COVID-19 outbreak analysis and predictions for future cases

Karan Gupta [1], Luv Dhamija[2],Ritin Behl[3]
Dept. of Information Technology [1], Dept. of Computer Science and Engineering [2], Dept. of Information Technology [3]

karan.17bit1077@abes.ac.in, luv.18bcs1148@abes.ac.in, ritin.behl@abes.ac.in

Abstract—The Coronavirus pandemic (COVID-19) is an infectious disease which has its origin from Wuhan, China in December 2019 and is now considered as a threat for people in this world , and due to this The Centre for Disease Control and Prevention in China has issued an notification on epidemic and assessment of risk on COVID-19 in January[1]. According to studies it is believed that disease was first spread through bats to humans leading to severe respiratory problems [2][3]. As of 22, April, COVID-19 has affected over 2.5 million people, causing about 170,000 deaths. Many studies suggest that the main reason for wide spread of this virus is due to air or water droplets being exposed from an already infected person. At current, there is no vaccine or treatment available for this disease, however continuous medical researches and clinical experiments have been conducted to find out the cure. This paper is an attempt to study and analyze how this virus is being spread across the globe with short span of time. This paper suggests the determine the patterns and trends using deep learning.

Index Terms— **COVID-19, Data visualization Deep Learning, Machine learning**

I. INTRODUCTION

During the month of December in 2019, the people residing in Wuhan, which is the capital city the province of Hubei and functions as the one of the major hubs of China for transport services had been visiting to hospitals who were facing pneumonia of severe condition. Some of the starting cases which were observed had a connection with the Huanan wholesale seafood market that use to buy and sell living animals like bats also since they were exposed to the condition. The system used for the surveillance that was put into the place after the outbreak of SARS was seen as active and patients sample were sent to laboratories to search for some evidence regarding the disease[4].Since study of the novel Coronavirus was unknown at that time it spread rapidly around the whole world by different mediums. With being officially sated as a pandemic the novel (COVID-19) has proved to be very fatal, compromising the lives of many people and consequentially damaging the World peace. The virus that the scientists believed to causes the COVID-19 disease and the one which was related to SARS in 2003 seems to emerge from a common family of viruses. According to the studies the virus is more dangerous to people who have weak immune system and are already diagnosed with some other lungs or heart disease. has become clear that though controlling the outbreak of the virus is somewhat difficult, but it is somehow possible. Countries like China have been able to control the outbreak whereas countries such as USA, Italy, Spain have had lost many lives and are still trying to figure out on how to stop the spreading of this virus. The rampant follows an exponential and the decline of the curve does not indicate a turning point positively rather it is due to the less number of people that are being tested and identified Chiang analyzed that though the median time of incubation was 7-day, the best effective way is to self-quarantine yourself for a period of 18-21 days. This practice is likely to be effective in preventing the further outbreak of disease and controlling it. has not only affected due health related problems but as a consequence of it many other things are also affected. world economy has declined, work areas are shut down, there is a shortage of supply of basic at some places. As bad as it seems there is an emergent need to take a rightful action to mitigate the risks of this virus and to do that we have to dig deeper on how this virus is spreading and what we can do from our side so that we can stop this exponential growth and flatten the pandemic curve. To discuss about the development structure of the COVID-19 virus i.e. RTP RdRp complex consisting of 1 nsp12, 1 nsp7 and 1 nsp8. However ,the 2nd nsp8 was unseen in the EM.[5]

With the fast progression in development curve of computer technology, it has been tried to being applied in vast multitude of fields of medicine, including segmentation of organ , its enhancement of its image and repair as well, as it provides support to diverse fields of medical diagnosis. Some of the researches describes effectiveness of a strategy, that proposes fast discovery of drug which in turn leads to medicinal potential in reply to new infectious diseases like this for which there is no particular drugs or vaccines are available[6]. Some of the technologies used in Deep learning, such as CNN and its robust ability of developing non-linear model and prediction, it has been extensive used for the same [7].

In this research we will analyze the impact of COVID-19 and provide observations and trends that are present globally and will try to find out which regions are most affected and we will build a deep-learning model for predictions on the amount of death cases and cases that were confirmed.

Some of the researches describes effectiveness of a strategy, that proposes fast discovery of drug which in turn leads to medicinal potential in reply to new infectious diseases like

Accepted 20th April 2020

this for which there is no particular drugs or vaccines are available.

