Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

Long Short-Term Memory Approach for Coronavirus Disease Prediction

Omar Ibrahim Obaid

Department of Computer Science, College of Education, AL-Iraqia University, Baghdad, Iraq. E-mail: alhamdanyomar23@gmail.com

Mazin Abed Mohammed*

*Corresponding Author, Ph.D., College of Computer Science and Information Technology, University of Anbar, Ramadi, 31001, Iraq. E-mail: mazinalshujeary@uoanbar.edu.iq

Salama A. Mostafa

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, 86400, Malaysia. E-mail: salama@uthm.edu.my

Abstract

Corona Virus (COVID-19) is a major problem among people, and it causes suffering worldwide. Yet, the traditional prediction models are not yet suitably efficient in catching the fundamental expertise as they cannot visualize the difficulty in the health's representation problem areas. This paper states prediction mechanism that uses a model of deep learning called Long Short-Term Memory (LSTM). We have carried this model out on corona virus dataset that obtained from the records of infections, deaths, and recovery cases across the world. Furthermore, producing a dataset which includes features of geographic regions (temperature and humidity) that have experienced severe virus outbreaks, risk factors, spatio-temporal analysis, and social behavior of people, a predictive model can be developed for areas where the virus is likely to spread. However, the outcomes of this study are justifiable to alert the authorities and the people to take precautions.

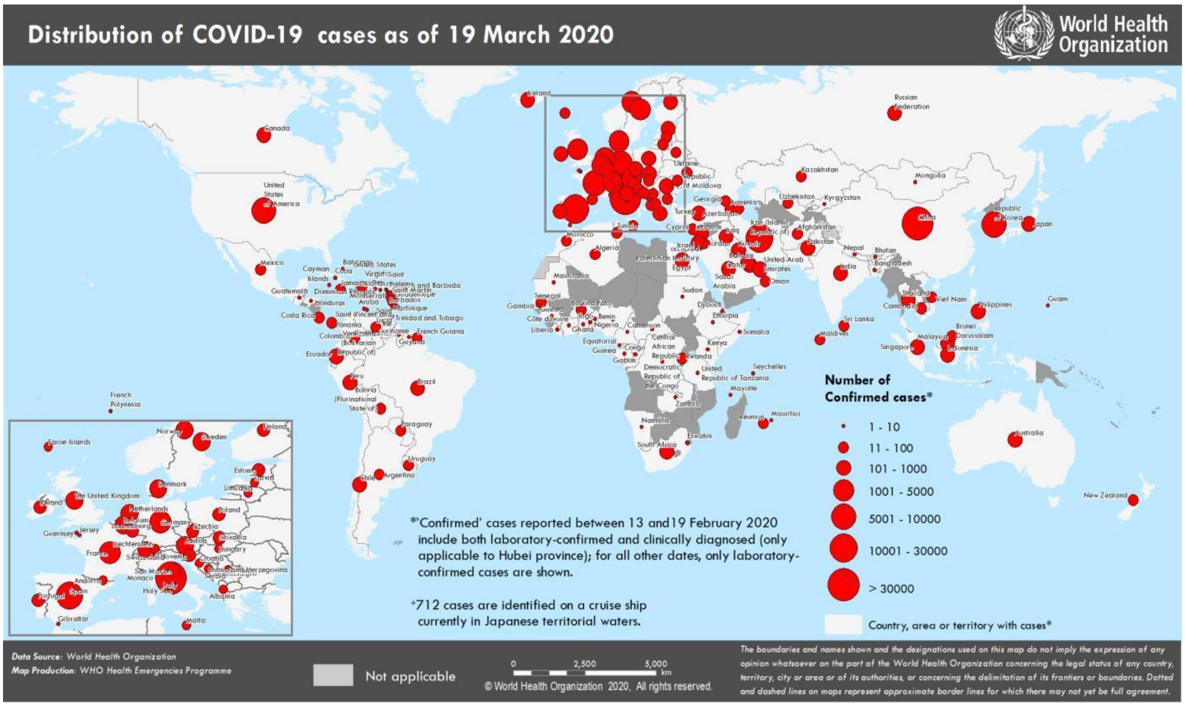
Keywords: Deep learning, LSTM, Prediction, COVID-19, Recurrent Neural Network (RNN).

DOI: 10.22059/jitm.2020.79187

Introduction

Some say that prevention is better than healing, and this is a global fact. Nowadays, number of coronavirus (COVID-19) infections were 428,405, and the deaths was over 19,000 at the time of writing this research paper in April according to the World Health Organization (WHO.,2020). In addition, (COVID-19) belongs to a big family of viruses which bring about sickness that is more dangerous than a popular cold as reported by (WHO.,2020).

