International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217532

COVID-19 Future Predictions Using Machine Learning Algorithms

R. Saradha Devi[1], Dr. J. G. R. Sathiaseelan[2]

[1]Department of Computer Science, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India

[2]Head, Department of Computer Science, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India

ABSTRACT

Article Info
Volume 7, Issue 5
Page Number: 102-110

Publication Issue :
September-October-2021

Article History
Accepted : 07 Oct 2021
Published : 17 Oct 2021

Corona Virus Infectious Disease (COVID-19) is an 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 year. COVID-19 was declared “Pandemic” by the World Health Organization (WHO) on March 11, 2020. This research proposed a method for confirming COVID-19 instances after doctors' diagnoses. The goal of this study is to see how similar the projected findings are to the original data in COVID-19 Confirmed-Negative-Released-Death situations using machine learning. This paper suggests a verification approach created on the Deep-learning Neural Network concept for this purpose. Long short-term memory (LSTM) and Gated Recurrent Unit (GRU) are also used in this framework to train the dataset. The outcomes of the forecast match those predicted by clinical doctors.

Keywords : COVID-19, Long short-term memory (LSTM), Gated Recurrent Unit (GRU), Recurrent Neural Network (RNN), Accuracy, RMSE

I. INTRODUCTION

The coronavirus disease 2019 was confirmed to have occurred in the Chinese city of Wuhan at the end of the year. On January 30, 2020, the World Health Organization (WHO) detected and labelled a novel coronavirus as "2019-nCoV," which was later proclaimed a Public Health Emergency of International Concern, and on March 11, 2020, Covid-19 was classified as a Pandemic. [1]. In the last two decades, epidemics of two beta coronaviruses, severe acute respiratory syndrome coronavirus (SARS-COV) and Middle East respiratory syndrome coronavirus (MERS-COV), have impacted over 10,000 people. [2]. This novel coronavirus shares some similarities with SARS-COV and MERS-COV, according to the Centers for Disease Control and Prevention (CDC). These infections are transmitted from one person to another via respiratory droplets. After a period of 2 to 14 days, symptoms such as fever, cough, and shortness of breath are noted as the disease's results. [3].

Human-to-human interaction is thought to be the cause of the disease's exponential spread in the population. As a result, people's social detachment must be followed in

order to combat COVID-19 on the front lines as well as in the backend. People's social separation will be implemented by government-imposed control mechanisms. Locking down countries, restricting and/or minimizing travel connections between countries and within cities, enforcing quarantine and hospitalization of sick individuals, suspending schools, offices, shopping centers, restaurants, and so on are all examples of preventive measures. The imposition of a curfew in the affected countries could result in significant economic losses. Because this disease spreads quickly, timing is critical for controlling it as soon as feasible. As a result, authorities will require a high level of monitoring from the start in order to manage the pandemic scenario. Scientists from all over the world are working 24 hours a day to find a vaccine for the disease.

This study proposes Machine Learning architecture to aid health planning for COVID-19. This will confirm the confirmed cases, negative cases, recovered cases, and death cases based on the instances in the dataset. The suggested prediction model assures that it follows the original result in this epidemic situation, allowing for massive economic loss, community spread, and a significant amount of social separation to be detected, as well as appropriate decision-making. This strategy will ensure that government authorities produce preventative measures based on our future research for forecasting the prevalence of this disease.

In this paper, data mining methods are used to forecast confirmed, negative, recovered, and death cases, with the use of a Recurrent Neural Network (RNN). On the basis of some preset metric, the real cases and prediction cases are compared. Finally, for training and testing, a hybrid model comprising of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) is applied to the dataset. A comparison of the performance of the suggested three models-LSTM-RNN, GRU-RNN, and LSTM-GRU-RNN-is made.

II. RELATED WORK

COVID-19 research work is going around the world. Some of them may be research work that can assist in finding different ways in recovering the patients or related to vaccines and drugs that can help patients to recover. Apart from India, various papers are published with different available models for other nations especially for China, Italy, the UK, and the USA as the number of infected patients has been high (Gambhir, Jain, Gupta, & Tomer, 2020) [4].

