COVID-19 Epidemic Trend Prediction using Long Short-term Memory Network

Tianren Zhang *

Department of Electrical Engineering, Hebei University of Technology, Hebei, China

* Corresponding author email: 191341@stu.hebut.edu.cn

Abstract. The COVID-19 pandemic is continuously spreading in various countries and different regions. It produces serious economic shock worldwide and negatively impacts the life and work of people. Although many control measures are conducted to contain its spread, it is still not known when the epidemic will end. Predicting the trend of COVID-19 accurately is extremely important. It can improve the resource allocation rate and make better preventive and control measures for the epidemic. In this paper, Long Short-term Memory (LSTM) models are leveraged for predicting the epidemic in different countries, including Germany, Japan, Russia, and Italy. The LSTM is a type of recurrent neural network (RNN), which is effective for predicting sequential data such as the time series. In this work, a visualization analysis is firstly conducted for demonstrating the trends of COVID-19 in various countries. Then the performances of the LSTM network are validated on the data of four countries.

Keywords: LSTM; COVID-19; Machine Learning; Deep Learning; Trend Prediction.

1. Introduction

Since the year 2019, the new coronavirus has ravaged the world and the momentum of the new coronavirus disease has not decreased. It is very important and valuable to be able to predict the direction of neo-coronavirus reasonably and correctly to arrange the restoration of production and life. On the one hand, medical and other resources could be pre-allocated according to the predicted results. In turn, resource utilization can be increased and resource waste can be reduced, such as allocating medical drugs, doctors, etc. On the other hand, further prevention and control of epidemics can be done based on the results.

Neural networks can perform classification tasks effectively. However, in traditional neural networks, the model only focuses on the state at the current moment at a time and does not record information that is relevant to subsequent predictions. Therefore, a traditional neural network has no memory, so it classifies the events that occur at each moment without using information that has already occurred before. Recurrent Neural Networks (RNNs) and other neural networks with memory functions can be used to make predictions. There are many works on the introduction of these networks [1,2]. Many experiments with predictions for various countries [3,4,5,6,7,8,9], also for global epidemics [10,11], and experiments using improved LSTM models [12].

This experiment processes the new crown data from Hopkinson University statistics and visualizes the cases in each country. Afterward, RNNs, details of the LSTM network [13], and its advantages over RNNs will be presented. Then, the results of the training fit of the LSTM time series prediction model to the data will be shown and analyzed. Previous work has focused on single countries, such as the United States, India, and some provinces and cities in China. Here validation was performed on data from several countries to test the prediction effectiveness of LSTM for different countries. Finally, some parameters were modified to test and discuss the performance of the LSTM.

2. Method

2.1 Data Preparation

The data of this experiment used the new crown data from Hopkinson University statistics. This data is more comprehensive, including data from January 22, 2020, to the present day for each country,

and has more categories, including latitude and longitude, Confirmed, Deaths, Recovered, Active, Incident Rate, Case Fatality Ratio, and many other statistics, which are convenient for the analysis and prediction of the epidemic.

2.2 Data Visualization

Data visualization is a very important step before data processing. Data visualization allows not only the timely detection of data anomalies or other data anomalies but also the understanding of data trends and structures to better analyze the data.

Figure 1 shows the daily data additions for several countries and Figure 2 displays the cumulative number of infections in countries worldwide.

