ISSN 2563-7568

RESEARCH ARTICLE

COVID-19 Pandemic: A Comparative Prediction Using Machine Learning

Rifat Sadik[1], Md Latifur Reza[1], Abdullah Al Noman[1], Shamim Al Mamun[2], M Shamim Kaiser[2], Muhammad Arifur Rahman[3*]

[1] *Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh*
[2] *Institute of Information Technology, Jahangirnagar University, Dhaka, Bangladesh*
[3] *Department of Physics, Jahangirnagar University, Dhaka, Bangladesh*

Abstract

Corona virus Disease 2019 or COVID-19 is an infectious disease that is declared as a pandemic by the World Health Organization (WHO) and has a noxious effect on the entire human civilization. Every day the number of infected people is going higher and higher and so the death toll. Many countries like Italy, the UK, and the USA were affected badly, yet since the identification of the first case, after a certain number of days, the scenario of infection rate has been reduced significantly. However, a country like Bangladesh could not keep the infection rate down. Several algorithms have been proposed to forecast the scenario in terms of the number of infections, recovery, and death toll. Here, in this work, we present a comprehensive comparison based on Machine Learning to predict the outbreak of COVID-19 in Bangladesh. Among Several Machine Learning algorithms, here we used Polynomial Regression (PR) and Multilayer Perception (MLP) and Long Short-Term Memory (LSTM) algorithm and the epidemiological model Susceptible, Infected and Recovered (SIR), projected comparative outcomes. This comparative study will help the policymakers of a densely populated country like Bangladesh and others to set up new policies to stop the spread, establish healthcare facilities, and vaccination strategies.

Key Words: *Pandemic; COVID-19; Machine learning; SIR; PR; MLP; LSTM*

****Corresponding Author****: Muhammad Arifur Rahman, Associate Professor, Department of Physics, Jahangirnagar University, Dhaka, Bangladesh, Tel: +88 017 2642 8888; E-mail: arif@juniv.edu*

***Received Date**: August 10, 2020, **Accepted Date**: September 04, 2020, **Published Date**: October 30, 2020*

***Citation**: Sadik R, Reza L, Al Noman A, et al. COVID-19 Pandemic: A Comparative Prediction Using Machine Learning. Int J Auto AI Mach Learn. 2020;1(1):01-16.*

1. Introduction

In December 2019 Corona virus disease 2019 or COVID-19 was first detected in Wuhan, which is the capital of the Hubei province of China and started to spread rapidly throughout the country. At present, the virus spread globally over 200 countries which caused the death of over 8.2 hundred thousand people and more than 24 million people with confirmed infected cases and the number is increasing geometrically every day. COVID-19 is an RNA [1] virus with complex nature and variation in its genome sequence making it difficult to invent any vaccine. The best way to fight against the virus is to predict its outbreak and ensure social distancing and isolate areas. To identify a more dangerous area and isolate those areas based on the acuteness of the virus, researchers are relying on artificial intelligence and Machine Learning algorithms to create a standard epidemiological model. This modeling strategy is based on the total number of deaths and confirmed cases in a particular region, age of the infected patients, their gender and previous medical history, traveling history, physical contacts, and so on. Being the most densely populated country in the world, Bangladesh is on the verge of the catastrophic effects of this virus and within a short time, it caused a large number of deaths and the rate of infected patients is growing rapidly. So, it is a must to establish a standard model that will predict the outbreak of the virus and aid the government to take necessary steps to control the outbreak.

A general epidemiological model such as Susceptible-Infected-Recover (SIR) [2] and Susceptible-Exposed-Infectious-Recovered (SEIR) [3] model is widely used to predict epidemics and pandemics but suffers from some limitations. Nowadays Machine Learning (ML) algorithms are widely used to model epidemics such as Ebola, Zika, H1N1, etc. and showed a significant result [4,5]. To deal with COVID-19 researchers are now using AI and Machine learning to detect the disease by analyzing chest x-ray and screening [6], predict the effect of antiviral drugs [7], predict and identify the genome sequence of the virus [8], analyzing socioeconomic effects [9] due to the pandemic and so on.

This comparative study will assist the policymakers of Bangladesh to establish an infrastructure that will monitor the effectiveness of the actions taken to control the spread. The spread of the virus is exponential, and it is quite impossible to identify the clusters of an affected community if there occurs a community transmission. In the case of community transmission, a comparative study will be helpful since it can provide a predictive estimation of the spread nationwide and policymakers can take necessary actions in advance to deal with the casualties. Policies such as social distancing, lockdown, relief plan for the poor, isolation hospitals set up, etc. can be estimated using this comparative study. Besides, this study approximates how long the Government should continue to back up these policies. Vaccination strategy can be set up depending on this study and approximately how many doses of vaccine are needed can be estimated from this study.

Here in this research work, we performed a comparative analysis between epidemiological models and different Machine Learning approaches that can be used to predict the outbreak of the COVID-19. We focused on three Machine Learning (ML) algorithms namely Long Short-Term Memory (LSTM), Polynomial Regression (PR), and Multilayer Perception (MLP) which individually predicts the number of deaths, infections, and recoveries due to the outbreak. As a case scenario, we have considered the situation of Bangladesh and predicted how far the calamity will impact on the coming days. We have considered the scenario of deaths, infections, and recoveries of the USA and Brazil for validation, where the pandemic started before Bangladesh. Furthermore, we studied how Machine Learning approaches can be used as a state of the art method for prediction outbreak compared to epidemiological model SIR. These prediction models will help the authority to optimize disease management facilities by analyzing causality and responses taken to control the pandemic.

2. Background Study

With the advancement of Machine Learning (ML) technology, its applications are also emerging in a discipline like viral epidemiology. Machine Learning is considered as a state of the art technology to detect early contamination of viruses, development of drugs and vaccines, risk assessments, identifying protein structures of viruses, etc. One of the prominent applications of Machine Learning in viral epidemiology is to build a statistical model or prediction model which is proved useful in the contamination of viruses such as Zika, Ebola, Dengue, Malaria, etc.

Yuhanis Yusof and Zuriani Mustaffa [10] proposed a prediction future dengue outbreak using Least Squares Support Vector Machine (LS-SVM) algorithm. The dataset used in this project contained data about the reported dengue cases and the amount of rainfall of five districts in Selangor, Malaysia. A total of 520 data was used among which 70% was for training and 30% for testing dataset. Decimal Point Normalization was used to preprocess both input and output data. The obtained average prediction accuracy was 86.84% which is better than the Neural Network Model whose accuracy was 65.58%. Naiyar Iqbal and Mohammad Islam [11] investigated the forecasting of dengue outbreak using 7 different ML algorithms. The dataset they collected from the test report of individual patients consisting of a total of 75 samples from which 36 samples were dengue negative report and rest were positive. The output result showed that the LogitBoost Ensemble model outperformed all the models acquiring accuracy of 92%. The computed sensitivity was 90% and specificity was 94% for the LogitBoost model which was higher than others.

