	95200
5	United Kingdom	130172	112156	17378	638
6	Turkey	95591	78414	2259	14918
7	Iran	84802	18540	5297	60965
8	China	83853	1418	4636	77799
9	Russia	52763	48434	456	3873
10	Brazil	43079	17347	2741	22991
11	Belgium	40956	25956	5998	9002
12	Canada	39401	37493	1908	0
13	Netherlands	34312	30309	3929	74
14	Switzerland	28063	7185	1478	19400
15	Portugal	21379	19700	762	917
16	India	20080	15460	645	3975
17	Peru	17837	10371	484	6982
18	Ireland	16040	6077	730	9233
19	Sweden	15322	13007	1765	550
20	Austria	14873	3411	491	10971
21	Israel	13942	9251	184	4507
22	Saudi Arabia	11631	9882	109	1640
23	Japan	11136	9633	263	1239
24	Chile	10832	5716	147	4969

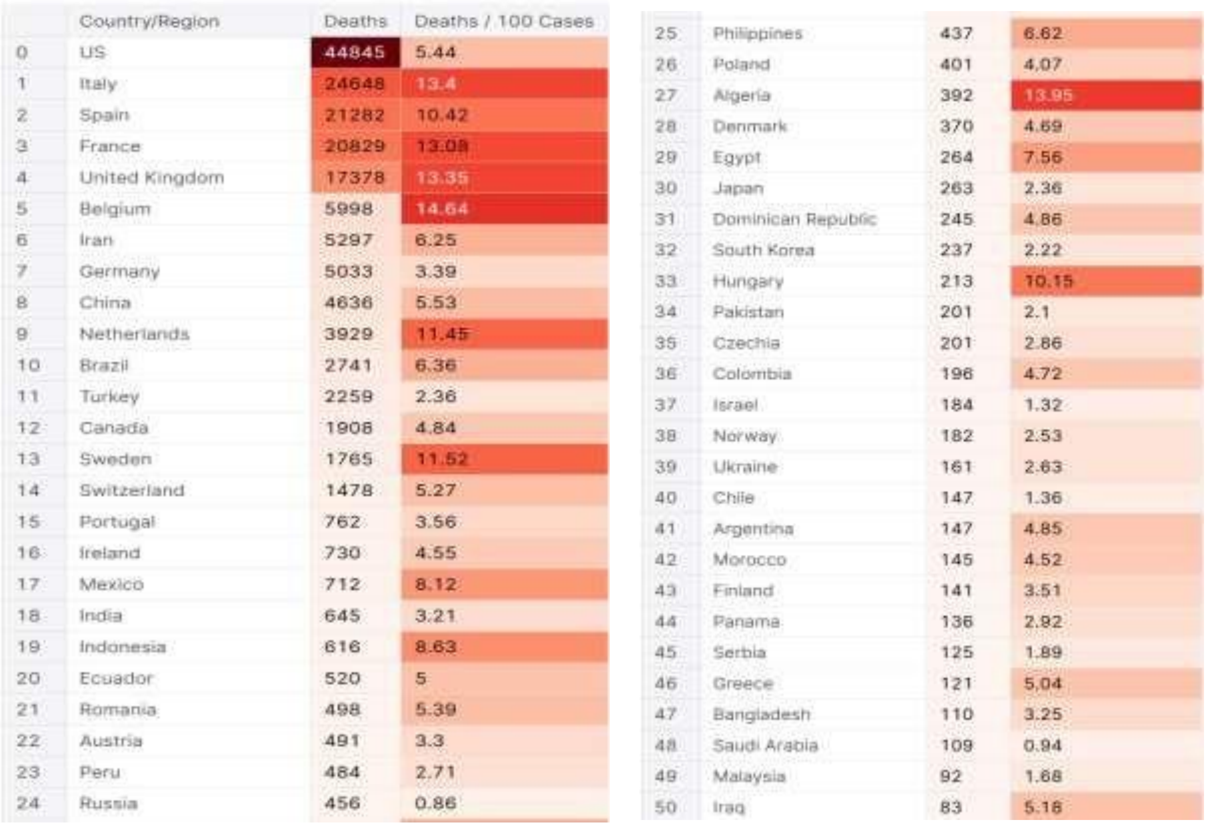
Figure 3.1 A Snippet of Data Set Representing Number of Confirmed, Active, Deaths and Recoveries

It is essential to predict the number of deaths per 100 cases that were recorded. Therefore, after writing a snippet of Python code as shown in figure 3.2, the following results for various places across the world were obtained as shown in figure 3.3.



```
temp_flg = temp_f[temp_f['Deaths']>0][['Country/Region', 'Deaths']]
temp_flg['Deaths / 100 Cases'] = round((temp_f['Deaths']/temp_f['Confirmed'])*100, 2)
temp_flg.sort_values('Deaths', ascending=False).reset_index(drop=True).style.background_gr
adient(cmap='Reds')
```

