Proceedings of the 2nd International Conference on Environmental Geoscience and Earth Ecology
DOI: 10.54254/2753-8818/37/20240154

Long short-term memory model for COVID-19 forecasting

Pengfei Lou

Rollins School of Public Health, Emory University, USA, Atlanta, 30306

plou3@emory.edu.cn

Abstract. Every time new variants of COVID-19, back to the pandemic, could bring massive loss to human society and traditional models have failed to catch the complexities of COVID-19. Thus, to handle the unpredictable challenges posed by the dynamic nature of COVID-19, Long Short-Term Memory (LSTM) models were utilized to make accurate short and long-term predictions about different variants and prepare the model for another variant or virus similar to COVID-19 by learning the data of different COVID-19 variants. The dataset, sourced from 'owid-covid-data,' is cleaned and divided into original, alpha, delta, and mixed variants in the United States from March 1, 2020, to April 30, 2022. MSE, RMSE, MAE, and R2 are used to compare the difference between real and predicted values to evaluate how accurately the model performs. Results demonstrate the model excels in long-term and short-term predicting COVID-19 cases and deaths for various variants, and mixed variants even promote accuracy. Thus, the proposed LSTM model shows promise for infectious disease forecasting, providing a foundation for anticipating future outbreaks to help policymakers make better decisions or early prevention.

Keywords: COVID-19, multivariate variants, Long Short-Term Memory, Time-series-forecast

1. Introduction

The global outbreak of COVID-19 [1,2] has not only claimed countless lives but has also caused profound societal losses, underscoring the critical need for effective forecasting and preventive measures. However, what we should focus on is the rapid and unpredictable evolution [3] of COVID-19 that intensified the challenge. The virus has shown an alarming ability to mutate into new variants, such as Delta and Alpha, each introducing additional complexities and posing unique threats. This swift evolution has led to unforeseen changes in the virus's behavior, making it crucial to adopt advanced technologies that can adapt to and understand these dynamic patterns.

Deep learning, and advanced models like Long Short-Term Memory (LSTM), have demonstrated remarkable efficacy in various medical fields [4], particularly in predicting and understanding the complexities of COVID-19 [5]. The unique architecture of LSTM models equips them to handle sequential data, irregular time intervals, and the inherent noise in COVID-19 datasets [6]. This makes them exceptionally well-suited for capturing both short-term fluctuations and long-term patterns in the data.

While current efforts focus on immediate response and control [7], it is equally crucial to prepare for the future return of the virus or the emergence of similar threats. By employing LSTM models for COVID-19 forecasting, the goal is not only to gain insights into the current pandemic but to build a

DOI: 10.54254/2753-8818/37/20240154

resilient model capable of adapting to different variations for future prevention and decision-making support in the face of evolving viral threats.

2. Dataset Information

2.1. Dataset Overview

Downloaded from the GitHub website [8], the "owid-covid-data" dataset is too massive to analyze. Thus, the data spanning from March 1, 2020, to April 30, 2022, providing a comprehensive set of 849 observations within the United States, capturing the evolution of the COVID-19 pandemic is used as the original dataset (Table1). After cleaning, the dirty dataset is divided into four clean datasets called "original variant", "alpha variant", "delta variant", and "mixed variants" based on four different periods only having two variables called "new_cases_smoothed_per_million" and "new_deaths_smoothed_per_million".

Table 1. Original dataset

#	Column	Non-Null Count	Dtype
0	iso_code	849 non-null	object
1	continent	849 non-null	object
2	location	849 non-null	object
3	date	849 non-null	int64
4	total_cases	832 non-null	float64
5	new_cases	849 non-null	int64
6	new_cases_smoothed	844 non-null	float64
7	total_deaths	792 non-null	float64
8	new_deaths	848 non-null	float64
9	new_deaths_smoothed	843 non-null	float64
10	total_cases_per_million	832 non-null	float64
11	new_cases_per_million	849 non-null	float64
12	new_cases_smoothed_per_million	844 non-null	float64
13	total_deaths_per_million	792 non-null	float64
14	new_deaths_per_million	848 non-null	float64
15	new_deaths_smoothed_per_million	843 non-null	float64

2.2. Significance of Dataset

The cleaned datasets serve as the foundation for LSTMs models, recording each day's new cases and new deaths which are perfect time series datasets for LSTMs models to learn, allowing users to train and evaluate its performance not only in forecasting COVID-19 trends but also prepare it to forecast the virus that is similar to the COVID-19 in the real world to help governments and scientists to make early prevention and better decision-making.