Dong Jiang et al. [4] mapped the Zika epidemic using Backward Propagation Neural Network (BPNN), Gradient Boosting Machine (GBM), and Random Forest (RF) algorithms. The study predicted high risked area based on the transmission of the virus. The experiment result showed that BPNN achieved an area under the curve (AUC) score of 0.966 which is better than GMB and RF (0.964, 0.963) for both training and testing datasets. It had been also seen that the prediction uncertainty was higher for BPNN. Chekol et al. [12] applied ML models namely Support Vector Regression (SVR) and Adaptive Neuro-Fuzzy Inference System (ANFIS) to predict the breakout of malaria in Ethiopia. A dataset containing malaria data ranged from the year 2013 to the year 2017 and attributed based on elevation, rainfall, humidity, lag malaria case, and positive confirmed case of the current month. The number of records in the training and testing set was 7516 and 3765 respectively. The experiment result showed that ANFIS performed better than SVR based on Regression Coefficient and Root Mean Square Error (RMSE).

Gustavo et al. [13] applied ML algorithms to predict the outbreak of Porcine Epidemic Diarrhea Virus (PEDV) on sow farms. The ML algorithms used in this experiment were Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting Machine (GBM) to predict the outbreak. The dataset was collected from the Morrison Swine Health Monitoring Project (MSHMP) from which 80% was used to train the model and 20% for the test scenario. The performance of the model was assessed based on different factors such as Area under the Curve (AUC) scores, accuracy, sensitivity, and specificity. From the study, it had been seen that the RF model outperformed others with ab accuracy of 83.0%, AUC score of 96%, the sensitivity of 65.2%, and specificity of 95%. Colubri et al. [14] Used Multivariate Logistic Regression to build a prognostic model for the Ebola virus outbreak. In this experiment, the EVD dataset comprised 470 EVD cases reported in five different locations in Sierra Leone and Liberia, and for external validation, 264 cases from two independent datasets were used. The prognostic model successfully performed the task of detection and contamination of the outbreak by achieving AUC ranges from 0.70–0.79 and accuracy range from 0.64–0.74.

Since COVID-19 is a new virus, scientists are working to build an effective epidemiological model to predict the outbreak of the virus. Pinter et al. [15] worked on a prediction model to analyze COVID-19 using time series data of infected persons and mortality using a hybrid Machine Learning approach composed of an Adaptive Network Based Fuzzy Inference System (ANFIS) and Multilayer Perception Imperialist Competitive Algorithm (MLP-ICA). The data was collected from the daily reported cases and the mortality rate in Hungary. The achieved RMSE for total cases for MLP-ICA is 187.88 and for ANFIS is 194.10 and for mortality rate achieved RMSE for total cases for MLP-ICA is 8.32 and for ANFIS is 15.25 which inferred that MLP-ICA outperformed ANFIS. Pandey et al. [16] proposed a model that predicts the outbreak of COVID-19 for India using the SEIR and Regression model and predicts the outbreak for the next 14 days. In this study, the calculated RMSE for the SEIR model is 1.52 and the Regression model achieved 1.75. Tuli et al. [17] proposed Machine Learning based models namely the Gaussian prediction model and the Robust Weibull model are used. For faster computation, these ML models had been employed in the Azure cloud platform where the FogBus [18] framework had used. The computed MSE is lower for the Robust Weibull model than the Gaussian model and MAPE for the Robust Weibull model is higher than the Gaussian model.

Sina et al. [19] worked on a model using Multilayered Perceptron (MLP) and Adaptive Network Based Fuzzy Inference System (ANFIS). Two scenarios were proposed to scale up the scope of the study based on the sampling of data processing (weekly and daily basis). For MLP, the number of neurons was adjusted, and different types of membership functions were used in the case of ANFIS. The experimental result showed that the performance of MLP was better than ANFIS for time series outbreaks in different countries. Yang et al. [20] proposed a modified Susceptible Exposed Infected Removed (SEIR) epidemiological model and LSTM model that predict the epidemic progression of COVID-19 in China. The SEIR model is modified by adding population migration data in the original SEIR equation. The epidemic outbreak is predicted using the modified SEIR within reasonable confidence and the LSTM model also showed impressive performance which is trained on past SARS datasets. It was inferred from the study that the peak time of the epidemic is February which eventually declined by the end of April. A study conducted by Kuniya [21] used the SEIR model for COVID-19 prediction in Japan. The study inferred the time when the epidemic will be at its peak. In this work, the least square-based method and Poisson noise were used to estimate the basic reproduction number. It has been seen that the SEIR model can effectively predict the Epidemic size, assess the necessary steps to improve the medical system.

3. Methodology

As stated in the background study section, many researchers have worked on different approaches to improve the methods used in modeling different epidemic and pandemic outbreak prediction. In this research work, we will compare outbreak prediction models based on three different

Machine Learning algorithms. These algorithms are Long Short-Term Memory (LSTM), Polynomial Regression (PR), and Multilayer Perceptron (MLP). We will also use an epidemiology prediction model called the SIR model for the outbreak prediction.

3.1. Susceptible, Infected and Removed (SIR) Model

SIR model is an epidemiological model to explain the outbreak of epidemics or pandemics [22]. Illustration of this model is given in Figure 1, where the total population of the area is divided into 3 groups. Susceptible (S) group contains people who could potentially be infected, Infected (I) group contains people who are currently infected and contagious and the last group is Removed (R) which contains a group of people who have recovered or died.

