PREDICTING COVID-19 TEST POSITIVITY RATE IN THE USA USING VANILLA LSTM, STACKED LSTM, AND BI-LSTM

O'la Hmoud Al-laymoun
Mu'tah University, Jordan
ola.allaymoun@mutah.edu.jo

ABSTRACT: *The COVID-19 pandemic, combined with uncertainties about the efficacy of current vaccines and medical procedures used in the fight against coronavirus, has disrupted all aspects of life worldwide, necessitating the urgent need for robust forecasts to assist humanity in bringing this global and unprecedented crisis under control. Using Google Trends data, this study employs three long short-term memory (LSTM) models: Stacked LSTM, Bi-LSTM, and Vanilla LSTM, to predict the test positivity rate of COVID-19 in the United States. The findings of the current study were promising, as the three models could closely capture the trend of the positive rate of infections over the study duration, which was consistent with the findings of other studies.*

Key words: COVID-19 prediction, test positivity rate, time series, LSTM

1. INTRODUCTION

Coronavirus disease 2019 (COVID-19) began in December 2019 in Hubei Province, China, and has since spread globally [1], disrupting not only economies but all aspects of human life in both developed and developing countries [2]. The World Health Organization (WHO) declared it a pandemic on March 11, 2020 [3]. With approximately 176 million cases and 3801 thousand deaths worldwide as of June 4, 2021, and the appearance of multiple mutant versions of the virus in different countries, the pandemic has created an urgent need for a global response and recovery strategy. Consequently, countries, organizations, and individuals worldwide have been devoting their resources and efforts to fight the newly discovered coronavirus and bring the situation under control. However, it appears that COVID-19's quick swamp and profound impact have put all stakeholders to the test. The healthcare system, for example, is currently under tremendous strain as the number of infected cases and deaths rises and alarmingly exceeds current capacities [4].

Indeed, this unique situation with high uncertainties has prompted a constantly adapting global response [5] as new information about the illness and its consequences emerges, prompting countries to implement a variety of measures to prevent, or at least mitigate, the spread of the coronavirus and promote healthy behaviors. To name a few, these measures include social isolation, lockdown, the use of facemasks and sanitizers, self-quarantine, travel restrictions, contact tracing [6], extensive gathering bans, and vaccines. The pandemic has also created an urgent need for reliable forecasts to aid in crisis management, policy development, decision making, and resource planning [5, 7]. Singh et al. [4] argue that using mathematical-epidemiological models to predict COVID-19 outbreaks and the factors influencing them is critical for preparing healthcare systems around the world to deal with the pandemic. Such models, for instance, can help in forecasting the potential number of infections, hospitalized patients, and deaths, among other things. This is especially true for real-time short-term estimates, which could help predict possible scenarios in the near future, resulting in better allocation of medical resources and equipment like ventilators [1, 6]. Accordingly, this study employs three variants of long short-term memory (LSTM) models, namely: Stacked LSTM, Bi-LSTM, and Vanilla LSTM, to contribute to the growing literature on COVID-19 prediction using time series analysis, which could provide decision-makers dealing with the pandemic with critical tools for facing this global crisis.

Furthermore, according to Santosh[6], the state-of-the-art forecasting models of the COVID-19 outbreak failed to consider some essential and unprecedented factors, such as daily test rates, hospital capacities, and demographics, among others. This study employs the COVID-19 test positivity rate, defined as the percentage of individuals who "tested positive in a given day" [8], rather than the absolute daily numbers of positive cases discovered. In the case of a pandemic, such as COVID-19, where it is impractical and impossible to test the entire population, the daily positive rate of infections may be a more reliable indicator of the disease's actual spread pattern than the absolute number of positive cases, which varies depending on the total number of tests performed. Indeed, a recent study discovered that the COVID-19 test positivity rate is a strong predictor of the number of critical coronavirus cases up to 12 days in advance [8].

Furthermore, this study uses Google Trends data on the top search terms on Google's popular search engine during the study period to predict the test positivity rate of COVID-19 spread in the United States (USA) during this unprecedented pandemic. Several studies have used Google Trends data to investigate various aspects of coronavirus disease, including predicting its incidence [3], people's interest in information about the loss of smell symptom [9], smoking cessation [10], and well-being-related topics in the United States and Europe during COVID-19 lockdowns [11]. The Google Trends website provides the most recent insights on people's actual search behavior around the world or in a specific country on various online channels, including YouTube and Google News, making it a handy online tool for researching trends in people's interests. Indeed, Google Trends data represents a promising unconventional tool for forecasting coronavirus outbreaks, as it has been successfully used in other contexts, such as influenza outbreak prediction [12].