II. Outbreak Analysis

The overall analysis is performed on Novel Coronavirus COVID-19 (2019-nCoV) Data Repository that is made available by Johns Hopkins CSSE[8] .The dataset is comprised of genuine latest records and is updated on daily basis. In this research paper we will be focusing on 3 major aspects of the COVID-19, death cases, confirmed cases and the recovered cases relating to the virus outbreak [9]. We have analyzed the spread of this virus across the globe as well as regional areas where the virus outbreak had an impact the most in accordance with time. The data that is being analyzed for this research is from 22nd January,2020 to 27th April,2020.

Fig(1) is a bar graph which shows number of countries that were affected by COVID-19 by date. As shown in Fig(1) we can observe that from 22nd Januray,2020,onwards the number of countries that have been affected increased rapidly and till now about 186 countries are found to be affected by the virus. From Fig(1) we can observe that the cases have increased in non-linear fashion where the last week of February comes out to be a turning point after which the spread of the virus kept on increasing at an alarming rate. By the end of March most of the countries were affected including some of the major countries of the world and the virus became a serious health issue on global scale. International travel and tourism could be one of the dominant factor which escalated the outbreak across the world.

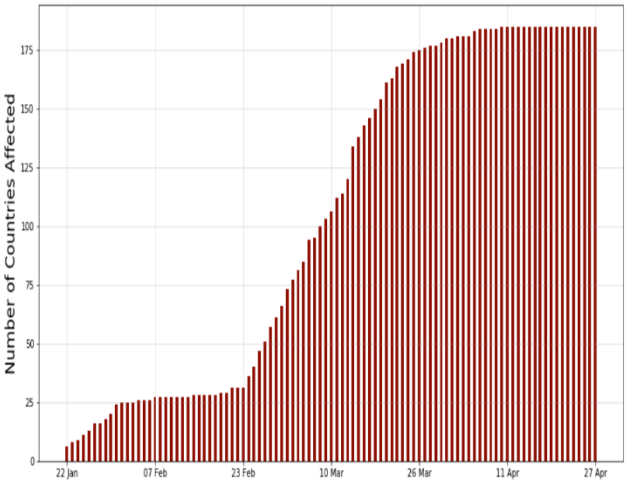
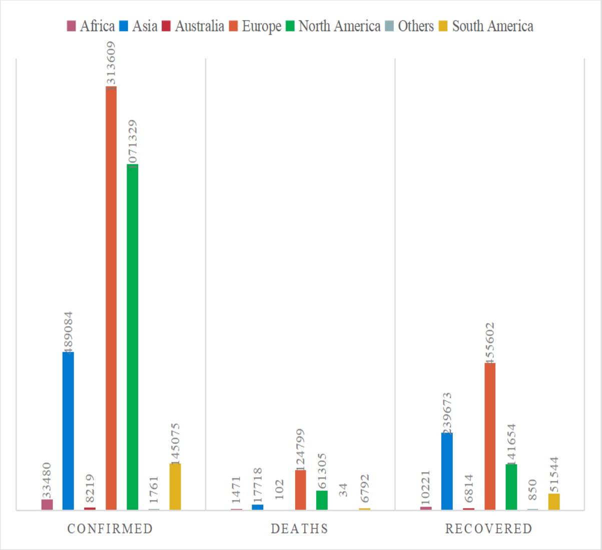


Fig (1). Number of countries affected due to COVID-19 by date
Source Data used from [10].

Fig(1) clearly represents of how the number of countries being affected by the virus increased but it does not provide a clear view of which continents got most affected and what was the recovery status and death records of those continents, Fig(2) represents a bar graph showing the records of different continents corresponding to their number of confirmed cases, number of death cases and recovered cases. From the graph it is obvious that the most affected continent is Europe and then we have North America after that Asia and so on. Although Asia was the first continent to get affected, Europe and North America are the other two continents that are more affected than Asia. Also, we can see that Europe has the greatest number of confirmed cases, it has a better recovery rate in comparison with North America. There could be various reasons for such type of anomalies such as number of people tested, availability of testing conditions, availability of health facilities, etc.

