A Real-time Model to Forecast the Outbreak of Covid-19 using LSTM

Omkar Bhoite
Modern Education Society's College of Engineering
Pune, India

Sohail Ahmad
Modern Education Society's College of Engineering
Pune, India

Saurabh Wagh
Modern Education Society's College of Engineering
Pune, India

Ketan Gaikwad
Modern Education Society's College of Engineering
Pune, India

Shalaka Deore
Modern Education Society's College of Engineering
Pune, India

Shubhangi Ingle
Modern Education Society's College of Engineering
Pune, India

ABSTRACT
Deep Learning based forecasting models have been in use for a long time and they have proven their significance in problems including time series forecasting and improve the accuracy and efficiency of the results for given problem. These models have long been utilization domains that required that identification and fiction of main factors of information. Based on review of work done in the field of forecasting, this study demonstrates the potential of LSTM algorithm to forecast the rise and fall in number of active cases and deaths of Covid 19 patients using real time input provided by John Hopkins University which is available on GitHub and updated on a daily basis. In short, a real time Covid 19 outbreak forecasting model implemented using long short term memory networks algorithm. The use of LSTM is suggested to improve the efficiency and accuracy of the presently available models and make predictions of 2 parameters including the number of active cases and the number of deaths for the upcoming 10 days. The goal was to analyze the algorithm by comparing the results of prediction and actual reports for a period of 60 days and forecast the number of newly confirmed and death cases of the disease for upcoming 10 days. To develop an algorithm faster than existing systems and use the most recently available data for a higher range of input and calculate the latest trend.

Keywords
LSTM, Optimizer, Activation function, MinMaxScaler, Dense Layer, Environment Variables

1. INTRODUCTION
This study reviews some of the state of art supervised machine learning regression models including least absolute linkage and selection operator (LASSO), supporter machine and exponential smoothening. These models have been trend using co-ed 19 station statistics data set provided by the John Hopkins University. The data has been preprocessed and divided into two subsets: training set and testing set. As of now, the dataset contains more than 850 entries out of which, the last 60 entries are reserved for testing the data. In the current human crisis, our main attempt is to develop a forecasting system for Covid 19 which will help in implementing several policies for the government and curtail its spread. The timely decisions taken by the government helped in controlling the speed of Covid 19 to a large extent. Despite these decisions the pandemic continued to spread.

Today, it seems we have restricted the spread but this system will be helpful to predict upcoming waves and take desired actions.

To improve your experience and increase the efficiency of the system we have developed a front end considering various UX design experience aspects. The design is simple but effective. It gives a general idea of the virus, the authors, provides latest covid-19 stats, predicts outbreak of 19 for the next 10 days and shows a graph of collected and predicted data for both the parameters.

2. MODULE IDENTIFICATION
Algorithm: LSTM algorithm

Input: Automatically downloaded CSV file from GitHub

Output: Prediction results and comparison graphs

3. LITERATURE SURVEY
In this paper [1], during this paper the algorithm used is simple regression. Prediction and outbreak are taken into account to be a regression problem and it's implemented using two regression models namely linear and polynomial regression. The Covid-19 India data set is used for implementation purposes it predicts the number of confirmed new cases, deaths and recoveries based on data available from 12th March 2:30 1st October 2020. for the forecasting purpose and analysis statistic forecasting approach is employed. The output of this implementation suggests that polynomial regression shows better results than simple regression. The forecasting was done using Tableau and also the results are satisfactory, but the results may be more accurate upon use of a much bigger dataset considered for an extended duration.

In the paper [2], Supervised Machine Learning Model algorithm is employed for the prediction in this paper. Machine learning (ML) based forecasting mechanisms have proved their significance in forecasting outcomes to improve the decision making on the future course of actions. Such models have been used for a protracted time in many applications which needed the identification and prediction of any sort of adverse factors or analyzing threats. Several prediction methods are being used to handle forecasting problems. This study demonstrates the potential of ML models to forecast the number of upcoming

patients affected by COVID-19 using Supervised Learning Models. In particular, four standard forecasting models, such as support vector machine (SVM), linear regression (LR), least absolute shrinkage and selection operator (LASSO), and exponential smoothing (ES) have been used in this study to forecast the outbreak of COVID-19. Three types of predictions are made by each of the models, such as the number of newly infected cases, the number of deaths, and also the number of recoveries in the next 10 days. The results produced by the study prove it is a promising mechanism to use these methods for the current scenario of the COVID-19 pandemic. The results prove that the ES performs best among all the used models followed by LR and LASSO which performs well in forecasting the new confirmed cases, death rate as well as recovery rate, while SVM performs poorly in all the prediction scenarios given the available dataset.