Figure 3.2 Snippet of Python Code



	Country/Region	Deaths	Deaths / 100 Cases
0	US	44845	5.44
1	Italy	24648	13.4
2	Spain	21282	10.42
3	France	20829	13.08
4	United Kingdom	17378	13.35
5	Belgium	5998	14.64
6	Iran	5297	6.25
7	Germany	5033	3.39
8	China	4636	5.53
9	Netherlands	3929	11.45
10	Brazil	2741	6.36
11	Turkey	2259	2.36
12	Canada	1906	4.84
13	Sweden	1765	11.52
14	Switzerland	1478	5.27
15	Portugal	762	3.56
16	Ireland	730	4.55
17	Mexico	712	8.12
18	India	645	3.21
19	Indonesia	616	8.63
20	Ecuador	520	5
21	Romania	498	5.39
22	Austria	491	3.3
23	Peru	484	2.71
24	Russia	456	0.86
25	Philippines	437	6.62
26	Poland	401	4.07
27	Algeria	392	13.95
28	Denmark	370	4.69
29	Egypt	264	7.56
30	Japan	263	2.36
31	Dominican Republic	245	4.86
32	South Korea	237	2.22
33	Hungary	213	10.15
34	Pakistan	201	2.1
35	Czechia	201	2.86
36	Colombia	196	4.72
37	Israel	184	1.32
38	Norway	182	2.53
39	Ukraine	161	2.63
40	Chile	147	1.36
41	Argentina	147	4.85
42	Morocco	145	4.52
43	Finland	141	3.51
44	Panama	136	2.92
45	Serbia	125	1.89
46	Greece	121	5.04
47	Bangladesh	110	3.25
48	Saudi Arabia	109	0.94
49	Malaysia	92	1.68
50	Iraq	83	5.18

Figure 3.3 Number Of Deaths Per 100 Cases In Various Countries/Regions

3.2 Time Series Forecasting

A sequence of metric observed and noted over regular time duration is said to be time series. Forecasting is nothing but predicting the future values using the past data as shown in figure 3.4. This forecasting in time series has its own importance and commercial value in industries.

By analysing this forecasted results, business driven insights, planning can be done. So, it's very crucial in forecasting the values accurately in order to save capital and time. Forecasting of time series is classified into two types, they are

a. Univariate Time Series Forecasting.

b. Multi-Variate Time Series Forecasting.

a. Univariate Time Series Forecasting :

Forecasting the values in future with the help of past values in time series is said to be Univariate Time Series Forecasting.

b. Multi-Variate Time Series Forecasting :

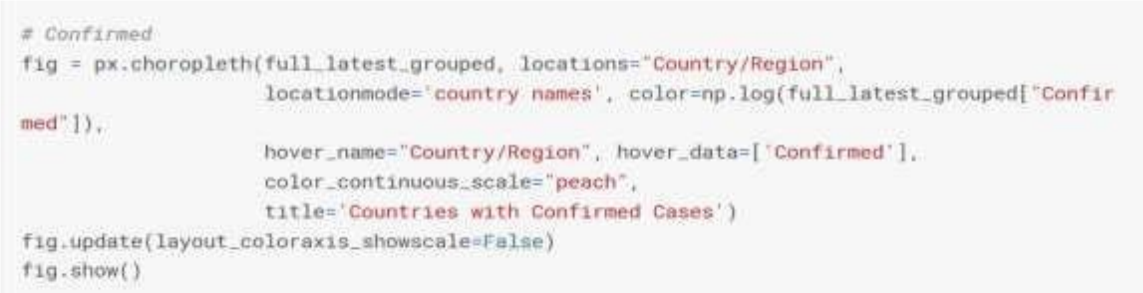
Each variable depends not only on its past values but also has some dependency on other variables as well. A Multivariate time series there is more than one-time dependent variable.

The prediction and analysis of COVID-19 is done based on "Univariate Time Series Forecasting" in this paper.

Figure 3.4 Time Series Analysis/Forecasting

"Geospatial Analysis" has been performed for various countries across the globe. An analysis technique wherein one gathers, displays and manipulates images, satellite photographs, GPS location and historical data, that is described by geographic coordinates i.e., latitude and longitude of a place can be termed as "Geospatial Analysis".

Here are some of the results that were obtained from the geospatial analysis shown in figure 3.6, by writing some Python code shown in figure 3.5,



```
# Confirmed
fig = px.choropleth(full_latest_grouped, locations="Country/Region",
                    locationmode='country names', color=np.log(full_latest_grouped["Confir
med"]),
                    hover_name="Country/Region", hover_data=['Confirmed'],
                    color_continuous_scale="peach",
                    title='Countries with Confirmed Cases')
fig.update(layout_coloraxis_showscale=False)
fig.show()
```

Figure 3.5 Snippet of Code for Geospatial Analysis

The following are the results that were obtained from the geospatial analysis, the intensity of the colour denotes the severity of the disease in that country.