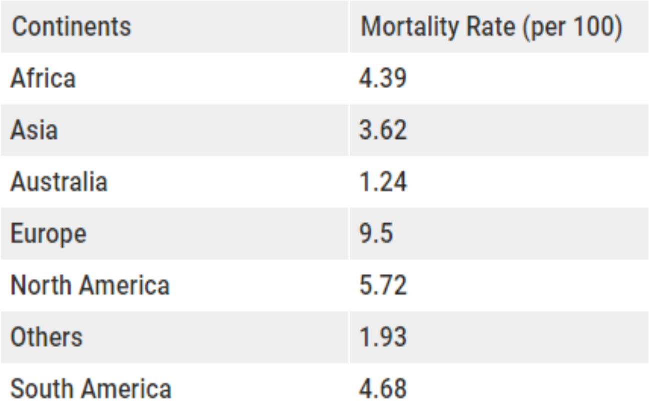


Fig(2). Number of Confirmed cases, death cases and recovered cases of

various continents
Source: Data used from [11]

The adversity of the virus across different continents cannot be just measured by estimating the number of people affected. To get a deeper understanding of the impact we should consider mortality rate could be a dominant factor which could play an important role in describing the impact of the virus across different continents. Mortality rate is a measure of number of deaths over a particular region of the population Table (1) shows the mortality rate in different continents. The value of the mortality rate (per 100) describes out of every 100 infected people what is the value of the number of deaths that occur due to the Virus We can observe that mortality rate is very high in Europe as compared to any other continents which is about 9.5(per 100) in Europe while the mortality rate for other continents generally lies around 5(per 100).

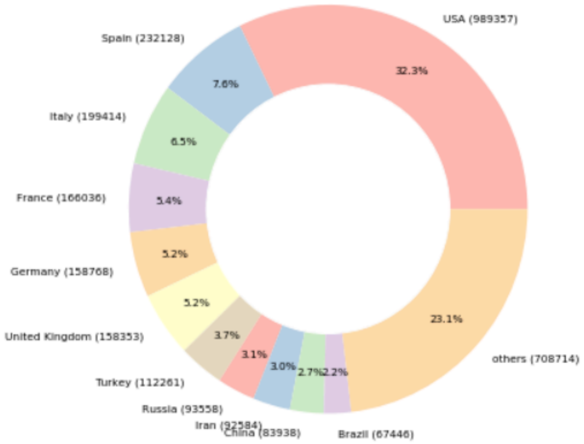
Table(1) Mortality rate(per 100) for different continents
Source: data taken from [11]



Continents	Mortality Rate (per 100)
Africa	4.39
Asia	3.62
Australia	1.24
Europe	9.5
North America	5.72
Others	1.93
South America	4.68

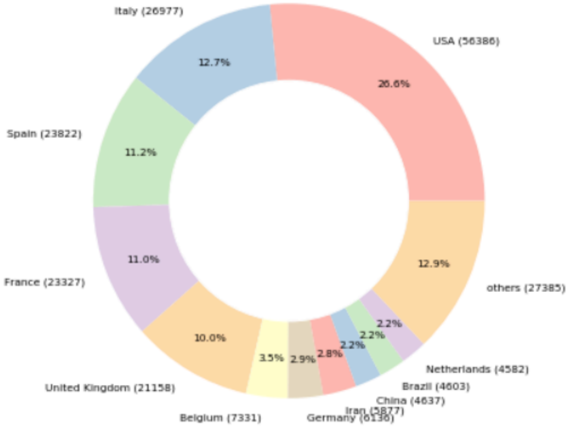
Till now we have analyzed the impact of COVID-19 for different continents across the world.

Now we will get into more depth and analyze the major countries currently under influence of the virus. We will be analyzing all the countries with their respective number of Confirmed cases, highest number of deaths and highest number of recovered cases. By analyzing so it becomes clearer about which countries are at higher risk and are susceptible of facing the outbreak on larger scale. Fig(3a) is a Pie chart representing different countries with their respective confirmed cases. Clearly, we can see that the most number of cases are identified in USA, following which we have Spain, Italy, France and Germany. These 5 countries contribute to about 50% of the total cases that have been identified till 27th April,2020.



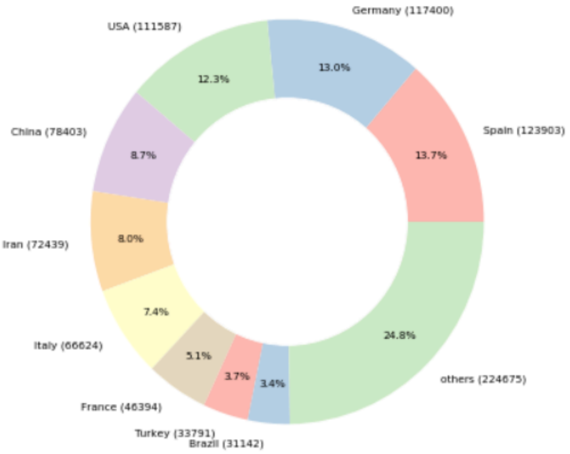
Fig(3a). Countries with total number confirmed cases
Source : Data taken from [11]

Fig(3b) is a Pie chart showing different countries of the world with their corresponding number of death cases with USA having the highest number of deaths consequently we have Italy being the second highest, Spain being third and France and Germany being fourth and fifth respectively. The number of death cases are pretty much higher in USA as compared to any other country which is just about double.



