Forecasting Rate of Spread of COVID-19

A report submitted in partial fulfilment of the
requirement for the award of certification of
IN-HOUSE SUMMER TRAINING
in
Machine Learning and Deep Learning

By

Ashwin Goyal

(Roll Number:01711502818)

(Branch:ECE-1)

Kartik Puri

(Roll Number:03811502818)

(Branch:ECE-1)

under the guidance of

Dr. Rachna Jain

Assistant Professor, Computer Science and Engineering

Electronics and Communication Engineering
Bharati Vidyapeeth's College of Engineering,
New Delhi – 110063, INDIA

May 2020-June 2020

Table of Contents

List of Figures

List of Tables

Certificate

I hereby certify that the work which is being submitted in this report titled **"Forecasting rate of spread of COVID19 "**, in partial fulfilment of the requirement for the award of certification of "In-House Summer Training in Machine Learning and Deep Learning" submitted in Bharati Vidyapeeth's College of Engineering, New Delhi, is an authentic record of my own work carried out under the supervision of "Name of the Supervisor" and refers to other researchers work which are duly listed in the reference section.

The matter presented in this report has not been submitted for the award of any other certificate of this or any other institution.

Ashwin Goyal

Roll No. 01711502818

Branch. ECE-1

Kartik Puri

Roll No. 03811502818

Branch. ECE-1

This is to certify that the statements made above by the candidate are correct and true to the best of our knowledge.

Dr. Rachna Jain

Assistant Professor

Computer Science & Engineering

BVCOE

New Delhi - 110063

Mrs Preeti Nagrath

Assistant Professor

Computer Science & Engineering

BVCOE

New Delhi - 110063

The Viva-Voce Examination of _____ has been held on _____.

Internal Examiner

External Examiner

Acknowledgement

I respect and thank Dr. Rachna Jain, Mrs Preeti Nagrath, Mrs Nikita Sharma, Mr Ashish Kumar, for providing me an opportunity to do the project work on topic ‘Forecasting the rate of spread of COVID-19’ and giving us all support and guidance which made me complete the project duly. I am extremely thankful to them for providing such a nice support and guidance, although he had busy schedule managing the corporate affairs.

The success and final outcome of this project required a lot of guidance and assistance and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them all.

Abstract

These days,Covid-19 coronavirus created a global health crisis and had impact on economic and social life of different countries of the world. The objective of this study determines if it is good to use machine learning and deep learning algorithms to evaluate how much prediction results are close to original data related to Confirmed-Recovered-Death cases of Covid-19.For This purpose, a verification method is proposed in this paper that uses the concept of Deep-learning Neural Network. In This framework, used are Long short-term memory (LSTM) And Linear Regression The model are trained using the training data and are compared to the original result to see the error and feasibility of the model. The prediction results are validated against the original data based on some predefined metric. The experimental results showcase that the proposed approach is useful in generating suitable results based on the critical disease outbreak.
It also helps doctors to recheck further verification of virus by the proposed method. The outbreak of Coronavirus has the nature of exponential growth and so it is difficult to control with limited clinical persons for handling a huge number of patients with in a reasonable time. So it is necessary to build an automated model, based on machine learning approach.
For corrective measure after the decision of clinical doctors. It could be a promising supplementary confirmation method for frontline clinical doctors. The proposed method has a high prediction rate and works fast for probable accurate identification of the disease. The performance analysis shows that a high rate of accuracy is obtained by the proposed method.

1.Introduction

Coronaviruses are a group of related RNA viruses that cause diseases in mammals and birds. It is now known that pneumonia occurred in the city of Wuhan, China, in December of 2019. The World Health Organization (WHO) identified and named novel coronavirus as “2019-nCoV” which was later declared as Public Health Emergency of International Concern on January 30, 2020 and further on March 11, 2020 this Covid-19 was characterized as Pandemic [1].