The researchers in (Amirhoshang Hoseinpour Dehkordi, 2020) [5] analyzed the transmission pattern of COVID-19 from China to different countries, based in light of every day reported cases, the observation strategy of China, South Korea, Japan, Spain and, Italy from the very first day of the outbreak, along with the different types of policies of the above-mentioned countries government in controlling the COVID-19 outbreak by the linear relation in their data.

In (Dutta, Usha, Anurag Sachan, 2020) [6] authors the study carried out in the field of corona virology describes coronavirus replication notes alongside the growth of coronavirus, and cells preparation, in addition to different analyzing techniques for the virus function, widely used in the genetic techniques of coronavirus. Titration techniques, and virus-cell fusion, identification of cellular receptors in addition with visualization of virus replication complexes of coronavirus, the life cycle virus in great depth.

As in (Kucharski, et al., 2020) [7] the individuals who have recovered are deeply analyzed, which can show some insight on how to manage the active cases. Data analysts and researchers around the globe are working hard in making sense of the data available and predicting for the near future. The discovery of trend patterns, the selection of features, the forecasting methods are developed in and out to conclude.

In the paper (Ardabili, et al., 2020) [8] To forecast the COVID-19 outbreak as a substitute to SIR and SEIR models, this paper provides a comparative study of machine learning and different soft computing models. After searching and testing a wide variety of machine learning models. two models were used that provided significant results.

World trade has already decreased in 2019. as per the World Trade Organization (Verma & Gustafsson, 2020) [9]. and now this pandemic has led to the global economic crisis. Early forecasts have projected those significant economies will lose about 2.4 to 3.0 percent of their gross domestic product (GDP) during 2020 because of the COVID-19 pandemic. The article (Verma & Gustafsson, 2020) [9] defines the existing research areas and provides a way forward, they provided a bibliometric analysis of COVID-19 and its effect on academics, business, and the executive space.

The model in (Mittal, 2020) [10] helps to predict national opinion with different behavior of the public. The political as well as economic impact of the virus. Different methods are used for analysis as EDA, in which cases, death and recovered are recorded and the second method of SEIR model, they also used statistical approach. In (Wang, Hu, Jiang, Lu, & Zhang, 2020) used particle swarm optimization, the current SIR model, they also offered the method for understanding the sentiment analysis and proposes the fake news detection method and researchers.

In (Latif, 2020) [11] also used different techniques related to the analysis of different papers. The Gaussian distribution theory was used in (Li, et al., 2020) [12] while replicating the propagation mechanism of COVID-19 to simulate the curves of mainly Hubei and Non-Hubei areas of China.

In (Furqan Rustom, 2020) [13] the researchers have taken four famous algorithms of Machine Learning and compared them based on evaluation results from R? Score, R? Adjusted, MSE, MAE & RMSE. From all those, they chose Exponential Smoothing as the best method, Linear Regression, and LASSO gave average or second-best results whereas, SVM (Support Vector Machine) gave the worst result out of the four models. It had taken a dataset from Johns Hopkins University and performed the tests.

After taking a look at the work of different researches related to the analysis, by using different techniques and data sets from different sources we used COVID-19 datasets from Johns Hopkins University and used linear regression model for our analysis because our datasets were continuous, and with the help of inbuilt python libraries like pandas, numpy, matplotlib helped in our analysis.

III. METHODS

The RNN approach is used in this paper to forecast confirmed cases, negative cases, released cases, and deceased cases of the Covid-19 corona virus. RNN is a type of neural network architecture that takes both sequential and parallel data processing into account. It is possible to imitate operations comparable to those of the human brain by including memory cells into neural networks.

Data is first pre-processed by removing missing values and irrelevant values. After that, data transformation operations are undertaken so that it may be fed into Deep Learning Models as input. Three models are implemented and applied on the dataset in this paper to test the supplied prediction findings in relation to the accessible data set. The prediction results are evaluated using performance indicators such as accuracy and root mean square error (RMSE). By selecting appropriate parameter values, the accuracy of these three models can be increased. The default settings may not deliver the best results. To boost the accuracy level, hyper-parameter tuning is required. The RMSE number, on the other hand, should be optimised to indicate a superior model. It should be emphasised that the dataset includes confirmed, negative, released, and deceased patients. In the presence of clinical doctors, it is evaluated in a clinical laboratory. This process is used to each of these circumstances on an individual basis. The workflow of the suggested methodology is depicted in Figure 1.

