Original article

# Machine learning approach for confirmation of COVID-19 cases: positive, negative, death and release

**Shawni Dutta**
Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India, India

**Samir Kumar Bandyopadhyay** 1954samir@gmail.com
Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India, India

Machine learning approach for confirmation of COVID-19 cases: positive, negative, death and release

Iberoamerican Journal of Medicine, vol. 2, núm. 3, pp. 172-177, 2020

Hospital San Pedro

**Recepción:** 24 Abril 2020
**Corregido:** 07 Mayo 2020
**Aprobación:** 09 Mayo 2020
**DOI:** https://doi.org/10.5281/zenodo.3822623

## Abstract:

Introduction: Corona Virus Infectious Disease (COVID-19) is the infectious disease. The COVID-19 disease came to earth in early 2019. It is expanding exponentially throughout the world and affected an enormous number of human beings starting from the last month. The World Health Organization (WHO) on March 11, 2020 declared COVID-19 was characterized as "Pandemic". This paper proposed approach for confirmation of COVID-19 cases after the diagnosis of doctors. The objective of this study uses machine learning method to evaluate how much predicted results are close to original data related to Confirmed-Negative-Released-Death cases of COVID-19.

Materials and methods: For this purpose, a verification method is proposed in this paper that uses the concept of Deep-learning Neural Network. In this framework, Long shrt-term memory (LSTM) and Gated Recurrent Unit (GRU) are also assimilated finally for training the dataset. The prediction results are tally with the results predicted by clinical doctors.

Results: The results are obtained from the proposed method with accuracy 87 % for the "confirmed Cases", 67.8 % for "Negative Cases", 62% for "Deceased Case" and 40.5 % for "Released Case". Another important parameter i.e. RMSE shows 30.15% for Confirmed Case, 49.4 % for Negative Cases, 4.16 % for Deceased Case and 13.72 % for Released Case.

Conclusions: 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 within 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.

**Keywords:**
Machine learning, LSTM, GRU, RNN, COVID-19.

## INTRODUCTION

It is now known that the coronavirus disease 2019 occurred in the city of Wuhan, China, at the end 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]. Epidemics of two betacoronaviruses, named as, severe acute respiratory syndrome coronavirus (SARS-COV) and Middle East respiratory syndrome coronavirus (MERS-COV) affected more than 10,000 people in cumulative order in the past two decades [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. 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 [3].

Human to human contacts are considered to be responsible for community spread of this disease in exponential growth. Hence social detachment of people should be followed necessarily to combat COVID-19 from the front line as well as in backend also. Control measures are to be imposed from the government side to implement social detachment of people. Preventive measures may include actions such as locking down the countries, closing and/or minimizing travel connections amongst the countries and within the cities as well, enforcing quarantine and hospitalization of infected individuals, suspending schools, offices, shopping malls, restaurants etc. Curfew imposed in the affected countries is facing potentially huge economical loss. As this disease spreads rapidly, timing is an important factor for controlling the disease as early as possible. Hence, from the initial stage an enormous level of monitoring is required for the authorities in order to handle the epidemic situation. All over the world Scientists are working 24 hours for finding the vaccine of the disease.

For assisting health planning for COVID-19, Machine Learning framework is proposed in this paper. This will confirm the confirmed cases, negative cases, recovered cases, and death cases considering the cases present in the dataset. The proposed prediction model ensures that it follows the original result regarding this epidemic situation so that enormous economic loss, community spread, amount of social detachment of people may be detected and also accurate decision can be taken accordingly. This method will ensure government authorities to yield preventive measures based on our next work for forecasting the occurrence of this disease in future.

In this paper, data mining concepts are exploited for obtaining prediction of confirmed, negative cases, recovered cases, and death cases where Recurrent Neural Network (RNN) is also employed. The real cases and prediction cases are compared based on some predefined metric. A combined model consisting of Long short-term memory (LSTM) and Gated Recurrent Unit (GRU) is applied to the dataset finally for training and testing purpose. A comparative study is drawn amongst the performance of proposed three models-LSTM-RNN, GRU-RNN, and LSTM-GRU-RNN.

### *RELATED WORKS*

A computation and analysis based on Suspected-Infected-Recovered-Dead (SIRD) model is provided by Anastassopoulou et. al. [3]. Based on the dataset available from January 11 to February 10 2020, it estimates of the main epidemiological parameters, i.e. the basic reproduction number (R0) and the infection, recovery and mortality rates, along with their 90% confidence intervals are provided. Computations on SIRD model, this R0 parameter value turn out to be 2.5. Experimental results forecast declining mortality rate that in turn help government authorities to impose safeguards.