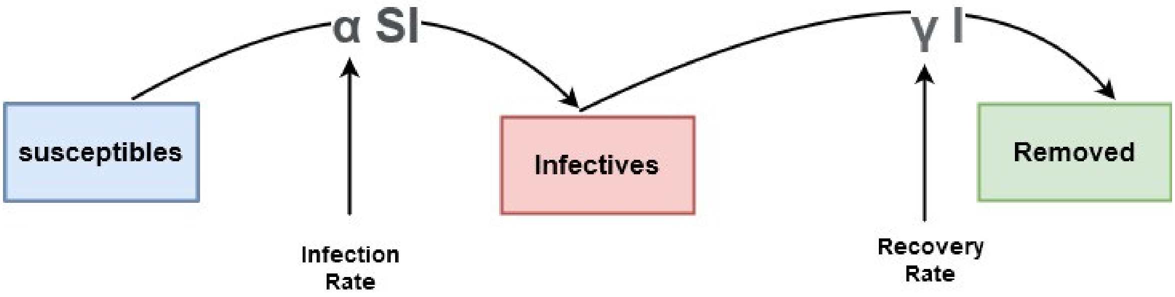

Figure 1: *SIR Model (Susceptible, Infected and Removed classes and their relation with respect to Infection rate, a and recovery rate, γ).*

The SIR model is based on some assumptions such as epidemics that should last for a short period, the total population (N) should remain constant and the increment rate of infected and recovered people should happen at a constant rate. Initially $S=S_0$, $R=R_0$ and $I=I_0$.

The mathematical form of the different groups is given by [23],

Rate of change over time of a number of people in a group,

$$S=\frac{dS}{dt}=-\alpha IS;$$

Where α is the rate of transmission. Rate of change over time of infected people in a group,

$$I=\frac{dI}{dt}=\alpha IS-\gamma I;$$

Where γ is the recovery rate per day. Rate of change over time of removed people in a group,

$$R=\frac{dR}{dt}=\gamma I$$

Here,

$\frac{dS}{dt}+\frac{dI}{dt}+\frac{dR}{dt}=0$ and $S_0 + I_0 + R_0 = N$ indicating closed population.

Contacts per infection are given by, $R_0 = \frac{\alpha}{\gamma}$,If $R_0>0$ then it implies that infection exists and if $R_0<0$ then it implies that there exists no more infection.

3.2. Polynomial Regression (PR) Model

Polynomial Regression (PR) [24] algorithm is used when the range of independent data fluctuates a lot, and the Linear Regression algorithm cannot predict the pattern of data well. The independent variable is modeled as an n^{th} degree polynomial of the dependent variable. Polynomial Regression is used to describe how diseases spread, pandemics or epidemics are spread. The algorithm is based on the following equation [25]

$$y=b_0+b_1x_1+b_2x_1^2+b_3x_1^3+......+b_nx_1^n$$

Where, x_i^j represents the input terms and y is the predicted output.

3.3. Long Short Term Memory (LSTM) Model

LSTM network is a variant of Recurrent Neural Network (RNN) [26] and is widely used to predict time series data [27]. LSTM consist of gates which gives the network a mechanism to regulate the flow of information. These gates can learn to distinguish between data in a sequence and chose the relevant information to predict from time-series data. The model is depicted in Figure 2 LSTM network maintains a cell called state vector.

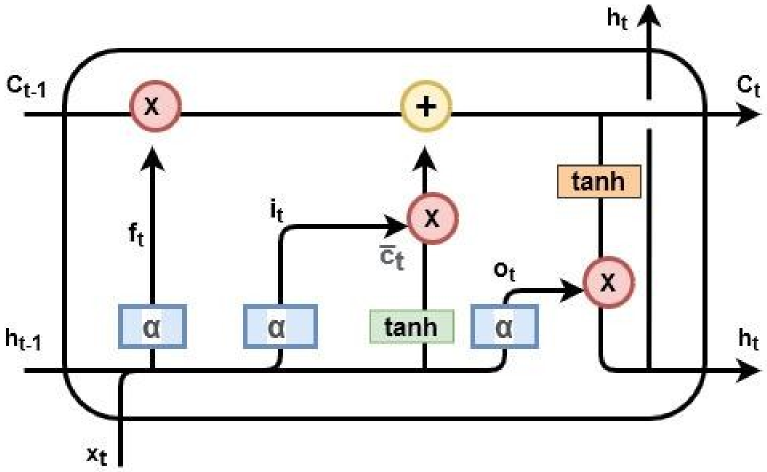

Figure 2: *Long Short Term Memory cell (Three gates namely input, output and forget gate to control the propagation of the cell with sigmoid and tanh activation function.*

Each time step the next LSTM cell can read or write from it. With the current input vector $x^{(t)}$ and current hidden layer vector $h^{(t)}$ and bias b , each LSTM unit has three gates of the same shape [28]. They are:

Input gate: controls whether the memory cell is updated.

$i^{(t)}=\sigma(W^i[h^{(t-1)},x(t)]+b^i)$; Where W is the recurrent weight.

Forget gate: controls if the memory cell is reset to 0.

$$f^{(t)}=\sigma(W^f[h^{(t-1)},x(t)]+b^f)$$

Output gate: controls the visibility of information of the current cell.

$$o^{(t)}=\sigma(W^o[h^{(t-1)},x(t)]+b^o)$$

All these three gates have sigmoid activation function which constitutes smooth curves in the range of 0 and 1. Vector $\bar{c}$ modifies the cell state.

$$\bar{c}^{(t)}=\tanh(W^c[h^{(t-1)},x^{(t)}]+b^c)$$

tanh activation is used that allows a longer flow of state information without vanishing or exploding. Each gate takes the hidden state and the current input x as inputs. After applying the input gates to $\bar{c}$, we obtain c.

$$c^{(t)}=f^{(t)}c^{(t-1)}+i^{(t)}\bar{c}^{(t)}$$

The output $h^{(t)}$of the LSTM cell is defined by:

$$h^{(t)}=\tanh(c^{(t)})*o^{(t)}$$

3.4. Multilayer Perception (MLP) Model

MLP is an artificial neural network method used to model univariate time series data [29]. So, it is a good choice to model pandemic data which contains temporal ordering of data. MLP consists of one input layer, one or more hidden layers, and one output layer. Every layer except the output layer includes a bias neuron that is fully connected to the next layer. This layering scheme for MLP is illustrated in Figure 3.

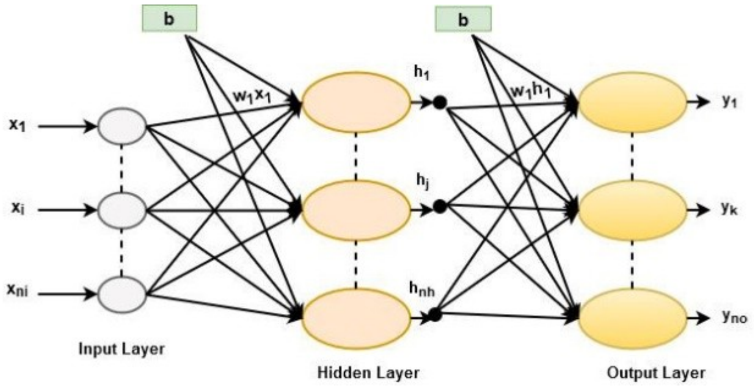

Figure 3: *Architecture of Multilayer Perception (Multiple hidden layers with bias b and weight w) and the non-linear activation function is used to make a difference with linear perception.*

Back Propagation training algorithm is used to train the MLP network. For each training instance, the algorithm feeds it to the network and computes the output of every neuron in each consecutive layer. Then it measures the network output error, then goes through each layer in reverse order to measure the error contribution from each connection and finally slightly tweaks connection weights to reduce the error. The computations of each layer follow certain equations [30].