3. Data Cleaning and Preprocessing

Missing data in eight columns, including total_cases, new_cases_smoothed, total_deaths, new_deaths, new_deaths_smoothed, total_cases_per_million, new_deaths_per_million, new_deaths_per_million was removed to enhance the dataset's reliability (Table1). Unnecessary data was identified and removed to streamline the dataset for efficient model training. Duplicated data was eliminated to prevent redundancy and ensure the uniqueness of observations in the dataset. By calculating cases per million, variables were smoothed to ensure consistent scaling.

DOI: 10.54254/2753-8818/37/20240154

The original dataset is divided into four clean datasets called "original variant", "alpha variant", "delta variant", and "mixed variants" respectively based on different periods, which are "March 1, 2020 to December 29, 2020", "December 29, 2020, to September 21, 2021", "June 15, 2021 to April 30, 2022" and "March 1, 2020 to December 1, 2021". These four datasets only have two variables called "new_cases_smoothed_per_million" and "new_deaths_smoothed_per_million".

4. Methodology

4.1. Importing and Processing Data

The initial step involved importing the cleaned COVID-19 dataset, "original variant", "alpha variant", "delta variant", and "mixed variants" dataset for the original, alpha, delta and mixed variants relatively, utilizing the 'pd.read_excel()' function from Pandas library. To ensure the suitability of data for the LSTMs models, the 'MinMaxScaler()' from the 'sklearn. preprocessing' module was employed. This scaling technique transforms the data into a specified range, enhancing the model's training performance.

4.2. Data Arrangement and Transformations

For efficient data arrangement and transformations, the NumPy library played a crucial role. The 'numpy' module facilitated critical tasks such as reshaping the data to meet the LSTMs models' input requirements. Functions like 'model().data.reshape()', 'numpy()', 'test_y.numpy()', and 'np.arange()' were utilized to structure and transform the dataset appropriately.

4.3. Model Training

The PyTorch library, known for its versatility in deep learning, was employed for training the LSTMs models. Utilizing the 'torch' and 'nn' (neural network) modules, the architecture of the model was defined. This included using 'nn. LSTM()' to create LSTMs layers and 'nn.Linear()' to implement linear layers. The training process involved defining a loss function, specifically the Mean Squared Error (MSE) loss, using 'nn.MSELoss()'. Model training was executed with meticulous consideration for optimal hyper-parameters settings. Adam optimizer is also used to train this LSTMs model.

4.4. Model Evaluation

The evaluation of the trained LSTMs models was conducted using metrics from the scikit-learn library. Beyond the standard metrics like Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, and R-squared, the analysis extended to different variants of COVID-19. This evaluation provides insights into the model's performance across varied scenarios, including the original, Alpha, Delta and mixed variants.

4.5. Plotting

The visualization of the model's predictions and actual data was facilitated by the Matplotlib library. Functions such as 'plt.plot()', 'plt.legend()', 'plt.title()', 'plt.xlabel()', and 'plt.ylabel()' were employed to create clear and informative plots. These visualizations aided in presenting the predicted trends in a comprehensible manner, contributing to the overall interpretability of the model's outputs.

5. Results

5.1. Short-Term Prediction

5.1.1. Original Variant. The complete cycle for original COVID-19 is March 1, 2020, to December 29, 2020, and 150 observations are chosen as the experiment dataset. 80% of the chosen data is used as a training set to make 30-day predictions on the unseen data, which can be seen in Figure 2. Table 2 shows specific hyper-parameters for predicting original variant's new daily cases and deaths.