The two types of coronaviruses, named as, severe acute respiratory syndrome coronavirus (SARS-COV) and Middle East respiratory syndrome coronavirus (MERS-COV) have affected more than 20,000 people in past decade[2]. According to the Centres for Disease Control and Prevention (CDC), this novel coronavirus has some similarities with SARS-COV and MERS-COV. These diseases are spread through respiratory droplets from one human being to other. Respiratory infections can be transmitted through droplets of different sizes: when the droplet particles are >5-10 μm in diameter they are referred to as respiratory droplets, and when then are <5μm in diameter, they are referred to as droplet nuclei. According to current evidence, COVID-19 virus is primarily transmitted between people through respiratory droplets and contact routes. In an analysis of 75,465 COVID-19 cases in China, airborne transmission was not reported. Symptoms as fever, cough, and shortness of breath after a period ranging from 2 to 14 days are observed as the outcomes of the disease. Detailed investigations found that SARS-CoV was transmitted from civet cats to humans in China in 2002 and MERS-CoV from dromedary camels to humans in Saudi Arabia in 2012. Several known coronaviruses are circulating in animals that have not yet infected humans.

For helping combat coronavirus machine learning and deep learning models are used 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.

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 real and predicted cases are compared on a predefined metrics. A comparative study is drawn to see the performance of LSTM and Linear regression model to see which model best fit for the data.

The report is divided into various sections.The literature review section talks about the related work done by the researcher and scientist on this topic as well as the methodology used by them.

The methodology used in the paper and the approach to handle the problem is also discussed. The Methods and models talks about the dataset used and its features. Since the classification is done worldwide so the data was processed to suit the needs of the model in use and a brief description of processed dataset is also provided.

Next, Evalutation metrics are discussed to understand and compare the result between the two models used.MAPE (Mean Absolute Percentage Error) 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, and plot the models.

2.Literature Review

In [3], an AI based approach is proposed as an alternative to epidemiological model for monitoring transmission dynamics for Covid-19. This AI based approach is executed by implementing modified stacked auto-encoder model.

In [4], an deep learning based approach is proposed to compared the predicted forcasting 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 [5],LSTM and Linear regression model was use 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 [6], Several linear and non-linear machine learning algorithms were trained and picked the best one as baseline, after that chose the best features using wrapper and embedded feature selection methods and finally used genetic algorithm (GA) to find 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 [13], an LSTM networks based approach is proposed for forecasting time series data of COVID19 in Canada. This paper uses LSTM network to overcome problems faced by linear model where algorthms assigns high probability and neglects temporal information leading to biased predictions.

In [14] a computation and analysis based on Suspected-Infected-Recovered-Dead (SIRD) model is provided. Based on the dataset, it estimates the parameters, i.e. the basic reproduction number (R0) and the infection, recovery and mortality rates, Computations on SIRD model, this R0 parameter value turn out to be 2.5.

In [17],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[18] ,a transmission network visualization (s) of COVID-19 in India was created and analysis was performed upon them. Using the transmission networks obtained there was an attempt to find the possible Super Spreader Individual (s) and Super Spreader Events (SSE) responsible for the outbreak in their respective regions.

In[20], it presents a comparison of day level forecasting models on COVID-19 affected cases using time series models and mathematical formulation. The forecasting models and data are used to suggest that the number of coronavirus cases grows exponentially in countries that do not mandate quarantines.

In[21],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, the epicenter of the epidemic, and for the overall trajectory in China.

3.Work carried out

3.1. Evaluation metrics

To identify best model, it is necessary to put concentration on comparison of measures of the algorithm's performance. In this report, following parameters are used for measuring algorithm's performance-

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

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

   Where y is true value and y' is predicted value

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

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

3.2. Models

In this report, a prediction of confirmed cases of COVID 19 Corona virus are obtained using a Recurrent Neural Network method(LSTM) and Linear Regression. Linear regression is a linear model, that assumes a linear relationship between the input variables (x) and the single output variable (y).When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, method is referred to as as multiple linear regression.

A Recurrent Neural Network (RNN) is kind of neural network architecture that considers both sequential and parallel information processing. Incorporating memory cells to neural network; it is possible to simulate the operations similar to human brain.There are alternatives from RNN depending on the gating units, such as Long Short Term Memory (LSTM) RNN and Gated Recurrent Unit (GRU) RNN.Traditional RNN lacks of considering context based prediction, which can be overcome by introducing Long short-term memory (LSTM). LSTM has a good potential to regulate gradient flow and enable better preservation of long-range dependencies [7].