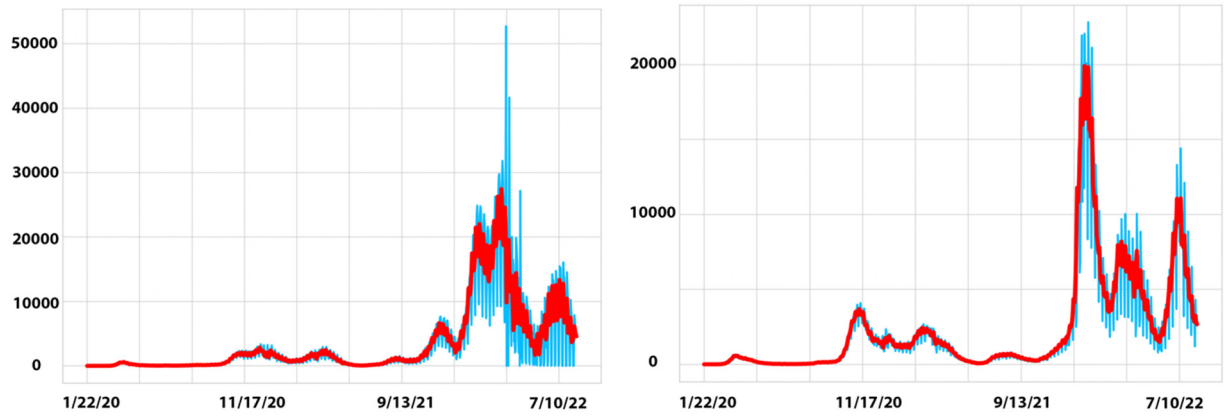


Fig 1. Daily confirmation data visualization for Germany(left) and Italy(right)

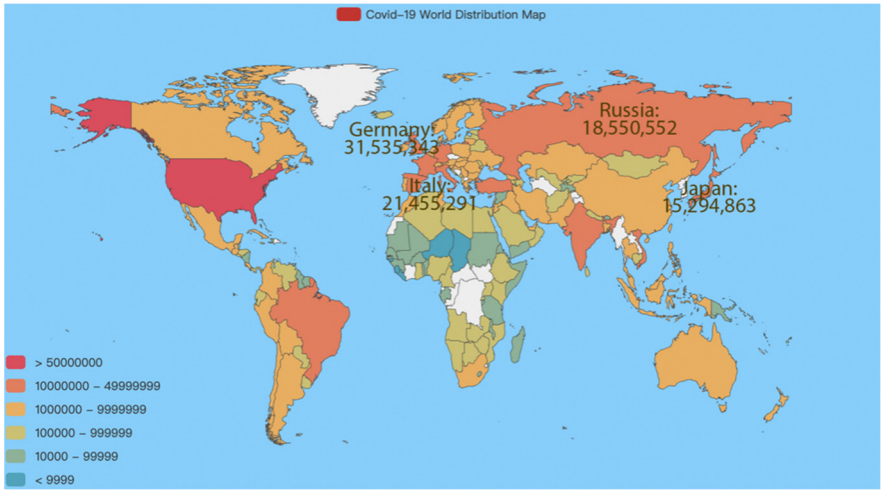


Fig 2. Map of the cumulative number of infections in countries worldwide

The blue line in Figure 1 shows the number of daily additions and the red line is the smoothed value. The visualization of the data for some countries shows that the daily additions are oscillating and may have multiple peaks, and this trend varies from country to country. However, at the end of 2021, there is a large increase in the number of new confirmed diagnoses in many different countries. It is hypothesized that the effect of mutated Omicron virus has led to an increase in the number of new diagnoses. The reason for this is the higher transmission rate of mutated viruses.

And the map shows that almost all countries have been affected by the outbreak to a lesser or greater extent. The United States has the highest number of confirmed cases, followed by Brazil, Russia, and India. A few countries are blank as no statistics are available.

2.3 Models

In this section, the reader will be introduced to some details of the model LSTM used in this experiment and its advantages compared to the traditional method RNN.

2.3.1 Recurrent Neural Networks

The RNNs are naturally leveraged given sequential data. Time series data reflects the state of change of something, phenomenon, etc. over time. It is important to note that there is a connection between adjacent data. For example, a prediction of what the next word of a sentence will be is desired, and the previous words are required. It is because the meaning of a word is dependent on its context words. RNNs get the name because of the recurrent behavior, where the output sequence is related to the previous output as well. This is shown by the fact that the network remembers the previous information and applies it to the computation of the current output.