This paper [3] has analyzed the COVID-19 progression in India and the three most affected Indian states (viz. Maharashtra, Tamil Nadu and Andhra Pradesh) till 29 August 2020 and developed a prediction and forecasting model to forecast the behavior of COVID-19 spread in the upcoming months. They have used time series data for India and applied the Susceptible-Infective-Removed (SIR) model and the FbProphet model to predict the peak infection cases and peak infection date, i.e., the date with the highest surge in cases for India and the three most Covid 19 affected states. In this paper, they have further performed the comparative analysis of the prediction results from SIR and FbProphet models. From this study, with the assumption that at least 5 percent of India's population might be infected by the covid-19, it is concluded that, the countrywide spread is forecasted to reach its high by the end of Nov 20. And till the time there is no vaccination, for the states that have already reached their peak and with festivals around the corner, there are more chances of increase in the number of cases, mainly if the social distancing and other control measures are not followed by the people in the upcoming months.

In the paper [4] the study of the spread of Covid 19 pandemic across vari- ous countries and regions of the world including Australia, Canada, Italy, Japan, Spain, UK and USA. Not identified as being exposed or infected, the group of pa-tients has become the key source of infectious hosts for the COVID-19 pandemic, triggering the re-emergence of outbreaks. To consider the impacts of movement of unidentified patients and the limited testing capacity, an augmented Susceptible- Exposed-Infectious-Confirmed-Recovered (SEICR) model along with intercity mi- gration data and testing capacity is developed to probe into the number of unidenti- fied COVID-19 infected patients. This model allows evaluation of the effectiveness of active cases, and better prediction of the pandemic outbreak in a country, region or area. A pseudo algorithm is adopted in the model to provide the estimation of the outbreak problem using a limited amount of historical data. The model is applied to 175 regions in Australia, Canada, Italy, Japan, Spain, the UK and USA to estimate the number of unconfirmed cases. Results showed that the actual number of infected cases could be 4.309 times as many as the official confirmed number. By implement- ing mass COVID-19 testing, the number of infected cases could be reduced by about 50 percent. The SEICR algorithm considered movement of people along the covid dataset for a better prediction.

In paper [5] Along with the overall trend analysis in India, this study also takes into account 5 most affected states of the country: Maharashtra, Andhra Pradesh, Tamil Nadu, Karnataka and Uttar Pradesh as the subjects of the research. ARIMA and Prophet time series forecasting models have been used to make three types of predic- tions: confirmed cases, deaths and recovered cases in India as well as in the adopted states. The effectiveness of the forecasting models is evaluated based on metrics such as Root Mean Squared Error, Mean Absolute Error, Mean Absolute Percent- age Error and Coefficient of Determination. The results suggest that the adopted models are promising mechanisms for forecasting COVID-19 trends. Our study also suggests that ARIMA model performs better than Prophet Model at this task of fore- casting the outbreak. The forecasts can be useful in increasing the preparedness level of government authorities, health facilities and hospitals to combat against massive spread of the virus.