The dataset used for predicting the value is taken from John Hopkin University which contains cases form 22 January 2020 to 8 June,2020.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 the prediction result are measured with respect to performance measures metrics such as MAPE and accuracy.The methodology performed for each of the step is shown in the figure 1 as follows:

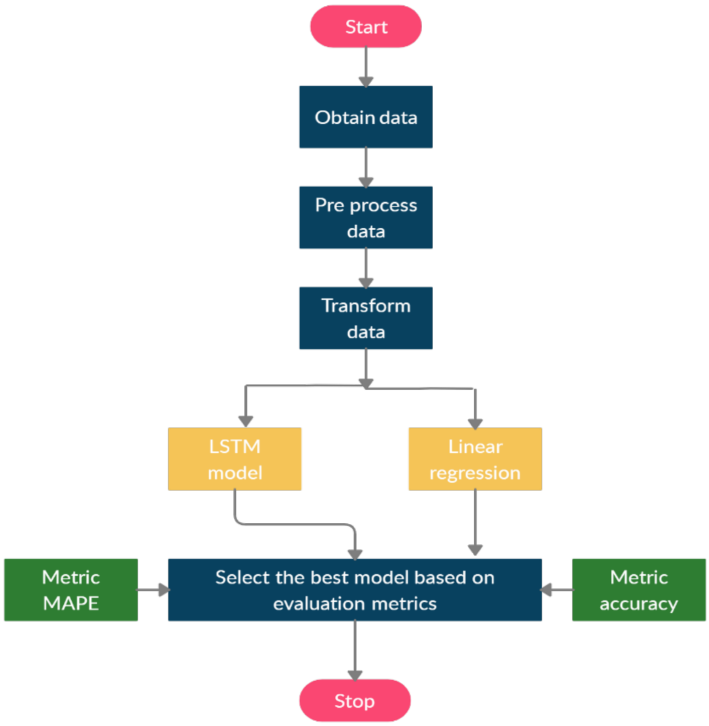

Figure 1:Flowchart for proposed methodology

3.3Proposed Linear Regression Model:

Linear regression is used for prediction tasks.In a problem wuth one predicting value this 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 model used is multiple linear regression which is used in this case of our study.

3.4Proposed LSTM Model:

Long Short term memory 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) and accuracy metrics.The proposed architecture of LSTM model is shown in the figure as:

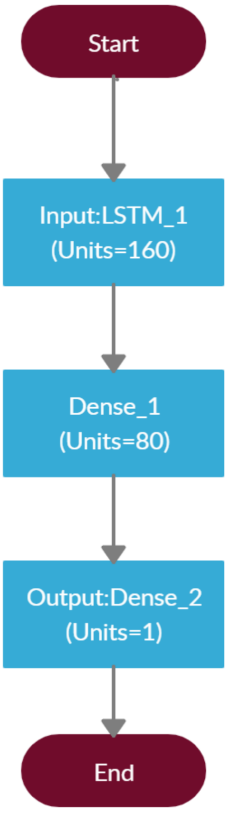

Figure 2:Architecture of LSTM model

4.Experimental Result

In LSTM model LSTM layers use sequence of 50 nodes. A 2 layered structure followed by a Dense Layer is used as LSTM model for verifying prediction result. The best hyper-parameters used is a batch size of 1.The prediction accuracy of the model is shown in table 1.