DOI: 10.54254/2753-8818/37/20240154

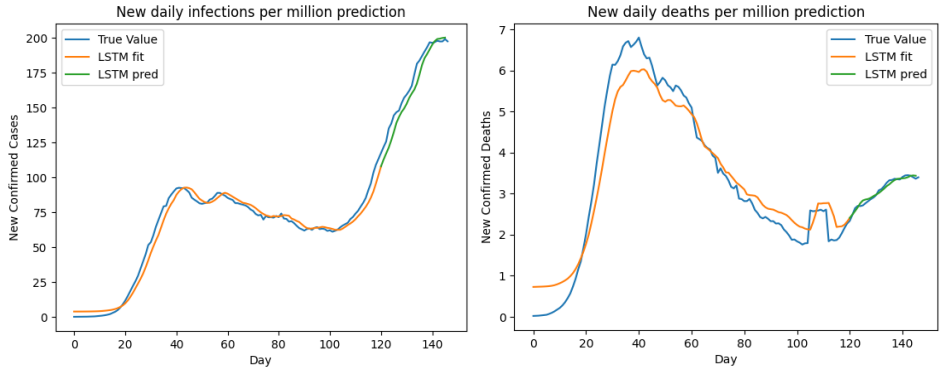

Figure 2. New daily cases and New daily deaths

Table 2. Hyper-parameters for original variant

Original	input size	hidden size	seq	epoch	learning rate	output size
cases	1	16	3	1000	0.0006	1
deaths	1	16	3	1000	0.001	1

5.1.2. Alpha Variant. The complete cycle for the alpha variant is December 29, 2020, to September 21, 2021, and 150 observations are chosen as the experiment dataset. 80% of the chosen data is used as a training set to make 30-day predictions on the unseen data, which can be seen in Figure 3. Table 3 shows specific hyper-parameters for predicting the alpha variant's new daily cases and deaths.

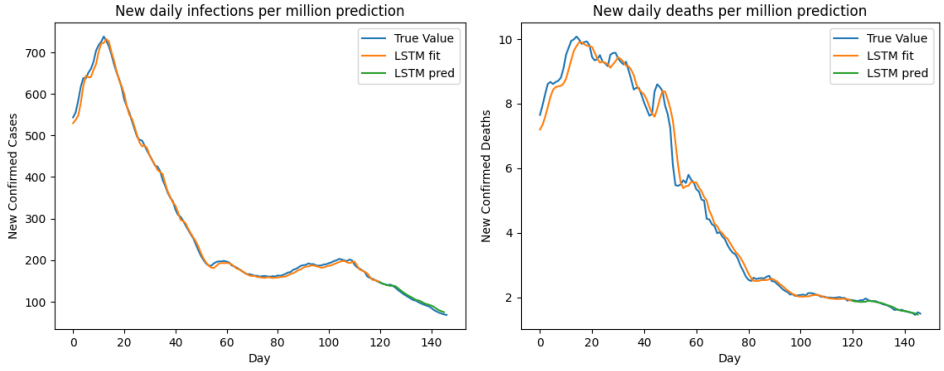

Figure 3. New daily cases and New daily deaths

Table 3. Hyper-parameters for alpha variant

Alpha	input size	hidden size	seq	epoch	learning rate	output size
cases	1	16	3	1000	0.005	1
deaths	1	16	3	1500	0.1	1

5.1.3. Delta Variant. The complete cycle for the delta variant is June 15, 2021, to April 30, 2022, and 150 observations are chosen as the experiment dataset. 80% of the chosen data is used as a training set to make 30-day predictions on the unseen data, which can be seen in Figure 4. Table 4 shows specific hyper-parameters for predicting delta variant's new daily cases and deaths.

The reason why making 30-day predictions is that governments and scientists may have no clues about what will happen in the first few months. Thus, it is normal to focus on a short period like one month to make quick decisions or early prevention.

