CNN-LSTM Model for Verifying Predictions of Covid-19 Cases

Shawni Dutta[1], Samir Kumar Bandyopadhyay[2*] and Tai-Hoon Kim[3]

[1]*Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India.*
[2]*The Bhawanipur Education Society College, Kolkata, India.*
[3]*School of Economics and Management, Beijing Jiaotong University, No.3 Shangyuancun,*
Haidian District, Beijing 100044, P. R. China.

Authors' contributions

This work was carried out in collaboration among all authors. Author SD designed the proposed method, coding and statistical work. Author SKB initiates the work and wrote the first draft. Author THK edited all processes before finalizing the manuscript. All authors read and approved the final manuscript.

Original Research Article

ABSTRACT

COVID-19 disease came to earth in December 2019 in Wuhan. It is increasing exponentially throughout the world and affected an enormous number of human beings. The World Health Organization (WHO) on March 11, 2020 declared COVID-19 was characterized as "Pandemic". Clinical Doctors have been working on it 24 hours in the entire world. These doctors are testing whether the particular human has been affected with the disease using testing kit and other related process. Researchers have been working day-night for developing vaccine for the disease. Since the rate of affected people is so high, it is difficult for clinical doctors to check such a large number of coronavirus detected humans within reasonable time.

This paper attempts to use Machine Learning Approach to build up model which will help clinical doctors for verification of disease within short period of time and also the paper attempts to predict growth of the disease in near future in the world. Two models were used for achieving this purpose-One is based on Convolutional Neural Network model where as another one consists of

_____

*Corresponding author: E-mail: 1954samir@gmail.com;

Convolutional Neural Network and Recurrent Neural Network. These two models are evaluated and compared for verifying the predicted result with respect to the original one. Experimental results indicate that the combined CNN-LSTM approach outperforms well over the other model.

1. INTRODUCTION

The whole world is now facing seriously for disease COVID-19 without any indication. It is known that the disease carries infection from infected humans who is affected by the disease. Some measures throughout the world have been taken as precautions and some countries declare lockdown so that the disease cannot be spread rapidly. However, the disease affects human so rapidly that a large number of peoples are dead due to these humans affected by the disease. The world is now indicated "Stages" of coronavirus affected countries [1].

The scientists all over the world have been doing researches to find the vaccine of coronavirus for reducing the death as well as trying to reduce the cases with coronavirus affected humans. These scientists are partially successful but they are unable to recover all cases. It is also noted that affected cases are growing so rapidly in the world. At the same time death due to affected coronavirus cases are rising exponentially. It is also difficult to diagnose since affected patients are needed to be isolated from others. So the world is now facing a difficult situation and this in turn creates problems on economy. Researches are going on till the invention of proper vaccine.

In countries separate hospitals are created to tackle the disease. Clinical doctors are working day to night for giving attention to the suspected humans with symptoms of the disease. Since the affected rate of the disease is so high so sometimes it creates problems for handling patients. The reason behind it that the disease is completely new and it is not yet prove the exact symptoms to be seen in the patient. The death of humans is not restricted to developed countries but also in under developed countries also. The affected patients are so high that validation and prediction of the disease are now essential in the entire world [2,3].

This paper uses Machine Learning approach for building model that will help clinical doctors. It is true that concluding decision by doctors is important although the model may help them. In the proposed framework two models were implemented that exploit the concept of deep learning. Convolutional Neural Network and Recurrent Neural Network are employed while forming the models. One model is based on Convolutional Neural Network while the other model comprises of Convolution Neural Network and Recurrent Neural Network as well. For implementing Recurrent Neural Network, Long Short-Term Memory (LSTM) is employed. These models are used to validate the actual result with respect to predicted result. Performance measure metrics, such as, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) are used for evaluating these models.

2. RELATED WORK

In [3], a methodology called Group of Optimized and Multisource Selection (GROOMS) is suggested 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 (PNN) with large dataset is employed in this paper. Later a corrective feedback (cf) is imposed on the PNN for forecasting purpose. Experimental results indicate that the combined approach PNN+cf outperforms well over other approaches in terms of Root Mean Squared Error (RMSE) as performance measure.

In [4], 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. 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.