The approach is made up of three models that can be utilized for both training and assessment. Each of these models is given a thorough description, as well as how they are implemented in the process. The explanation details the performance of all of these models in terms of categories like confirmed, negative, deceased, and released cases, which are shown one by one.

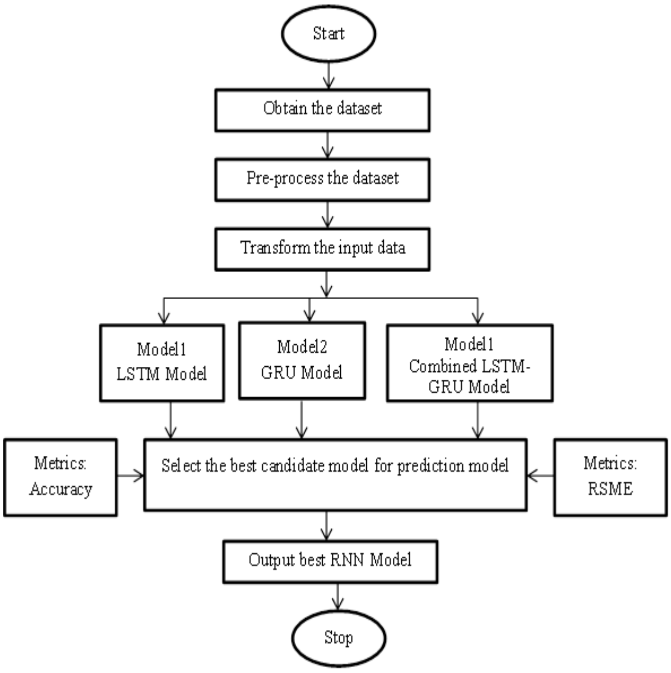


Fig 1: Diagrammatic Approach for the workflow of proposed methodology.

A. MODEL 1

The LSTM layers in this model use a 50-node sequence. For verifying prediction results, an LSTM model with a total of four layers and a Dense Layer is used. A dropout rate of 0.2 and a batch size of 32 are the optimum hyper-parameters to utilise. Table 1 depicts the accuracy and LSTM of all models in relation to each case. The accuracy and RMSE indicated in Table 1and 2 may be improved if model 2 is applied.

1) Long short-term memory (LSTM)

Figure 2 shows the usual structure of LSTM NN cells. The four gates that make up a standard LSTM NN cell are the input gate, input modulation gate, forget gate, and output gate. The input gate receives a fresh input point from the outside and processes the data. The output of the LSTM NN cell in the previous cycle is fed into the memory cell input gate. The forget gate determines when the output results should be forgotten, and so determines the best time lag for the input sequence. The output gate collects all calculated results and generates an output for the LSTM NN cell. In most language models, a soft-max layer is used to decide the NN's final outputIn our traffic flow prediction model, a linear regression layer is applied to the output layer of the LSTM cell.

Let us denote the input time series as $X = (x_1, x_2 \dots x_n)$ hidden state of memory cells as $H = (h_1, h_2, \dots h_n)$ output time series as $Y = (y_1, y_2 \dots y)$. LSTM NNs do the computation as follows:

$$h_t = H(W_{hx}x_t + W_{hh}h_{t-1} + b_h)$$
$$p_t = W_{hy}y_{t-1} + b_y$$

Where weight matrices are denoted as W, and bias vectors denoted as b. The hidden state of memory cells is computed in the following formulas:

$$i_t = \sigma(W_{ix}x_t + W_{hh}h_{t-1} + W_{ic}c_{t-1} + b_i)$$

$$f_t = \sigma\left(W_{fx}x_t + W_{hh}h_{t-1} + W_{fc}c_{t-1} + b_f\right)$$
$$c_t = f_t^* c_{t-1} + i_t^* g(W_{cx}x_t + W_{hh}h_{t-1} + W_{cc}c_{t-1} + b_c$$
$$o_t = \sigma(W_{ox}x_t + W_{hh}h_{t-1} + W_{oc}c_{t-1} + b_o)$$
$$h_t = o_t^* h(c_t)$$

where σ stands for the standard sigmoid function defined in Eq. (8), * stands for the scalar product of two vectors or matrices, g and h are the extends of stand sigmoid function with the range changing to [-2, 2] and [-1, 1]

$$\sigma(x) = \frac{1}{1+e^x}$$

For objective function we use square loss function given by the following formula:

$$e = \sum_{t=1}^{n}(y_t - p_t)^2$$

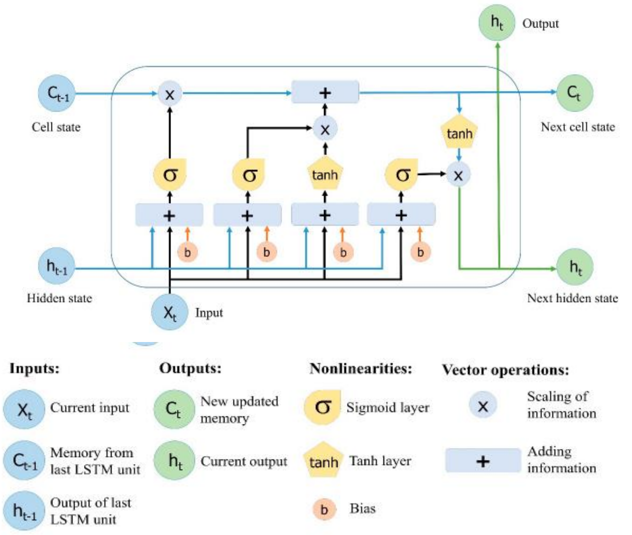


Fig. 2. Structure of LSTM NN cells

B. MODEL 2

GRU layers are used in this model, which has a 50-node sequence. The GRU model uses a total of four layers, followed by a Dense Layer, for the purpose of confirming prediction results. A dropout rate of 0.2 and a batch size of 32 were the optimal hyper-parameters. The model's accuracy can be found in Table 1. In comparison to Model 1, Model 2 performs better in terms of accuracy and RMSE. Then, to improve the outcome obtained with model 3, another model is provided.

1) Gated Recurrent Unit (GRU)

Cho et al. [14] proposed GRU in 2014, which is similar to LSTM but easier to compute and implement. The reset gate r and the update gate z are the two gates that make up a typical GRU cell. The hidden state output at time t is computed using the hidden state of time t-1 and the input time series value at time t, as in the LSTM cell (11).

$$h_t = f(h_{t-1}, x_t)$$

The function of reset gates is similar to that of the LSTM forget gates. We won't go into detail about the formula because GRU NNs are comparable to LSTM NNs in various ways. Cho et al. [14] are recommended to those who are interested in this topic. We apply the same regression portion and optimization strategy for GRU NNs as we do for LSTM NNs in this paper.

C. MODEL 3

To anticipate the above-mentioned scenarios, a Recurrent Neural Network is used, with LSTM and GRU assimilated together. Figure 4 depicts this. The altered input is fitted into LSTM and GRU in order to improve the performance of prediction outcomes. Model 3 has a greater impact on the verification of the prediction result when compared to the original dataset, as shown in Table 1 and 2.

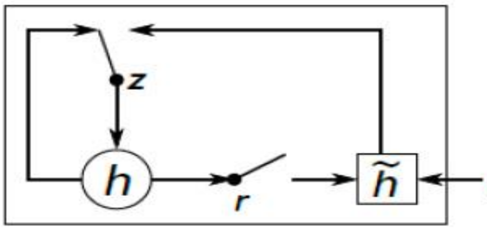


Fig. 3. Structure of GRU cells.