RNN introduces this parameter of hidden state h to model the sequential problem. h allows for the extraction of features from sequence-shaped data, followed by conversion to output. The process of computing hidden state h1 is shown first.

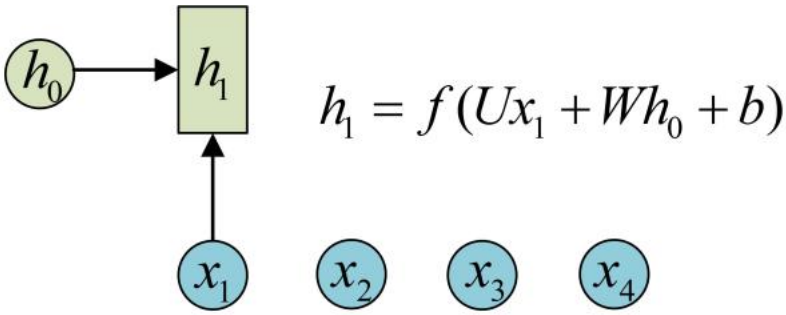


Fig 3. Calculation of hidden layer h1. Figure reproduced from [14]

The calculation process of h2 is similar to that of h1, but there are some things to note. The parameters U, W, and b are the same at each step, which means that the parameters are shared at each step. In this way, more hidden states are computed, and the process can continue indefinitely. The output can be obtained directly by using h after calculation. As mentioned before, an arrow represents the transformation of the corresponding vector of f(Wx+b). This arrow in Figure 4 indicates that obtaining the output y1 can be achieved by a transformation of h1. The other outputs are similar, using the same parameters V and c as y1.

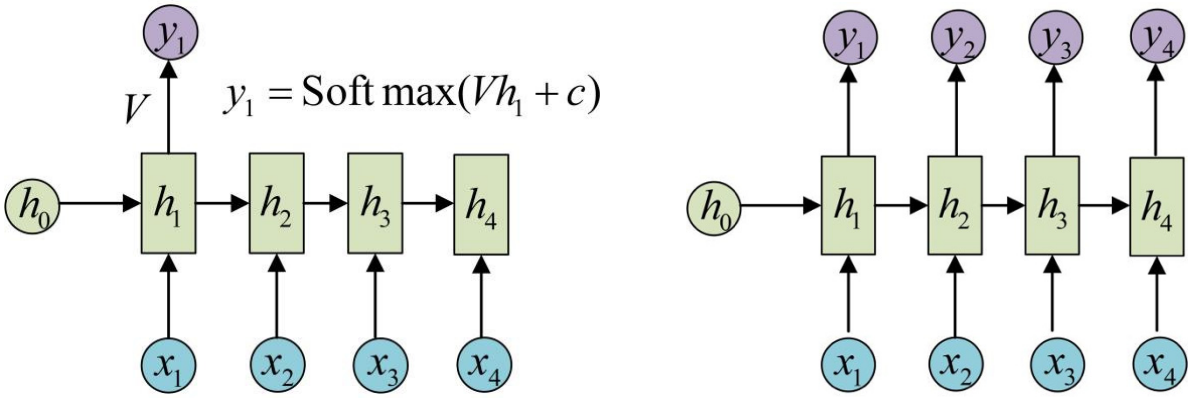


Fig 4. The diagram on the left shows how to calculate y1, while the diagram on the right shows the flow chart of the structure of RNN. Figure reproduced from [14]

This is the most classical RNN construction, where the inputs are x_1, x_2, ... x_n, and the outputs are y_1, y_2, ...y_n. This means that the input length of a sequence and the length of the output sequence are equal.

But RNN also has its shortcomings, and one of them is the long-term dependency problem. When there is a small distance between the target information and the relevant information, RNNs can better use the previous information. However, this is not enough. When it comes to complex situations the RNNs become inaccurate and efficient. This indicates that there is a very large interval between the relevant information and the current predicted partial information.