Hu et. al. [4] proposed an AI based approach which is 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. This model performs real-time forecasting of the confirmed cases of Covid-19 across China. This model is applied on the dataset collected from January, 11 to February 27, 2020 given by World Health Organization (WHO). Use of latent variables in the auto-encoder and clustering algorithms helps in investigating the transmission procedure by grouping the provinces/cities.

Fong et. al. [5] suggested a methodology called Group of Optimized and Multisource Selection (GROOMS) which ensembles a collection of five groups of forecasting methods. Classical time-series forecasting methods as well as machine learning methods are implemented where small available dataset are passed from top to down through optimization processes. This will prepare the best winning models for panel section with lowest error. A polynomial neural network and corrective feedback (PNN+cf) is assimilated and implemented in this paper. Experimental results indicate that the combined approach PNN+cf outperforms well over other approaches in terms of Root Mean Squared Error (RMSE) performance measure parameter.

### *PERFRORMANCE MEASURE METRICS*

To identify best candidate model from its peers, it is necessary to put concentration on comparison of measures of the algorithm's performance. In this paper, following parameters are used for measuring algorithm's performance:

- Accuracy identifies the overall effectiveness of the algorithm. It is formulated as follows [6]: Accuracy= (tp+tn)/(tp+tn+fn+fp).

  Where tp denotes true positive, fp – false positive, fn – false negative, and tn – true negative counts

- Root-mean-square-error (RMSE) is a standard performance measure used for time series forecasting purpose. It is formulated as follows [7]: RMSE= √( 2/ N )

  Where Xi is the real value and Xi' is the predicted value.

## MATERIALS AND METHODS

In this paper, a prediction of confirmed cases, negative cases, released, and deceased cases of Covid-19 corona virus are obtained using RNN method. 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 [7].

Another RNN called Bidirectional RNN (BRNN) accesses future and past context in both directions. There are alternatives from RNN depending on the gating units, such as LSTM-RNN and GRU-RNN.

Traditional RNN lacks of considering context based prediction, which can be overcome by introducing LSTM. LSTM has a good potential to regulate gradient flow and enable better preservation of long-range dependencies [8].

GRU is quite similar to LSTM, where the gating units of GRU control the flow of information inside the unit, without considering separate memory cells [9]. Like LSTM, GRU lacks of having memory cells in it and it has a lesser number of gates are required and the gates are activated using current input as well as previous output. As compared to LSTM, GRU has better convergence rate due to the reduction of parameters and in some cases GRU outperforms well over LSTM models [10] .

For predicting confirmed, negative, released, deceased cases in Covid-19, dataset from kaggle [11] contains cases from 20th January 2020 to 12th March 2020, are used for training and testing three models.

At first, Data are pre-processed by eliminating missing values, irrelevant values. Then data transforming operations are performed so that it can be given as input to the Deep Learning Models. In this paper, three models are implemented and applied on the dataset for verifying the given prediction results with respect to available data set. The prediction results are measured with respect to performance measure metrics such as-accuracy and RMSE. The accuracy of these three models can be improved by choosing proper parameter values. The default parameters may not provide the maximum performance. Hyper-parameter setting is necessary to improve the accuracy level. However, the RMSE value should be optimised as to signify a better model. It is to be noted here that the dataset contains cases for confirmed, negative, released, dead patients. It is tested in clinical laboratory in presence of clinical doctors. This methodology is performed for each of these individual cases separately. The Figure 1 denotes the workflow of proposed methodology.

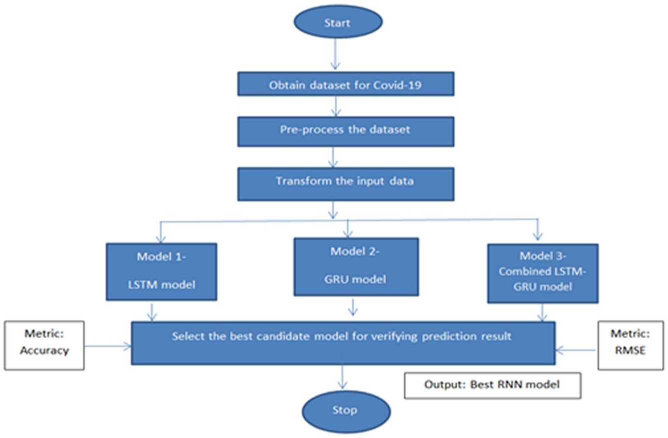

Figure 1
**Diagrammatic Approach for the workflow of proposed methodology.**

The method consists of three models, which are used for training as well as testing purpose. A detailed explanation with respect to each of these models is provided along with their implementation in the process. The explanation specifies performance of all these models with respect to categories such as, confirmed, negative, deceased, released cases are also presented one after another.