COVID-19 represents a novel strain which has not formerly recognized in humans before. This virus spreads through humans and the popular marks of contagion contain symptoms of a respiratory system like Pneumonia, coughing, and difficulty of breathing (Mohammed, Abdulkareem, Al-Waisy, et al., 2020). The virus has showed up at first in Wuhan, China of 2019, they have found it in humans around the cities of China and other 24 countries in the world. It is very difficult to be diagnosed, as its symptoms are similar to those of the flu or even bad cold, so a laboratory test is required to confirm the diagnosis (Al-Dhief et al., 2020). Figure 1 shows the confirmed cases globally as reported by World Health Organization as of 01 March 2020 (E.V. et al., 2020).



*'Confirmed' cases reported between 13 and19 February 2020 include both laboratory-confirmed and clinically diagnosed (only applicable to Hubei province); for all other dates, only laboratory-confirmed cases are shown.

+712 cases are identified on a cruise ship currently in Japanese territorial waters.

Data Source: World Health Organization
Map Production: WHO Health Emergencies Programme

© World Health Organization 2020, All rights reserved.

The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Figure 1. Distribution of COVID-19 (E.V. et al., 2020)

Figure 2 shows the confirmed cases of COVID-19 as reported outside China by time of report and WHO zone with entire days of notification during the duration between 30 Dec 2019 to 16 Mar 2020 (E.V. et al., 2020).

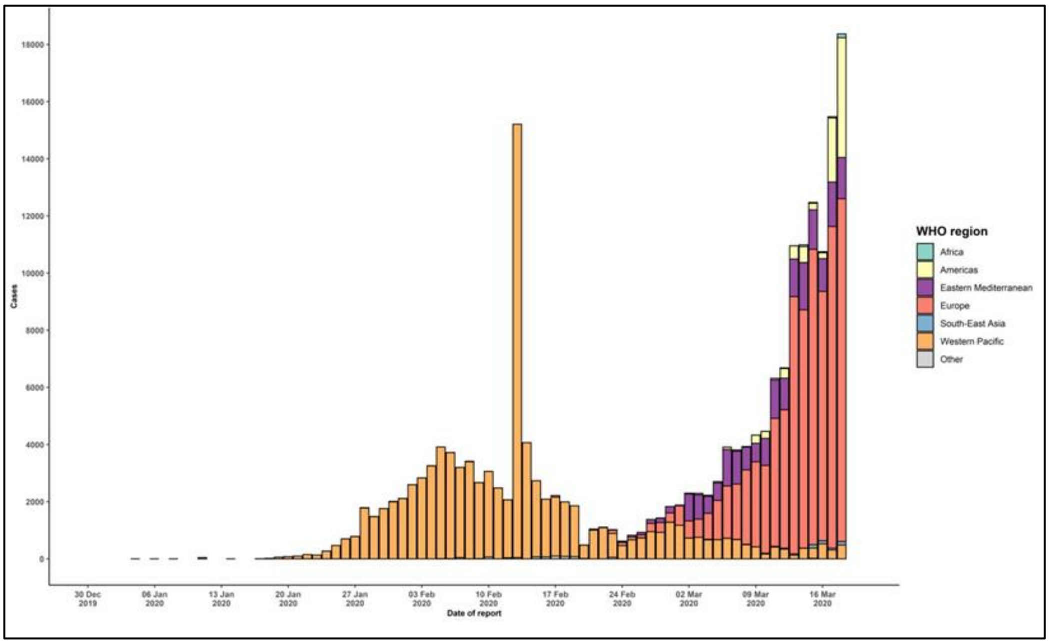


Figure 2. Confirmed Cases of COVID-19 outside China (E.V. et al., 2020)

The heath is the substantial factor in the life, and hence there is a necessity for prediction of illness (CHEN et al., 2019). In the last decade, many researchers around the world have applied deep learning approaches to predict illness based on the datasets in the medicine field. Recently, best outcomes have obtained by using deep learning algorithms for prediction of illness. Recurrent Neural Networks (RNNs) have utilized vastly in predicting diseases. They need the necessity for a technique with exceptionally top accuracy, given that the medical prediction is regarded a considerable mission which requires to executed accurately and effectively.

