Predicting the number of new cases of COVID-19 in India using Survival Analysis and LSTM

Aarathi S
Department of Computer Science and Engineering
G.Pulla Reddy Engineering College
Kurnool, India
aarthis1@gmail.com

Rithika F Johnson
Department of Computer Science and Engineering
CMR Institute of Technology
Bangalore, India
rithikajohnson@gmail.com

RajaPraveen.k.N
Department of Computer Science and Engineering
Jain (Deemed-to-be University)
Bangalore, India
p.raja@jainuniversity.ac.in

Mahesh T R
Department of Computer Science and Engineering
Jain (Deemed-to-be University)
Bangalore, India
t.mahesh@jainuniversity.ac.in

Vivek V
Department of Computer Science and Engineering
Jain (Deemed-to-be University)
Bangalore, India)
Bangalore, India
v.vullikanti@jainuniversity.ac.in

2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) | 978-1-6654-2642-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/I-SMAC52330.2021.9640899

***Abstract*—COVID-19 has been the cause of death for thousands of people across the globe. The goal of this paper is to forecast the new COVID-19 cases in India. The other methods used to forecast COVID-19 cases fail to give results with good accuracy when they try to predict the new cases number for a long time period or when the count of daily cases reported is large since the population of a country is large. The proposed study overcomes the challenge by firstly customizing the dataset. Second, the survival analysis has been utilized to choose appropriate factors, and third, the data will be integrated into the Long Short-Term Memory Network (LSTM). With a mean absolute percentage error of 5.79 percent, data from the 30th of January, 2020, to the 16th of June, 2021, was used to determine the new cases number of every day for the next 21 days.**

Keywords-Covid-19, forecasting, India, Survival analysis, LSTM

I. INTRODUCTION

Covid-19 pandemic, which was caused by the SARS-CoV-2 corona virus, is an infectious respiratory disease that is growing exponentially. Therefore controlling its spread and its devastating outcomes, has become one of the major concerns of the government.

Many countries including India have either experienced or are experiencing a second wave of this deadly virus. Some countries like Germany have experienced a third wave. The virus constantly keeps mutating. For instance, on 14th December, 2020, the authorities of the United Kingdom and Northern Ireland, reported the presence of the new variant, B.1.1.7, which spreads more readily than its former and slightly reduces the potency of the vaccine. B.1.617 variant, that is normally called as double mutant strain which was identified in India, was one of the major variants responsible for the second wave that hit India. The most recently discovered Delta plus variant i.e, B.1617.2 variant found in India after the second wave was found in states like Tamil Nadu, Punjab and so on. Doctors still have no information regarding the right treatment for this variant. Hence there is no permanent cure for this disease as of now.

Forecasting is a quantitative measure that aids in the process of predicting the future trends based on past and present data. Time series forecasting methods such as ARIMA and other techniques of forecasting such as SVM, linear regression etc. have been used in the past for forecasting COVID-19 cases but it has been observed that larger the count of COVID-19 cases that are identified per day greater is error produced by these models. In [1], where ARIMA was used to predict the new cases number every day, within a span of 10 days the difference between the actual and the predicted number of cases is approximately twenty thousand on the 10th day. In another study [2], where multiple linear regression was used to predict the count of cases for 30 days, the difference in the total count of cases was almost one hundred thousand.

LSTM was proved to be more effective in comparison with these models but again was efficient only for a short time period. When it came to a longer time span, the error percentage of the model increased. This study aims to tackle this problem by using a more efficient way of forecasting new cases in India using LSTM neural-network for a span of 21 days.

II. METHOD

A. Dataset

The dataset provided by our world in data was used to collect information on new cases per day, total number of cases and population density. The other features used in this study were day, new cases/ total number of cases, growth rate and status. Day is the number of days after the 23rd of January, 2020.

It is observed that when we take the ratio of new cases/total number of cases , as we approach peak, the ratio varies between 0.01 to 0.08(most of the time the ratio stays between 0.01 to 0.03) and as the cases decrease this value decreases. This helps in the reduction of error that occurs while forecasting the new number of cases for a long period of time and helps in normalizing data to give predictions with good accuracy when a country has a large population.

B. Survival Analysis

Cox regression also called as proportional hazards regression is a technique used for investigation of the effect of various variables upon the time a specified event takes to happen. Here we are using Cox proportional hazard to check the effect of the covariates used with respect to the increase and decrease of new cases found. The data from 13 different

countries was used to choose the best covariates that have the most influence in the number of new cases observed.

```
---
                coef  exp(coef)   se(coef)   coef lower 95%  coef upper 95%  exp(coef) lower 95%  exp(coef) upper 95%
covariate
avg_new/total 117.62   1.21e+51      17.45           83.42          151.83             1.69e+36             8.65e+65
growth rate     8.46    4744.71       2.62            3.34           13.59             28.20              7.98e+05
Pden           -0.36       0.70       0.23           -0.81            0.10              0.45                1.10

                  z     p   -log2(p)
covariate
avg_new/total  6.74 <0.005      35.88
growth rate    3.24 <0.005       9.69
Pden          -1.54  0.12       3.02
---
Concordance = 0.86
Partial AIC = 618.14
log-likelihood ratio test = 133.42 on 3 df
-log2(p) of ll-ratio test = 93.03
```