2. LITERATURE REVIEW

2.1 Long Short-Term Memory (LSTM)

Since the outbreak, a growing body of research has been conducted in many disciplines to shed light on this unprecedented situation, and the field of data science is no exception, as researchers in this field are racing against the clock to develop COVID-19 prediction models. For example, Ray et al. [5] used ensemble forecasts to predict the total number of COVID 19 deaths in the United States.

Furthermore, Singh et al. [4] investigated the pandemic in India using the basic Susceptible, Infected, and Recovered (SIR) model and the Holt-Winters model. Roosa et al. [1] also used the Richards growth model, the generalized logistic growth model, and a sub-epidemic wave model to generate short-term forecasts of cumulative cases in Guangdong and Zhejiang, China. Another study by Magesh et al. [2] employed the SIR mathematical model to classify COVID-19 cases into three categories: suspected, infected, and recovered cases, as well as the recurrent neural network (RNN) with LSTM to predict COVID-19 confirmed cases.

LSTM has emerged as an appropriate forecasting technique to predict trends in COVID-19 data, which is sequential in nature during the current pandemic. In essence, the strength of LSTM networks stems from their ability to generate cutting-edge results by addressing the limitations of traditional time series prediction models and techniques in dealing with non-linear problems, as is the case with COVID-19 datasets [13]. Hochreiter and Schmidhuber [14] proposed LSTM as a particular type of RNN with a "gradient-based learning algorithm" to solve the limitation of conventional RNNs in the form of vanishing or exploding error signals while back-propagating through time [14]. LSTM is made up of memory blocks [15]. Each memory block is composed of memory cells that remember “the temporal state of the network through self-connections" as well as input, forget, and output gates that control information flow [15]. The following equations represent the LSTM gates as stated by Chimmula and Zhang [13]:

$$J_t = \text{sigmoid}(w_J [h_{t-1}, k_t] + b_J) \quad (1)$$
$$G_t = \text{sigmoid}(w_G [h_{t-1}, k_t] + b_G) \quad (2)$$
$$P_t = \text{sigmoid}(w_P [h_{t-1}, k_t] + b_P) \quad (3)$$

Where: J_t = function of input gate
G_t = function of forget gate
P_t = function of output gate
W_x = coefficients of neurons at gate (x)
H_{t-1} = result from previous time step
k_t = input to the current function at time-step t
b_x = bias of neurons at gate (x)

Researchers are increasingly interested in using LSTM to aid in response to the coronavirus pandemic. For example, Arora, Kumar, and Panigrahi [16] used a variety of LSTM variants, including deep LSTM, convolutional LSTM, and bidirectional LSTM (Bi-LSTM), to forecast new positive daily and weekly cases in 32 locations across India. The proposed model had high accuracies, with mean absolute percentage error (MAPE) of less than 3% for daily forecasts and less than 8% for weekly predictions. Another study by Chimmula and Zhang [13] used LSTM to predict COVID-19 cases in Canada and found that the growth rate of transmission in Canada was linear, as opposed to the exponential trends in Italy and the United States. The short-term predictions of the proposed LSTM model were 93.4 percent accurate with root mean square error (RMSE) of 34.83, while the long-term predictions were 92.67 percent accurate with RMSE of 45.7. A third study by Shahid, Zameer, and Muneeb [17] compared the performance of five time series analysis models, namely: autoregressive integrated moving average (ARIMA), Gated recurrent unit (GRU), support vector regression (SVR), LSTM, and Bi-LSTM, in predicting the number of positive, recovered, and death cases in ten countries: Brazil, China, India, Israel, Italy, Russia, Spain, UK, and the USA. To compare the performance of the five models, three performance indices were used: mean absolute error (MAE), RMSE, and r2 score. The models were ordered from best to worst in terms of performance: Bi-LSTM, LSTM, GRU, SVR, and ARIMA. Last but not least, Achterberg, Prasse, Ma, Trajanovski, Kitsak, and Van Mieghem [18] proposed a Network Inference-based Prediction Algorithm (NIPA) and compared its accuracy in forecasting the daily cases of COVID-19 in Hubei, China, and the Netherlands up to six days against the accuracy of LSTM, Sigmoid curve variants, and some modified forms of NIPA. The researchers used the Symmetric Mean Absolute Percentage Error (sMAPE) to compare the prediction accuracy of the various forecasting models and discovered that the basic NIPA outperformed the rest of the algorithms.