A computation and analysis based on Suspected-Infected-Recovered-Dead (SIRD) model is provided in [5]. 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 (R_0) and the infection, recovery and mortality rates, along with their 90% confidence intervals are provided. Computations on SIRD model, this R_0 parameter value turn out to be 2.5. Experimental results may forecast declining mortality rate which will help government authorities to impose safeguards subject to the best safety measure.

3. PROPOSED METHODOLOGY

In this paper, a deep-learning based approach is employed which confirms the prediction based on Confirmed, Negative, Released, Death cases. In this context, a Convolution Neural Network (CNN) and Recurrent Neural Network (RNN) are employed.

Convolutional Neural Networks (CNNs) is a class of deep neural network that comprises of neurons and they self-optimize themselves through learning [6]. A Recurrent Neural Network (RNN) is a type of neural network architecture that processes both sequential and parallel information. Similar operations like human brain can be simulated by incorporating memory cells to the neural network [7]. Long short-term memory (LSTM) is a kind of RNN that implements context based prediction which is not considered in traditional RNN. LSTM has a good potential to regulate gradient flow as well as better preservation of long-range dependencies.

In this framework, two models are employed. One is based on Convolutional Neural Network and the second model is a combined model that stacks CNN layers and LSTM layers as well [8-10]. For obtaining confirmation on the prediction of the confirmed, negative, released, deceased cases in Covid-19 dataset [11] containing data from 1st February 2020 to 11th May 2020, are used for training and testing above mentioned models. After that, preprocessing of that data is obtained by eliminating irrelevant data, missing values. Next, preprocessed data are transformed and provided to the specified model 1 and model 2. The purpose of these two models is to provide prediction on the different cases for Covid-19. Next, based on performance measure metric RMSE, MAE these models are compared and the best model is selected. For maximizing the

performance of these models, default parameters may not be sufficient enough. Hyper-parameter adjusting is necessary to increase the performance level. Hence, the performance measure metric RMSE, MAE value should be optimised as to signify a better model. Fig. 1 provides diagrammatic approach of our proposed framework.
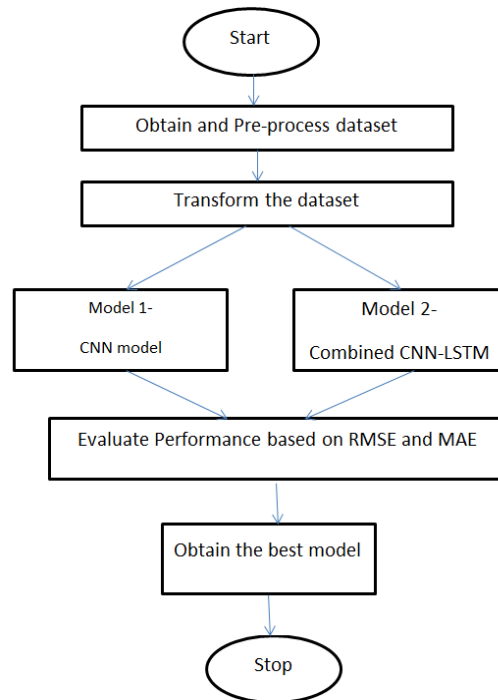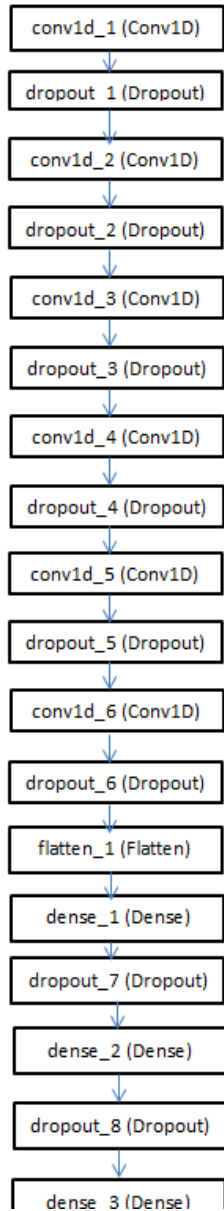