Here w is the weight between input neurons and hidden layer neurons and b_j is the bias weight. The Output of the hidden layers at an i^{th} neuron is h.

$$hj=\varphi\left(net\left(h_j\right)\right)=\frac{1}{1+e^{-net\left(h_j\right)}}$$

The total input to the output layer at k^{th} neuron from hidden with bias b_k is

$$net\left(y_k\right)=\sum_{j=1}^{n}\left(w_k.h_j+b_k\right)$$

k^{th} neuron of the output layer produces value y_k

$$y_k=\varphi\left(net\left(y_k\right)\right)=\frac{1}{1+e^{-net\left(y_k\right)}}$$

4. Results and Analysis

4.1. Dataset

Since the COVID-19 is a newly emerging virus, there exist a few numbers of globally accepted datasets. We gather the statistical data from Worldometer [31] and IEDCR [32] which provides global COVID-19 data. Our dataset contains time series data of the number of total confirmed cases, total recovery cases, total death cases. We fit these data in our Machine Learning models to predict the outbreak of the pandemic for Bangladesh. A glimpse of our dataset is illustrated using the following figure 4.

4.2. Environment Specification

We have trained our models using GPU provided by Google Colab. It is a free Jupyter notebook that requires no setup and runs entirely in the cloud for deep learning and scientific computations. The cloud service provides 4992 CUDA [33] cores and a memory bandwidth of 480GB/sec (240GB/sec per GPU).

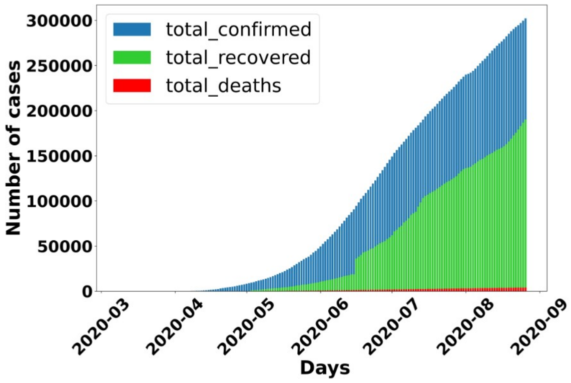

Figure 4: *COVID-19 situation in Bangladesh (Infection, Death, and Recovery cases recorded from 8th of March to 26th of August 2020).*

4.3. Results

The spread of the COVID-19 virus is very unpredictable since its spread fluctuates with public awareness and Government policies. The government is changing its policies such as lockdown, social distancing in public places, testing criteria, and hospitalization continuously. Our model was designed to give time to time insight into the spread so that the Government can update its policy to control the outbreak. For this reason, the initial assumption is assumed to be a short period.

We built a SIR model based on the given equations. The timeline is selected for the next 60 days starting from the 27th of August. By this time the infection spread enormously. The value of effective contact rate and recovery rate for our model is 0.65 and 0.43 respectively [34]. The following figure 5 illustrated a whole scenario for the COVID-19 outbreak.

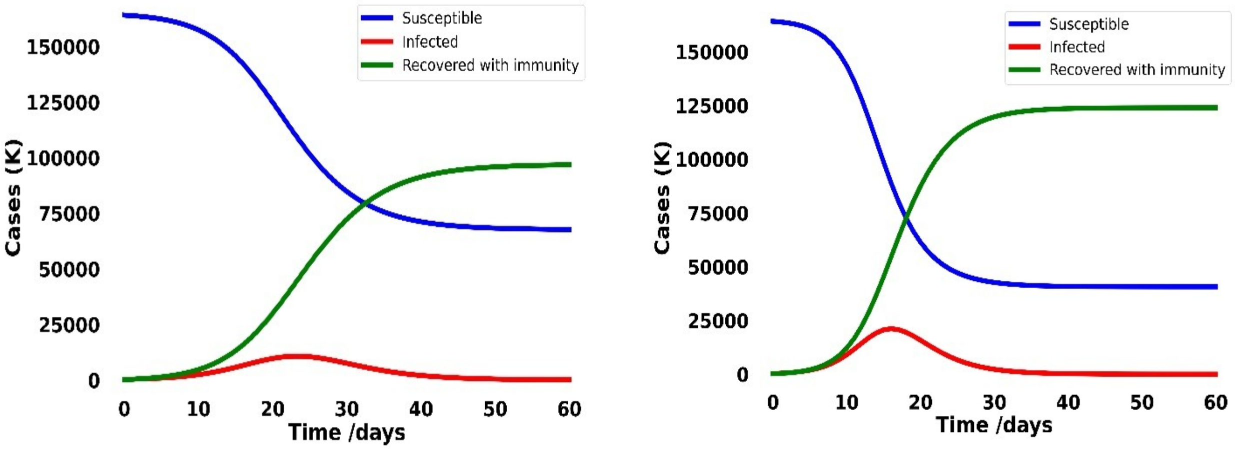

Figure 5: *SIR model to prediction result for the COVID-19 outbreak: (left) SIR model (The total population is 164582631. The data till the 27th of August reported that the number of recovered = 193458, infected = 304583 and death =4127. Calculated effective contact rate, a=0.65 and recovery rate, γ=0.43(Right) and SIR model for worse condition (no lockdown, or social distancing is conducted, effective contact rate (a) = 0.80 and recovery rate (γ) = 0.43).*

Due to the highly-dense population and socio-economic perspective, a large scale of lock down and social distancing may not be possible to impose. For that reason, the assumption is based on the fact that 60% of the population can be brought under the lock down and social distancing strategies. For this reason, the effective contact rate and recovery rate varies from time to time. Besides a poor number of testing may create confusion in estimating the contact rate and recovery rate. By taking pre-fixed values of effective contact rate and recovery rate, we can get a stable model to an extent. From time to time, when policymakers need to update their strategy, they can update these values also to inspect the impact of the current policies regarding stop the spread of the virus. This strategy can resolve the problems in the case of under-fitting.