Fig(3b) Countries with total number of death cases
Source : Data taken from [11]

Fig(3c) is a Pie chart which shows all the major countries which have the highest number of recovery cases. It can be inferred that Italy has the highest number of recovered cases after that we have Germany, USA, China, Spain on position second, third, fourth and fifth respectively.



Fig(3c). Countries with total number of recovered cases.

Source : Data taken from[11]

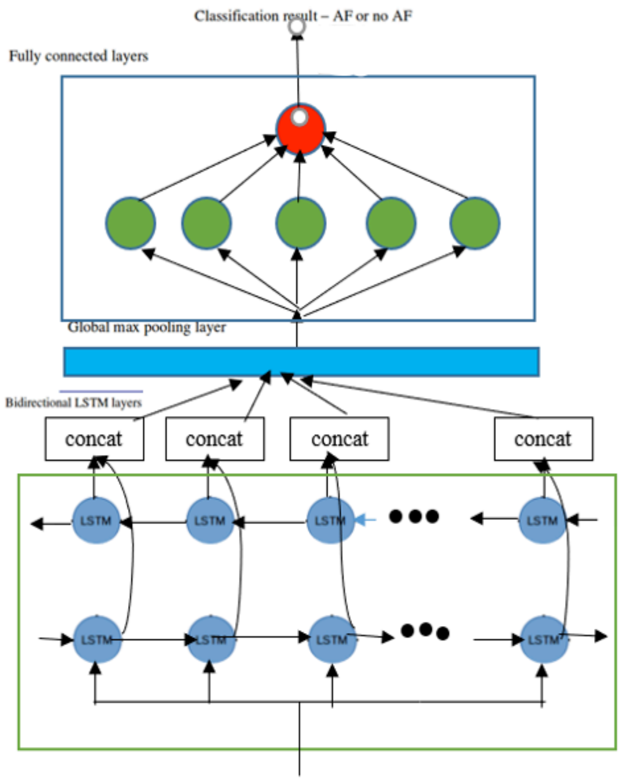
III. Proposed Model for Future Prediction

The COVID-19 is highly contagious disease, as it can spread out very quickly. Whenever a person sneezes or cough the surrounding people may get affected by the virus easily as it releases air droplets which can contaminate the surrounding. Since the incubation period of this virus has been believed to be of 14 days, the infected person may pass on this disease to more people without even knowing that they are affected which causes exponential growth in Corona virus patients. For forecasting the future results of the outbreak, many different statistical models, machine learning models and deep learning models have been proposed [12][13][14][15][16].

Some statistical techniques can be used for forecasting like Simple Moving Average(SMA), exponential smoothing and Autoregressive Integration and moving Average(ARIMA), however these statistical techniques have the pre-assumption of linearity in the data and these models may not be useful for non-linear data[17].Since COVID-19 is a non-linear time dependent data, we will use neural networks for forecasting like Artificial Neural Networks(ANN) for non-linear data[18], however they are not able to extract descriptive patterns from the sequence of data. In our predictive model we used Bidirectional LSTM [19][20],which are capable of learning the input sequence backwards and forward as well and then combine the result of the two interpretations.

For COVID-19 data we developed a univariate time-series forecasting Bidirectional LSTM model. As the name suggest Bidirectional LSTMS operates in both the directions. Bidirectional LSTMs are helpful in a way that they are able to retrieve sequences from the past as well as future time series patters because it involves 2 Layers one the original input sequence and second being the reversed copy of the first[21].