Fig. 1. Workflow of proposed framework

The workflow comprised of two models those are used for obtaining the predicted result and comparing the original result with the predicted result. 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.
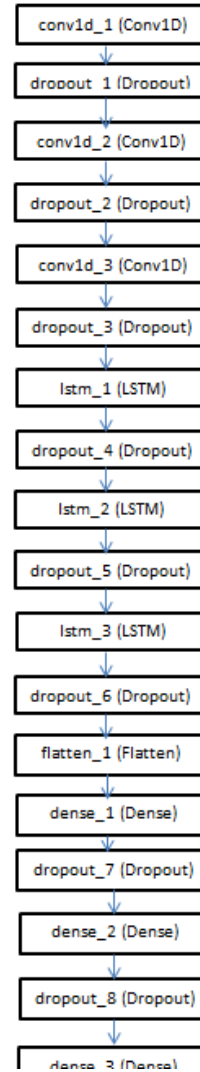
3.1 Model 1

For confirming the prediction results, 1-dimensional CNN layers are stacked and complied for obtaining the CNN-based model. The transformed input in fed into this model and final prediction is compared with respect to actual result. The following Fig. 2 summarizes the description of this model.

Table 1 shows that model 1 performs well. However, this performance may be improved if model 2 is imposed and applied.

conv1d_1 (Conv1D)

dropout_1 (Dropout)

conv1d_2 (Conv1D)

dropout_2 (Dropout)

conv1d_3 (Conv1D)

dropout_3 (Dropout)

conv1d_4 (Conv1D)

dropout_4 (Dropout)

conv1d_5 (Conv1D)

dropout_5 (Dropout)

conv1d_6 (Conv1D)

dropout_6 (Dropout)

flatten_1 (Flatten)

dense_1 (Dense)

dropout_7 (Dropout)

dense_2 (Dense)

dropout_8 (Dropout)

dense_3 (Dense)

Fig. 2. Neural network structure for CNN model

conv1d_1 (Conv1D)

dropout_1 (Dropout)

conv1d_2 (Conv1D)

dropout_2 (Dropout)

conv1d_3 (Conv1D)

dropout_3 (Dropout)

lstm_1 (LSTM)

dropout_4 (Dropout)

lstm_2 (LSTM)

dropout_5 (Dropout)

lstm_3 (LSTM)

dropout_6 (Dropout)

flatten_1 (Flatten)

dense_1 (Dense)

dropout_7 (Dropout)

dense_2 (Dense)

dropout_8 (Dropout)

dense_3 (Dense)

Fig. 3. Neural network structure for combined CNN-LSTM approach

3.2 Model 2

In this case, CNN layers and LSTM-RNN layers are stacked and later they are compiled in order to obtain a final combined model. The transformed data as fed as input to this model and fitting to this model is obtained. Later prediction from this model is obtained and that result is compared with the original result to decide how much the output result is related to original result. The following Fig. 3 depicts the summary of the model.

In our demonstration, six 1-dimensional CNN layers use sequence of 256 nodes. Each CNN layer is followed by recurrent dropout layer. The layered structure is followed by flatten layer and Dense Layers along with recurrent dropout layer. However, for the last layer only dense layer is applied without imposing dropout layer. The best hyper-parameters used are a dropout rate of 0.2 and a batch size of 32. The evaluation of the model with respect to RMSE is shown in Table 1.

While implementing this model, three 1-dimensional CNN layers are given sequence of 256 nodes. Each CNN layer is followed by recurrent dropout layer. These layers were followed by three LSTM layers with recurrent dropout layers. Finally, the layered structure is followed by flatten layer and Dense Layers along with recurrent dropout layer. However, for the last dense layer no dropout is applied. A dropout rate of 0.2 and a batch size of 32 are used the best hyper-parameters. The evaluation of the model with respect to RMSE and MAE is shown in Table 2.

From the Table 1, it is observed that combined approach CNN-LSTM outperforms well over model 1 in terms of MAE and RMSE. Following Figs. 4 and 5 summarize the overall performance of the specified models with respect to performance measure metrics MAE and RMSE. The following Figs. 6 to 9 depict the overall performance of these two above mentioned models along with their cases with respect to RMSE and MAE.