There are some reasons for this problem after an in-depth study of the problem. RNNs are influenced by short-term memory. It is difficult for RNNs to remember information related to it from long ago when predicting the current information. If a long text prediction is desired, the RNN will not perform well because it forgets the previous information. And RNNs also have the problem of gradient disappearance when backpropagating.

2.3.2 LSTM Networks

One of the reasons why LSTM is a very successful and widely used neural network model is that it is specifically designed to solve the long-term dependence problem. Compared with RNN networks, the most important feature of LSTM is that it can keep only relevant information for prediction and does not learn irrelevant data. This ability to remember important information and forget irrelevant information is critical, especially when memory capacity is limited. Similar to RNN, LSTM has some repetitive units. These include four interacting layers, three sigmoid and one tanh layers, and interact in a very specific way. Figure 5 shows some of the symbols that will be used next.

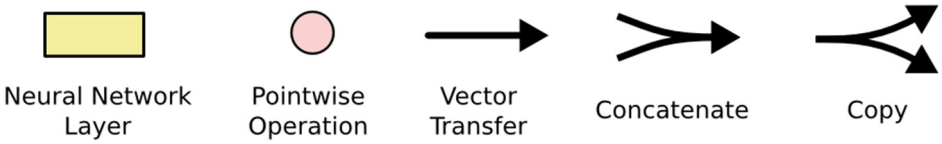


Fig 5. Some of the graphical symbols that will be used below

The core of the LSTM is the cell state, which runs above all the cyclic cells. It is similar to a conveyor belt, with only a few interactions, and is more conducive to the flow of information over it while remaining unchanged. At the same time, a clever structure exists to comfortably allow information to be selected for passage, which contains a sigmoid neural network layer and a nonlinear module for multiplication. The sigmoid function maps the input values between [0,1], representing the proportion of information retained by this gate.

The LSTM network has three types of gate structures, Figure 5 and Figure 6 show the basic structure of the LSTM

The first one is called the "forget gate layer", which determines the information to be lost. This gate reads the output h_{t-1} of the previous cell and the current input x_t. After a Sigmoid nonlinear mapping, it outputs a vector f_t, where each dimension has a value between [0,1]. Finally, f_t is multiplied with the cell state C_{t-1}, thus dropping some information for each dimension, as shown below.

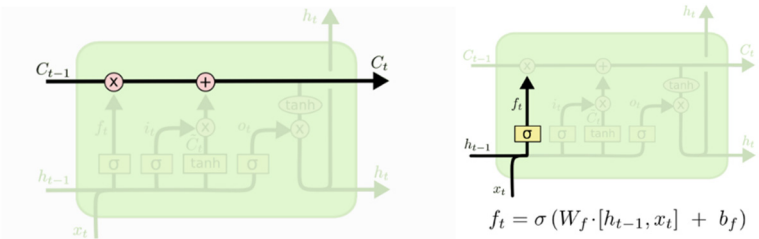


Fig 6. The left side shows the cell state of the LSTM, while the right side shows the first basic gate structure 'forget gate', and the computational equation. Figure reproduced from [13]

The second type of gate structure is made up of two parts. The sigmoid layer on the left is named the "input gate layer". This layer decides the information be updated. The *tanh* layer on the right generates a new vector C_t, which is the current condition, and it will be added to the state.

After the input passes through these two layers, the old state of the cell state is updated from C_{t-1} to C_t. After multiplying the old cell state with f_t, i.e., dropping some information (forget gate layer), $i_t * C_t$ is added and the result is the new candidate value.

The last type of gate structure is the output gate, which decides the output based on the cell state. First, the input is processed through a sigmoid layer for determining the output part of the cell state. Afterward, the updated C_t is processed through a *tanh* layer, which maps the values in the vector between [-1,1]. Finally, it is multiplied by the output of the sigmoid gate to get the information of the determined portion of the output.