The related studies on COVID-19 identification based data analytics very limited due most of the researchers focus on medical images such as X-ray, and CT scan images. Some studies works on data predication like (Car et al., 2020) by Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron. Another study conducted by (Muhammad et al., 2020) they using data mining techniques for COVID-19 data. they used different methods such as support vector machine, K-nearest neighbor algorithm, decision tree, random forest, naive Bayes, and logistic regression on data have been collected from South Korea with patents have infected cases. The results shows the decision tree algorithm achieved a best accuracy of 99.85%. According to (Mondal et al., 2020) they discussed different aspects of COVID-19. This study presented the role of COVID-19 data analytics dataset from Johns Hopkins University of USA and Hospital Israelita Albert Einstein of Brazil. The results achieved by XGBoost, multilayer perceptron and logistic regression have successfully classified COVID-19 cases with a good accuracy of 91%

Due to (RNNs) has two major issues: a gradient disappearing and the gradient crashing, that most make it useless, Long Short-Term Memory (LSTM) is introduced to fix this case. Therefore, LSTM is used in this paper as it is preferable, especially when the datasets are of time series type.

Materials and Methods

We use a novel coronavirus 2020 dataset in this paper. We gained the CSV file of this dataset from (Github covid-19 datasets., 2020). The dataset has 8 columns and 28920 rows of cases for the period from 19 January until 22 March in the entire world. We have drawn all figures using a python library called (plotly). Figure 3 shows the confirmed infections within this period. Figure 4 illustrates the rate of deaths and recovery within this period. The rate of recovery cases raised up at the beginning of February 2020.

Figure 3. Confirmed infections cases

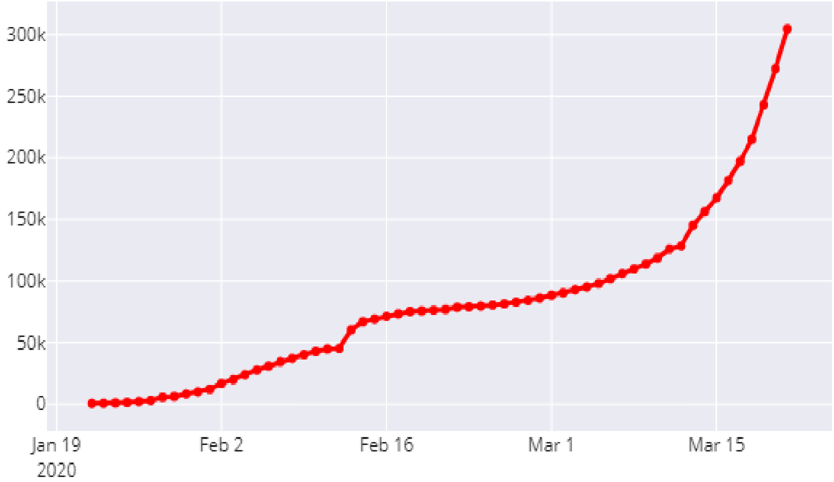


Figure 4. The rate of deaths and recovery

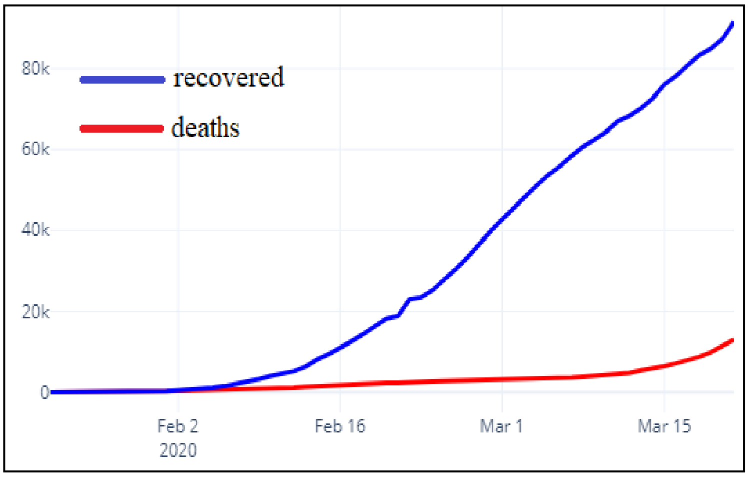


Figure 5 illustrates the rate of infections, recovered, and deaths. The deaths rate increased between 24 and 26 January before is decreased in 27 January. Infections rate has raised up slightly while recovered rate is increased at the end of January as well. Then, deaths rate is decreased continuously at the beginning of February while the recovered rate has decline as well but it still higher than the deaths rate until 17 February.