**Table 1. Performance evaluation of model 1**

Performance measure metric	Confirmed case	Negative case	Deceased case	Recovered case
MAE	24.197971032	93.431761	2.20603625	7.710796667
RMSE	301.99	12074.736	4.306	123.8794

Table 2. Performance evaluation of model 2

Performance measure metric	Confirmed case	Negative case	deceased case	Recovered case
MAE	14.6809	74.96420	1.35783	6.80861
RMSE	218.81535	7259.61834	1.45994	66.7537
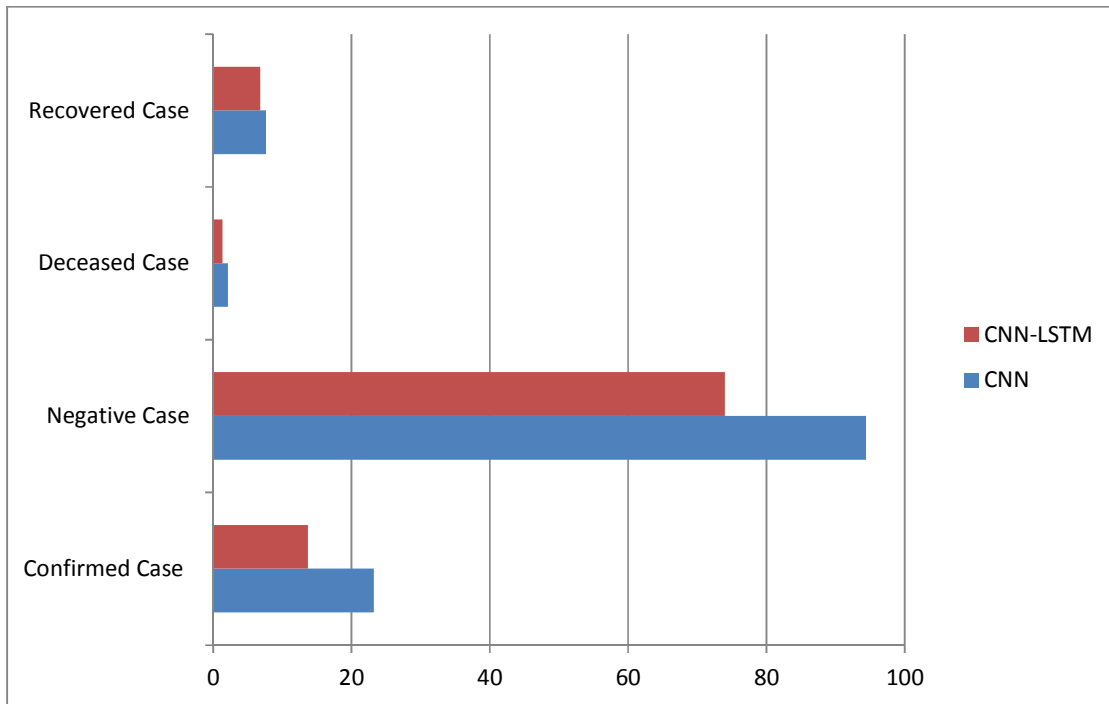
Fig. 4. Performance comparison between CNN and CNN-LSTM with respect to MAE
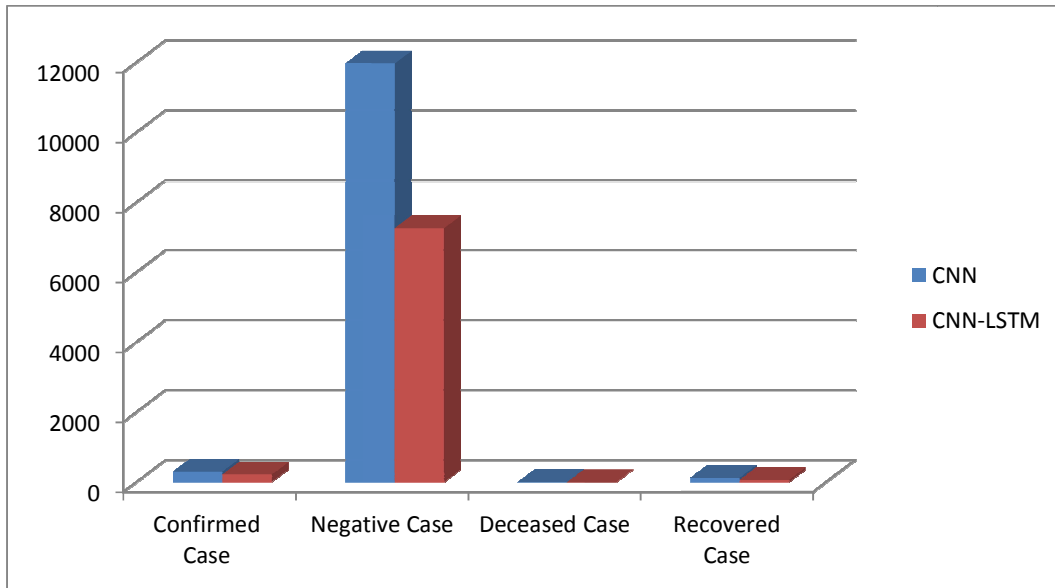
Fig. 5. Performance comparison between CNN and CNN-LSTM with respect to RMSE