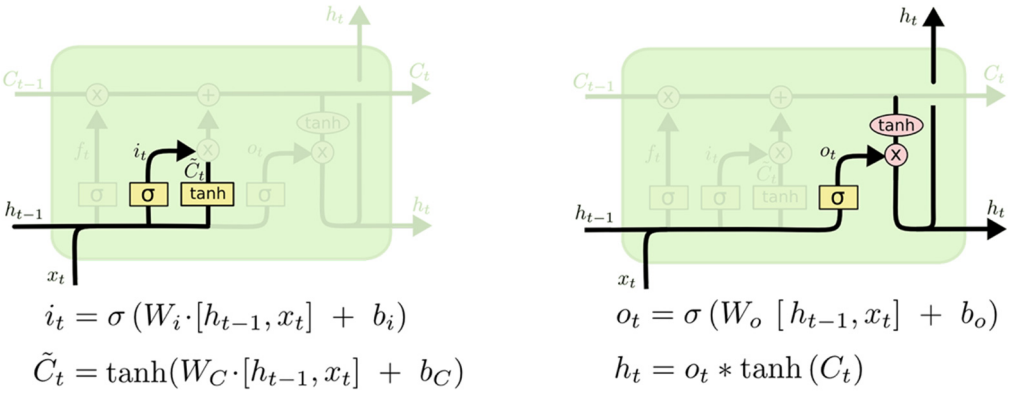


Fig 7. Structural diagrams of the second (left) and third (right) gate structures and equations. Figure reproduced from [13]

3. Results

In the experiments, the data from Japan, Germany, Italy, and Russia were processed and predicted using LSTM models. The PyTorch library with version 1.10.2 was used in the experiments to build the basic LSTM network structure.

First, the training data and test data are separated from the dataset according to a certain ratio. The ratio of training data used in the experiments is 0.8. If LSTM is used to predict sequence data, then it is necessary to set the size of the sequence length. This means that the LSTM will predict the next data based on the sequence length of the data. If it is set to 50, the LSTM will predict the output of the next day according to the data input of the previous 50 days. In general, when the sequence length is larger, the complexity of the computation is larger, but the prediction accuracy will be improved. In the experiment, the sequence length is set to 50.

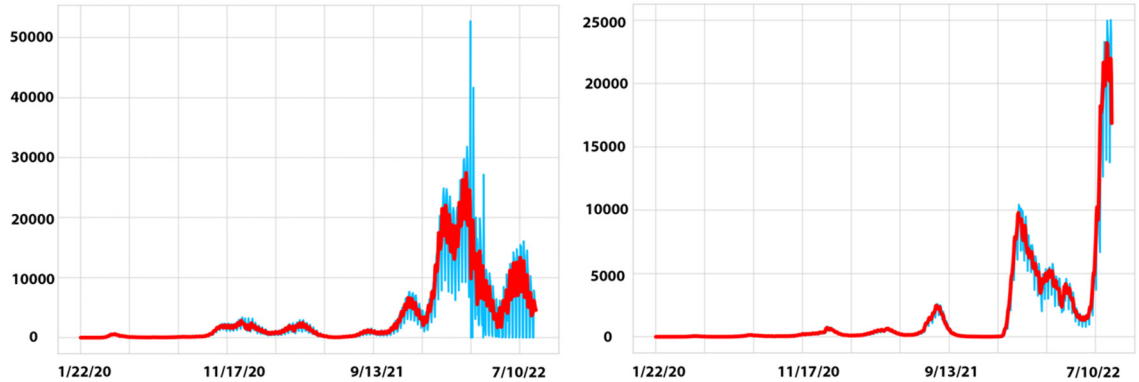


Fig 8. Number of confirmed daily cases in Germany (left) and Japan (right)