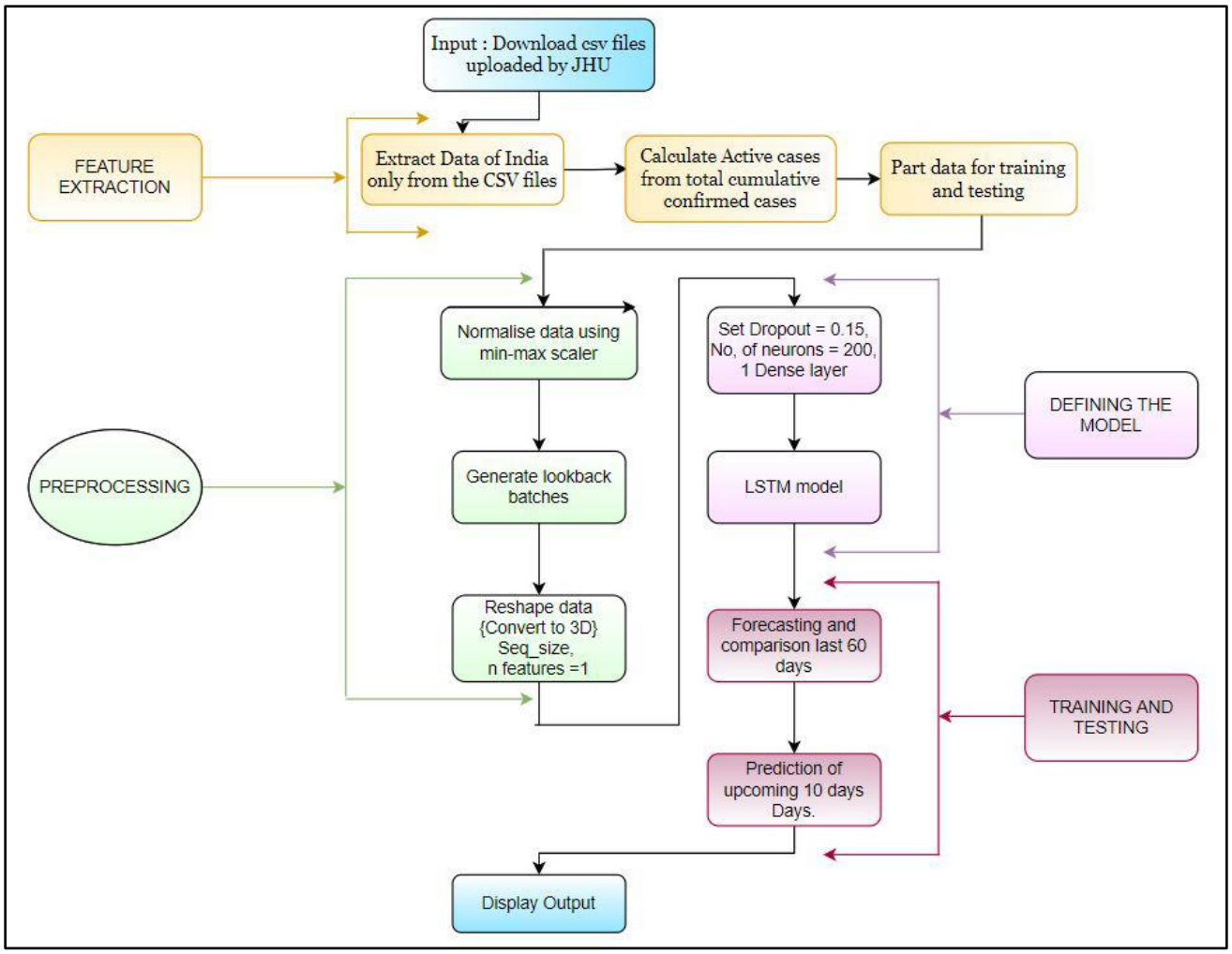
4. SYSTEM ARCHITECTURE

INPUT: Download the CSV file uploaded by John Hopkins University from GitHub and load the CSV for further use. This CSV file will be served as input to the system.

FEATURE EXTRACTION: The file contains Covid data of all countries and the globe which is irrelevant to us. Extract data where country=India from the csv.

The file has a cumulative count of all past days for a certain entry, we need to calculate the change in the cases for all days in particular. Subtract the days data from its constitutive previous entry to get the count of new cases and deaths per day.

Part the data for training and testing by reserving the last 60 days for testing purpose. This will further be used to compare with the results of the system.



PREPROCESSING:

The data we have has numbers in range of lakhs and thousands whereas the LSTM function needs it to be in range of 0 to 1. We apply the min max scaler to it and change it to required format. LSTM function operates in sets of data rather than testing the system all at once, so, we create lookback batches of 30 days each. Also, is a 3-dimensional algorithm which requires 3D data but our data is 2 dimensional as it is in the form of a csv. Add a third parameter which will not have any impact on the output but will help in feeding it to the system. For converting to 3D set n_features=1 as the third parameter. We set it to a constant because our input is univariate nor will it hamper our results.

DEFINING THE MODEL: Once our data is ready to be fed, during model definition, we set dropout to 0.15 to decrease the complexity of the algorithm. The optimizer used is Adam and the activation function is relu. We set the number of neurons to 200 and dense layer for output to 1. The LSTM model is now ready.

TRAINING AND TESTING: We fit the model to our data and make the prediction for a total 70 days. The prediction results of the last 60 days are compared with the actual data to calculate the accuracy of the model. We also forecast the outbreak for the next 10 days.

DISPLAY OUTPUT: We display the prediction result as well as the comparison for 60 days. Graphs based on the prediction of active cases as well as deaths is plotted for better visualization. A table for forecast of the upcoming 10 days is displayed.