DOI: 10.54254/2753-8818/37/20240154

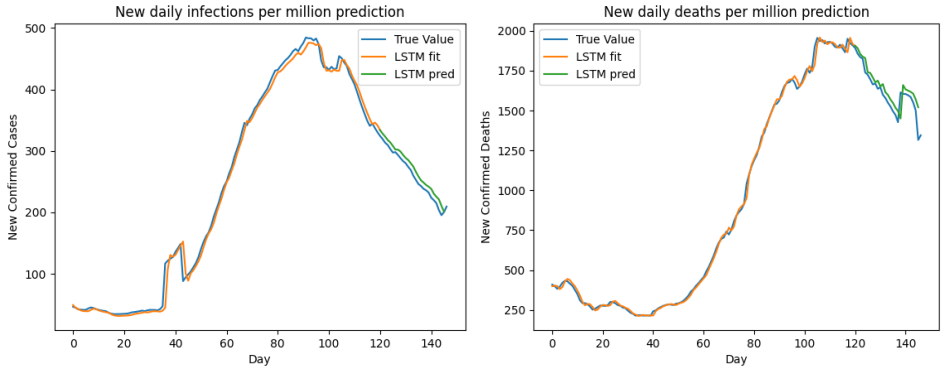

Figure 4. New daily cases and New daily deaths

Table 4. Hyper-parameters for delta variant

Delta	input size	hidden size	seq	epoch	learning rate	output size
cases	1	16	3	2000	0.005	1
deaths	1	16	3	2500	0.005	1

5.2. Long Term Prediction

The period for long-term prediction is March 1, 2020 to December 1, 2021 and 700 observations are chosen as the experiment dataset. This long-term prediction learns the mixed patterns of the original COVID-19 and alpha variant to forecast the delta variant. 85% of the chosen data is used as a training set to make a 100-day prediction on the unseen data, which can be seen in the Figure 5. Table 5 shows specific hyper-parameters for predicting mixed variants' new daily cases and deaths.

The prediction value is close to the true value, which provides evidence that different versions of variants do promote the model learning ability.

After learning nearly one year and a half of data, making a 100-day prediction can help the government and scientists determine whether the turning point is coming.

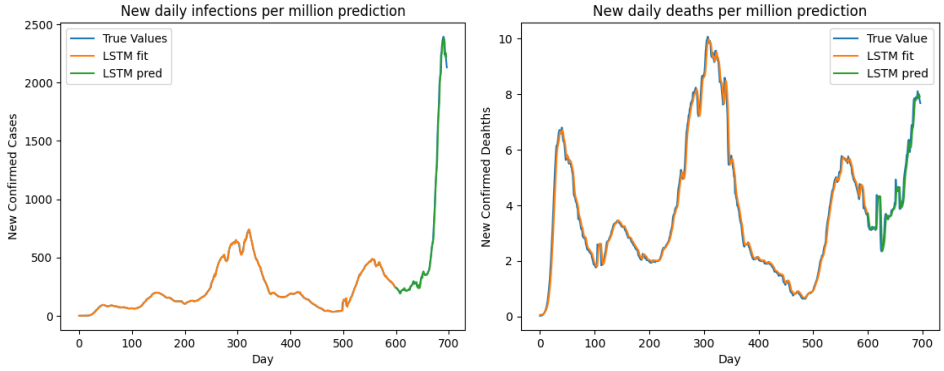

Figure 5. New daily cases and New daily deaths

Table 5. Hyper-parameters for mixed variants

Mixed	input size	hidden size	seq	epoch	learning rate	output size
cases	1	16	3	2000	0.006	1
deaths	1	16	3	2000	0.005	1

DOI: 10.54254/2753-8818/37/20240154

6. Model Evaluation

Table 6. Statistic values for new daily cases

Cases	MSE	RMSE	MAE	R2
Original variant	0.0001	0.0072	0.006	0.9791
Alpha variant	0.0000	0.0042	0.0037	0.9714
Delta variant	0.0001	0.0108	0.0102	0.9644
Mixed variants	0.0015	0.0388	0.0234	0.9973

Table 7. Statistic values for new daily deaths

Deaths	MSE	RMSE	MAE	R2
Original variant	0.0000	0.0001	0.0001	0.9689
Alpha variant	0.0000	0.0000	0.0000	0.9452
Delta variant	0.0036	0.0598	0.0432	0.8782
Mixed variants	0.0000	0.0004	0.0002	0.947

Reflecting a diminished average error magnitude, smaller values for MSE, RMSE, and MAE will signify a more precise prediction. Conversely, R2 thrives on higher values, indicating a model that fits the data well by explaining a larger proportion of the variance in the dependent variable. These evaluation methods collectively guide the assessment of a model's performance, allowing for a better understanding of its predictive accuracy and fitting capabilities [9]. Table 6 and Table 7 show the detail of model evaluation of each variant.