2.2 Google Trends

According to Ayyoubzadeh et al. [7], search engines provide valuable data from populations, which may help analyze epidemics. Google Trends is a Google website that analyzes the popularity of search terms on Google over a given time. "Google trends values are not absolute numbers of searches; rather, they represent the number of searches normalized between 100 and 0, where 100 represents the peak of search frequency, 50 indicates that the keyword is half as popular, and 0 indicates that there is insufficient data to calculate the search frequency"[7, 11]. Google trends searchers have been used successfully in predicting the spread of influenza. For example, Zhang et al. [12] used Google Trends and ambient temperature data to predict seasonal influenza outbreaks in Brisbane and the Gold Coast between January 1, 2011, and December 31, 2016. The researchers used the time-series cross correlation analysis, temporal risk analysis, regression tree model, and seasonal autoregressive integrated moving average (SARMIA) model. During the study period, the weekly cases of influenza infections were significantly correlated with Google trends with a lag of 1-7 weeks in both areas. Another study by Liu et al. [15] used LSTM to predict influenza trends in six cities in Georgia state, USA. The study employed data from virologic surveillance and influenza geographic spread and Google Trends, air pollution, and climatic data collected between 2012 and 2018. The researchers discovered that increasing the sample size and including other variables in addition to Google Trends data improved the model's prediction performance.

In coronavirus prediction, Ayyoubzadeh et al. [7] used the daily incidents in Iran from February 15, 2020, to March 18, 2020, and Google Trends searchers to apply Linear regression and LSTM models to predict the number of daily positive COVID-19 cases in the country. The results showed that the RMSE for the LSTM was 27.187 ± 20.705, and for the linear regression model was around 7.562 ± 6.492. The researchers discovered that the following factors were the most significant: previous day incidence of COVID-19, previous day frequency of searches for: "corona, covid-19, coronavirus, antiseptic buying, handwashing, hand sanitizer,

and antiseptic topic". Another study by Brodeur *et al*.[11] used Google Trends to investigate the potential effects of COVID-19 and associated lockdowns on people's mental health in America and Europe. The researchers used a regression discontinuity design and difference-in-differences to conduct their study. The findings revealed a significant increase in search interest for boredom, loneliness, worry, and sadness in both countries. On the contrary, there has been a decrease in interest in suicide, stress, and divorce.

3. RESEARCH METHODOLOGY AND ANALYSIS

3.1 Data Collection

For this study, the Google Trends website was searched for the most frequently searched terms related to the COVID-19 pandemic on the web in the United States. A variety of terms associated with the current pandemic were investigated, and the most frequently searched terms were used in this research paper: Corona, coronavirus, COVID-19, and symptoms. Generally speaking, the trend in people's searches for these terms decreased over time, except for the search term "symptoms," which increased, albeit at a slow rate.

Furthermore, the daily time-series data of the number of positive COVID-19 cases and the number of daily tests in the United States were extracted from the Worldometers.info website, which provides real-time COVID-19 statistics worldwide. The two variables were used to calculate the daily COVID-19 test positive rate in the United States which was normalized to be between 0-100 to be consistent with the scale used for Google Trends data. The data was collected between June 1, 2020 and March 23, 2021, yielding 263 data records.

3.2 Research methodology

As part of the data preprocessing stage, a flat file containing all the collected data from the two mentioned sources was created to prepare the data for time series data analysis. The dataset was then transformed into a supervised learning problem, with the test positivity rate and Google Trends data from the previous two days used to predict the test positivity rate of COVID-19 in the United States at day(x). Several time lags were investigated, and the one with the lowest root mean squared error (RMSE), which represents the square root of the difference between the actual and predicted values, was chosen, resulting in a two-day time lag.