Fig 1: Result of Cox regression

Cox model is exhibited by the hazard function noted as h(t).

$$h(t)=h_{0...}(t)\exp(b\text{-}_1X\text{----}_1 + b\text{-}_2X\text{----}_2 +\ldots+ b\text{-}_pX\text{----}_p)$$

where,

t denotes survival time, **h(t)** -→ hazard function found by a set of p covariates ($X\text{--}_1, X_2, \ldots, X_p$), t**he coefficients ($b\text{-}_1, b_2, \ldots, b_p$)** measure the impact (i.e., the effect size) of covariates,
h_0 -→ baseline hazard

If, Hazard-Ratio = 1: then No effect,
Hazard-Ratio < 1: then Reduction in the hazard
Hazard-Ratio > 1: then Increase in Hazard

Therefore, from Fig 1, we can see that the hazard ratio(exp(coef)) for growth rate and avg_new/total is greater than 1, it means that the risk in the increase in the number of cases strongly depends on these values. Also the p value of growth rate and avg_total cases is less than 0.005, that is, they are statistically significant. Hence we can choose the data of growth rate and avg_new/total to train the LSTM model.

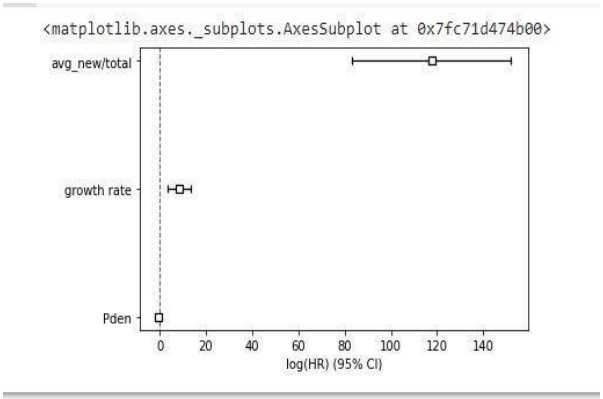

Fig 2: visualization of the significant difference between growth rate and avg_new/total with confidence interval of 95%

Fig 3: Graphs showing the variation of new cases, ratio of new cases/ total number of cases and growth rate over the time period used in the study

A. C. LSTM Architecture

LSTM consists of an input gate, forget gate and an output gate. The function of the input gate is to compute the latest information that set foot in memory cell. The function of the output gate is to decide what l activations of the cell in order to clean at the output end. The duty of the forget gate is to aid the LSTM n/w to forget its old input dataas well as to reset its memory cell. Equations of the gates implemented in LSTM are shown below.

$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1}, + b_i)$

$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1}, + b_f)$

$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1}, + b_o)$

where i_t→input gate

f_t→forget gate

o_t→output gate and c_t→memory cell.

The LSTM architecture used has a LSTM network with 50 hidden nodes and a dense layer. Adam is used as optimizer and mean square error is being used as the loss function. This model is actually trained for 200 epochs along with a batch size which is 25.

III. RESULTS

The total training process was approximately 13 minutes using the GPU provided by Google Colab. Fig 4. shows a graph that was plotted for training loss against validation for each epoch. Validation loss for the last epoch was 0.0192.

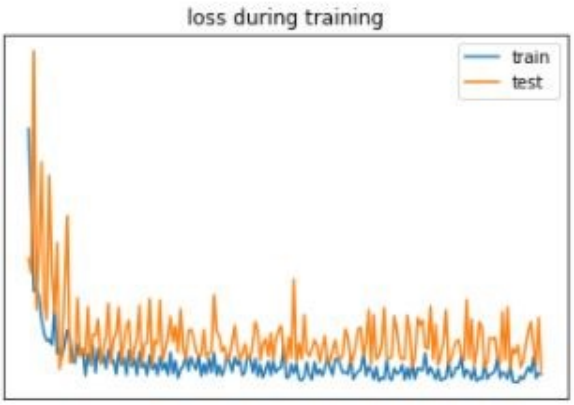

Fig 4: training loss v/s validation loss

The number of new cases on the last 21 days (May 27th 2021 to June 16th 2021) in the dataset was forecasted by the model and was compared with the actual number of new cases reported on those days. Figure 5 shows a graph comparing the actual value to the forecasted value of new cases reported upto 21 days.