IV. MEASURE METRICS

To distinguish the top candidate model from its peers, comparisons of algorithm performance measures must be prioritized. The following parameters are used to evaluate the performance of the method in this paper:

a) Accuracy

The percentage of correct predictions for the test data is known as accuracy. It is simple to calculate by dividing the number of right forecasts by the total number of predictions.

$$Accuracy = \frac{correctpredictions}{allpredictions}$$

b) Root Mean Square Error (RMSE)

The square root of the mean of the squared discrepancies between actual and predicted outcomes is used to calculate the RMSE. The square root of the mean squared error returns the error metric to its original units for comparison, and squaring each error causes the numbers to be positive.

$$RMSE = sqrt(\, sum(\, (predicted\_i - actual\_i)\hat{}2\,) \,/\, total\, predictions)$$

V. EXPERIMENTAL RESULTS

In terms of accuracy and RMSE metrics, experimental results show that the combined approach LSTM-GRU-RNN outperforms LSTM-RNN and GRU-RNN. Figures 4 and 5 show the overall accuracy as well as the RMSE for all of the above models as well as the specified examples. The model is preferable if it has a higher accuracy and a lower RMSE.

Figure 4 and 5 shows that the Model 3 outperforms the Model 1 and Model 2 in terms of performance. As a result, the model that combines LSTM and GRU RNN performs the best. Figure 6 shows the charting of the real and predicted results for all of the LSTM-GRU RNN situations.

Table 1. Prediction Accuracy of Models 1, 2 and 3

Models used	Confirmed case	Negative case	Deceased case	Released case
LSTM Model		38.8%	59.2%	32.06%
GRU Model	77.9%	77.9%	77.9%	77.9%
Combined LSTM-GRU Model	88%	68.9%	63%	41.5%

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

Models used	Confirmed case	Negative case	Deceased case	Released case
LSTM Model	54.48	72.3	4.87	81.20
GRU Model	31.85	71.8	5.117	81.13
Combined LSTM-GRU Model	31.25	48.9	4.25	14.81