For our machine learning models, the data-set that is used to train our models contain some data which is considered static. This may create a problem of over-fitting [35]. To overcome over-fitting we used Cross-validation techniques [34] that splits our training data into multiple train-test split. We also added Dropout [36] in our neural network models to tune our models.

In our experiment, for predicting, PR is simulated with different degrees such as 3,4,5,6. MLP is simulated with activation function 'relu', Dropout of 0.05, optimizer RMSprop with learning rate 0.01 and LSTM is simulated using activation function named 'relu', Dropout of 0.05, and optimizer named 'adam'. The following figure 6 gives us the insight of Polynomial Regression, Multilayer Perception, LSTM model to predict the Infections, recoveries, and deaths.

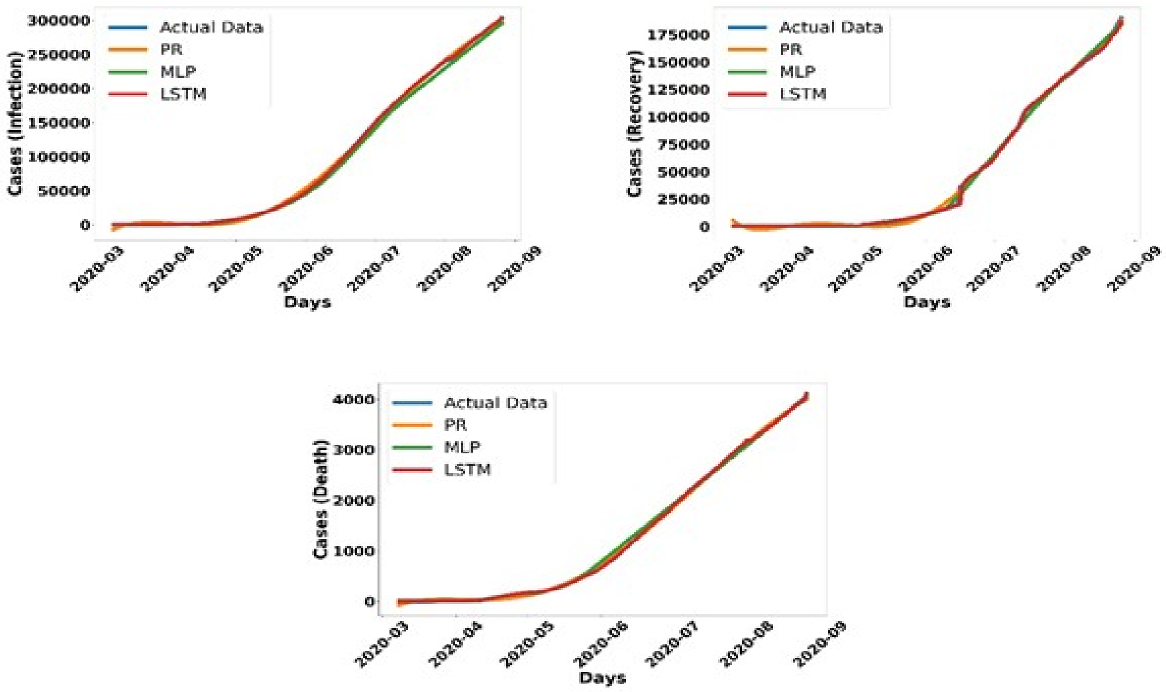

Figure 6: *Reported actual cases for Bangladesh and outcomes from different Machine Learning models. (top-left) confirmed cases of infection, (top-right) confirmed cases of recovery (bottom) confirmed cases of death.*

We have predicted the outbreak for Bangladesh for the next 60 days. In this time frame, we estimated the total number of infected people, deaths, and recoveries for Bangladesh. Figure 7 illustrates the futures cases for infection, recovery, and deaths. Tabular representation is given in table 1.

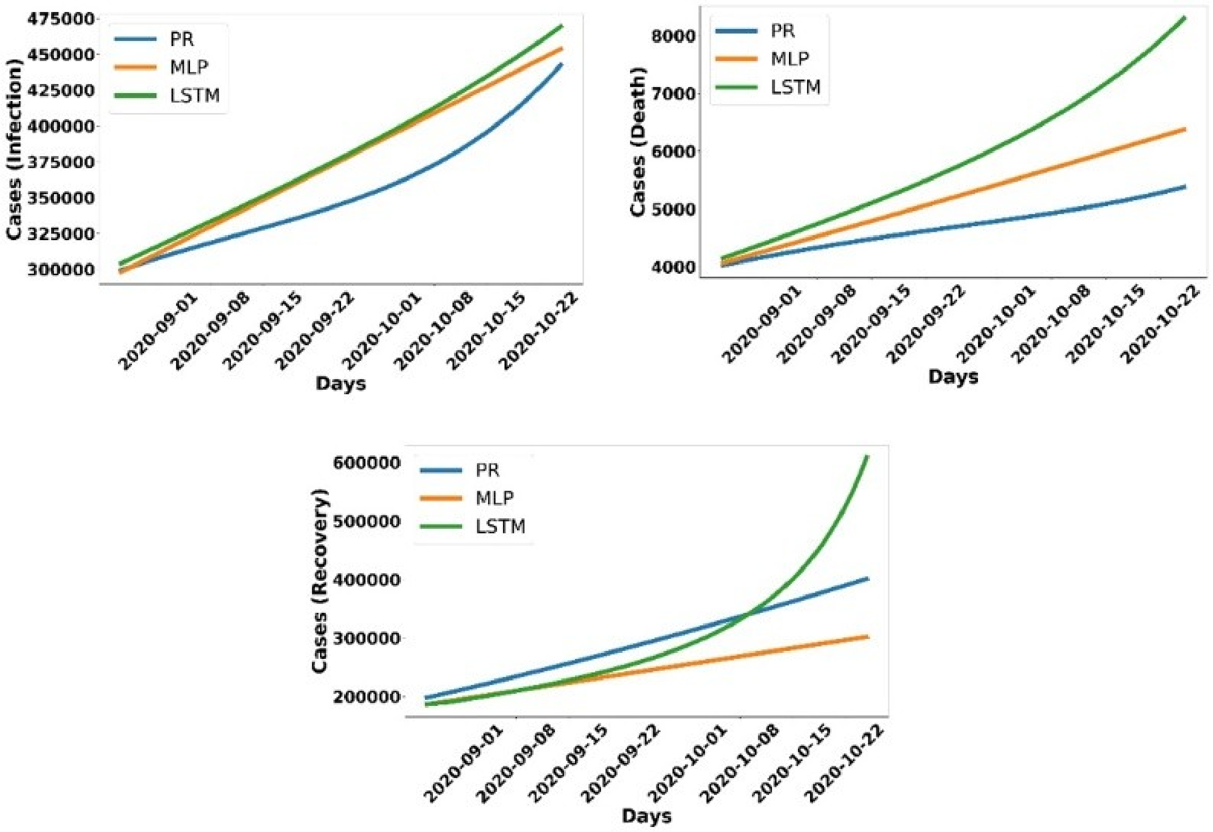

Figure 7: *Predicting future cases for Bangladesh using different Machine Learning models- (top-left) Infection prediction for the next 60 days, (top-right) Recovery prediction for next 60 days (bottom) Death prediction for the next 60 days.*

Table 1:
Prediction of future cases for Bangladesh using different Machine Learning models (August 26, 2020, to October 25, 2020).