TABLE I: NUMBER OF NEW CASES FOR 21 DAYS, ACTUAL AND FORECASTED VALUES

Day	*Actual Value*	*Forecast value*
1	186364	192836
2	173790	179610
3	165553	168461
4	152734	159211
5	127510	148303
6	132788	130458
7	134154	125232
8	132364	122524
9	120529	120404
10	114460	113915
11	100636	108742
12	86498	99759
13	92596	88871
14	93463	86983
15	92291	85857
16	84332	84864
17	80834	80694
18	70421	77527
19	60471	70988
20	62224	63174
21	67208	60249

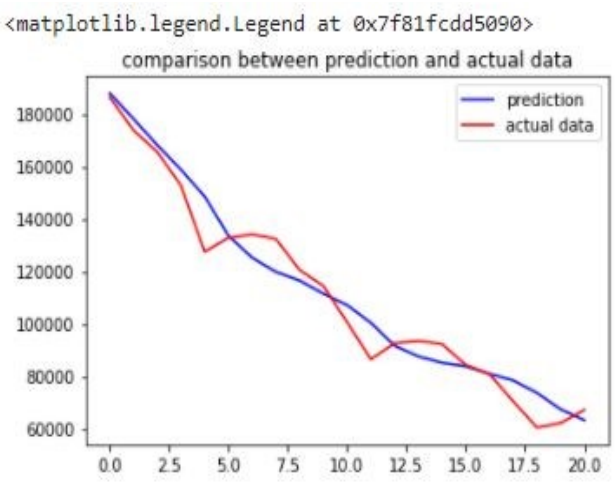

Fig 5: A graph comparing the actual new cases number with the forecasted new cases number for 21 days.

The mean-absolute-percentage-error (MAPE) is a statistical method of measuring accuracy of a forecast system. The formula to calculate mean absolute percentage error is:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{(A_t - F_t)}{(A_t)}\right| \times 100$$

Where n is the total number of days, A_t is the actual value of new cases on the t^{th} day and F_t is the forecast value of the new cases on the t^{th} day.

The MAPE for this LSTM model is 5.79%.

IV. CONCLUSION

In this work, we forecasted the number of new COVID-19 cases for the 21 days in India, by choosing suitable covariates using cox regression and feeding their data into the LSTM model. The same can be used for forecasting the number of new cases in other countries as well. In this method of forecasting we can observe that the MAPE is low when compared to other studies, when the number of new cases was high or predicted the number of new cases for a very long duration.

V. LIMITATIONS AND FUTURE SCOPE

The limitation observed in this method is that since the total number of cases increases constantly, there will come a time when the ratio of cases that are new to the total count of cases will become insignificant. To overcome this, we can reset the total count of cases to zero at the close of a COVID-19 wave and train the model with data collected a month prior to the reset, assuming that COVID-19 outbreak began the previous month. A better method can also be suggested to tackle this problem.

REFERENCES

[1] Narayana Darapaneni,Deepak Reddy, Anwesh Reddy Paduri, Pooja Acharya and Nithin H S, "Forecasting of COVID-19 in India Using ARIMA Model"

[2] Rajani Kumari, Sandeep Kumar*, Ramesh Chandra Paonia, Vijander Singh, Linesh Raja,Vaibhav Bhatnagar, and Pankaj Agarwal,

"Analysis and Predictions of Spread, Recovery, and Death Caused by COVID-19 in India"

[3] Hadeel I. Mustafa and Noor Y. Fareed, "COVID-19 Cases in Iraq; Forecasting Incidents Using Box - Jenkins ARIMA Model"

[4] Selahattin Serdar HELLİ, Çağkan DEMİRCİ, Onur ÇOBAN and Andaç HAMAMCI, "Short-Term Forecasting COVID-19 Cases In Turkey Using Long Short-Term Memory Network"

[5] Sohini Sengupta, Sareeta Mugde and Dr Garima Sharma, "Covid-19 pandemic data analysis and forecasting using machine learning algorithms"

[6] FURQAN RUSTAM , AIJAZ AHMAD RESHI , ARIF MEHMOOD, SALEEM ULLAH , BYUNG-WON, WAQAR ASLAM, AND GYU SANG CHOI, "COVID-19 Future Forecasting Using Supervised Machine Learning Models"machine learning forecasting model for COVID-19 pandemic in India"

[7] Vinay Kumar Reddy Chimmula and LeiZhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks"

[8] PAPASTEFANOPOULOS, Vasilis; LINARDATOS, Pantelis;KOTSIANTIS, Sotiris. COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population. Applied Sciences, 2020, 10.11: 3880.

[9] ELSWORTH, Steven; GÜTTEL, Stefan. Time Series Forecasting Using LSTM Networks: A Symbolic Approach. arXiv preprintarXiv:2003.05672, 2020.

[10] HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. Neural computation, 1997, 9.8: 1735-1780.

[11] Jayanthi Devaraj, Rajvikram Madurai Elavarasan, Rishi Pugazhendhi, G.M. Shafiullah, Sumathi Ganesan, Ajay Kaarthic Jeysree, Irfan Ahmad Khan and Eklas Hossain,"Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant?"

[12] Chen, Joy Iong-Zong. "Design of Accurate Classification of COVID-19 Disease in X- Ray Images Using Deep Learning Approach" Journal of ISMAC 3, no. 02 (2021): 132- 48.

[13] Manoharan, J. S. (2020), "Early diagnosis of lung cancer with probability of malignancy calculation and automatic segmentation of Lung CT scan images", Journal of Innovative Image processing, 2(4): 175 – 186.