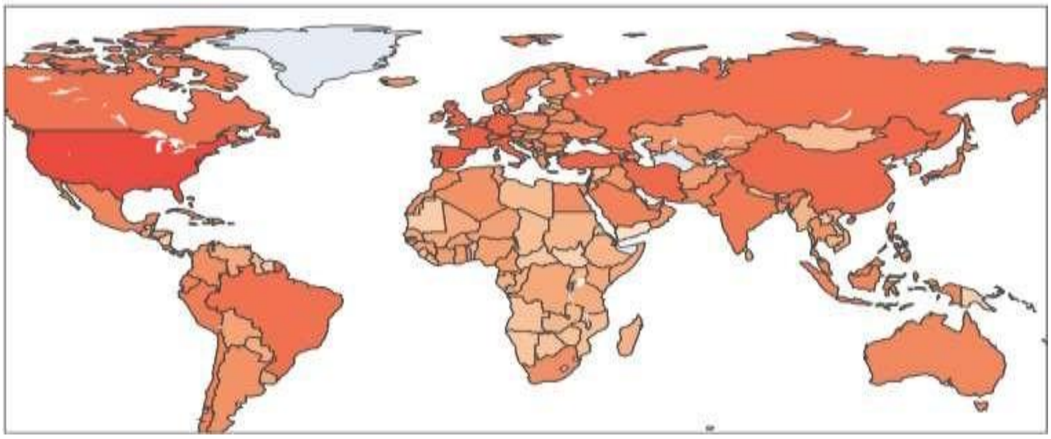


Figure 3.6 Geospatial Analysis Across the World (Visualization)

From the figure 3.6, one can infer that most of the world's population is affected by the virus. There are only few regions / countries like Greenland, Madagascar, Papua New Guinea etc.

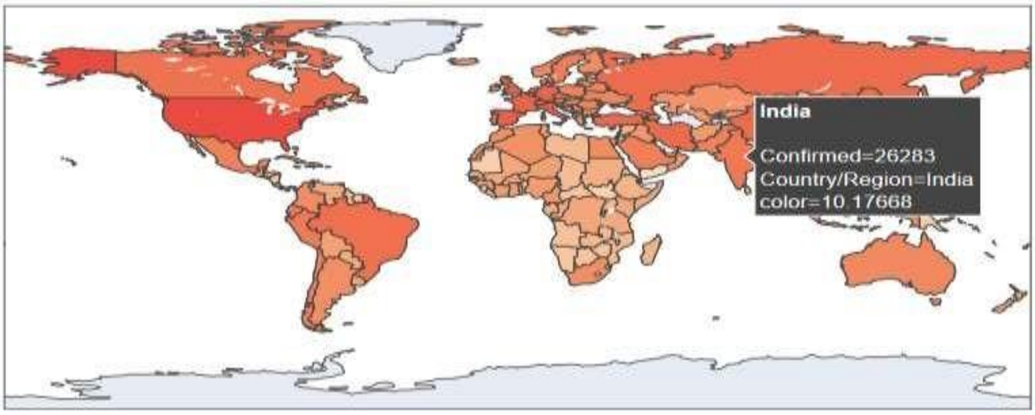


Figure 3.7 Geospatial Analysis Showing Confirmed Cases in India

As On 26 April 2020

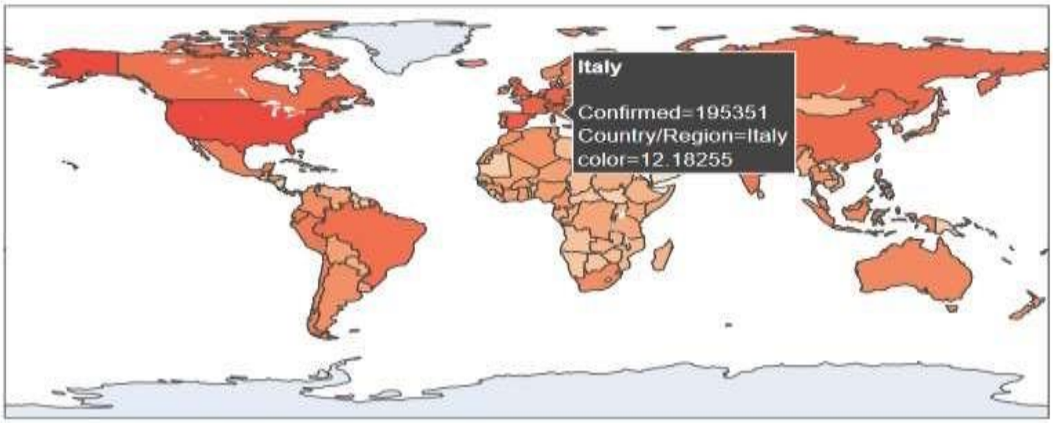


Figure 3.8 Geospatial Analysis Showing Confirmed Cases in Italy

As On 26 April 2020

The figures 3.7 and 3.8 show the results of the geospatial analysis for the countries India and Italy respectively. It shows the number of confirmed cases, the name of the country and intensity of the colour denotes the severity of the disease spread.