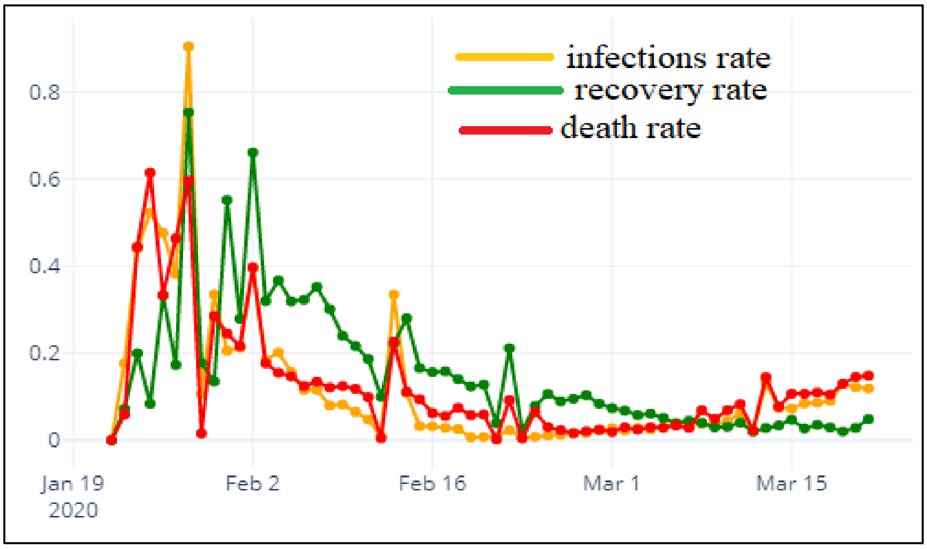


Figure 5. Rate of infections, recovered, and deaths

Google Collaboratory (colab) is used in this study, colab has created as google research project to support both educational and research of machine learning. It runs completely in the cloud so it need not be setup (Carneiro et al., 2018). Python 3.7 has used in this environment which is based on Jubyter notebook along with its most important libraries, especially TensorFlow and Keras which we use them in this research. TensorFlow is an open source library which has used for machine learning, it gives a strong experiment for an academic research. It allows us to train the model and then deploy it readily. Further, Keras is a library of neural networks at high levels and it runs on top of TensorFlow. The main use of Keras is for deep learning as it is easy and makes rapid prototype (Shanmugamani, 2018).

Recurrent Neural Networks (RNNs) is applied in many fields like prediction of time series and machine translation. Due to (RNNs) have two major issues: a gradient disappearing and the gradient crashing, that most make it useless, long short-term memory (LSTM) approach is solved this problem, and we use it in this paper as it is most suitable technique for time series problems. LSTM term was first introduced by Sepp Hochreiter in 1997 (Hochreiter & Schmidhuber, 1997). Figure 6 shows the simple (RNN) with LSTM cell (Donahue et al., 2015). The architecture of LSTM differs from RNN architecture. LSTM integrates the units of memory to authorize the temporal dynamics of learning. In addition, bidirectional architecture of LSTM has solved the problem of dependency of outcomes based on prior inputs (Graves & Schmidhuber, 2005). The units permit learning when to pass prior hidden states and also update them when they get new data. The main aim of the gates is to set the interactions between the cell of memory and its environment (Rassem et al., 2017).