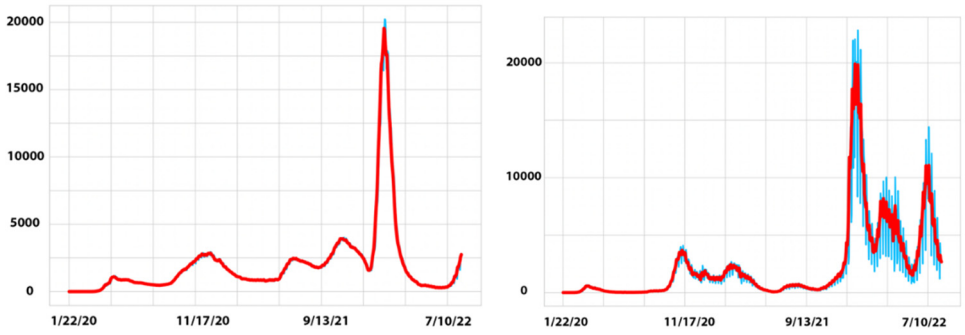


Fig 9. Number of confirmed daily cases in Russia (left) and Italy(right)

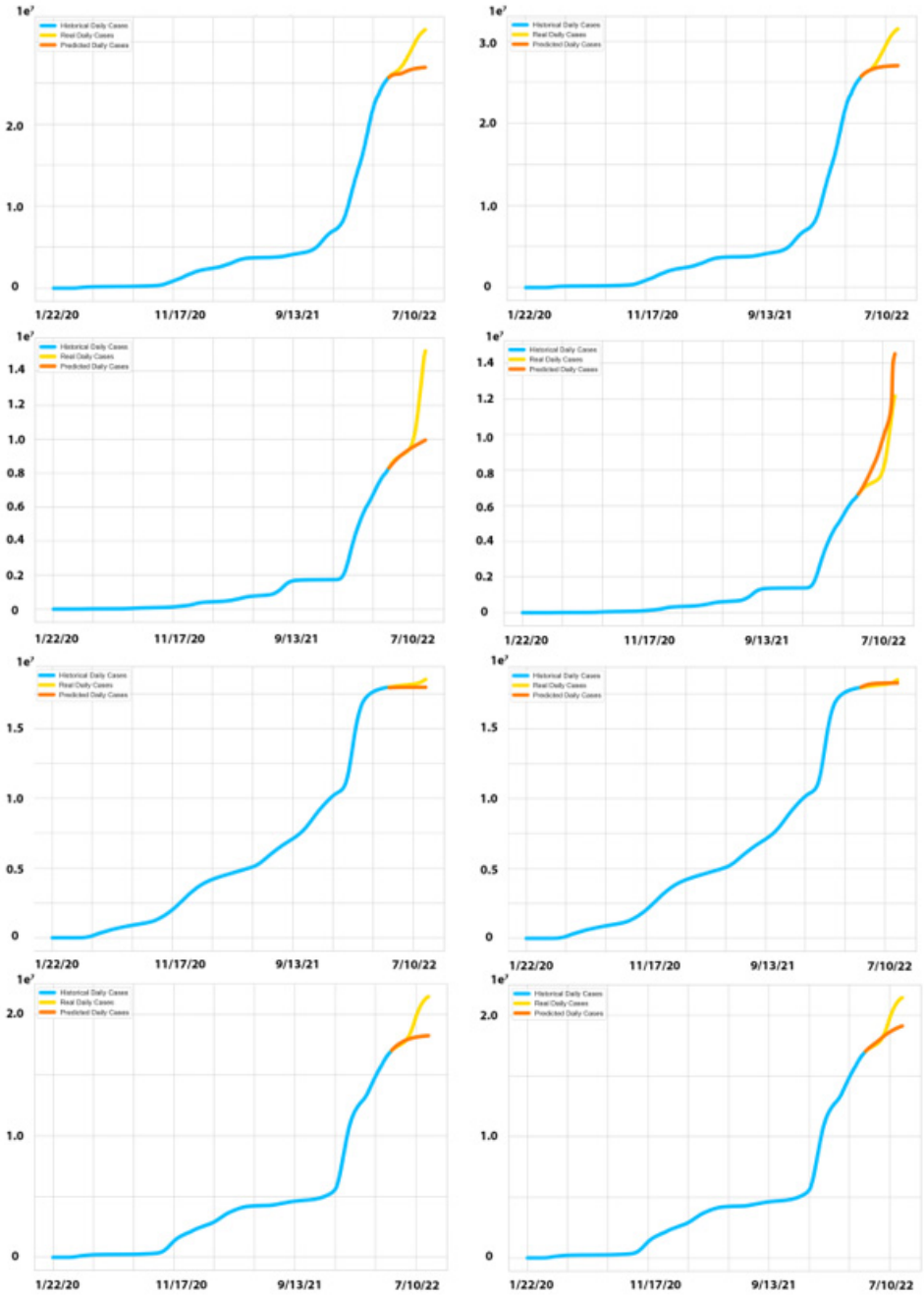


Fig 10. Comparative experiments with sequence lengths of 50 and 100 in various countries. From top to bottom are Germany, Japan, Russia, and Italy. The seq-length on the left is 50 and on the right is 100

From the real daily confirmed data in Figure 7 and Figure 8, the data fluctuations and wave positions of the four countries are not the same. Germany's data fluctuates more obviously, but Russia's data is smoother. But without exception, the number of daily confirmed in each country has a sharp rise after the end of 2021. The speculation is that the effect should be caused by viral ideation. New coronavirus mutations have produced the more infectious beta virus and, more recently, the Omicron virus. This uncertainty also poses an obstacle to accurate prediction.

The data from the above countries were trained using the LSTM model with 3000 epochs. The initial learning rate is configured as 1e-3 and the learning rate adjustment strategy is cosine decay, the loss function is set to MSE, sequence length is set to 50. Moreover, to investigate how the model performs with different sequence lengths, the second experiment sets the sequence length to 100 and compares the experimental results. The predicted results are shown here, where the red line denotes the data predicted by the model, and the yellow line is the real data.

To quantify the quality of the results, the predicted results are now normalized and squared against the real data to obtain the relative error values.

Table 1. Results of error for four countries with different seq-lengths

Seq-length	Germany	Japan	Russia	Italy
50	0.640	1.589	0.014	0.558
100	0.531	1.755	0.003	0.294