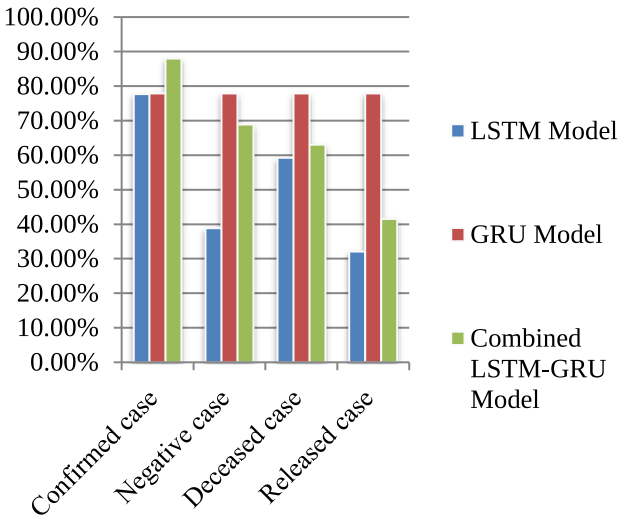


Fig 4. Prediction Accuracy of Models 1, 2 and 3

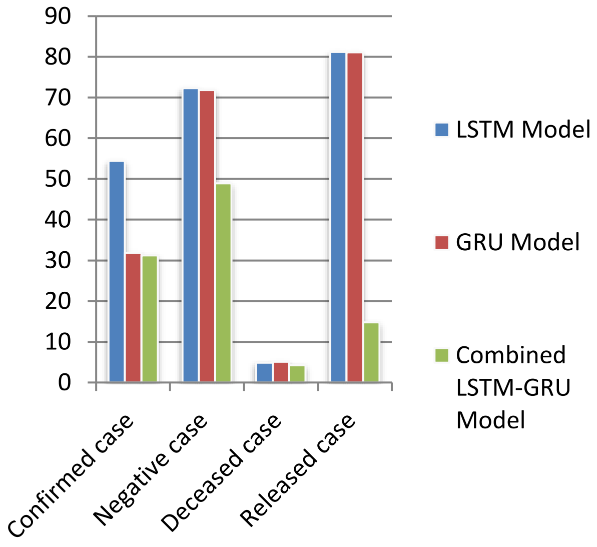
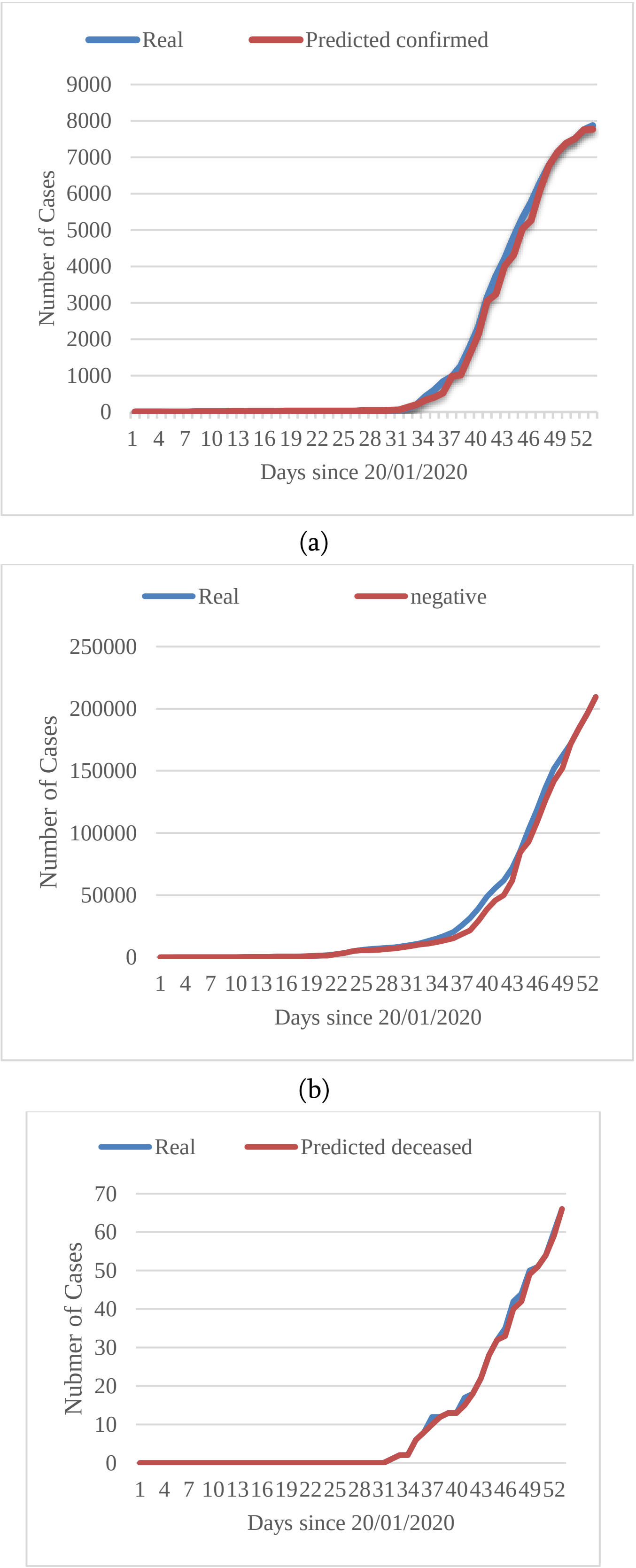


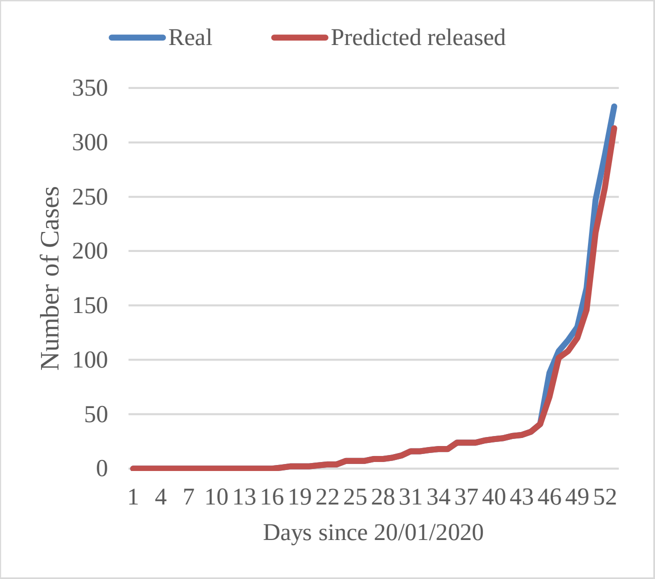
Fig 5. Prediction Efficiency of Models 1, 2 and 3