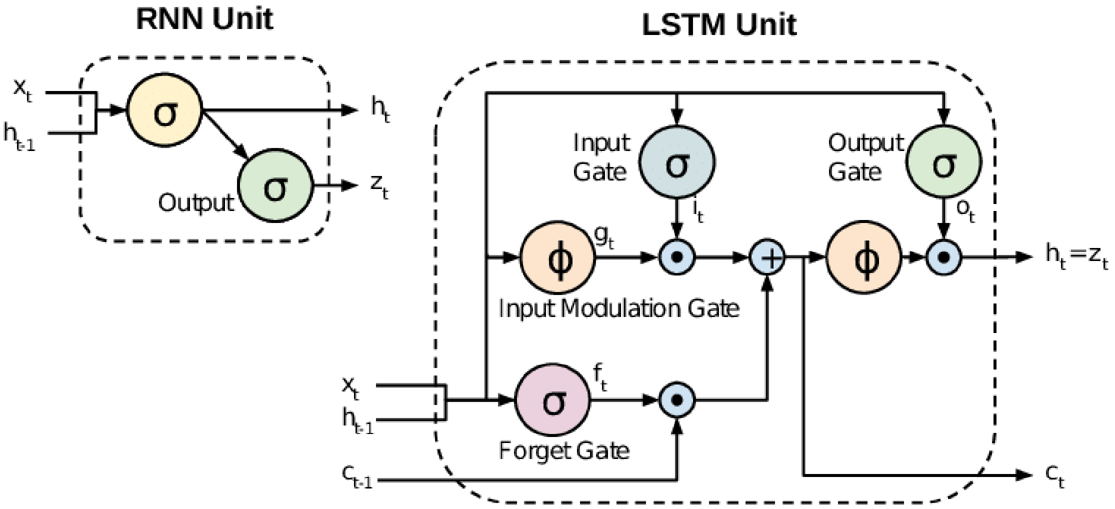


Figure 6. Simple RNN and LSTM cell (Donahue et al., 2015)

LSTM approach is used in this research and we have applied it on the dataset of coronavirus COVID-19. Python 3.7 is used in google colab environment, and TensorFlow with a version of 1.15 is used as back end so the Keras library runs on top of it. A neural network model called (Sequential) was used along with 8 LSTM cells.

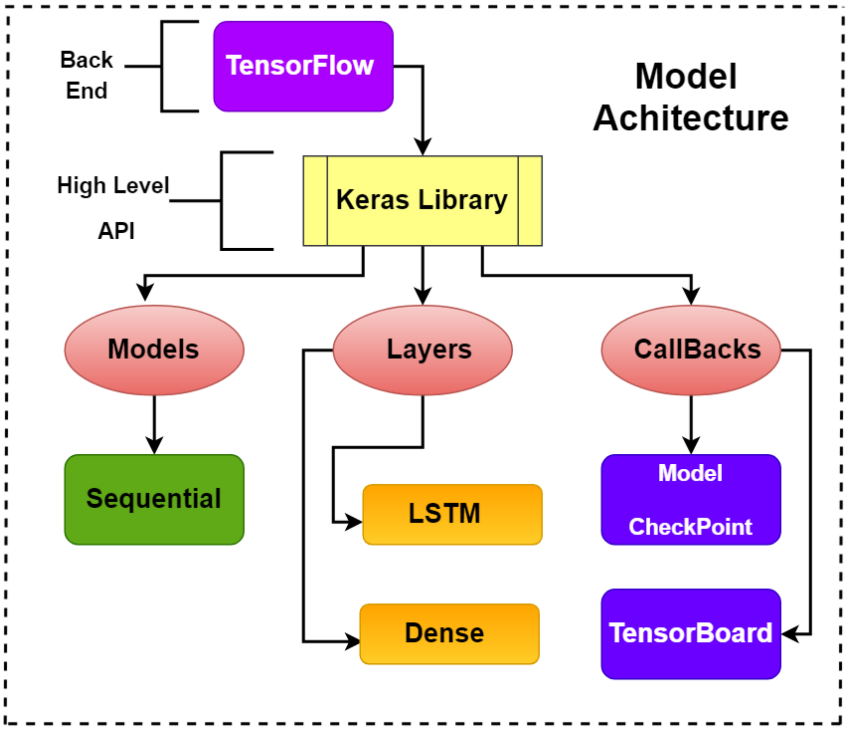


Figure 7. Proposed Model Architecture

In addition, a fully connected layer called (dense) and one input shape were used as well. An error toleration technique called (model checkpoint) was used along with tensor board. We used Adam optimizer with its default values as an approach for computing the learning rate for each parameter. Figure 7 shows the architecture of the proposed model and figure 8 shows the workflow of this research.