6.1. Original Variant

The LSTM model demonstrates excellent performance in predicting COVID-19 cases and deaths. The low MSE (0.0001) and RMSE (0.0072) values show a close match between predicted and actual values. A small MAE (0.0060) further confirms the accuracy of predictions. The high R-squared value of 0.9791 indicates that the model explains a significant portion of the variance in the data.

For death predictions, the model exhibits exceptional accuracy with extremely low MSE, RMSE, and MAE values (close to zero). The R-squared value of 0.9689 underscores the model's high explanatory power, affirming its reliability in capturing the details of COVID-19 deaths.

6.2. Alpha Variant

The model maintains exceptional accuracy with minimal errors (MSE, RMSE close to zero) and high R2 values for both case predictions and deaths. For short-term predictions of COVID-19 cases associated with the Alpha variant, low MSE (0.0000) and RMSE (0.0042) values indicate minimal errors. The small MAE (0.0037) emphasizes the precision of short-term predictions. The high R-squared value of 0.9714 confirms a strong correlation, showing the model's ability to capture immediate trends associated with the Alpha variant accurately.

In forecasting deaths related to the Alpha variant for the short term, the LSTM model continues to exhibit exceptional performance. Low MSE, RMSE, and MAE values (close to zero) indicate precise predictions of fatality trends. The strong R2 value reflects the model's ability to provide valuable insights into mortality trends associated with the Alpha variant over a brief time horizon.

6.3. Delta Variant

The LSTM model for the Delta variant shows notable performance in predicting both cases and deaths. Low MSE (0.0001) and RMSE (0.0108) values indicate a close match between predicted and actual case values. The MAE (0.0102) suggests reasonable accuracy, and the high R2 value of 0.9644 indicates a strong correlation between the model's predictions and observed data.

DOI: 10.54254/2753-8818/37/20240154

For forecasting deaths related to the Delta variant, the model performs satisfactorily. MSE (0.0036), RMSE (0.0598), and MAE (0.0432) values indicate reasonable accuracy in predicting fatality trends. The R2 value of 0.8782 reflects substantial explanatory power, highlighting the model's effectiveness in capturing the variance in death data for the Delta variant.

Although the model's precision in predicting deaths for the Delta variant is slightly lower than for cases, the MSE, RMSE, and MAE values remain at acceptable levels. The strong R2 value indicates a continued ability of the LSTM model to provide valuable insights into mortality trends associated with the Delta variant.

6.4. Mixed Variant

The LSTM model for the Mixed variant demonstrates robust performance, providing accurate long-term predictions for both cases and deaths. Low MSE (0.0015) and RMSE (0.0388) values indicate a close match between predicted and actual cases over an extended period. The small MAE (0.0234) underscores the accuracy of long-term predictions. The high R-squared (R2) value of 0.9973 signifies a strong correlation, highlighting the model's proficiency in capturing prolonged trends associated with the Mixed variant.

In forecasting deaths for the Mixed variant in the long term, the LSTM model performs exceptionally well. Precise predictions are indicated by low MSE (0.0000), RMSE (0.0004), and MAE (0.0002) values. The R2 value of 0.9470 reflects substantial explanatory power, emphasizing the model's effectiveness in capturing the variance in death data for the Mixed variant over an extended period.

In summary, the LSTM model consistently shows remarkable accuracy in predicting COVID-19 cases and deaths across different variants, with low error values and high R2 values validating its reliability in providing valuable insights.