Day	**Infected**				**Death**			**Recovered**		
	SIR	**PR**	**MLP**	**LSTM**	**PR**	**MLP**	**LSTM**	**PR**	**MLP**	**LSTM**
1	378231	299356	298090	304062	4026	4066	4150	197983	186506	187167
6	1094750	307746	311289	316289	4161	4256	4391	212786	196301	194413
12	3522799	317026	327127	330943	4308	4485	4688	231162	208054	207539
18	8261354	325853	342966	345741	4440	4715	4997	250196	219808	222297
24	10696265	334787	358804	360844	4562	4947	5323	269878	231561	239628
30	7524667	344541	374643	376416	4676	5181	5673	290199	243315	260803
36	3670749	355991	390481	392650	4788	5417	6058	311150	255068	287834
42	1521633	370184	406319	409770	4906	5656	6489	332721	266822	324205
48	591345	388357	422158	428041	5037	5895	6986	354902	278575	376629
54	244225	411942	437996	447793	5191	6136	7576	377685	290329	459489
60	84237	442581	453835	469445	5381	6376	8301	401060	302082	608165

4.4. Performance Evaluation

Root mean square error or RMSE [37] measures the error of Machine Learning models while predicting time-series data. The less RMSE the better fitted and better predictive time model.

The following equation shows the general formula for measuring the RMSE -

$$\sqrt{\frac{\sum_{i=1}^{n}\left(predicted_i - actual_i\right)^2}{total\ number\ of\ observations\,(N)}}$$

Where N is the total number of observations. Calculated RMSE for our Machine Learning models is given in table 2.

Table 2:
Calculated RMSE of ML models while predicting cases for Bangladesh.

Models	Infected	Deaths	Recovery
Polynomial Regression	467.42	175.60	1679.66
MLP	1475.20	48.03	2393.27
LSTM	161.61	10.56	13.68

As a part of the validation, how our used ML models can predict the pandemic, we made comparisons between the predicted outcome of the models and already reported actual cases (marked as original in figure 8 and figure 9 for USA and Brazil (from June 17, 2020, to July 16, 2020) respectively. Here we predicted the results using PR, MLP, LSTM, and compared the outcome with the actual reported COVID-19 cases for 30 days. Figures 8 and 9 illustrate these situations.

5. Discussions and Findings

Figure 5(left) and figure 5(right) provide insights about the SIR model for the overall prediction timeline of the outbreak. Figure 5(left) illustrates a general case with a calculated exposure rate, α=0.65, and recovery rate, γ=0.43. In figure 5(right) a worst-case scenario is depicted by increasing the value of α (0.80), which implies that no strict lock down or social distancing is conducted. But the model has shortcomings. A pandemic is an unpredictable situation and policy making to fight against it is not fixed and the government imposes different approaches, such as imposing strict lockdown, social distancing, etc. These cases will impact the model as the value of α and γ changes overtime. The value of α has a positive effect on R0 and α has a negative effect on R0. In order to keep the disease under control, we need to keep the rate of transmission low and recovery rate per day high.