3.3 ARIMA Model

“Auto Regressive Integrated Moving Average” in short ARIMA explains a time series based on its past values. So, the equation obtained can be used to forecast future results.

To predict the confirmed cases and fatalities in up-coming weeks ARIMA model was used. ARIMA is a technique in time series analysis. As the cases and fatalities were changing from time to time and to understand the past data better to predict future trends, ARIMA model is used in implementing the project.

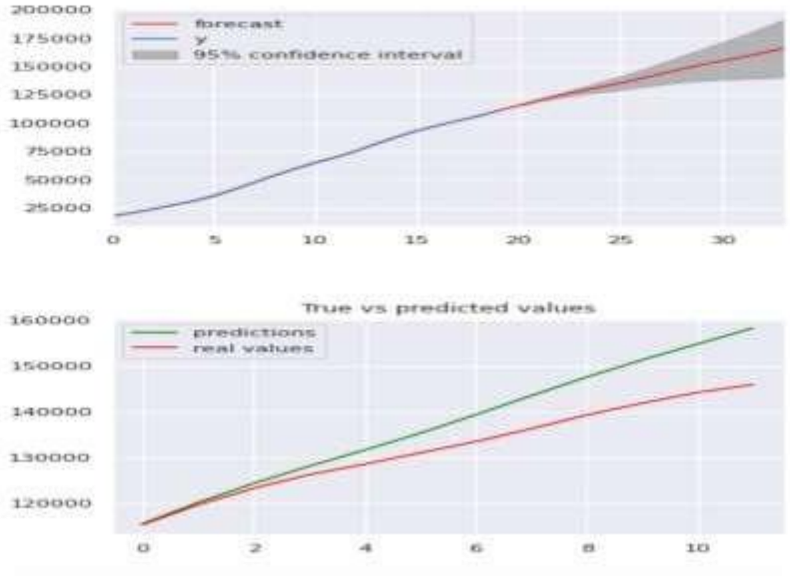


Figure 3.9 Graph Illustrating the Predicted Number Of Confirmed Cases in Upcoming Weeks

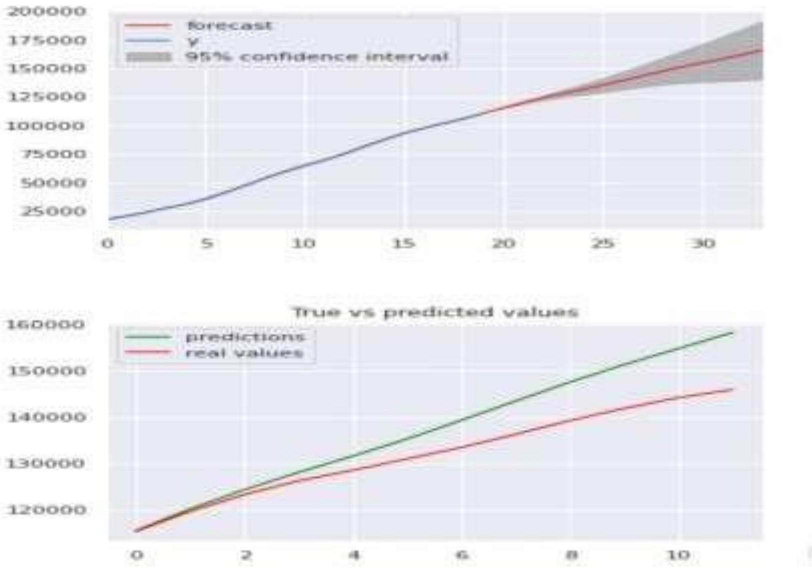


Figure 3.10 Graph Illustrating the Predicted Number Of Deaths in Upcoming Weeks

The above graphs in the figures 3.9 and 3.10 illustrate how close the predicted values of confirmed cases and deaths respectively, are going to be in the upcoming weeks. It can be observed that the predicted and real values are close enough.

4. Results

By using the concepts of machine learning, Python and efficient models like ARIMA, the following appreciable results were achieved as shown in the figures below. The accuracy of the model generated is also good enough to carry out the prediction and tracking tasks for COVID-19 pandemic across the world.

