Proceedings of the 2nd International Conference on Biological Engineering and Medical Science (ICBioMed 2022), Part II
DOI: 10.54254/2753-8818/4/20220607

Time-series and Machine Learning Scenarios for COVID-19 Infection Prediction

Yuze Du[1, a,†] Zixuan Huang[2, b*,†] and Donglizhen Shi[3, c,†]

[1]Department of biomedical science, Monash University, Melbourne, 3168, Australia
[2]Jinan University-University of Birmingham Joint Institute, Jinan University, Guangzhou, 510000, China
[3]International College Beijing, China Agricultural University, Beijing,100083, China/ Ferguson College of Agriculture, Oklahoma State University, Stillwater, 74078, USA

[a] yduu0003@student.monash.edu
[b*]corresponding author: huangzixuan@stu2020.jnu.edu.cn
[c] murph.shi@okstate.edu
†These authors contributed equally

Abstract. COVID-19 is an infectious respiratory disease spread around the world. In past three years, more than 600 million people had been impacted and many countries healthcare system is taking a huge hit. The economy is affected by many epidemic policies such as quarantine policy, social distance, etc. Many scientists use different mathematical models to forecast trend of the confirmed cases quantity of epidemics. This article topic is mainly about the comparisons of different models for COVID-19 predictions. This paper focuses the different method based on COVID-19 infection prediction. We show the different methods' experiment results by many metrics but not under same datasets. Many methods are proposed, but few people discuss the time properties and interpretability of model. To be specific, this paper selects ARIMA, SVM and LSTM as target to offer the comparison between the model. By studying and analyzing the accuracy of different models for predicting COVID-19, we can find better models for predicting COVID-19 development. Eventually, we can find more suitable methods to control the development of COVID-19. These results shed light on guiding further exploration of models selection for COVID-19 infection prediction.

Keywords: COVID-19 prediction, machine learning, ARIMA, LSTM Model.

1. Introduction

COVID-19 is one of the deadliest and highly contagious diseases which caused by severe acute respiratory syndrome [1]. It primarily transmitted by exposure to the infectious respiratory droplets and particles. Due to its mode of transmission, quarantine, wear mask, maintain social distance and regularly disinfection these are all effective epidemic prevention policy. The global economy had knock-on effects because of the prolonged lockdown and many people loss job. Coronavirus continues to wreak havoc and virus has many different types of mutated variants leads the second or third waves outbreaks in many countries [1]. Since covid-19 is global health emergency, vaccine development time is limited and didn't process the lengthy clinical trial, many people not trust the vaccination

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/4/20220607

safety and worries about unknown side effects. Major types of COVID-19 vaccine include mRNA, inactivated vaccine, viral vector vaccine etc. Different types of vaccines might have different side effects based on people past medical history, age, family history and other factors. Even though the COVID-19 vaccine development and the robust global mass vaccination efforts, the mutated virus will limit the protection.

Contemporarily, COVID-19 has swept globally, many data scientists decided to use different models for COVID-19 infection predictions. Real-time prediction is important for the COVID-19 prevention and control. Many reports about long-lasting coronavirus are rising but people don't know about the prevalence, risk factor, and predict the disease course in early stage. The whole world is interested in the mechanism of COVID-19 development and spread. In this article, we will mainly talk about three models for COVID-19 infection predictions. Lots of previous studies have proposed various models to predict infection confirm cases. To be specific, some research uses public data to study COVID pandemic properties in order to develop hybrid model [2]. This model composed by two parts, eliminating variations, and improving accuracy as well as performances of the approaches. Finally, this fantastic model had been proved to accurately predict COVID patient infection, recover and death in order to help government policy makers to implement certain issues [2].

In order to use different models to predict COVID-19 confirmed cases, we aim to separate into three sections. The reminder of the paper will be presented as following. The Sec. 2 will offer a description of the situation and cases for COVID-19. The Sec. 3 will mainly be talking about ARIMA model based on the theory and actual cases prediction in different countries during the pandemic. The other two section (i.e., the Sec. 4 and Sec. 5) will follow by LTSM and machine learning this order. The Sec. 6 will give discussion for limitations for current state-of-art approaches and scenarios as well as give the proposal for further study. Eventually, a summary of the whole study will be demonstrated in Sec. 7.