### *MODEL 1*

In this model LSTM layers use sequence of 50 nodes. A total 4 layered structure followed by a Dense Layer is used as LSTM model for verifying prediction result. The best hyper-parameters used are a dropout rate of 0.2 and a batch size of 32. The model is shown in Figure 2. The Accuracy and RSTM of all models with respect to each case are shown in Table 1.

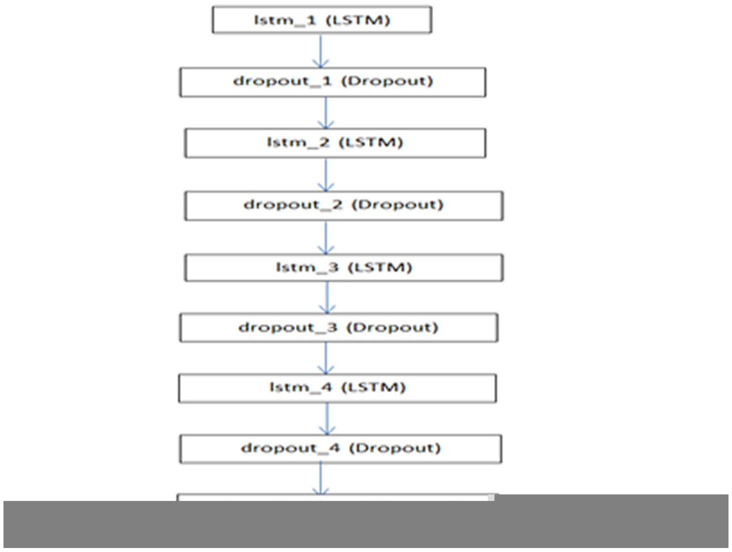

Figure 2
**Neural network structure for model 1.**

**Table 1**
**Prediction Efficiency of Models 1, 2 and 3**

**Table 1. Prediction Efficiency of Models 1, 2 and 3**

| **Models used** | **Metrics** | **Confirmed case** | **Negative case** | **Deceased case** | **Released case** |
|---|---|---|---|---|---|
| **LSTM Model** | Accuracy | 76.6% | 37.7% | 59.2% | 32.06% |
| | RMSE | 53.38 | 71.2 | 4.97 | 80.20 |
| **GRU Model** | Accuracy | 76.9% | 42.6% | 60.38% | 32.08% |
| | RMSE | 30.95 | 70.7 | 5.218 | 80.12 |
| **Combined LSTM-GRU Model** | Accuracy | 87% | 67.8% | 62% | 40.5% |
| | RMSE | 30.15 | 49.4 | 4.16 | 13.72 |

LSTM: Long short-term memory; GRU: Gated Recurrent Unit; RMSE: Root Mean Squared Error.

The Accuracy and RMSE depicted in the Table 1 may be improved if the model 2 is imposed.

### *MODEL 2*

This model has GRU layers and it uses sequence of 50 nodes. A total 4 layered structure followed by a Dense Layer is used as GRU model for prediction result verifying purpose. The best hyper-parameters used were a dropout rate of 0.2 and a batch size of 32. The Accuracy of the model can be found in Table 1. A structure of this model is shown in Figure 3.

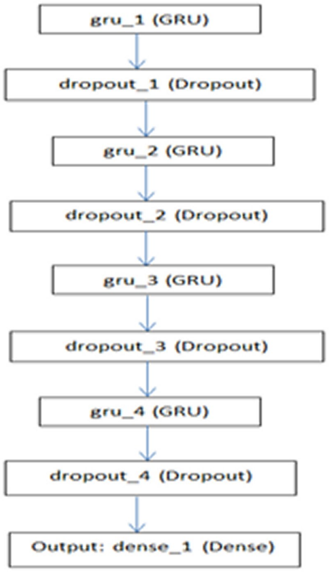

Figure 3

**Neural network structure for model 2.**

With Respect to the Model 1, Model 2 gives better result in terms of Accuracy as well as RMSE. Next, another model is proposed to improve the result obtained using model 3.