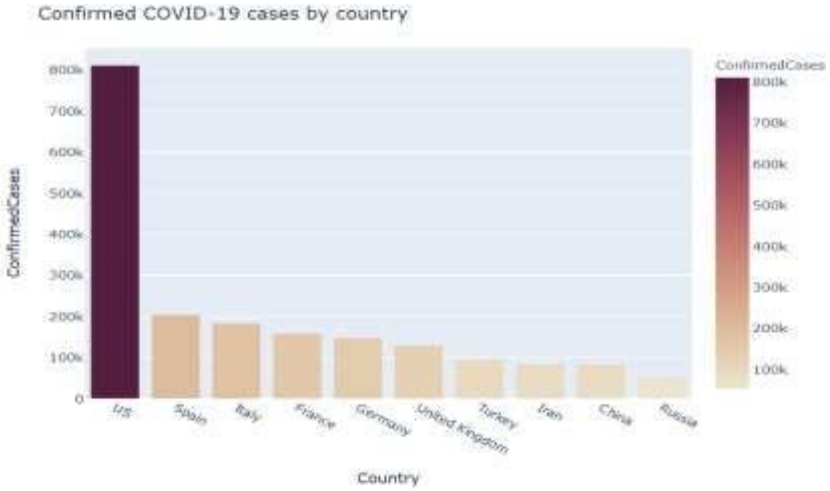


Figure 4.1 Graph Depicting the Number of Confirmed Cases in Various Countries

From the graph in the figure 4.1, one can infer that though the primary hub for the virus was in Wuhan, China, the number of confirmed cases were more in countries of U.S, Spain and Italy. This clearly shows that the city of Wuhan is one of the main cities of China, attracting people from various countries for various reasons like education, business etc.