5. LSTM ALGORITHM

Long short-term memory(LSTM) is one of the most widely used artificial neural network algorithms in the field of deep learning and machine learning. Its feedback connections are a major advantage over feed forward neural networks. This is useful in processing entire sequence of data including a speech or a video unlike other neural networks which can process only single data point at a time.

Algorithm Implementation:

Step 1: Start

Step 2: Scale input using MinMaxScaler

Step 3: Import time series generator from Keras preprocessing library

Step 4: Set seq_size = 30

Step 5: Set n_features = 1

Step 6: train_generator=Time series generator()

Step 7: Import LSTM from keras library

Step 8: Set model = Sequential

Step 9: Set number of neurons = 200

Step 10: Set activation function = relu

Step 11: Set input shape = sequence size, n_features

Step 12: Set Dropout = 0.15

Step 13: Set Dense layer to 1

Step 14: Set optimizer = adam, loss = mse

Step 15: Fit model with epoch = 50, steps per epoch= 10

Step 16: Plot the training and validation accuracy

Step 17: Set prediction= []

Step 18: Set future = 10

Step 19: for i in range length of test+future

Step 20: Set current_prediction = predict

Step 21: append prediction result to prediction variable

Step 22: Rescale results

Step 23: for k from 0 to future

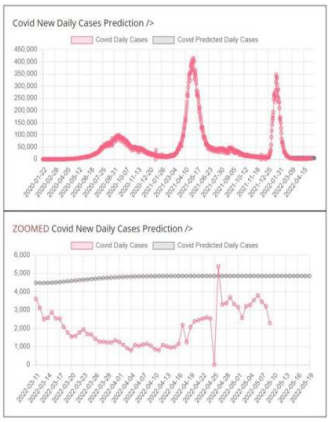
Step 24: Display prediction

Step 25: STOP

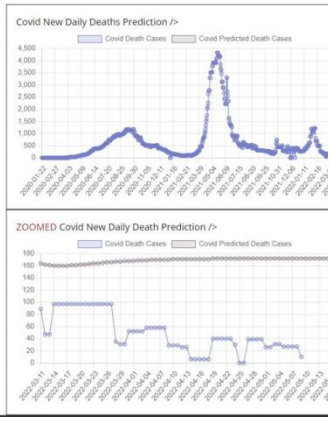
6. RESULTS

11th May, 2022

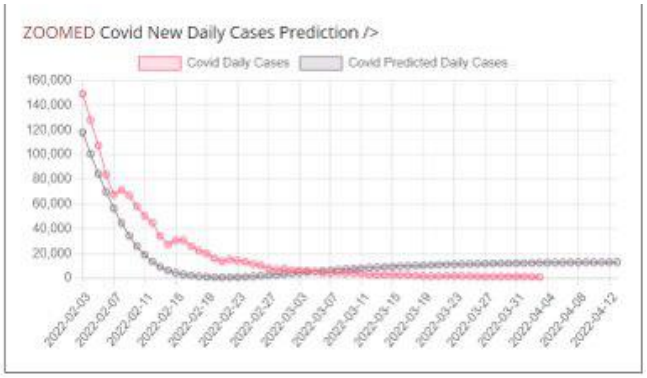
Date	Covid Predicted Daily Cases By LSTM
2022-05-10	4862.0
2022-05-11	4862.0
2022-05-12	4862.0
2022-05-13	4862.0
2022-05-14	4862.0
2022-05-15	4862.0
2022-05-16	4862.0
2022-05-17	4862.0
2022-05-18	4862.0
2022-05-19	4862.0

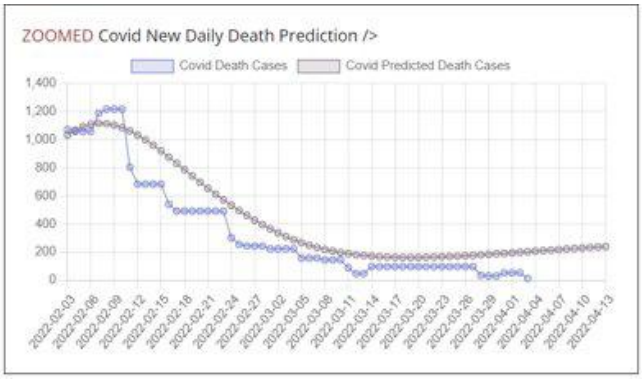


Date	Covid Predicted Daily Deaths By LSTM
2022-05-10	172.0
2022-05-11	172.0
2022-05-12	172.0
2022-05-13	172.0
2022-05-14	172.0
2022-05-15	172.0
2022-05-16	172.0
2022-05-17	172.0
2022-05-18	172.0
2022-05-19	172.0