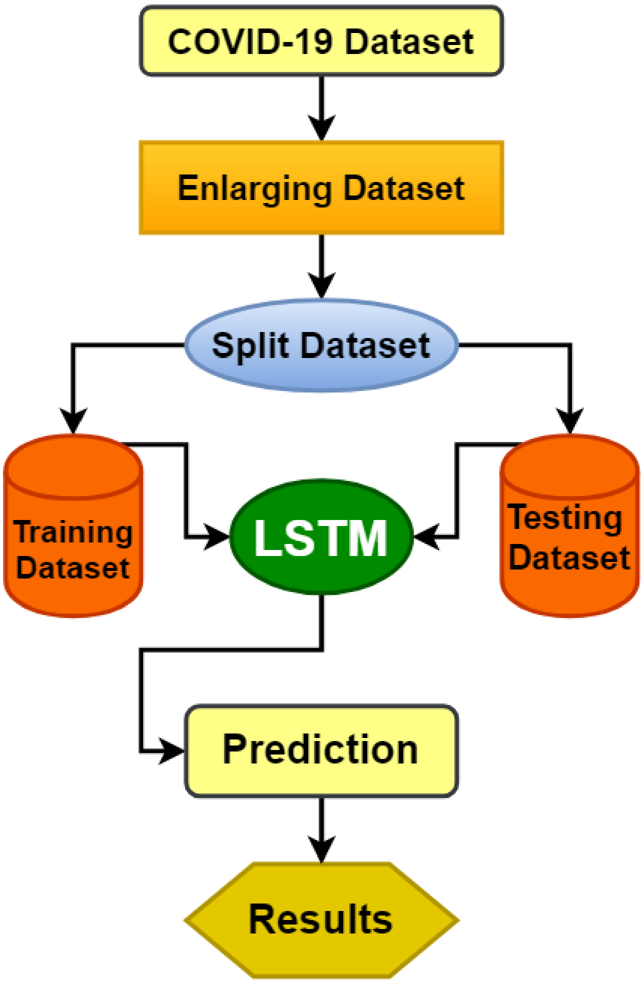


Figure 8. Overall flow of prediction model

A scikit learn library (Pedregosa et al., 2011) is used for pre-processing and metrics. For the pre-processing, MinMaxScaler was used to convert each value into a range from zero to one, also mean squared error is used as a metrics. A huge dataset is a necessity for constructing effective deep learning projects. But, since we have small dataset for the aim of model training, we have enlarged the dataset by splitting it into extra samples and for each value, we will multiply it 30 times to do this purpose (Abd Ghani et al., 2020), (Mohammed, Abdulkareem, Mostafa, et al., 2020) and (Abed Mohammed et al., 2018). The dataset is trained on 1595 samples with 100 epochs.

Results

The prediction model is created based on 61 days of data and we have carried it out on 22 March 2020. The dataset is enlarged to construct effective deep learning approach and thus the dataset is split into extra samples and for each row has multiplied 30 times. Figure 9 shows the dataset after enlarging.

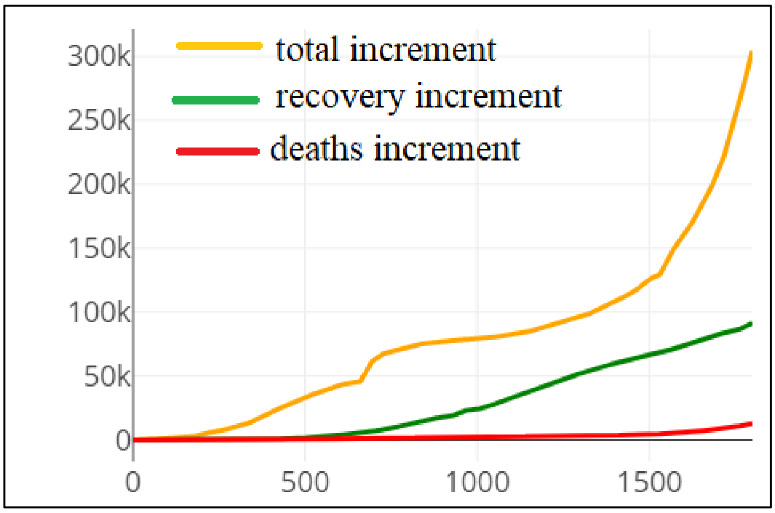


Figure 9. Dataset enlarging process

Root Mean Square Error (RMSE) for training score was 0.03, and for testing score was 0.00 which means the higher accuracy of prediction and it shows that there is a perfect match between the factual and predicted values. Figure 10 shows that the recovery rate has raised up and, it was higher than death rate during that period.