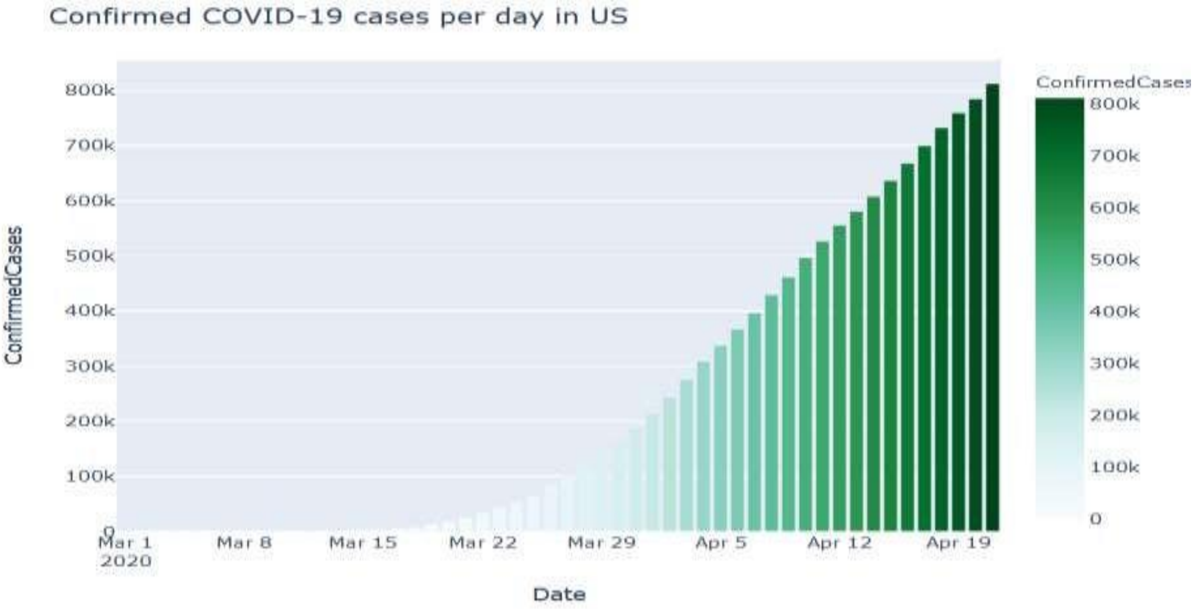


Figure 4.2 Histogram Depicting the Number of Confirmed Cases

Per Day in US

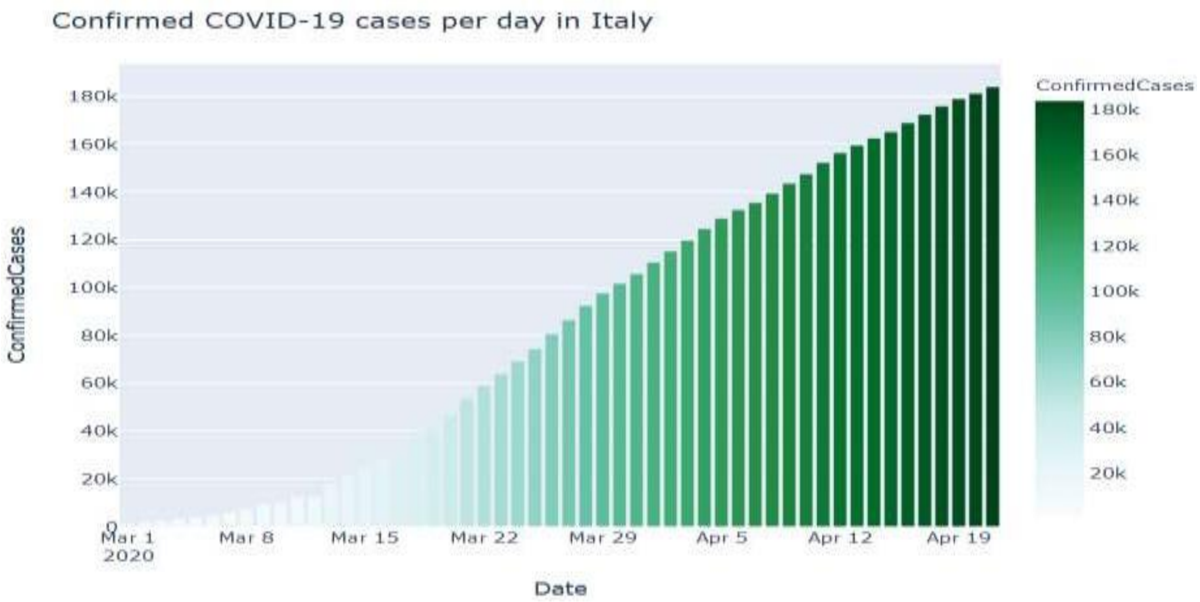


Figure 4.3 Histogram Depicting the Number of Confirmed Cases Per Day in Italy

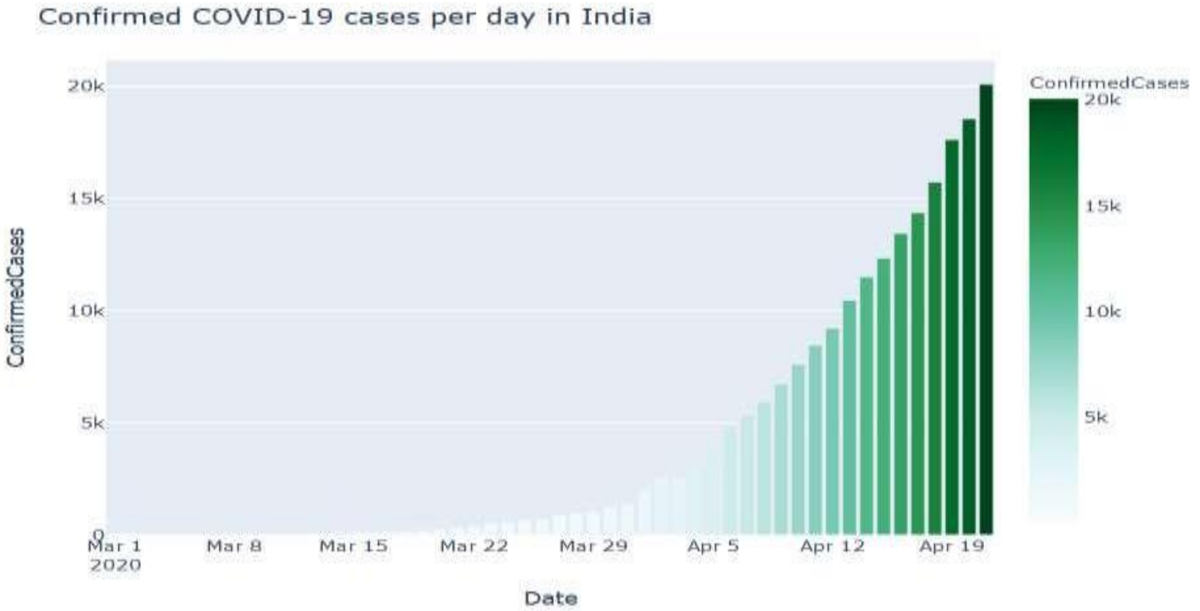


Figure 4.4 Histogram Depicting the Number of Confirmed Cases Per Day in India

One can infer from the histograms in the figures 4.2, 4.3, 4.4, for the example countries considered, which are the U.S, Italy and India respectively, how the cases are varying from day-to-day. It is observed that USA has a steep increase in the number of cases starting on from early days in the month of March. Upon looking into the number of cases being confirmed in the country of Italy, there no much difference in the situation to that of USA. The only difference being that the fatality rate of Italy is high when compared any other country in the world. India has a gradual increase in the number of cases that will be reported in the further days to come.