The dataset was divided into two parts: training and testing. The first 190 days of data were used to train three LSTM algorithms: Vanilla LSTM, Stacked LSTM, and Bi-LSTM. Only in the training set was the min/max scaler used to normalize the data to be between 0 and 100. The remaining days were part of the test set. The performance of the models was assessed by comparing their RMSE values.

Python 3.9 was used to conduct the time series analysis. The input size, which corresponds to the number of features used, was five, representing the Google trends data of four search terms (corona, coronavirus, COVID-19, symptoms) and the COVID-19 test positivity rate during the study period. The output size was set to 1 to represent the predicted COVID-19 test positive rate two days ahead. In the stacked LSTM model, the number of the hidden layers was set to two. Each hidden layer in all models had 100 neurons, and each of the three LSTM models was trained for 100 epochs with a batch size of 72. Finally, the Adam optimizer was the optimizer function. It is worth noting that the number of hidden layers, neurons, epochs, and batch size were all determined by trial and error.

4. RESULTS AND DISCUSSION

This study responds to today's critical need for forecasting coronavirus spread rate, not only to improve the healthcare sector's readiness and ability to face the epidemic and avoid deaths [19] successfully, but also to maximize the flexibility of all sectors in their response and accommodation to the current, fluctuating epidemiological situation. This study specifically aimed to predict the test positivity rate of COVID-19 incidents in the United States using Google Trends data. Three LSTM model variants, namely Vanilla LSTM, Stacked LSTM, and Bi-LSTM, were applied to a dataset containing Google trends data on four search terms: corona, coronavirus, COVID-19, and symptoms, in addition to the positive rate of COVID-19 tests in the United States.

The actual test positivity rate COVID-19 is depicted in Figure 1- Figure 6 below by the orange lines, while the predicted values are depicted in blue. As shown below, the LSTM time series analysis results show that the predictions of the three algorithms were able to closely capture the trend in the data, albeit not very accurately. The RMSE values for Stacked LSTM, Vanilla LSTM, and Bidirectional LSTM were 0.040, 0.037, and 0.036, respectively. In general, those findings are consistent with previous research that examined and reported the usefulness of employing LSTM models in studying health pandemics and their evolution over time [7, 15, 20], particularly when dealing with small datasets [21]. Indeed, there is a surge in research using artificial intelligence models, such as LSTM, to surveillance the COVID-19 epidemic, exploring its potential short- and long-term impacts, and predicting the future mutations and evolution behavior of the virus [20, 21].

Furthermore, the current study's findings provide academics and practitioners with insights into the important role of online search queries conducted on Google and other search engines as a nonconventional real data source on the trends and topics people are thinking about, particularly in novel and urgent situations with high uncertainties, such as pandemics, where humankind is still gradually accumulating and developing knowledge about the situation while being under pressure to find quick and reasonable answers and solutions to keep it under