7. Discussion

The LSTMs excel in capturing short and long-term patterns in sequential data, irregular time intervals, and the ability to handle noisy data but the LSTMs are sensitive to the change of training data size. Thus, it is important to figure out the specific percentage of data that is used as training set to avoid over-fitting. For short-term prediction, 80% of observations are used as a training set. For long-term prediction, 85% of observations are used as a training set.

During analysis, all the missing data are removed for they only occur at the beginning of COVID-19, which may be the reason that people are not paying enough attention to the outbreak of COVID-19 to record daily data. Thus, the missing data was removed to enhance the dataset's reliability. Due to the character that the virus has a latent period [10], unnecessary values like 0 values occur in the 'new_deaths' variable in the first few months. Learning a bunch of 0 values will negatively affect the reliability of the model to catch the whole pattern of data. Thus, the unnecessary data is removed for a cleaner and more reliable dataset. The reason why this study just uses the data between 2020.3.1 to 2022.4.30 is that these data cover complete cycles [11] for three different versions of variants, that are recorded by the CDC, from the initial appearance to no longer being predominant [12].

8. Conclusion

In the study, Long Short-Term Memory (LSTM) models are used to make predictions about COVID-19. The results demonstrate the model's efficacy in providing accurate short and long-term forecasts, even when confronted with different virus variants. The different patterns of variants help the model to understand the whole data better and make more precise predictions, which shows great potential to fit the real-world situation and forecast another variant or virus similar to COVID-19.

While the LSTM model proved effective, it is essential to acknowledge its limitations. In future research, more advanced measures, such as Bayesian optimization, should be employed to search for optimal hyper-parameters. This approach can enhance the model's performance and stability by automating the process of parameter tuning, leading to improved forecasting accuracy. Future research endeavors should include more countries or regions to enhance the generalization to be applied to

DOI: 10.54254/2753-8818/37/20240154

various circumstances and explore the integration of more advanced deep learning architectures and ensemble methods to further enhance prediction accuracy. Additionally, the inclusion of real-time data and continuous model updates can contribute to the model's responsiveness to emerging trends and evolving virus characteristics.

References

[1] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020 Apr;5(4):536-544.

[2] Bari A, Khubchandani A, Wang J, Heymann M, Coffee M. COVID-19 early-alert signals using human behavior alternative data. Soc Netw Anal Min. 2021;11(1):18.

[3] Fernandes Q, Inchakalody VP, Merhi M, et al. Emerging COVID-19 variants and their impact on SARS-CoV-2 diagnosis, therapeutics and vaccines. Ann Med. 2022 Dec;54(1):524-540.

[4] Heidari A, Jafari Navimipour N, Unal M, et al. The COVID-19 epidemic analysis and diagnosis using deep learning: A systematic literature review and future directions. Comput Biol Med. 2022, Feb; 141:105141.

[5] Ghany KKA, Zawbaa HM, Sabri HM. COVID-19 prediction using LSTM algorithm: GCC case study. Inform Med Unlocked. 2021; 23: 100566.

[6] Sepp Hochreiter, Jürgen Schmidhuber; Long Short-Term Memory. Neural Comput1997; 9 (8): 1735–1780.

[7] Adamidi ES, Mitsis K, Nikita KS. Artificial intelligence in clinical care amidst COVID-19 pandemic: A systematic review. Comput Struct Biotechnol J. 2021; 19:2833-2850.

[8] Our World in Data. (2023, Nov,17). COVID-19 Data. https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv

[9] Rania Echrigui, Mhamed Hamiche. Optimizing LSTM Models for EUR/USD Prediction in the context of reducing energy consumption: An Analysis of Mean Squared Error, Mean Absolute Error and R-Squared E3S Web Conf., 412 (2023) 01069:1-8

[10] Iqbal M, Al-Obeidat F, Maqbool F, et al. COVID-19 Patient Count Prediction Using LSTM. IEEE Trans Comput Soc Syst. 2021 Feb 19;8(4):974-981.

[11] Centres for Disease Control and Prevention. (2023, Dec,3). SARS-CoV-2 Variant Classifications and Definitions. https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html

[12] Abbasimehr H, Paki R. Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization. Chaos Solitons Fractals. 2021 Jan, 142:110511.