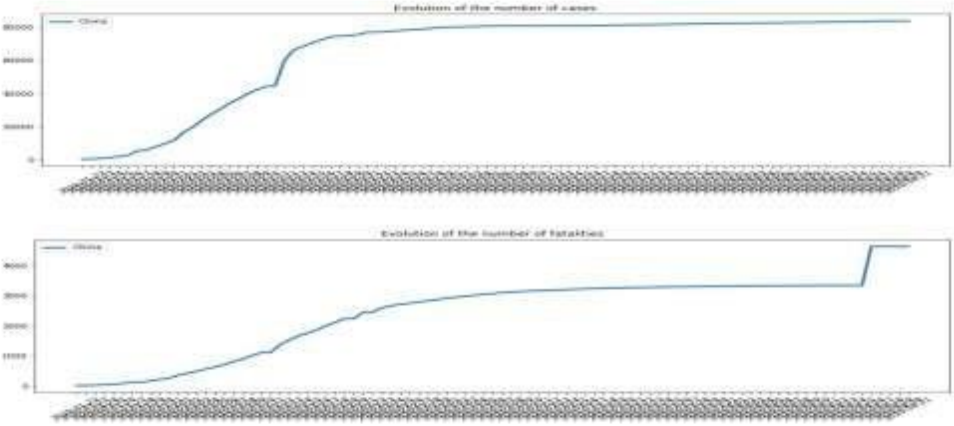


Figure 4.5 Time Series of China

The above figure 4.5 illustrates the evolution of cases and deaths in China over the time and is called time series of China.

The following result as shown in figure 4.6 is obtained on comparing the countries which have similar trends in the number of cases reported and number of fatalities.

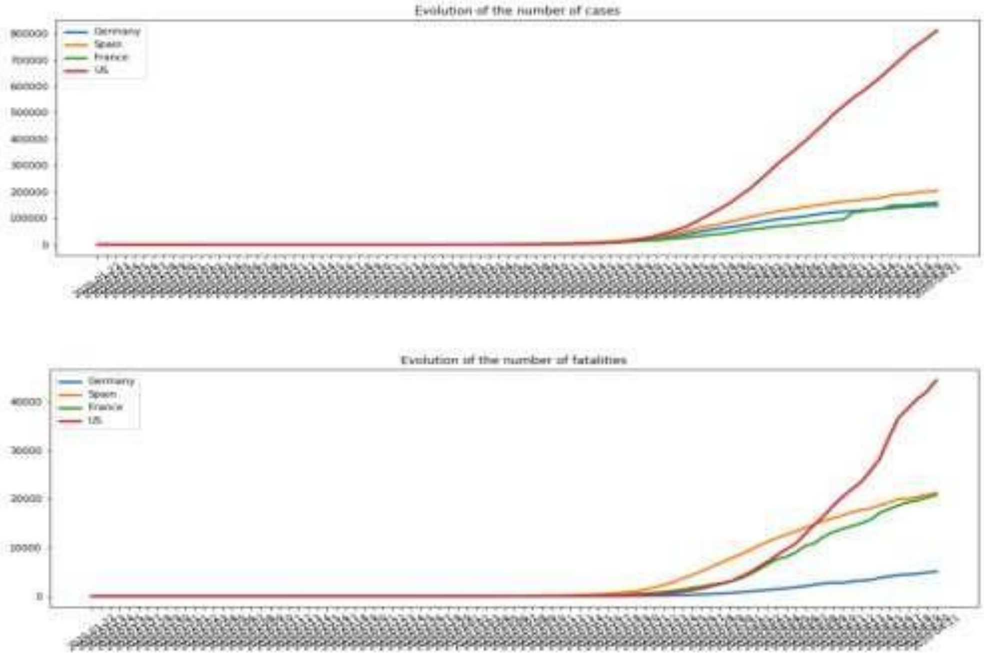


Figure 4.6 Graphs Depicting the Number of Cases and the Number of Fatalities

in the Countries of Germany, Spain, France and USA

5. Future Scope

This model presented in this paper will help the government of a country to take measures to stop the virus from spreading among the society, but doesn't provide any major support to the health facility of the country. During such times, the health facilities require more support and guidance from technology.

Therefore, in order to support the healthcare resource of any country, a model that is capable of detecting the patients, who have come into contact with the virus and have got infected is the need of the hour. So, one can use the X-ray of the lungs of the patients who have been bought for a test of COVID-19 to detect whether they are infected by the SARS CoV-2 virus or is it some other health issue. The X-ray of the lungs of a patient who has come in for a COVID-19 test can be compared with the X-ray of a healthy person lungs to detect whether the person is infected or not. This process can be done effectively and efficiently by a machine learning model.