(a)

(b)

(c)



(d)

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.

VI. CONCLUSIONS

When it comes to predicting specific cases of Covid-19 disease, the combined method of Deep-Learning models outperforms. The method given here will aid in the development of an automated predictive tool. If new data is included into this model, the accuracy of this tool may improve. This proposed method's structure could be changed to speed up the prediction system. In terms of predicting confirmed, released, negative, and death cases on the data, the combined LSTM-GRU based RNN model performs significantly better. This research offered a novel method for automatically rechecking COVID-19 situations that had occurred previously. The data-driven RNN-based model can provide an automated tool for confirming, estimating the present position of the pandemic, measuring the severity, and aiding government and health personnel in making policy decisions. For frontline clinical practitioners, it could be a beneficial additional rechecking method. It is currently necessary for improving the detection process' accuracy.

VII. 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]. Gambhir, Ekta, Ritika Jain, Alankrit Gupta, and Uma Tomer. "Regression analysis of COVID-19 using machine learning algorithms." In 2020 International conference on smart electronics and communication (ICOSEC), pp. 65-71. IEEE, 2020.

[5]. Hoseinpour Dehkordi, Amirhoshang, Majid Alizadeh, Pegah Derakhshan, Peyman Babazadeh, and Arash Jahandideh. "Understanding epidemic data and statistics: A case study of COVID-19." Journal of medical virology 92, no. 7 (2020): 868-882.

[6]. Dutta, Usha, Anurag Sachan, Madhumita Premkumar, Tulika Gupta, Swapanjeet Sahoo, Sandeep Grover, Sugandhi Sharma et al. "Saving and Supporting Health Care Workers During the COVID-19 Pandemic in a Developing Country Using a Multidimensional Healthcare Personnel Centric Policy." (2020).

[7]. Kucharski, Adam J., Timothy W. Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M. Eggo et al. "Early dynamics of transmission and control of COVID-19: a mathematical modelling study."

The lancet infectious diseases 20, no. 5 (2020): 553-558.

[8]. Ardabili, Sina F., Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R. Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, and Peter M. Atkinson. "Covid-19 outbreak prediction with machine learning." Algorithms 13, no. 10 (2020): 249.

[9]. Verma, Surabhi, and Anders Gustafsson. "Investigating the emerging COVID-19 research trends in the field of business and management: A bibliometric analysis approach." Journal of Business Research 118 (2020): 253-261.

[10]. Mittal, Rajat, Rui Ni, and Jung-Hee Seo. "The flow physics of COVID-19." Journal of fluid Mechanics 894 (2020).

[11]. Kanniah, Kasturi Devi, Nurul Amalin Fatihah Kamarul Zaman, Dimitris G. Kaskaoutis, and Mohd Talib Latif. "COVID-19's impact on the atmospheric environment in the Southeast Asia region." Science of the Total Environment 736 (2020): 139658.

[12]. Huang, Chaolin, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang et al. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China." The lancet 395, no. 10223 (2020): 497-506.

[13]. James, Jubin. "Covid-19 Future Predictions using Machine Learning Algorithms." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12, no. 11 (2021): 6292-6302.

[14]. K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, D. Bahdanau, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

Cite this article as :

R. Saradha Devi, Dr. J. G. R. Sathiaseelan, "COVID-19 Future Predictions Using Machine Learning Algorithms", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 5, pp. 102-110, September-October 2021. Available at doi : https://doi.org/10.32628/CSEIT217532
Journal URL : https://ijsrcseit.com/CSEIT217532