Forecasting Rate of Spread of Covid19 using Linear Regression and LSTM

Kartik Puri
kartikpuri99@gmail.com

Ashwin Goyal
ashwingoyal180@gmail.com

Dr. Rachna Jain
rachnajain.bvcoe@bvp.edu.in

Dr. Preeti Nagrath
preeti.nagrath@bharatividyapeeth.edu

Department of Electronics & Communication Engineering
Bharati Vidyapeeths College Of Enginnering
9th March 2021

Abstract

COVID-19 virus, knows as novel coronavirus, spread across the world. The World Health Organisation (WHO), marked 11th March, 2020 as the day when COVID19 was declared as pandemic. It, was first originated in Wuhan, China. In recent days, Covid19 impacted various social and economic fields in the world. It is necesssary to quantify its spread and make predictions on how it is going to affect the world in coming months. In this paper, our aim is to use linear regression and LSTM algorithms to forecast of Covid19 spread. The objective of this study is to determine if spread can be forecasted to better accuracy using linear regression and LSTM algortihms.

KeyWords: Machine Learning, Linear Regression, LSTM, Mean Absolute Error, COVID-19

1 Introduction

The spread of COVID19, from the sars-cov2 virus occurred in Wuhan, China, is on the rise and has shaken the world. The World Health Organization christened the illness as COVID-19 when the first case of this virus was reported.

The Global spread of COVID19 affected every major nation and was defined as a pandemic by the WHO in March 2020.

This paper tracks the spread of the novel coronavirus, also known as the COVID-19. COVID-19 is a contagious respiratory virus that first started in Wuhan December 2019. [1]

The two types of coronaviruses, named as, "severe acute respiratory syndrome coronavirus" and "Middle East respiratory syndrome" have affected more than 20,000 individuals in last ten years [2].

The coronavirus can spread by various means.However some of the common means through which the infection can occur are:

1. airborne or aerosol transmission
2. direct or indirect contact with another human
3. and lastly through droplet spray transmission

However a person can protect himself from these transmission modes.Close contact can be avoided and a minimum distance of 1.8 metres should be maintained to avoid contact with a person as well as respiratory droplets.However for airborne transmission a minimum of 4 metre should be maintained to avoid contact.Symptoms of COVID 19 are coughing ,extreme fever,tiredness or weakness and pain in some joints of the body.

So for helping combat coronavirus, the use of artificial intelligence techniques such as machine learning and deep learning models were studied and implemented in this paper.These model will gives us a rough estimate as to how the disease will spread in the upcoming days how many more people will be effected.It will a rough estimate to the government of various countries about how the spread and will enable them to be prepared well in advance for the epidemic.

Most of the data driven approaches used in previous studies [3] have been linear models and often neglects the temporal components of the data.

In this report data preprocessing techniques are applied on the confirmed cases data and then the preprocessed data is applied to two models i.e. LSTM and Linear Regression .The actual and forecast values of cases are compared on a predefined metrics. A comparison is made between the performance of LSTM and Linear regression model to see which model best for the data.

The section **Literature Review** talks about similar work done by other researchers on this topic and talk about the model and approach used by them.

The methodology used in the paper and the approach on how to handle this problem is also discussed.

The section **Methods and models** talks about the dataset used and and its features. Since classification is done worldwide, so the data was processed to suite the needs of the models in use and a brief description of the processed dataset was also provided.

Next, Evaluation metrics are discussed to understand and compare the result between the two models used. MAPE and Accuracy were used to compare the result and were used to draw conclusions.

Also the models of Linear regression and LSTM network are explained demonstrating our approach.

In the end **Experiment Result** are shown. Evaluation metrics are used to compare the result.

2 Literature Review

In [4],an machine learning based alternative to transmission dynamics for Covid-19 is used. This AI based approach is executed by implementing modified stacked auto-encoder model.

In [5], an deep learning based approach is proposed to compared the predicted forecasting value of LSTM and GRU model. The Model was prepared and tested on the data and a comparison was made using the predefined metrics.