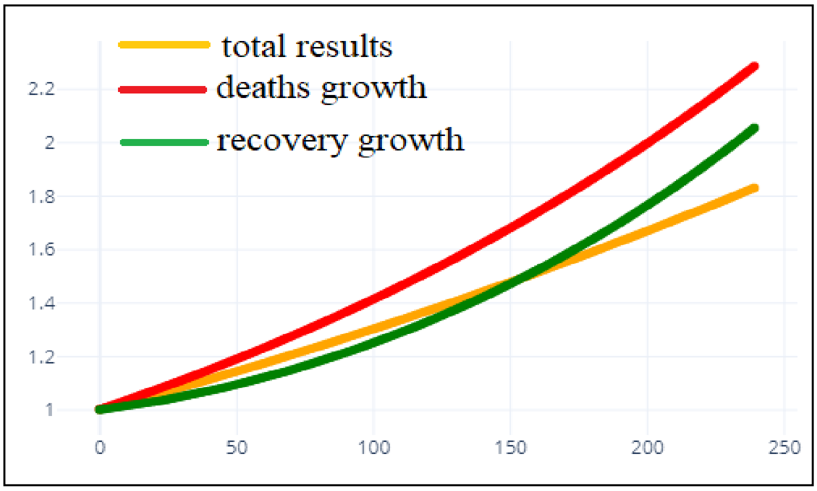


Figure 10. Growth of cases

Figure 11 illustrates the real expected numbers over the next 10 days. Unfortunately, the rate of infections keeps rising, but the good thing is that the rate of recovery is higher than the deaths ones. The infections rate was around **597,694** and the recovery rate was around **188,000** and the deaths rate was **29,640** over the next 10 days.

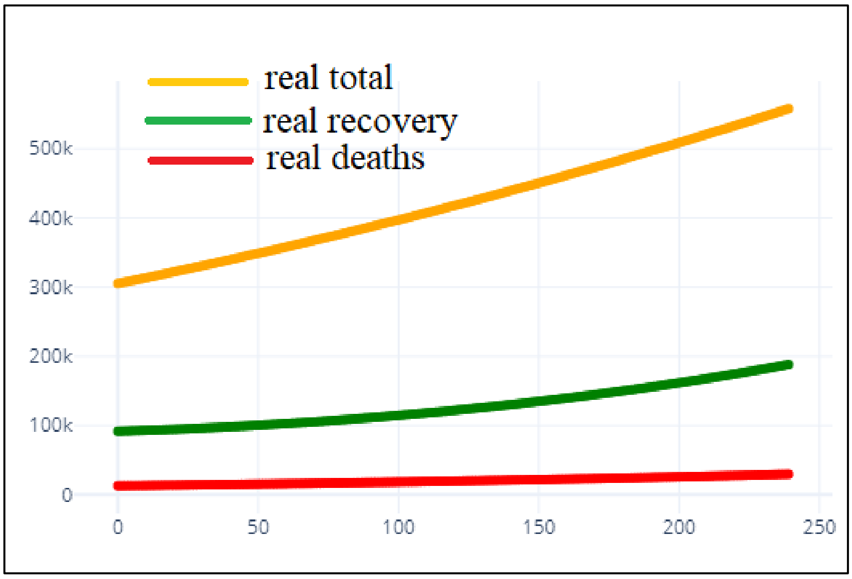


Figure 11. Real predicted cases

Discussion

The prediction has done on the available data that contains statistics which use deaths, infections and recoveries. Thus, developing a model based on deep learning provides nothing significant for forecasting since there are no powerful characteristics in the dataset, so that the model was not based on a robust and vital indicator. In addition, producing a dataset which includes features of geographic regions (temperature and humidity) that have experienced severe virus outbreaks, risk factors, spatio-temporal analysis, and social behaviour of people, a predictive model can be developed for areas where the virus is likely to spread, expecting to warn the authorities and people there to take the precautions. Lastly, other features other than heat and humidity can be derived which are related to the region's suitability for the virus to spread dramatically.

Conclusion

In this paper, we have used Long Short-Term Memory (LSTM) as it is a preferable method for time series prediction to predict COVID-19. This technique uses the dataset got from the records of infections, deaths, and recovery cases across the world. The outcomes are justifiable to alert the authorities and the people to take precautions. In the future, we will use a dataset which includes features of geographic regions (temperature and humidity) that have experienced severe virus outbreaks, risk factors, spatio-temporal analysis, and social behaviour of people in case it will be available to develop a powerful predictive model.