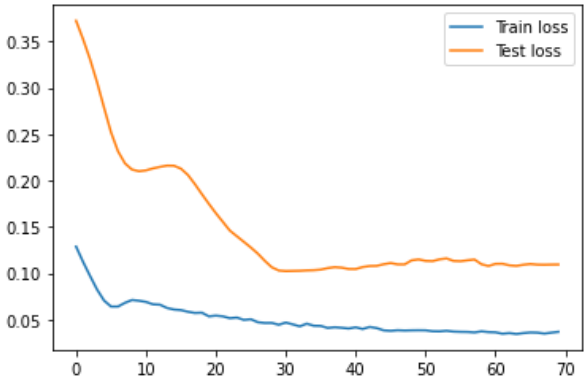

Figure 1: Training and testing loss of Stacked LSTM

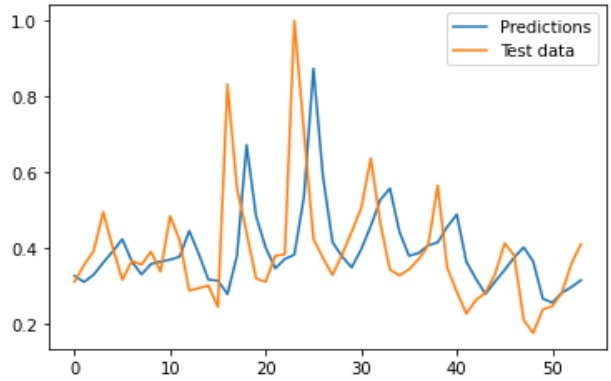

Figure 2: Actual and predicted positive test rate of COVID-19 using Stacked LSTM

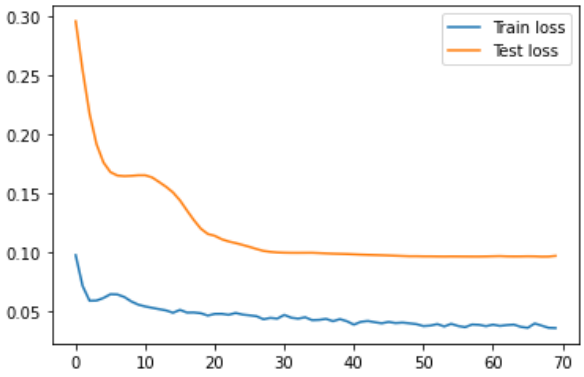

Figure 3: Training and testing loss of Vanilla LSTM

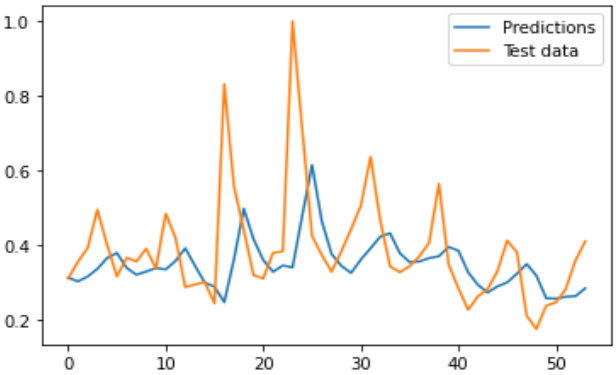

Figure 4: Actual and predicted positive test rate of COVID-19 using Vanilla LSTM

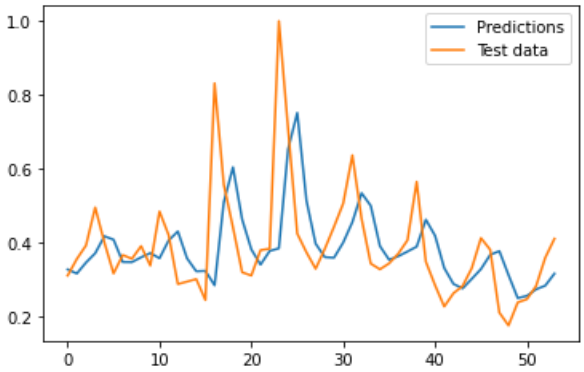

Figure 5: Training and testing loss of Bi- LSTM

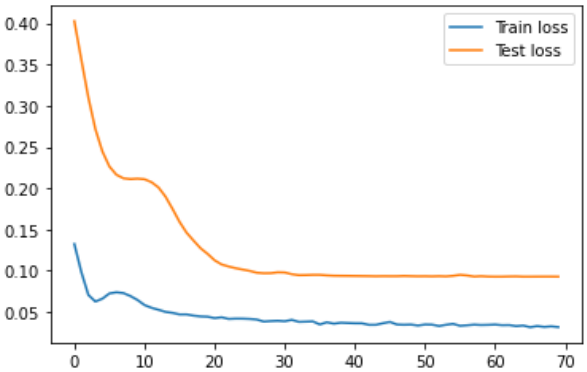

Figure 6: Actual and predicted positive test rate of COVID-19 using Vanilla LSTM

control. The findings of this study, as well as other studies in this context, may pave the way for new avenues of research that use the wealth of data on search engines on people's online behaviors to further investigate individuals' awareness of the disease and its symptoms, among other important topics. For example, the data could help us predict which home remedies people use to treat their coronavirus infections, their motivation to accept or refuse the vaccine, any health problems or issues that arise after taking the vaccine, and mental health issues related to virus fear, to name a few.

5. CONCLUSION AND FUTURE WORK

This study used Stacked LSTM, Bi-LSTM, and Vanilla LSTM to predict the positive test rate of COVID-19 in the United States using Google Trends data. The results were promising because the three models could capture the desired trend throughout the study closely. Thus, as indicated by the current research findings, both Google Trends data and LSTM could help stakeholders anticipate coronavirus outbreaks and track their spread. This is especially true as increasing amounts of data related to this disease accumulate over time; there is still plenty of room for improving the

performance of the prediction models by, for example, using larger sample sizes, taking into account the timing of some critical events such as imposing or ending closures and travel bans and incorporating other valuable data such as vaccination rates.