In [6], LSTM and Linear regression model was used to predict the COVID-19 incidence through Analysis of Google Trends data in Iran. The Model were compared on the Basis of RMSE metrics.

In [7], an LSTM networks based approach is proposed for forecasting time series data of COVID-19. This paper uses Linear short Term memory network to overcome problems faced by linear model where algorithms assigns high probability and neglects temporal information leading to biased predictions.

In [8], temporal dynamics of the corona virus outbreak in China, Italy, and France in the span of three months are analyzed.

In [9], a variety of linear and non-linear machine learning algorithms approaches were studied and the best one as baseline, after that the best features were chosen, using wrapper and embedded feature selection methods and genetic algorithm (GA) was used to determine optimal time lags and number of layers for LSTM model predictive performance optimization.

In [10], temporal dynamics of the corona virus outbreak in China, Italy, and France in the span of three months are analysed.

In [11],a modeling tool was constructed to aid active public health officials to estimate healthcare demand from the pandemic.The model used was SEIR compartmental model to project the pandemic's local spread.

In [12],a transmission network based visualization of COVID-19 in India was created and analyzed. The transmission networks obtained were used to find the possible Super Spreader Individual and Super Spreader Events (SSE).

In [13], comparison of day level forecasting models on COVID-19 affected cases using time series models and mathematical formulation. The study concluded exponential growth in countries that do not follow quarantine rules.

In [14],phenomenological models that have been validated during previous outbreaks were used to generate and assess short-term forecasts of the cumulative number of confirmed reported cases in Hubei province.

2.1 Our Work

In our report, the confirmed cases of corona virus are studied from the start of the epidemic and the two approaches of Linear Regression and LSTM networks are used, and an report is presented stating which of the above stated model works best these type of data on the basis of Mean Absolute Error.

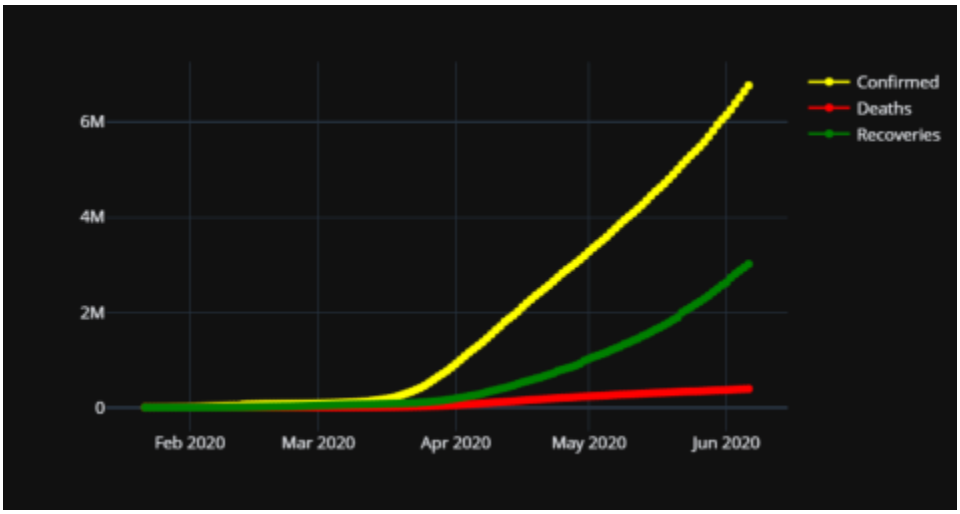

Figure 1: Number of cases around the world

3 Methods and models

3.1 Data

The dataset used was the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) for COVID-19.

It consist of 3 dataset each of Death, Confirmed, Recovered cases of 188 countries datewise. The number of date columns are 138 starting from 22 Jan,2020 to 8 June,2020.Out of this about 85% are used as training data and the rest used as testing and validating data. So the model would be predicting next 15% data value.

The prediction would not be made on a specific country rather it will be worldwide.