Model	Accuracy	MAPE
LSTM model	96.90%	3.092%

Table 1:Accuracy and MAPE of LSTM model

The prediction result is shown in figure below:

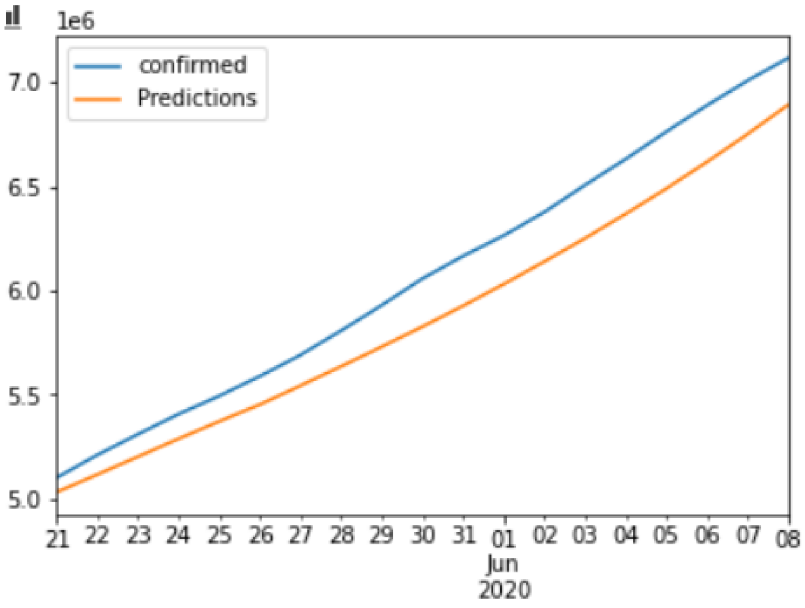

Figure 3:Comparison of predicted and true value using LSTM model

Linear regression was also applied on the time series data and the date columns were taken as input and the 18 days data was predicted.The model was fit with exponential data and the prediction and accuracy were calculated.The prediction accuracy of the model is shown in the table 2 below:

Model	Accuracy	MAPE
Regression model	93.57%	6.421%

Table 2:Accuracy and MAPE of Regression model

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

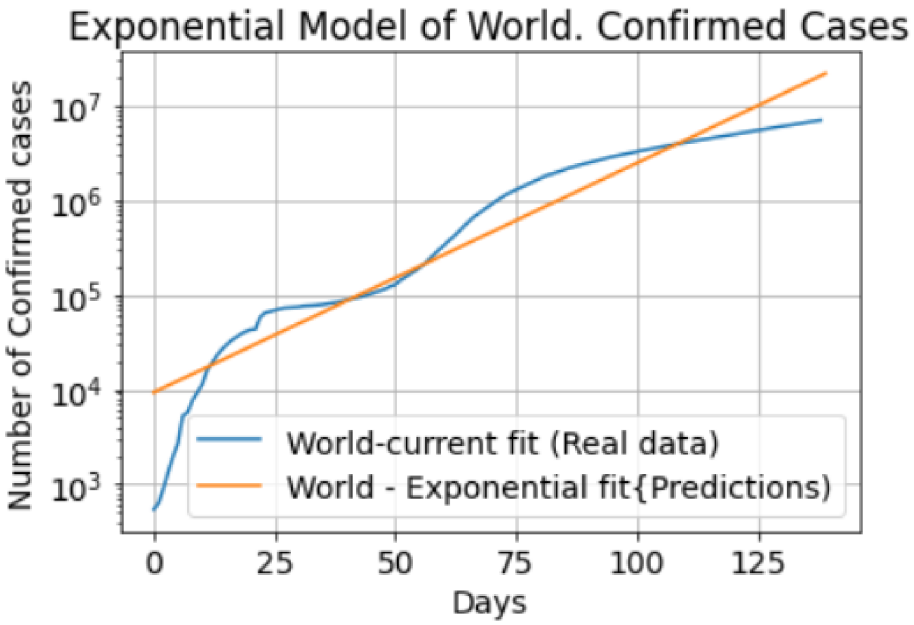

Figure 4: Comparison of predicted and true value using Linear Regression model

4.1 Comparing with other studies

In [3], they used a multi-step forecasting system on the population of china, and the estimated average errors are as show in table 3

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

Table 3: Result [3]: Method and Average Errors

In [13], LSTM netwoks are used to on Canadian population, the result are shown in table 4:

Model	RMSE	Accuracy
LSTM	34.63	93.4%

Table 4: Results [13]: Canadian Datasets

In [4], an deep learning based approach is pro-posed to compared the predicted forecasting value of LSTM and GRU model is used the result are shown in table 5:

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

Table 5: Results [13]: Canadian Datasets

5. Conclusion and Future Scope

The experimental analysis showed that the LSTM model has a accuracy of 96.90% whereas regression model has accuracy of 93.57%.The comparison between Regression and LSTM model signifies that LSTM provides a comparatively better results in terms of prediction of confirmed, released, negative, death cases on the data. This paper presented a novel method that could check occurred cases of COVID-19 manually. 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. It could be a promising supplementary rechecking method for frontline

clinical doctors. It is now essential for improving the accuracy of detection process. In conclusion, the data mining models could help policymakers and health managers to plan health care resources and control the prevention of an epidemic outbreak. 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 Pneumonia Cases in Wuhan, China, 2020.*

[2] C. Huang *et al.,* "Clinical features of patients infected with 2019 novel coronavirus in Wuhan China," *Lancet,* vol. 395, no. 10223, pp. 497–506, 2020, doi: 10.1016/S0140-6736(20)301835

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

[4] Bandyopadhyay, Samir Kumar, and Shawni Dutta. "Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release." *medRxiv* (2020).

[5] Ayyoubzadeh, Seyed Mohammad, et al. "Predicting COVID-19 incidence through analysis of google trends data in iran: data mining and deep learning pilot study." *JMIR Public Health and Surveillance* 6.2 (2020): e18828.

[6] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, "Optimal deep learning LSTM model for electricload forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches," *Energies*, vol. 11, no. 7, 2018, doi: 10.3390/en11071636.

[7] Tomar, Anuradha, and Neeraj Gupta. "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures." *Science of The Total Environment* (2020): 138762.

[8] Pal, Ratnabali, Arif Ahmed Sekh, Samarjit Kar, and Dilip K. Prasad. "Neural network based country wise risk prediction of COVID-19." *arXiv preprint arXiv:2004.00959* (2020).

[9] Pandey, Gaurav, Poonam Chaudhary, Rajan Gupta, and Saibal Pal. "SEIR and Regression Model based COVID-19 outbreak predictions in India." *arXiv preprint arXiv:2004.00958* (2020).

[10]Yang, Zifeng, Zhiqi Zeng, Ke Wang, Sook-San Wong, Peng Liu et al. "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions." *Journal of Thoracic Disease* 12, no. 3 (2020): 165.

[11] Zheng, Nanning, Shaoyi Du, Jianji Wang, He Zhang, Wenting Cui, Zijian Kang, Tao Yang et al. "Predicting covid-19 in china using hybrid AI model." *IEEE Transactions on Cybernetics* (2020).

[12]C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, “Data-Based Analysis, Modelling andForecasting of the novel Coronavirus (2019-nCoV) outbreak,” *medRxiv,* no. February, p. 2020.02.11.20022186, 2020, doi: 10.1101/2020.02.11.20022186.

[13]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.

[14] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, “Data-Based Analysis, Modelling and Forecasting of the novel Coronavirus (2019-nCoV) outbreak,” *medRxiv*, no. February, p. 2020.02.11.20022186, 2020, doi: 10.1101/2020.02.11.20022186.

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

[16]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].

[17]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].

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

[19]A. Koubaa, *Understanding the covid19 outbreak: A comparative data analytics and study*, 2020. arXiv: 2003 . 14150 [q-bio.PE].

[20]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.

[21]K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. Hyman, P. Yan and G. Chowell, "Real-time forecasts of the covid19 epidemic in china from february 5th to february 24th, 2020", *Infectious Disease Modelling*, vol. 5, pp. 256–263, 2020.

[22]S. Boccaletti, W. Ditto, G. Mindlin and A. Atangana, "Modeling and forecasting of epidemic spreading: The case of covid19 and beyond", *Chaos, Solitons, and Fractals,* vol. 135, p. 109 794, 2020.

Appendix

1.Link to the code: https://github.com/kartikpuri99/forecasting-rate-of-spread-of-COVID-19