From Table 1, the LSTM model has different predictions for the daily confirmed for each country. The model's prediction for Russia is closer to the real one, and the relative error is only 0.0037 when the sequence length is 100. For Italy and Germany, the model can accurately predict the trend of the first part, but the error of long-term prediction is larger.

After increasing the sequence length, the accuracy of daily confirmation for all countries except Japan is improved by different magnitudes. The prediction for Japan has a relatively large error of 1.755, but after increasing the sequence length, the model succeeds in predicting the subsequent upward trend, although there is some error relative to the true value.

4. Discussion

From the results, the model is better at predicting future changes in the short term. In the experiment, the model predicts the first month of each country very accurately. However, from the first month or so, the accuracy of the model starts to drop significantly, which means that the model does not accurately predict the trend over a longer period. The longer the predicted outcome, the less accurate the model is. In general, the longer the sequence length used by the model, the better the prediction. Increasing the size of the sequence length from 50 to 100 increases the accuracy for each country to varying degrees. The prediction for Russia is the most accurate, with an error value of 0.003. The prediction for Japan is larger for the longer sequence length, but the shorter sequence length does not predict the future trend. On the contrary, the model predicted the future outbreak trend after increasing the size of the sequence length.

Both the training and test loss values decreased and stabilized with training, indicating that the training process was relatively normal. However, some of the loss curves fluctuate greatly and tend to rise and fall repeatedly, probably because the sequence length is short and the model is not able to learn the trend well.