*MODEL 3*

A Recurrent Neural Network is employed where LSTM and GRU are assimilated together in order to predict the above mentioned cases. It is shown in Figure 4. For improving the performance of prediction results LSTM and GRU are assimilated and the transformed input is fitted into it. Table 1 signifies that model 3 imposes better impact in verification of the prediction result with respect to its original dataset.

Neural network structure for model 3.

Figure 4

**Neural network structure for model 3.**

Experimental Results shows that the combined approach LSTM-GRU-RNN provides better result over LSTM-RNN, GRU-RNN in terms of Accuracy, RMSE metrics. Figure 5 (A) and Figure 5 (B) depict the overall accuracy as well as RMSE for all these above mentioned model along with the given cases. Higher the accuracy and lower the RMSE indicates that the model is superior.

Accuracy Comparisions (A) and RMSE Comparisions of models (B).

Figure 5

**Accuracy Comparisions (A) and RMSE Comparisions of models (B).**

## RESULTS

From the above Figure 5 it is quite clear that the Model 3 provides better result over Model 1 and Model 2. So the model combines LSTM and GRU RNN shows best performance. Figure 6 depict the plotting of real result and predicted result for all the cases obtained for LSTM-GRU RNN.

Figure 6

**Real V/S Predicted Cases using LSTM-GRU RNN model for (a) Confirmed Cases (b) Negative Cases (c) Deceased Cases (d) Released Cases**

## DISCUSSIONS AND CONCLUSIONS

The combined approach of Deep-Learning models outperforms well while predicting given cases of Covid-19 disease. The approach presented here will assist in obtaining an automated predictive tool. This tool may achieve higher accuracy if more amounts of data are incorporated to this model. Structural changes to this proposed method may accelerate the prediction system.

The combined LSTM-GRU based RNN model 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 recheck occurred cases of COVID-19 automatically. The data driven RNN based model is capable of providing automated tool for confirming, estimating the current position of this pandemic, assessing the severity, and assisting government and health workers to act for good decision making policy. It could be a promising supplementary rechecking method for frontline clinical doctors. It is now essential for improving the accuracy of detection process.

## REFERENCES

1. *World Health Organization. WHO Statement Regarding Cluster of Pneumonia Cases in Wuhan, China. Available from: https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china.*

2. *Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395(10223):497-506. doi: 10.1016/S0140-6736(20)30183-5.*

3. *Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-Based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS One. 2020;15(3):e0230405. doi: 10.1371/journal.pone.0230405.*

4. *Hu Z, Ge Q, Li S, Jin L, Xiong M. Artificial Intelligence Forecasting of Covid-19 in China. 2020;1–20. Available from: https://arxiv.org/ftp/arxiv/papers/2002/2002.07112.pdf.*

5. *Fong SJ, Li G, Dey N, Gonzalez-Crespo R, Herrera-Viedma E. Finding an Accurate Early Forecasting Model from Small Dataset: A Case of 2019-nCoV Novel Coronavirus Outbreak. Int J Interact Multimed Artif Intell. 2020;(6(1):1322-40.*

6. *Sokolova M, Japkowicz N, Szpakowicz S. Beyond Accuracy , F-Score and ROC: A family of discriminant measures for performance evaluation. In: Sattar A, Kang BH, editors. AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings. Heidelberg: Springer;2006:1015-21.*

7. *Tjandra A, Sakti S, Manurung R, Adriani M, Nakamura S. Gated Recurrent Neural Tensor Network. doi: 10.1109/IJCNN.2016.7727233.*

8. *Bouktif S, Fiaz A, Ouni A, Serhani MA. Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. Energies. 2018;11(7: 1636. doi: https://doi.org/10.3390/en11071636.*

9. *Chung J, Gulcehre C, Cho K, Bengio Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. aRxiv. 2014. Available from: https://arxiv.org/abs/1412.3555.*

10. *Tang Y, Huang Y, Wu Z, Meng H, Xu M, Cai L. Question detection from acoustic features using recurrent neural network with gated recurrent unit. Available from: 10.1109/ICASSP.2016.7472854.*

*11. Datarist. Coronavirus-Dataset Version 1[Data file]. KCDC (Korea Centers for Disease Control and Prevention). Available from: https://www.kaggle.com/kimjihoo/coronavirusdataset-old.*

## Notas de autor

1954samir@gmail.com