Table 1: World Dataset of Corona virus spread with confirmed, death, and recovery rates

	Confirmed	Recoveries	Deaths	Confirmed Change	Recovery Rate	Growth Rate
count	1.390000e+02	1.390000e+02	139.000000	138.000000	139.000000	138.000000
mean	1.918547e+06	6.817390e+05	123264.726619	50666.268116	0.286331	0.076081
std	2.170725e+06	8.911273e+05	138597.907312	42526.463980	0.143922	0.117824
min	5.400000e+02	2.800000e+01	17.000000	89.000000	0.017598	0.005032
25%	7.862450e+04	2.747150e+04	2703.000000	2957.500000	0.207790	0.021193
50%	8.430870e+05	1.738930e+05	44056.000000	67738.000000	0.288055	0.032183
75%	3.546736e+06	1.142438e+06	249918.000000	84446.500000	0.395898	0.085793
max	6.992485e+06	3.220219e+06	397840.000000	130518.000000	0.544809	0.951446

Table [1] show the world data of Corona virus spread with confirmed, death and recovery rates.

3.2 Evaluation Metrics

For the selection of better performing model, it is necessary to use some kind of performance/evaluation metrics to evaluate the algorithm's performance. In this paper, MAPE and Accuracy are used to measure model's performance:

1. **Mean Absolute Percentage Error**: It is defined by the following formula:

$$MAPE = \frac{100\%}{n} \sum \left| \frac{y - y\prime}{y} \right| \tag{1}$$

   Where y' is true value and y' is predicted value.

2. **Accuracy**: It is defined by the following formula:

$$Accuracy = (100 - MAPE)\% \tag{2}$$

3.3 Method

The prediction of confirmed cases due to COVID-19 are evaluated using Recurrent Neural Network method(LSTM) and Linear Regression.

Linear regression is a statistical model, that works with values where the input variable (x) and output variable (y) have a linear relationship, for single input the model is known as simple linear regression.

A recurrent neural network is a special kind of Artificial neural network which has memory of the previous inputs i.e it remembers the previous inputs. In these neural networks the output of previous neuron is fed as input to the next neuron.It is generally used in problems like when it is required to predict the following word in a sentence or in time-series data.However a main problem associated with RNN is gradient vanishing and exploding.In this the gradient starts vanishing as we go deeper into the layers due to which the model stops updating weights.This problems can be solved using special RNN like Long Short Term Memory(LSTM) RNN and Gated Recurrent Unit(GRU).These have a much better gradient flow and perform better than traditional RNN and are generally used. [5].

The dataset used for predicting the value is taken from John Hopkin University which contains cases form 21 Jan 2020 to 8 June,2020. The training and testing of both the models is done on this dataset.It contains 138 date columns out of which 120 are used for training and the rest 18days are used for testing data or for forecasting it.At first the data is preprocessed by converting the date columns into datetime object and also eliminate the missing values. The preprocessed data is then transformed in the required shape to be put into the model.The models are trained and the test data is predicted and prediction result are quantified using performance measures metrics such as MAPE and accuracy.The methodology performed for each of the step is shown in the figure 2 as show.