References

Abd Ghani, M. K., Mohammed, M. A., Arunkumar, N., Mostafa, S. A., Ibrahim, D. A., Abdullah, M. K., Jaber, M. M., Abdulhay, E., Ramirez-Gonzalez, G., & Burhanuddin, M. A. (2020). Decision-level fusion scheme for nasopharyngeal carcinoma identification using machine learning techniques. *Neural Computing and Applications*, *32*(3), 625–638. https://doi.org/10.1007/s00521-018-3882-6

Abed Mohammed, M., Khanapi Abd Ghani, M., Mostafa, S., Taha Al-Dhief, F., Ibrahim Obaid, O., Mostafa, S. A., & Taha AL-Dhief, F. (2018). Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer Performance evaluation of Wisconsin Breast Cancer. *Article in International Journal of Engineering and Technology*, *7*(4), 160–166. https://doi.org/10.14419/ijet.v7i4.36.23737

Al-Dhief, F. T., Latiff, N. M. azzah A., Malik, N. N. N. A., Salim, N. S., Baki, M. M., Albadr, M. A. A., & Mohammed, M. A. (2020). A Survey of Voice Pathology Surveillance Systems Based on Internet of Things and Machine Learning Algorithms. *IEEE Access*, *8*, 64514–64533. https://doi.org/10.1109/ACCESS.2020.2984925

Carneiro, T., Da Nobrega, R. V. M., Nepomuceno, T., Bian, G. Bin, De Albuquerque, V. H. C., & Filho, P. P. R. (2018). Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access*, *6*, 61677–61685. https://doi.org/10.1109/ACCESS.2018.2874767

Car, Z., Baressi Šegota, S., Anđelić, N., Lorencin, I. and Mrzljak, V., 2020. Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron. Computational and Mathematical Methods in Medicine, 2020.

CHEN, M., HAO, Y., HWANG, K., WANG, L., & WANG, L. (2019). Disease Prediction by Machine Learning from Healthcare Communities. *International Journal of Scientific Research in Science and Technology*, 29–35. https://doi.org/10.32628/ijsrst19633

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., & Saenko, K. (2015). Long-term recurrent convolutional networks for visual recognition and description. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June-2015*, 2625–2634. https://doi.org/10.1109/CVPR.2015.7298878

E.V., S., Konradi, A.O., Arutyunov, G. P., Arutyunov, A. G., Bautin, A. E., Boytsov, S. A., Villevalde, S.V., Grigoryeva, N. Y., Duplyakov, D. V., Zvartau, N. E., & Koziolova, N. A. (2020). Guidelines for the diagnosis and treatment of circulatory diseases in the context of the COVID-19 pandemic. *Russian Journal of Cardiology*, *25*(3), 3801.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5–6), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 1735–1780.

Mohammed, M. A., Abdulkareem, K. H., Al-Waisy, A. S., Mostafa, S. A., Al-Fahdawi, S., Dinar, A. M., Alhakami, W., Baz, A., Al-Mhiqani, M. N., Alhakami, H., Arbaiy, N., Maashi, M. S., Mutlag, A. A., Garcia-Zapirain, B., & De La Torre Diez, I. (2020). Benchmarking Methodology for Selection of Optimal COVID-19 Diagnostic Model Based on Entropy and TOPSIS Methods. *IEEE Access*, *8*, 99115–99131. https://doi.org/10.1109/ACCESS.2020.2995597

Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Ghani, M. K. A., Maashi, M. S., Garcia-Zapirain, B., Oleagordia, I., Alhakami, H., & Al-Dhief, F. T. (2020). Voice pathology detection and classification using convolutional neural network model. *Applied Sciences (Switzerland)*, *10*(11), 1–13. https://doi.org/10.3390/app10113723

Muhammad, L.J., Islam, M.M., Sharif, U.S. and Ayon, S.I., 2020. Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients Recovery.

M. Rubaiyat Hossain Mondal, Subrato Bharati, Prajoy Podder, Priya Podder, Data analytics for novel coronavirus disease, Informatics in Medicine Unlocked,Vol 20,100374, 2020.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Mathieu, B., Peter, P., Ron, W., & Vincent, D. (2011). Scikit-learn. *GetMobile: Mobile Computing and Communications*, *19*(1), 29–33. https://doi.org/10.1145/2786984.2786995

Rassem, A., El-Beltagy, M., & Saleh, M. (2017). *Cross-Country Skiing Gears Classification using Deep Learning*. 1–14. http://arxiv.org/abs/1706.08924

Raw.githubusercontent.com. 2020. [online] Available at: https://raw.githubusercontent.com/datasets/covid-19/ master/time-series-19-covid-combined.csv [Accessed 22 March 2020].

Shanmugamani, R. (2018). *Deep Learning for Computer Vision: Expert techniques to train advanced neural networks using TensorFlow and Keras*

Bibliographic information of this paper for citing:

Ibrahim Obaid, Omar; Abed Mohammed, Mazin & Mostafa, Salama A. (2020). Long Short-Term Memory Approach for Coronavirus Disease Prediction. *Journal of Information Technology Management*, Special Issue, 11-21.