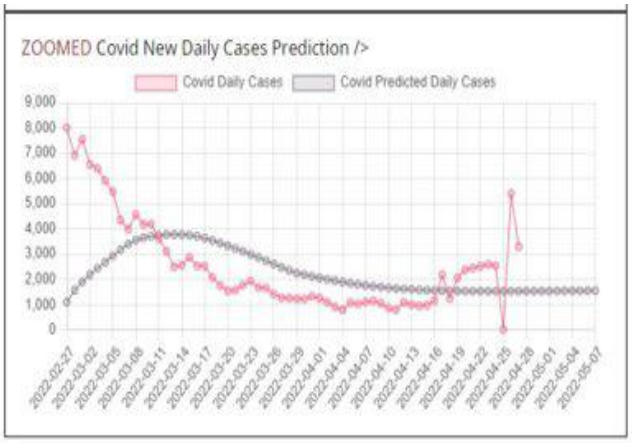


6th April, 2022





9th May, 2022



7. ADVANTAGES

Algorithm was able to successfully predict the outbreak of Kuwait 19 for the new confirmed cases and death cases for the upcoming 10 days.

LSTM shows better accuracy than other algorithms.

The algorithm is fast and precise.

The data set used is most recently available, so the results are accurate and can be updated as per change in the covid stats.

8. LIMITATIONS

The algorithm can follow the trend and predict the cases and deaths but it cannot predict waves.

Due to limited in resources, factors such as migration and vaccination have not been considered.

Sudden rise and fall in the numbers can affect the accuracy of the algorithm.

9. CONCLUSION

In this project, we have successfully implemented the LSTM algorithm to forecast the outbreak of Covid 19. The results show that LSTM has the highest accuracy among other available deep learning algorithms. The problem of gathering recently updated stats has been solved using real-time data. The model is faster and precise. RMSE Values: Confirmed Cases: 5336, Death Cases: 14

10. FUTURE SCOPE

Corona Future Forecasting demonstrates the complexity of controlling COVID- 19 outbreaks that how and when to take what level of interventions.

The feasibility of this system increases with the increase in outbreak of covid cases.

Analyzing the continuous changes in the covid dataset will increase the efficiency of the model.

Requirement of the system will increase in case of a major rise in cases.

Future scope of the system can include consideration of vaccination as an important factor in the prediction.

The data for vaccination is not available currently, but with increased resources it can be accessed from the respective sources. Once we have implemented the vaccination feature, we can opt for considering the mutation of Covid Variants, we can study which variant has higher death casualties and which one has a faster spread.

11. REFERENCES

[1] S. Shaikh, J. Gala, A. Jain, S. Advani, S. Jaidhara and M. Roja Edinburgh, ”Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting,” 2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence), 2021, pp. 989-995, doi: 10.1109/Confluence51648.2021.9377137.

[2] F. Rustam et al., ”COVID-19 Future Forecasting Using Supervised Machine Learning Models,” in IEEE Access, vol. 8, pp. 101489-101499, 2020, doi: 10.1109/ACCESS.2020.2997311.

[3] N. Darapaneni, P. Jain, R. Khattar, M. Chawla, R. Vaish and A. R. Paduri, ”Analysis and Prediction of COVID-19 Pandemic in India,” 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 291-296, doi: 10.1109/ICACCCN51052.2020.9362817.

[4] Zhan C, Tse CK, Gao Y, Hao T. Comparative Study of COVID-19 Pandemic Progressions in 175 Regions in Australia, Canada, Italy, Japan, Spain, U.K. and USA Using a Novel Model That Considers Testing Capacity and Deficiency in Confirming Infected Cases. IEEE J Biomed Health Inform. 2021 Aug;25(8):2836-2847. doi: 10.1109/JBHI.2021.3089577. Epub 2021 Aug 5. PMID: 34129512.

[5] S. Chordia and Y. Pawar, ”Analyzing and Forecasting COVID-19 Outbreak in India,” 2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence), 2021, pp. 1059-1066, doi: 10.1109/Confluence51648.2021.9377115.

[6] Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM. COVID-19 Outbreak Prediction with Machine Learning. Algorithms. 2020; 13(10):249. https://doi.org/10.3390/a13100249