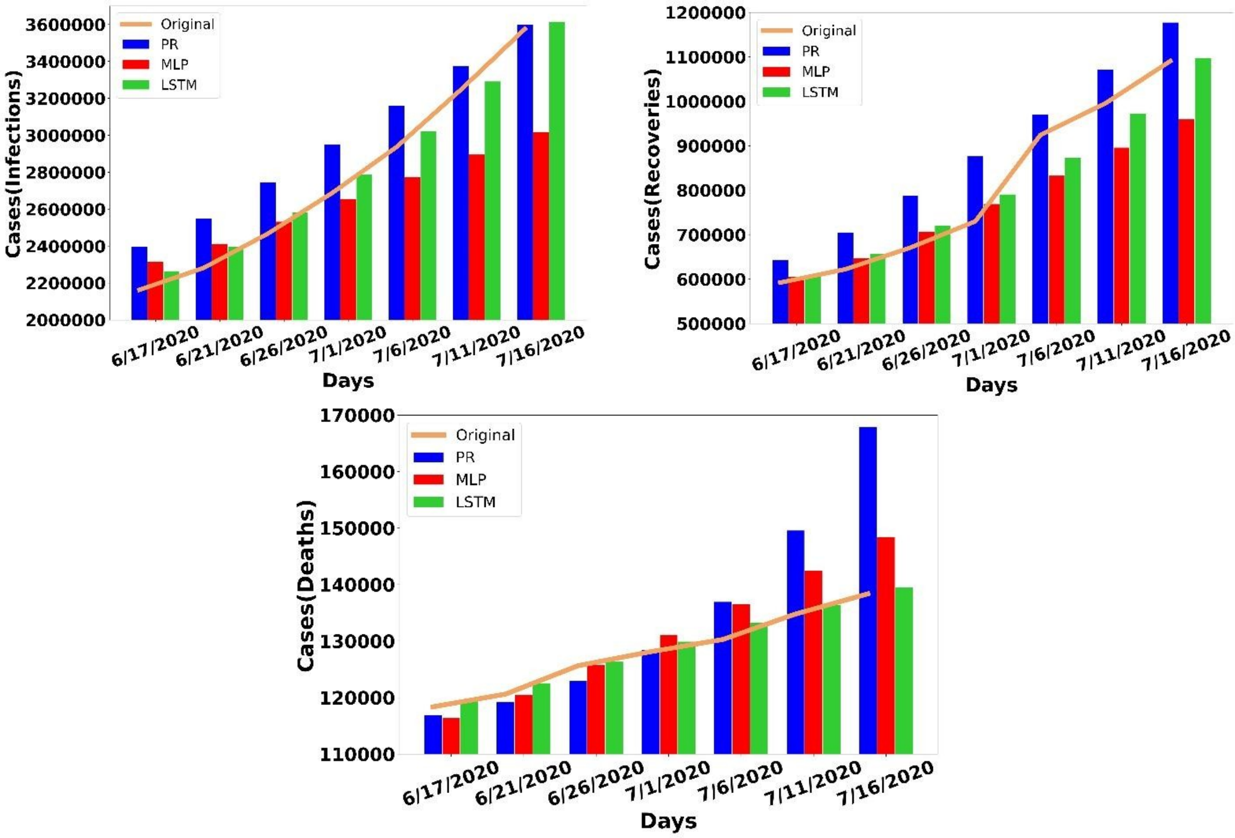

Figure 8: *As a part of the validation, comparison between the predicted and actual cases (marked as original) in the USA (from June 17, 2020, to July 16, 2020). Here we predicted the results using PR, MLP, LSTM, and the compared the outcome with the actual reported.*

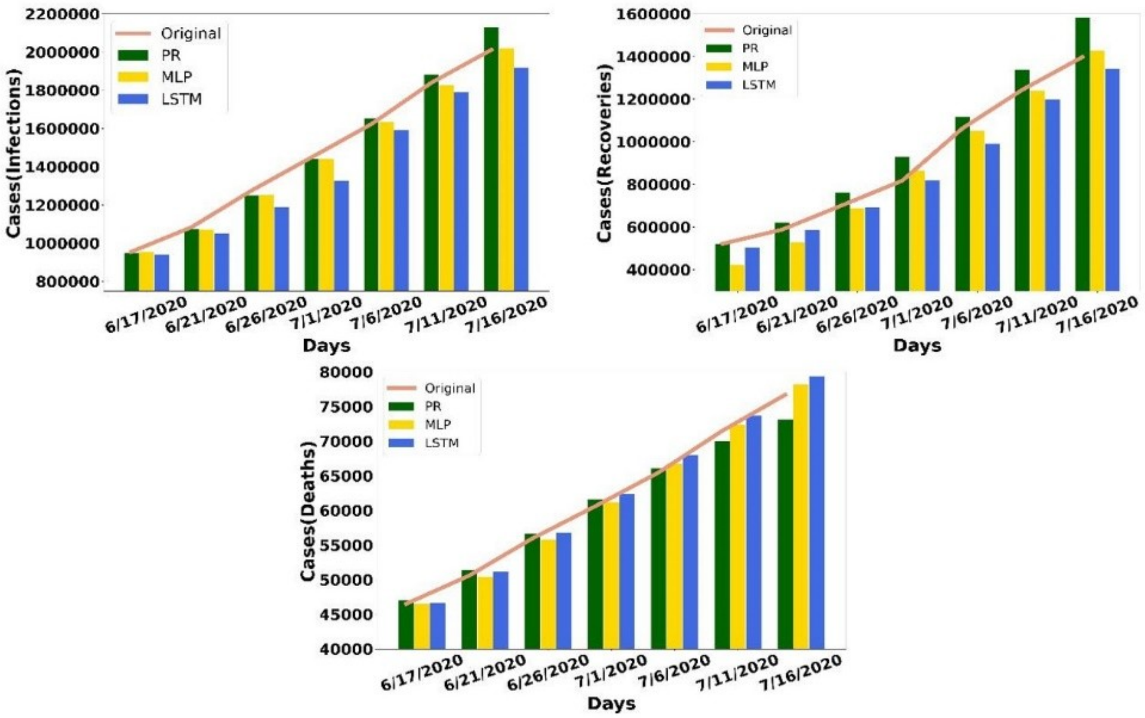

Figure 9: *As a part of the validation, comparison between the predicted and actual cases (marked as original) in Brazil (from June 17, 2020, to July 16, 2020). Here we predicted the results using PR, MLP, LSTM, and the compared the outcome with the actual reported.*

By applying Machine Learning algorithms it can be seen from figure 6 that the trend of the outbreak is well predicted using Polynomial Regression, Multilayer Perceptron, and Long Short Term Memory algorithm. Figure 6 has illustrated that all these three algorithms give a well-fitted prediction curve with respect to the real data. To predict the total confirmed cases, deaths and recoveries PR and LSTM prediction curve is smoother than MLP which implies that both PR and LSTM are well suited to predict the outbreak than MLP. From figure 7 the future outbreak of the disease for the next 60 days can be described. By this time the highest number of infections is given by LSTM which is 469445, the highest number of deaths is predicted using the LSTM model which is 8301and the highest number of recovery is predicted by LSTM which is 608165. To evaluate the model we calculated the root mean square error. From the table, we find that, measured RMSE is quite low for LSTM for all cases (infection: 161.61, deaths: 10.56, recovery: 13.68) which implies that LSTM should be the best model for fitting the prediction curve. RMSE of PR is also low compared to MLP. Despite of giving the highest number of infections and recovery MLP is not suited well for the predicted curve fitting. LSTM outperforms all of them.

The First Corona Virus case in Bangladesh was reported in the first of March, 2020, which is almost 2 months after its first report. During time period, most of the developed countries as well as the developing countries faced serious consequences due to this virus in terms of causalities. To predict the outbreak of this virus, researches all around the world worked on AI-based prediction models deploying Machine Learning and Deep Learning Algorithms. But those are the early prediction based on a small amount of data. For this reason, those models could not give satisfactory results then. In the case of Bangladesh, it the most densely populated country and its economy are still burgeoning. In the meantime, when corona virus impacted in Bangladesh, its behavior of spread was known, and predicting model can be build based on a huge number of data. Besides infection rate in the USA and Brazil was very high, which in case of Bangladesh is much lower and at a steady rate. This situation gives us the advantages to track the spread of the virus better using the LSTM model for Bangladesh.

6. Conclusion

COVID-19 pandemic has become an issue of grave concern for its menacing effects on humankind. In this paper, a comparative analysis of different methods to predict the outbreak of the virus is conducted for Bangladesh. First, we used a well-known mathematical model, SIR, to predict the outbreak for 60 days. The result showed that this model is not suitable for long term prediction due to the inconsistency with affecting factors such as effective contact rate and recovery rate. Later we used three Machine Learning models (Polynomial Regression (PR), Multilayer Perception (MLP), and Long Short Term Memory (LSTM)) to predict the number of infections, deaths, and recoveries. In spite of the high uncertainty of the effects of this virus, these three Machine Learning models showed promising results in predicting outbreak by estimating infections, deaths, and recoveries. Furthermore, the prediction accuracy for the models with respect to calculated RMSE values indicates that the LSTM algorithm outperformed other models for prediction.