6. Conclusion

The concepts of machine learning have paved their way into every possible existing fields these days. The field of healthcare is no exception. During the outbreak of such pandemic, the medical fraternity is also struggling to serve the affected people. Therefore, there is an immediate need of machine learning concepts to ease the work of the healthcare workers and the government in decision making for the well-being of the people.

This paper presents the work done in analyzing the number of cases and fatalities that are increasing around the globe. The number of deaths per 100 cases all around the world has been predicted, comparison of the number of positive cases, deaths and recoveries in various countries has been done and analyzed the trend across various territories using concepts like Univariate time series forecasting, geospatial analysis and ARIMA model. The results thus obtained are till 26th April 2020. This helps us in taking precautions on how to stop the spread of the virus. This will help the governments to take necessary measures such as arranging sufficient medical equipment in the hospitals and also warning the citizens about the situation if proper care is not taken.

Therefore, it is expected that the work carried out and communicated by this paper will be a great help to the medical fraternity and governments of countries of the world in eradicating the deadly virus.

References

[1] Lifang, Li., Qingpeng, Zhang., Xiao, Wang., Jun, Zhang., Tao, Wang., Tian-Lu, Gao., Wei Duan., Kelvin, Kam-fai, Tsoi., Fei-Yue Wang., 2020. Characterizing the Propagation of Situational Information in Social Media During COVID-19 Epidemic: A Case Study on Weibo. IEEE Digital Library.

[2] Feng Shi ; Jun Wang ; Jun Shi ; Ziyan Wu ; Qian Wang ; Zhenyu Tang ; Kelei He ; Yinghuan Shi ; Dinggang Shen., 2020. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19. IEEE Digital Library.

[3] Rae Yule Kim., 2020. The Impact of COVID-19 on Consumers: Preparing for Digital Sales. IEEE Engineering Management Review. IEEE Digital Library.

[4] Donghyeon Kim ; Soyoung Hong ; Sungwoo Choi ; Taeseon Yoon ., 2016. Analysis of transmission route of MERS coronavirus using decision tree and Apriori algorithm. 2016 18th International Conference on Advanced Communication Technology (ICACT). IEEE.

[5] Theo Wibisono ; Dionne M. Aleman ; Brian Schwartz., 2008. A non-homogeneous approach to simulating the spread of disease in a pandemic outbreak. 2008 Winter Simulation Conference. IEEE.

[6] Ummadi Janardhan Reddy, Pandluri Dhanalakshmi, Pallela Dileep Kumar Reddy Image Segmentation Technique Using SVM Classifier for Detection of Medical Disorders Ingénierie des Systèmes d'Information, Vol. 24, No. 2, pp. 173-176, April 2019, https://doi.org/10.18280/isi.240207 (Scopus) ISSN: 1633-1311 (print); 2116-7125 (online) Impact Factor : 0.409

[7] Singamaneni Kranthi Kumar, Pallela Dileep Kumar Reddy, Ramesh G, Venkata Rao Maddumala (2019). Image transformation technique using steganography methods using LWT technique. Traitement du Signal, Vol. 36, No. 3, pp. 233-237. June 2019, https://doi.org/10.18280/ts.360305, (WOS, SCI- E) (UGC Care List) ISSN: 0765-0019 (print); 1958-5608 (online) Impact Factor : 0.387

[8] J. Somasekar a, , G. Ramesh , Gandikota Ramu, P. Dileep Kumar Reddy, B. Eswara Reddy e, Ching-Hao Lai, "A dataset for automatic contrast enhancement of microscopic malaria infected blood RGB images", Data in brief, Elsevier, https://doi.org/10.1016/j.dib.2019.104643,2352-3409/© 2019. (WOS, E-SCI)

[9] Kuo-Yuan Hwa ; Wan Man Lin ; Yung-I Hou ; Trai-Ming Yeh., 2007. Molecular Mimicry between SARS Coronavirus Spike Protein and Human Protein. 2007 Frontiers in the Convergence of Bioscience and Information Technologies. IEEE.

[10] Hao Zhang ; Chunpu Zou ; Fengfeng Shao ; GuoZheng Li., 2013. Effects of climate factorson bacillary dysentery epidemic in Harbin City, China. 2013 IEEE International Conference on Bioinformatics and Biomedicine. IEEE.

[11] Ming Yan ; Zhan Zhou., 2009. An Infectious Disease Model on Risk Diffusion from the Perspective of Business Gene's Vulnerability. 2009 International Conference on Management and Service Science. IEEE. Sensing Symposium Proceedings. Remote Sensing - A Scientific Vision for Sustainable Development. Vol 4. IEEE.