ARCHITECTURE OF BIDIRECTIONAL-LSTM NETWORK



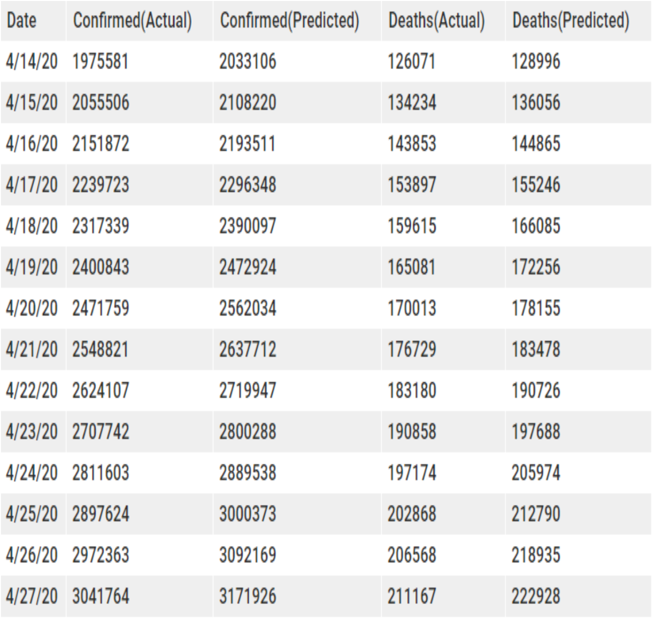
Fig(4) Architecture of Bidirectional LSTM [23]

Some of the RNN architectures for example Bi-LSTM which is also referred to as Bidirectional LSTM is generally applied areas where problem of learning is known to sequential in nature. Bidirectional LSTM's are famous in nature as they try to learn how and when to forget and at which time they are not required to use their gates that are present in the architecture. In previous RNN models, the problem of vanishing gradients was huge and it caused these nets not to learn. For being able to use Bidirectional LSTM's, you are required to feed the algorithm associated with learning, the data once from start to end and vice-versa.[24]

IV. RESULTS

In our model we used data from 22nd January,2020 to 13th April,2020 as training data for the model and rest was used for performing predictions and testing the model's performance. We divided the dataset into a 3 day timesteps using the sliding window algorithm[25] such that given the values of the previous 3 days our model will try to predict the value for the next day. Table(2) describes the results of the actual values and predicted values for number of confirmed cases and also for death cases from 14th,April 2020 to 27th,April,2020. The predicted values shown in the table comes out to be somewhat close to the actual values, however we have used only a timestep of 3 days which means for our predictions we have considered only values of previous 3 days, this is because creating a more robust model for prediction with more timesteps require greater amount of data. But the data used here consists of only records from 22nd January ,2020 to 27th Aril, 2020.

Table(2) Results of actual and predicted values for confirmed cases and death cases
Source : Actual data taken from [10][22].



Date	Confirmed(Actual)	Confirmed(Predicted)	Deaths(Actual)	Deaths(Predicted)
4/14/20	1975581	2033106	126071	128996
4/15/20	2055506	2108220	134234	136056
4/16/20	2151872	2193511	143853	144865
4/17/20	2239723	2296348	153897	155246
4/18/20	2317339	2390097	159615	166085
4/19/20	2400843	2472924	165081	172256
4/20/20	2471759	2562034	170013	178155
4/21/20	2548821	2637712	176729	183478
4/22/20	2624107	2719947	183180	190726
4/23/20	2707742	2800288	190858	197688
4/24/20	2811603	2889538	197174	205974
4/25/20	2897624	3000373	202868	212790
4/26/20	2972363	3092169	206568	218935
4/27/20	3041764	3171926	211167	222928

V. LIMITATIONS

The proposed model is used for univariate single-step forecasting. The dataset used contains only data from past 3 months and to create a predictive model with high accuracy requires large amount of data due to which Multi-step forecasting cannot be done on this data using Bidirectional LSTM network. Multi-step forecasting on such a data might not produce accurate results .

VI. CONCLUSION

We analyzed the COVID-19 dataset made available by John Hopkins University and created some meaningful insight and make some inferences about which continents are most affected currently and which countries are at greater risk. We also found out the mortality rate for different continents. Finally, we developed a deep learning architecture having bidirectional LSTM network for performing predictions on the dataset. We found out the results we close to the actual

Accepted 20th April 2020

data reported but it could provide only single-step forecasting.

VII. REFERENCES

[1] Chinese Center for Disease Control and Prevention. Epidemic update and risk assessment of 2019-nCoV

[2] Peng Zhou, Xing-Lou Yang, Zheng-Li Shi , A pneumonia outbreak associated with a new coronavirus of probable bat origin. (2020).