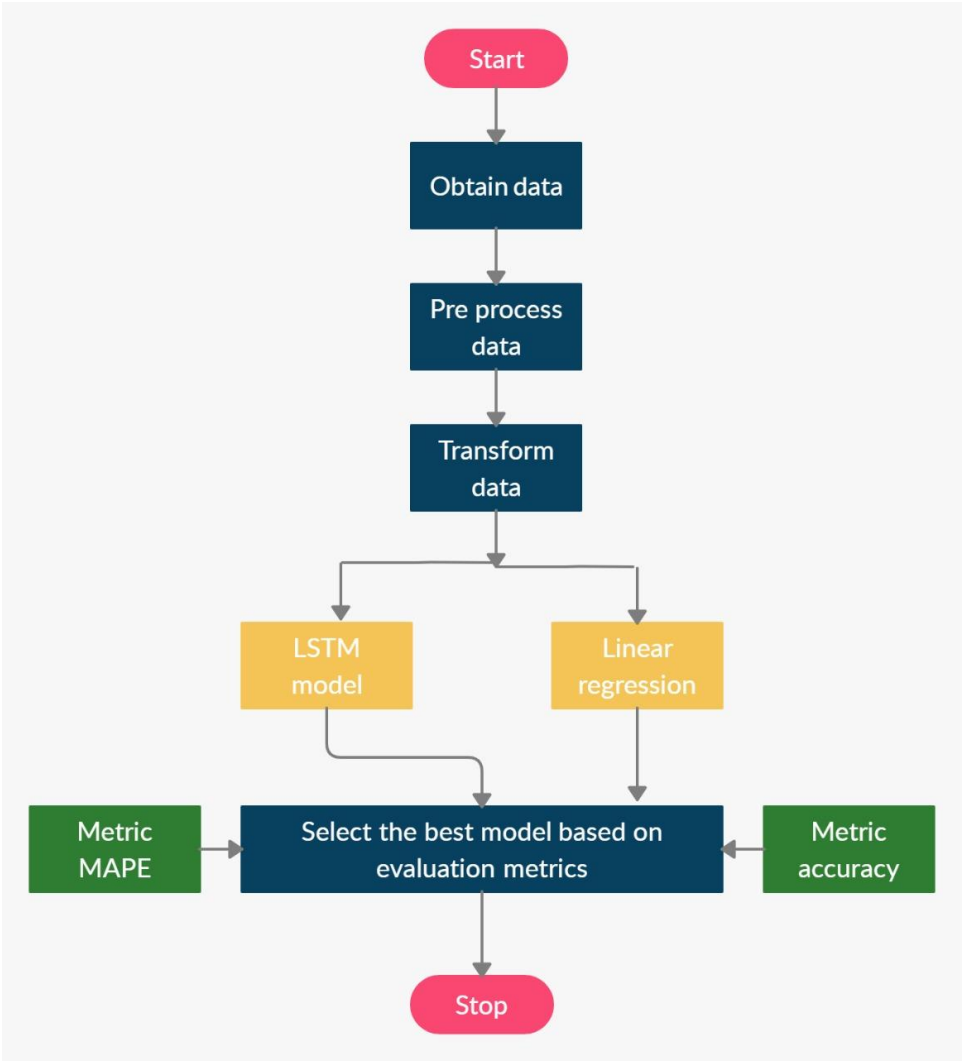

Figure 2: Flowchart for proposed methodology

3.3.1 Linear Regression

Linear regression based models are generally used for prediction tasks. The technique is used which tries to best fit the value to a linear line.This line can be used to relate both the predicting and predicted value.When there is more than one value then the

In case of exponential relations, linear regression can not be directly used. But after transformation to a linear expression, even exponential relations can be predicted using linear regression. For example,

$$y = \alpha e^{\beta x} \tag{3}$$

Taking the log on both sides of the equation, we get:

$$\ln y = \ln \alpha + \beta x \tag{4}$$

This expression is of the form of a linear regression model:

$$y\prime = \alpha\prime + \beta x \tag{5}$$

3.3.2 LSTM Model

Long Short term memory (LSTM) is an recurrent neural network which is most effective for time series prediction.The model used in this case is sequential.As the data was time series and we needed to predict the best positive corona cases so this model was best for our study.The model was build using tensorflow keras framework and the models performance was evaluated on the mean absolute error percentage (MAPE). The proposed architecture of LSTM model is depicted in the figure 3 as:

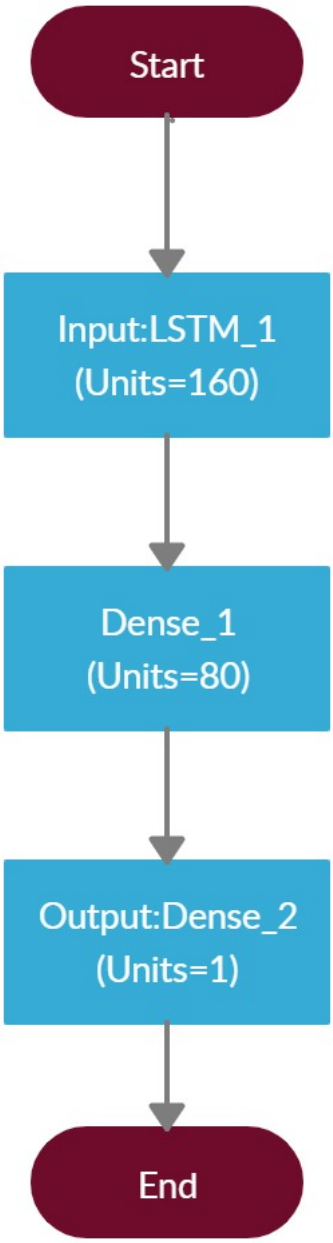

Figure 3: Architecture of LSTM model

4 Experiment Result

In LSTM prediction, LSTM layers use sequence of 180 nodes. Single layered structure followed by 2 Dense Layers with 60 nodes in the first layer and single node in the output layer is used as LSTM model for verifying prediction result. The best hyperparameters used is a batch size of 1. The result of the model is as shown 2

Table 2: Accuracy and MAPE of LSTM model

Model	Accuracy	MAPE	Middle East respiratory syndrome
LSTM model	96.90%	3.092%	

The prediction result is shown in figure below:

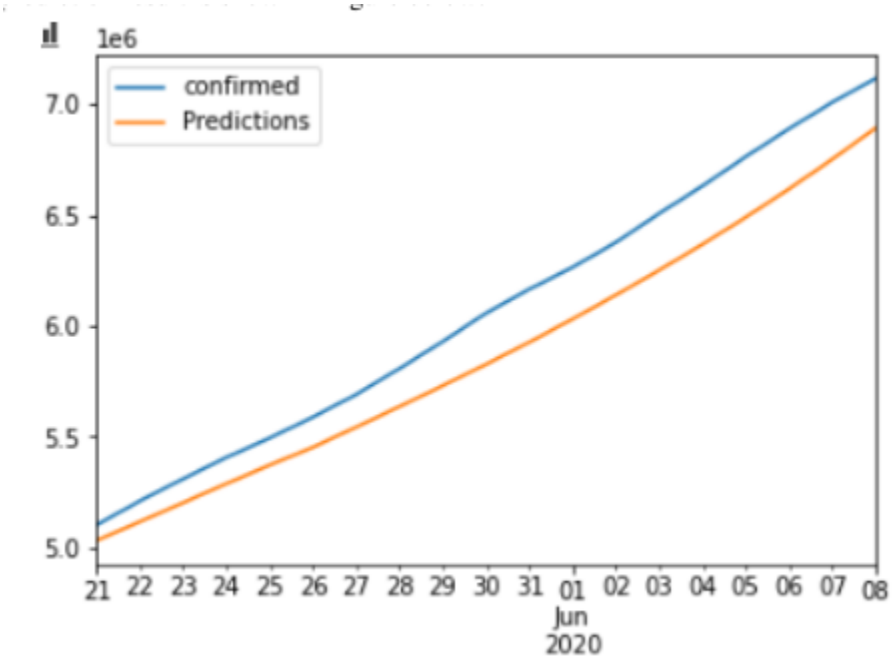

Figure 4: Comparison of predicted and true value using LSTM model

Linear regression model was used on the time series data and the date columns were taken as input and the 18 days data was predicted. The exponential fit of the model was fit and the result of the model is as shown 3

Table 3: Accuracy and MAPE of regression model

Model	Accuracy	MAPE
Linear model	93.57%	6.421%

The prediction result of comparing the test data predicted data is show below:

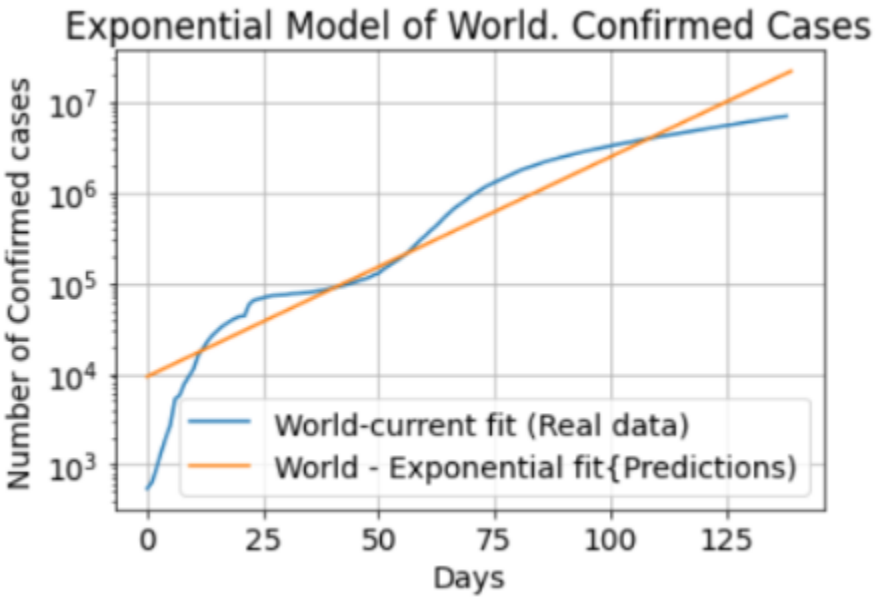

Figure 5: Comparison of predicted and true value using Linear Regression model

4.1 Comparing with other studies

In [4], they used an multi-step forecasting system on the population of china, and the estimated average errors are as show in 4

Table 4: Result [4]: Method and Average Errors

Model	Error
6-Step	1.64%
7-Step	2.27%
8-Step	2.14%
9-Step	2.08%
10-Step	0.73%

In [7], LSTM netwoks are used to on Canadian population, the reuslt are show is table 5

Table 5: Results [7]: Canadian Datasets

Model	RMSE	Accuracy
LSTM	34.63	93.4%

In [5], an deep learning based approach is proposed to compared the predicted forcasting value of LSTM and GRU model is used the result are as show in table 6:

Table 6: Results [7]: Canadian Datasets

Model	RMSE	Accuracy
LSTM	53.35	76.6%
GRU	30.95	76.9%
LSTM and GRU	30.15	87%

5 Conclusion and Future Scope

The comparison between Regression and LSTM model signifies that using LSTM yields better results for the forecasting the spread of confirmed cases. showcases a method that checks occurred cases of COVID-19. However it could be made automated to train on the updated data every week and see the predicted value. Also the model is trained only on confirmed cases same could be done for both recovered and death cases and predicted values could be found. The model shows only the worldwide cases however the dataset also provides country wise statistics so it can be used by different country to forecast the future outcome of the pandemic and take necessary preventive measures to be safe from this worldwide pandemic. A conclusion is drawn that shows forecasting models could be used by medical and government agencies to make better policies for controlling the spread of pandemic. The comparison between the 2 models allows them to choose the better suited model for the required task. The availability of high- quality and timely data in the early stages of the outbreak collaboration of the researchers to analyze the data could have positive effects on health care resource planning.

References

[1] *World health organization. who statement regarding cluster of pnemonia cases in wuhan, china, 2020.*

[2] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in wuhan, china," *The lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[3] G. M. Knight, N. J. Dharan, G. J. Fox, N. Stennis, A. Zwerling, R. Khurana and D. W. Dowdy, "Bridging the gap between evidence and policy for infectious diseases: How models can aid public health decision-making," *International journal of infectious diseases*, vol. 42, pp. 17–23, 2016.

[4] Z. Hu, Q. Ge, L. Jin and M. Xiong, "Artificial intelligence forecasting of covid-19 in china," *arXiv preprint arXiv:2002.07112*, 2020.

[5] S. K. Bandyopadhyay and S. Dutta, "Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release," *medRxiv*, 2020.

[6] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi and S. R. N. Kalhori, "Predicting covid-19 incidence through analysis of google trends data in iran: Data mining and deep learning pilot study," *JMIR Public Health and Surveillance*, vol. 6, no. 2, e18828, 2020.