6. REFERENCES

[1] K. Roosa *et al.*, "Short-term forecasts of the COVID-19 epidemic in Guangdong and Zhejiang, China: February 13–23, 2020," *J. Clin. Med.*, vol. 9, no. 2, p. 596, 2020.

[2] S. Magesh, V. R. Niveditha, P. S. Rajakumar, and L. Natrayan, "Pervasive computing in the context of COVID-19 prediction with AI-based algorithms," *Int. J. Pervasive Comput. Commun.*, 2020.

[3] Y. Ortiz-Martínez, J. E. Garcia-Robled, D. L. Vásquez-Castañeda, D. K. Bonilla-Aldana, and A. J. Rodriguez-Morales, "Can Google® trends predict COVID-19 incidence and help preparedness? The situation in Colombia," *Travel Med. Infect. Dis.*, 2020.

[4] R. K. Singh *et al.*, "Short-term statistical forecasts of COVID-19 infections in India," *Ieee Access*, vol. 8, pp. 186932–186938, 2020.

[5] E. L. Ray *et al.*, "Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US," *MedRXiv*, 2020.

[6] K. C. Santosh, "COVID-19 prediction models and unexploited data," *J. Med. Syst.*, vol. 44, no. 9, pp. 1–4, 2020.

[7] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and P. S RNK, "COVID-19 incidence using Google Trends and data mining techniques: A pilot study in Iran," *JMIR Public Heal. Surveill*, 2020.

[8] L. Fenga and M. Gaspari, "Predictive capacity of covid-19 test positivity rate," *Sensors*, vol. 21, no. 7, p. 2435, 2021.

[9] A. Walker, C. Hopkins, and P. Surda, "Use of Google Trends to investigate loss-of-smell–related searches during the COVID-19 outbreak," in *International forum of allergy & rhinology*, 2020, vol. 10, no. 7, pp. 839–847.

[10] C. Heerfordt and I. M. Heerfordt, "Has there been an increased interest in smoking cessation during the first months of the COVID-19 pandemic? A Google Trends study," *Public Health*, vol. 183, p. 6, 2020.

[11] A. Brodeur, A. E. Clark, S. Fleche, and N. Powdthavee, "COVID-19, lockdowns and well-being: Evidence from Google Trends," *J. Public Econ.*, vol. 193, p. 104346, 2021.

[12] Y. Zhang, H. Bambrick, K. Mengersen, S. Tong, and W. Hu, "Using Google Trends and ambient temperature to predict seasonal influenza outbreaks," *Environ. Int.*, vol. 117, pp. 284–291, 2018.

[13] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons & Fractals*, p. 109864, 2020.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] L. Liu, M. Han, Y. Zhou, and Y. Wang, "Lstm recurrent neural net works for influenza trends prediction," in *International Symposium on Bioinformatics Research and Applications*, 2018, pp. 259–264.

[16] P. Arora, H. Kumar, and B. K. Panigrahi, "Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India," *Chaos, Solitons & Fractals*, vol. 139, p. 110017, 2020.

[17] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM," *Chaos, Solitons & Fractals*, vol. 140, p. 110212, 2020.

[18] M. A. Achterberg, B. Prasse, L. Ma, S. Trajanovski, M. Kitsak, and P. Van Mieghem, "Comparing the accuracy of several network-based COVID-19 prediction algorithms," *Int. J. Forecast.*, 2020.

[19] N. Kumar and S. Susan, "Covid-19 pandemic prediction using time series forecasting models," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–7.

[20] A. M. A. Haimed, T. Saba, A. Albasha, A. Rehman, and M. Kolivand, "Viral reverse engineering using Artificial Intelligence and big data COVID-19 infection with Long Short-term Memory (LSTM)," *Environ. Technol. Innov.*, vol. 22, p. 101531, 2021.

[21] Q.-V. Pham, D. C. Nguyen, T. Huynh-The, W.-J. Hwang, and P. N. Pathirana, "Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts," *arXiv Prepr. arXiv2107.14040*, 2021.