[3] Jie Cui ,Fang Li,Zheng-Li Shi , Origin and evolution of pathogenic coronaviruses (2019)

[4] Tanu Singhal A Review of Coronavirus Disease-2019 (COVID-19)

[5] Wanchao Yin, Chunyou Mao, Xiaodong Luan, Dan-Dan Shen, Qingya Shen, Haixia Su, Xiaoxi Wang, Fulai Zhou, Wenfeng Zhao, Minqi Gao, Shenghai Chang, Yuan-Chao Xie, Guanghui Tian, He-Wei Jiang, Sheng-Ce Tao, Jingshan Shen, Yi Jiang, Hualiang Jiang, Yechun Xu, Shuyang Zhang, Yan Zhang, H. Eric Xu Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir

[6] Zhenming Jin, Xiaoyu Du, Yechun Xu, Yongqiang Deng, Meiqin Liu, Yao Zhao, Bing Zhang, Xiaofeng Li, Leike Zhang, Chao Peng, Yinkai Duan, Jing Yu, Lin Wang, Kailin Yang, Fengjiang Liu, Rendi Jiang, Xinglou Yang, Tian You, Xiaoce Liu, Xiuna Yang, Fang Bai, Hong Liu, Xiang Liu, Luke W. Guddat, Wenqing Xu, Gengfu Xiao, Chengfeng Qin, Zhengli Shi, Hualiang Jiang, Zihe Rao & Haitao Yang Structure of Mpro from COVID-19 virus and discovery of its inhibitors.

[7] Xiaowei Xu ; Xiangao Jiang, Chunlian Ma; Peng Du; Xukun Li; Shuangzhi Lv, Liang Yu; Yanfei Chen; Junwei Su ; Guanjing Lang,Yongtao Li, Hong Zhao ; Kaijin Xu, Lingxiang Ruan ;Wei Wu Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia

[8] https://github.com/CSSEGISandData/COVID-19 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE Accessed on 28/04/2020 8:00AM

[9] https://systems.jhu.edu//research/public-health/ncov/Mapping 2019-nCoV.

[10] https://github.com/CSSEGISandData/COVID-19/blob/bda67e3db0e8dca4540297633d431a8021c035c8/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv (28th April,2020, 8:00 IST)

[11] https://github.com/CSSEGISandData/COVID-19/blob/e6645fae67850899a8e31f973cfd76bcb7c2a29f/data/cases_country.csv (28th April,2020, 21:00 IST)

[12] Toshikazu Kuniya , Prediction of the Epidemic Peak of Coronavirus Disease in Japan, 2020

[13] Binti Hamzah FA, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, et al. CoronaTracker: Worldwide COVID-19 Outbreak Data Analysis and Prediction. [Submitted]. Bull World Health Organ.

[14] Han Li1 & Fengzhu Sun,Comparative studies of alignment, alignment-free and SVM based approaches for predicting the hosts of viruses based on viral sequences

[15] H. AL-NAJJAR, N. AL-ROUSAN,A classifier prediction model to predict the status of Coronavirus CoVID-19 patients in South Korea

[16] Fotios PetropoulosID1 *, Spyros MakridakisID2,Forecasting the novel coronavirus COVID-19,2020.

[17] Shruti Kaushik1*, Abhinav Choudhury1, Pankaj Kumar Sheron1, Nataraj Dasgupta, Sayee Natarajan2, Larry A. Pickett2 and Varun Dutt,AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures 2020.

[18] Tealab, A., Hefny, H., & Badr, A. (2017). Forecasting of nonlinear time series using ANN. Future Computing and Informatics

[19] Sima Siami-Namini,Neda Tavakoli,Akbar Siami Namin,A Comparative Analysis of Forecasting Financial Time Series Using ARIMA, LSTM, and BiLSTM

[20] Alex Graves and Jurgen Schmidhuber,Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures.

[21] https://en.wikipedia.org/wiki/Bidirectional_recurrent_neural_networks

[22] https://github.com/CSSEGISandData/COVID-19/blob/bda67e3db0e8dca4540297633d431a8021c035c8/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv (28th April,2020, 8:00 IST)

[23] Oliver Faust, Alex Shenfield, Murtadha Kareem, Ru San Tan Automated detection of atrial fibrillation using long short-term memory network with RR interval signals

[24] https://datascience.stackexchange.com/questions/25650/what-is-lstm-bilstm-and-when-to-use-them.

[25] P iyush Kapoor and Sarabjeet Singh Bedi,Weather Forecasting Using Sliding Window Algorithm,2013.