More factors affect the number of daily confirmed. The fluctuation of the number of DAILY CONFIRMED can be influenced to a large extent by the different governmental measures, the different infectious capacities of the virus, and the different number of people's exposure. More factors may need to be considered, such as transmission rate, mortality rate, number of infected, number of recovered, number of normal, etc. Or try to use other prediction models, such as SEIR,

GRU, etc., to verify how well more models perform for predicting new crown outbreaks to find a way to make predictions with higher accuracy.

5. Conclusion

In this paper, a deep learning model is presented for trend prediction of epidemic data from different countries using LSTM. The model shows high performance in time series forecasting. This paper first visualizes the covid-19 data, followed by an introduction of RNNs networks and the advantages of LSTM over them. Finally, the LSTM model is used to predict the daily number of covid-19 additions in Germany, Japan, Russia, and Italy. The results show that the LSTM model has different performances for different countries. Among them, the best prediction of daily additions is for Russia. For other countries, the prediction accuracy is higher in the short term and lower in the long term. Increasing the sequence length of the LSTM can increase the performance of the LSTM to some extent and bring accuracy improvement. After increasing the sequence length, it can predict the trend of daily new data in Japan better, although there is some error with the real value.

In conclusion, the LSTM model was better in predicting the number of daily confirmed for only some countries. Since there are more factors affecting the spread of covid-19, considering more factors, such as infectivity, the number of people's exposure, or optimizing the structure of the model, may improve the prediction accuracy.

References

[1] Staudemeyer, Ralf C., and Eric Rothstein Morris. Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. ArXiv preprint, 2019.

[2] Zachary C. Lipton, John Berkowitz, Charles Elkan. A Critical Review of Recurrent Neural Networks for Sequence Learning. ArXiv preprint, 2015.

[3] Rohitash Chandra, Ayush Jain, Divyanshu Singh Chauhan. Deep learning via LSTM models for COVID-19 infection forecasting in India. ArXiv preprint, 2021.

[4] Absar N, Uddin N, Khandaker MU, Ullah H. The efficacy of deep learning-based LSTM model in forecasting the outbreak of contagious diseases. Infect Dis Model, 2022, 7(1):170-183.

[5] Nikparvar, B., Rahman, M.M., Hatami, F. et al. Spatio-temporal prediction of the COVID-19 pandemic in US counties: modeling with a deep LSTM neural network. Scientific reports, 2021, 11(1), 1-12.

[6] Vinay Kumar Reddy Chimmula, Lei Zhang. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals, 2020, 109824.

[7] Shashank Reddy Vadyala, Sai Nethra Betgeri, Eric A. Sherer, Amod Amritphale. Prediction of the number of COVID-19 confirmed cases based on K-means-LSTM. Array, 2021, 100085.

[8] Kumar, S., Sharma, R., Tsunoda, T. et al. Forecasting the spread of COVID-19 using LSTM network. BMC Bioinformatics, 2021, 22 (Suppl 6), 316.

[9] Myung Hwa Kim, Ju Hyung Kim, Kyoungjin Lee, Gwang-Yong Gim. The Prediction of COVID-19 Using LSTM Algorithms. International Journal of Networked and Distributed Computing, 2021, 2211-7946.

[10] Shastri S, Singh K, Kumar S, Kour P, Mansotra V. Deep-LSTM ensemble framework to forecast Covid-19: an insight to the global pandemic. Int J Inf Technol, 2021, 13(4):1291-1301.

[11] Shastri, S., Singh, K., Kumar, S. et al. Deep-LSTM ensemble framework to forecast Covid-19: an insight to the global pandemic. Int. j. inf. tecnol. 2021, 13, 1291–1301.

[12] Farah Shahid, Aneela Zameer, Muhammad Muneeb. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. Chaos, Solitons & Fractals, 2020,110212.

[13] Colah. Understanding LSTM Networks. URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs /, 2015, 2022-09-11.

[14] July. Understanding LSTM from RNN. URL: https://blog.csdn.net/ yanqianglifei/ article/ details/ 80486678, 2019, 2022-09-11.