3.3 Performance Measure Metrics

For distinguishing the best candidate model from its peers, it is necessary to employ comparison of measures of the algorithm's performance. In this paper, following parameters are used for measuring algorithm's performance:

1. **Mean Absolute Error (MAE)** is formulated as follows [12]-

$$MAE= (\sum_{i=0}^{N}(Xi - Xi') / N) \qquad (1)$$

2. **Root-mean-square-error (RMSE)** is a standard performance measure used for time series forecasting purpose. It is formulated as follows [13]-

$$RMSE= \sqrt{(\sum_{i=0}^{N}(Xi - Xi')^2/ N)} \qquad (2)$$

where Xi is the real value and Xi' is the predicted value and N is the number of samples.

4. EXPERIMENTAL RESULTS

From the above performance comparison, it is quite clear that combined approach performs well. Figs. 6 to 9 depict the plotting of actual result and predicted result for all the specified cases.
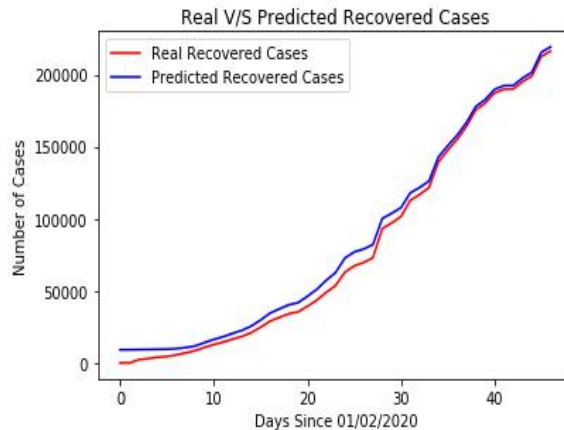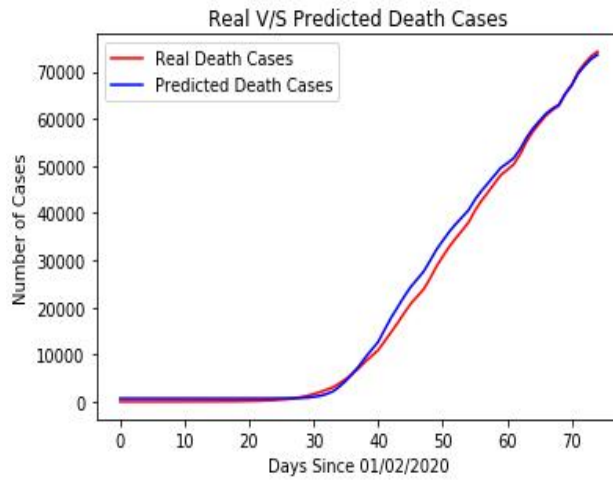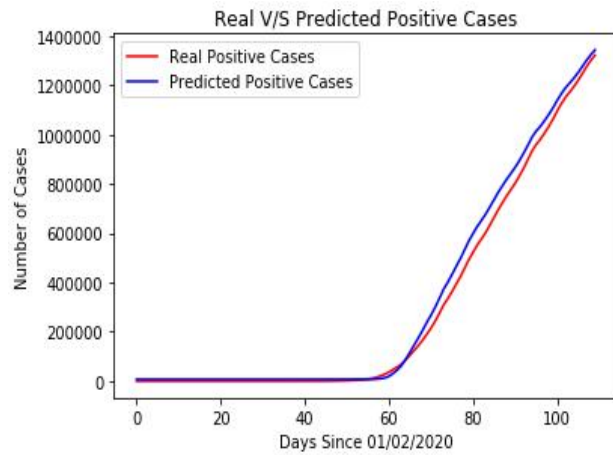