[7] V. K. R. Chimmula and L. Zhang, "Time series forecasting of covid-19 transmission in canada using lstm networks," *Chaos, Solitons & Fractals*, p. 109 864, 2020.

[8] D. Fanelli and F. Piazza, "Analysis and forecast of covid-19 spreading in china, italy and france," *Chaos, Solitons & Fractals*, vol. 134, p. 109 761, 2020.

[9] S. Bouktif, A. Fiaz, A. Ouni and M. A. Serhani, "Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches," *Energies*, vol. 11, no. 7, p. 1636, 2018.

[10] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai *et al.*, "Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions," *Journal of Thoracic Disease*, vol. 12, no. 3, p. 165, 2020.

[11] G. Rainisch, E. A. Undurraga and G. Chowell, *A dynamic modeling tool for estimating healthcare demand from the covid19 epidemic and evaluating population-wide interventions*, 2020. arXiv: `2004.13544 [q-bio.PE]`.

[12] R. Singh and P. K. Singh, *Connecting the dots of covid-19 transmissions in india*, 2020. arXiv: 2004.07610 [cs.SI].

[13] H. H. Elmousalami and A. E. Hassanien, "Day level forecasting for coronavirus disease (covid-19) spread: Analysis, modeling and recommendations," *arXiv preprint arXiv:2003.07778*, 2020.

[14] K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. Hyman, P. Yan and G. Chowell, "Real-time forecasts of the covid-19 epidemic in china from february 5th to february 24th, 2020," *Infectious Disease Modelling*, vol. 5, pp. 256–263, 2020.

[15] K. Aritra, B. Tushar and A. Roy, "Detailed study of covid-19 outbreak in india and west bengal," vol. 5, Jan. 2020. DOI: 10.5281/zenodo.3865821.

[16] A. Tomar and N. Gupta, "Prediction for the spread of covid-19 in india and effectiveness of preventive measures," *Science of The Total Environment*, p. 138 762, 2020.

[17] S. Tuli, S. Tuli, R. Tuli and S. S. Gill, "Predicting the growth and trend of covid-19 pandemic using machine learning and cloud computing," *Internet of Things*, p. 100 222, 2020.

[18] R. Salgotra, M. Gandomi and A. H. Gandomi, "Time series analysis and forecast of the covid-19 pandemic in india using genetic programming," *Chaos, Solitons & Fractals*, p. 109 945, 2020.

[19] G. S. Randhawa, M. P. M. Soltysiak, H. El Roz, C. P. E. de Souza, K. A. Hill and L. Kari, "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study," *PLOS ONE*, vol. 15, no. 4, Apr. 2020. DOI: 10.1371/journal.pone.0232391. [Online]. Available: https://doi.org/10.1371/journal.pone.0232391.

[20] R. Salgotra, *Covid-19: Time series datasets india versus world*, May 2020. [Online]. Available: http://dx.doi.org/10.17632/tmrs92j7pv.1.

[21] Tathagatbanerjee, *Covid-19 analytics india*, Apr. 2020. [Online]. Available: https://www.kaggle.com/tathagatbanerjee/covid-19-analytics-india.

[22] A. Palladino, V. Nardelli, L. G. Atzeni, N. Cantatore, M. Cataldo, F. Croccolo, N. Estrada and A. Tombolini, *Modelling the spread of covid19 in italy using a revised version of the sir model*, 2020. arXiv: 2005.08724 [physics.soc-ph].

[23] A. Koubaa, *Understanding the covid19 outbreak: A comparative data analytics and study*, 2020. arXiv: 2003.14150 [q-bio.PE].

[24] S. Boccaletti, W. Ditto, G. Mindlin and A. Atangana, "Modeling and forecasting of epidemic spreading: The case of covid-19 and beyond," *Chaos, Solitons, and Fractals*, vol. 135, p. 109 794, 2020.

[25] C. Anastassopoulou, L. Russo, A. Tsakris and C. Siettos, "Data-based analysis, modelling and forecasting of the covid-19 outbreak," *PloS one*, vol. 15, no. 3, e0230405, 2020.