Forecasting Rate of Spread of Covid19 using Linear Regression and LSTM

Kartik Puri
kartikpuri99@gmail.com

Ashwin Goyal
ashwingoyal180@gmail.com

Department of Electronics & Communication Engineering
Bharati Vidyapeeths College Of Enginnering
12th June 2020

Abstract

The 2019 novel coronavirus also known as SARS-Cov-2, generally knows as COVID-19 virus has spread all over the world. On March 11^{th} 2020, World Health Organisation (WHO) declared the 2019 novel corona virus as global pandemic. Corona virus, was first originated in Wuhan, China. In recent days, Covid19 has been an immense impact on social, economic fiels 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, the respiratory disease origination from coronavirus occured in Wuhan, China, is on the rise and has shaken the world. The World Health Organization named the disease COVID-19 when the first case of this virus was reported.

The Global spread of COVID19 has affected most countries 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 contagagious respiratory virus that first started in Wuhan December 2019. [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$> 5-10\mu m$ in diameter they are referred to as respiratory droplets, and when then are $< 5\mu m$ 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. Droplet transmission occurs when a person is in in close contact (within 1 m) with someone who has respiratory symptoms (e.g., coughing or sneezing) and is therefore at risk of having his/her mucosae (mouth and nose) or conjunctiva (eyes) exposed to potentially infective respiratory droplets. 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.

Most of the data driven approaches used in previous studies [15] 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 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 for the data.

The section **Literature Review** talks about similar work done by other reseachers 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 [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 [7], an deep learning based approach is proposed to campared 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 predifined metrics.

In [8], 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 [4], 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 [10], temporal dynamics of the corona virus outbreak in China, Italy, and France in the span of three months are analyzed.

In [23], 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 [24], temporal dynamics of the corona virus outbreak in China, Italy, and France in the span of three months are analysed.

In [25], 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,

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

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.

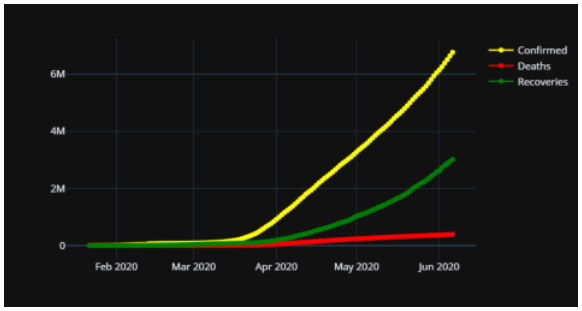

Figure 1: Number of cases around the world

3 Methods and models

3.1 Data

The dataset used was the 2019 Novel Coronavirus Dataset operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE).

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 predictingnext 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

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 Error Percentage Error**: It is defined by the following formula:

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

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

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

$$Accuracy = (100 - MAPE)\% \qquad (2)$$

3.3 Method

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

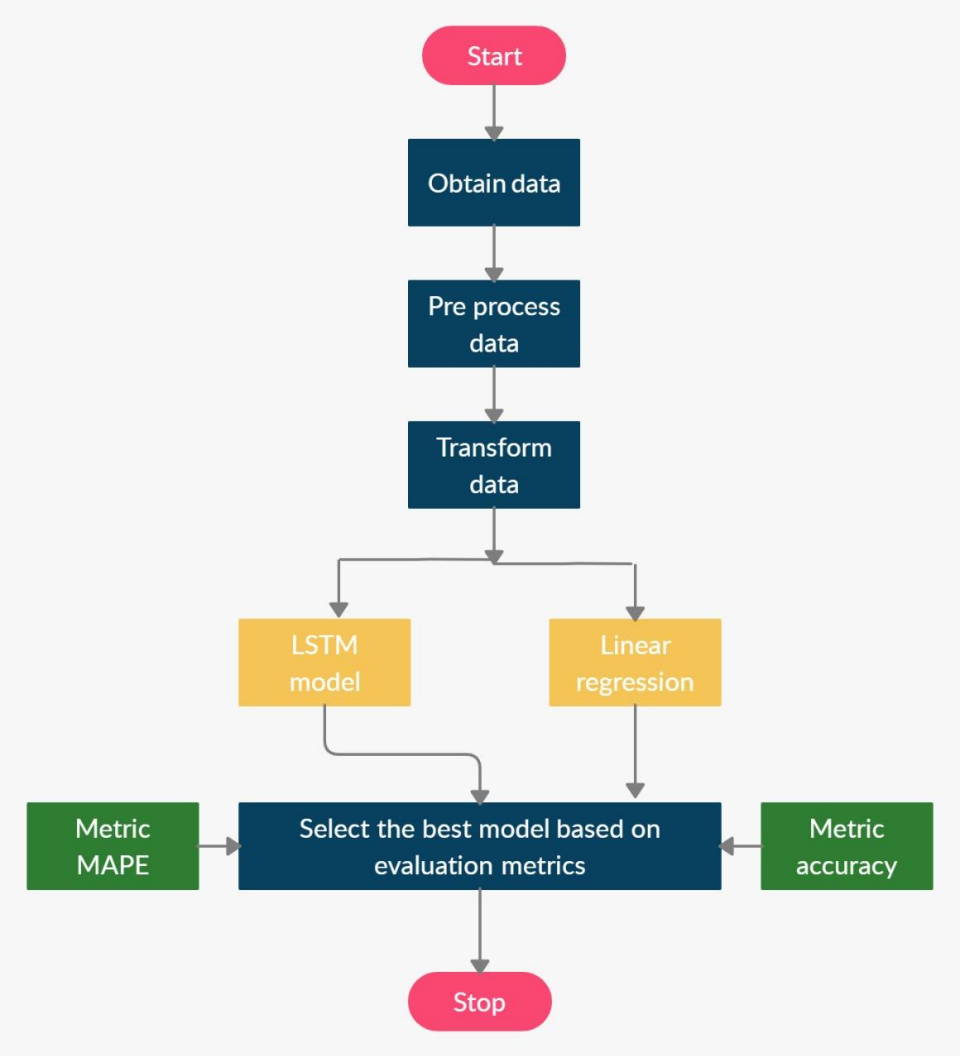

Figure 2: Flowchar for proposed methodology

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 [6].

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 2 as show.

3.3.1 Linear Regression

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

Sometimes linear regression can be used with relationships which are not inherently linear, but can be made to be linear after a transformation. In particular, we consider the following exponential model:

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

Taking the natural log or both sides of the equation, we have the following equivalent equation:

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

This equation has the form of a linear regression model:

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

3.3.2 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:

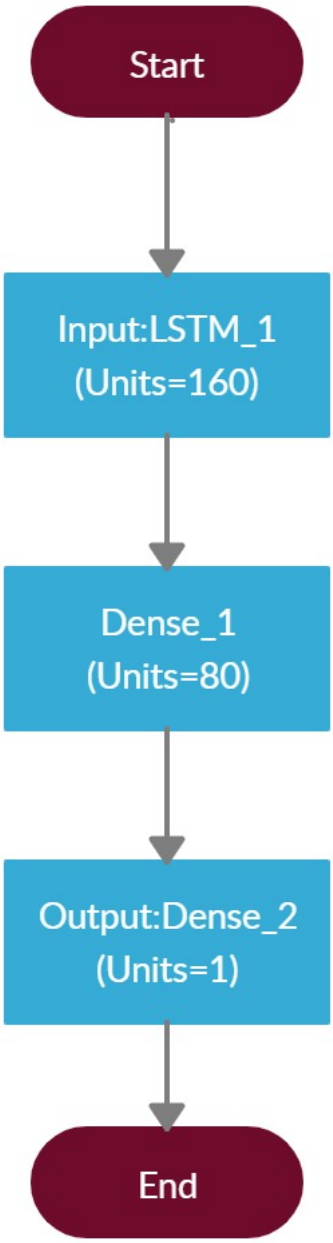

Figure 3: Architecture of LSTM model

4 Experiment 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 hyperparameters used is a batch size of 1.The prediction accuracy of the model is shown in table 2

Table 2: Accuracy and MAPE of LSTM model

Model	Accuracy	MAPE
LSTM model	96.90%	3.092%

The prediction result is shown in figure below:

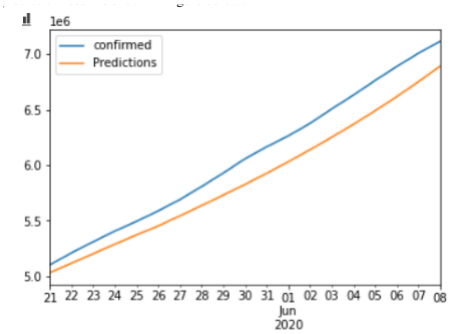

Figure 4: Comparison of predicted and true value using LSTM model

Linear regression was also applied 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 prediction accuracy of the model is shown in table 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:

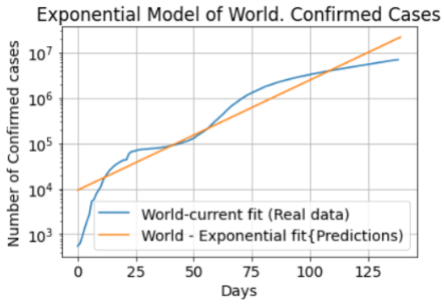

Figure 5: Comparison of predicted and true value using Linear Regression model

4.1 Comparing with other studies

In [3], 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 [3]: 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 [4], LSTM netwoks are used to on Canadian population, the reuslt are show is table 5

Table 5: Results [4]: Canadian Datasets

Model	RMSE	Accuracy
LSTM	34.63	93.4%

In [7], an deep learning based approach is proposed to campared the predicted forcasting value of LSTM and GRU model is used the result are as show in table 6:

Table 6: Results [4]: 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 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 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] Z. Hu, Q. Ge, L. Jin and M. Xiong, "Artificial intelligence forecasting of covid-19 in china", *arXiv preprint arXiv:2002.07112*, 2020.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[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 covid-19 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 covid-19 and beyond", *Chaos, Solitons, and Fractals*, vol. 135, p. 109 794, 2020.

[23] 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.

[24] 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.

[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.