Machine Learning models could be used as an alternative for epidemiological models without prior assumptions. Our proposed work can aid the Government of Bangladesh to establish policies by tracking the spread of the virus and estimating the total number of infections, deaths, and recoveries. These comparative outcomes can also be used for other countries where this pandemic prolongs for a longer time and predicts the forthcoming scenario. Furthermore, people can understand the severity of the virus and take precautions to be safe. This comparative study can be used as a guide for the Government to set up policies regarding the prevention of the virus such as social distancing, lockdown, vaccination, the establishment of healthcare facilities. For future studies, the LSTM model can be optimized using cloud services such as Azure and Amazon Web Service (AWS) and integrating with other models such as Gated Recurrent Units (GRUs). We can also the model to estimate the impact of post Corona virus effects on our socio-economic sectors and this can suggest new policies to overcome the damages caused by the virus. More advanced and robust models consist of the Gaussian model and Adaptive Neuro-Fuzzy Inference System (ANFIS) model, Deep Learning models can be applied to handle uncertainty and produce more accurate prediction results. Besides analyzing the socio-economic impact of COVID-19 on stock-market, education, mental health, garments sector, remittances, etc.

Conflict of interests: The authors declare that there is no conflict of interest.

References

1. Liu C, Zhou Q, Li YZ, et al. Research and development on therapeutic agents and vaccines for COVID-19 and related human coronavirus diseases. ACS Cent Sci. 2020;6:315-31.
2. Beretta C, Takeuchi Y. Global stability of an SIR epidemic model with time delays. J Math Biol. 1995;33:250-60.
3. Li MY, Smith HL, Wang L. Global dynamics of an SEIR epidemic model with vertical transmission. Siam J Appl Math. 2001;62:58-69.
4. Jiang D, Hao M, Ding F, et al. Mapping the transmission risk of zika virus using machine learning models. Acta tropica. 2018;185:391-9.
5. Zhang P, Chen B, Ma Ling, et al. The large-scale machine learning in an artificial society: prediction of the Ebola outbreak in Beijing. Comput Intel Neurosc. 2015;2015.
6. https://arxiv.org/abs/2003.10849
7. Wang J. Fast identification of possible drug treatment of coronavirus disease-19 (COVID-19) through computational drug repurposing study. J Chem Inf Model. 2020;60:3277-86.
8. https://arxiv.org/abs/2003.11336
9. Qiu Y, Chen X, Shi W. Impacts of social and economic factors on the transmission of coronavirus disease (COVID-19) in China. medRxiv. 2020.
10. Yusof Y, Mustaffa Z. Dengue outbreak prediction: A least squares support vector machines approach. Int J Comput Theory Eng. 2011;3:489.
11. Iqbal N, Islam M. Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers. Informatica. 2019;43.
12. Chekol BE, Hagras H. Employing machine learning techniques for the malaria epidemic prediction in ethiopia. 2018 10th Computer Science and Electronic Engineering (CEEC), Colchester, United Kingdom. 2018.
13. Machado G, Vilalta C, Mendoza MR, et al. Identifying outbreaks of porcine epidemic diarrhea virus through animal movements and spatial neighborhoods. Sci Rep. 2019;9:1-12.
14. Colubri A, Hartley MA, Siakor M, et al. Machine-learning prognostic models from the 2014-16 Ebola outbreak: data-harmonization challenges, validation strategies, and mHealth applications. E Clinical Medicine. 2019;11:54-64.

15. Pinter G, Felde I, Mosavi A, et al. COVID-19 pandemic prediction for hungary; a hybrid machine learning approach. Mathematics. 2020;8:890.
16. https://www.medrxiv.org/content/10.1101/2020.04.01.20049825v1
17. Tuli S, Tuli S, Tuli R, et al. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. Internet of Things. 2020;11:100222.
18. Tuli S, Mahmud R, Tuli S, et al. Fogbus: A blockchain-based lightweight framework for edge and fog computing. J Syst Software. 2019;154:22-36.
19. https://www.medrxiv.org/content/10.1101/2020.04.17.20070094v1
20. Yang Z, Zeng Z, Wang K, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis. 2020;12:165-74.
21. Kuniya T. Prediction of the epidemic peak of coronavirus disease in Japan, 2020. J Clin Med. 2020;9:789.
22. Allen LJS. An Introduction to mathematical biology. Pearson/Prentice Hall, 2007.
23. Hethcote HW. The mathematics of infectious diseases. Siam Rev. 2000;42:599-653.
24. Chen TH, Chen YC, Chen JL, et al. Flu trend prediction based on massive data analysis. 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China. 2018.
25. Ostertagova E. Modelling using polynomial regression. Procedia Eng. 2012;48:500-6.
26. Sherstinsky A. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D. 2020;404:132306.
27. Gers FA, Eck D, Schmidhuber J. Applying LSTM to time series predictable through time-window approaches. Neural Nets WIRN Vietri-01, 2002;pp.193-200.
28. Tao F, Liu G. Advanced LSTM: A study about better time dependency modeling in emotion recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada. 2018.
29. Koskela T, Lehtokangas M, Saarinen J, et.al. Time series prediction with multilayer perceptron, FIR and Elman neural networks. Proceedings of the World Congress on Neural Networks. Citeseer, 1996.
30. Shiblee Md, Kalra PK, Chandra B. Time series prediction with multilayer perceptron (MLP): a new generalized error-based approach. Advances in Neuro-Information Processing. 2018;pp.37-44.
31. https://www.worldometers.info/coronavirus/
32. https://www.iedcr.gov.bd/h
33. Holm HH, Brodtkorb AR, Sætra ML. GPU computing with python: performance, energy efficiency and usability. Computation. 2020;8:4.
34. Rahman MM, Ahmed A, Hossain KM, et al. Impact of control strategies on COVID-19 pandemic and the SIR model based forecasting in Bangladesh. medRxiv, 2020.
35. Peng Y, Nagata MH. An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. Chaos Soliton Fract. 2020;139:110055.
36. Qian L, Hu L, Zhao L, et al. Sequence-dropout block for reducing overfitting problem in image classification. IEEE Access. 2020;8:62830-40.
37. Wang W, Lu Y. Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. Iop Conf Ser Mater Sci Eng. 2018;324:012049.
38. Santos MS, Soares JP, Abreu PH, et al. Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. IEEE Comput Intell M. 2018;13:59-76.