Fig. 6. Released cases

30

Fig. 7. Death cases
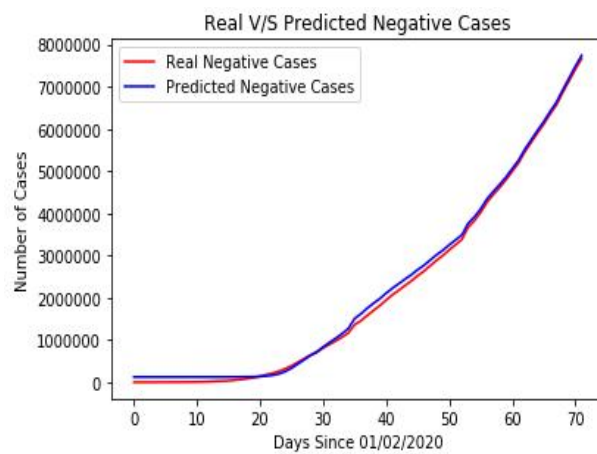

Fig. 8. Positive cases

Fig. 9. Negative cases

5. CONCLUSIONS

Experimental results have shown that the combined CNN-LSTM model obtains a good impact on confirmation of predicted result with respect to actual result. Performance measure metrics are used for verifying confirmation of all the specified cases with respect to Covid-19. The method shows the disease affected people may recover and proper vaccine will be developed to reduce the death rate. The proposed method is forecasting the disease pattern and helps doctors to make a study for prevention at the earliest. Clinical doctors have trying utmost for success and it is believed that they will win in near future.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. 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).
    DOI: 10.3390/en11071636
2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W. China novel coronavirus I. Research T. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020;8:727-733.
3. 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;1.
    DOI: 10.9781/ijimai.2020.02.002
4. Hu Z, Ge Q, Li S, Jin L, Xiong M. Artificial intelligence forecasting of Covid-19 in China. 2020;1–20.
5. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the novel coronavirus (2019-nCoV) outbreak. medRxi; 2020.
    DOI: 10.1101/2020.02.11.20022186
6. O'Shea K, Nash R. An introduction to convolutional neural networks; 2015.
7. Tjandra A, Sakti S, Manurung R, Adriani M, Nakamura S. Gated recurrent neural tensor network. Proc. Int. Jt. Conf. Neural Networks. 2016;448–455.
    DOI: 10.1109/IJCNN.2016.7727233
8. Liu Z, PM, Seydi O, Webb G. Predicting the cumulative number of cases for the COVID-19 epidemic in China from early data; 2020.
9. Guo Q, Li M, Wang C, Wang P, Fang Z, Tan J, Wu S, Xiao Y, Zhu H. Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm; 2020.
10. Metsky HC, Freije CA, Kosoko-Thoroddsen TSF, Sabeti PC, Myhrvold C. CRISPR-based surveillance for COVID-19 using genomically-comprehensive machine learning design; 2020.
11. SRK. COVID-19 in USA. Version 57 [Data file]; 2020.
    (Retrieved May 11, 2020)
    Available:https://www.kaggle.com/sudalairajkumar/covid19-in-usa
12. Chai T, Draxler RR. Root Mean Square Error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 2014; 7(3):1247–1250.
    DOI: 10.5194/gmd-7-1247-2014
13. Yildiz B, Bilbao JI, Sproul AB. A review and analysis of regression and machine learning models on commercial building electricity load forecasting. Renew. Sustain. Energy Rev. 2017;73(2016): 1104–1122.
    DOI: 10.1016/j.rser.2017.02.023

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/57148