2. Description of COVID-19

As more and more people become infected after the outbreak of COVID-19, scholars have begun to study the factors that influence the spread of COVID-19. However, we have not yet determined the source of transmission of COVID-19. Analysis of early cases speculates that the Wuhan Huanan Seafood Wholesale Market may be a source of transmission of the virus in this outbreak. Nevertheless, another study analyzing genomic data from 93 COVID-19 specimens found that early case sample genotypes may have come from outside the Huanan Seafood Wholesale Market. It is speculated that the Huanan Seafood Wholesale Market was not the only source of transmission [3]. In addition, although bats are likely hosts of the virus, a study found that pangolins may also be intermediate hosts [4]. Thus, the location of the source of infection is unknown, creating some uncertainty for the prevention and control of the disease. According to Li et al., saliva is an important factor in the transmission of the virus. The virus is transmitted from person to person mainly through droplet transmission such as coughing and sneezing [5]. Droplet transmission allows for widespread and rapid spread of the virus from person to person. The relationship between confirmed cases and AQI was investigated and evaluated in terms of collecting data from 33 sites in China [6].

3. ARIMA

The main goal of this ARIMA model is to reduce difference between observed value and model produced value which close to zero [7, 8]. The use of ARIMA model methodology contains three phases. Time series analysis ARIMA can be used as an alternative way to improve disease managements. It also can be used to manage infectious and non-infectious disease. China use ARIMA model to report more than 90% of hemorrhagic fever with kidney syndrome cases [8].

DOI: 10.54254/2753-8818/4/20220607

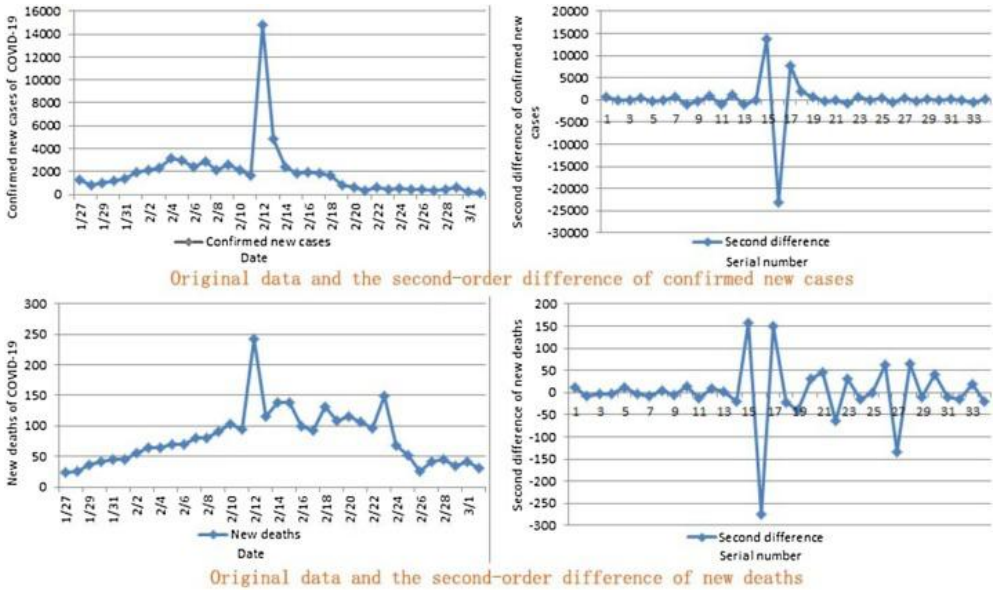


Figure 1. Original data and the second-order difference results.

COVID-19 pandemic poses the severe threat to the global health, many countries confirmed cases were dramatically increased. COVID-19 virus is high contagious, long incubation period. The fitting error of ARIMA model is 52.3107 in India and average absolute errors is 94.22 in Brazil [7]. Yang's group taken China as example to analyze COVID-19 condition in Italy. In Yang's study, Hubei China confirmed new cases had been used as original data for modelling in order to predict Italy next 10 days cases [9]. ARIMA model used Hubei new cases data (2020 01.27- 2020.03.10) to model and 2020.03.03- 2020.03.17 data to validate. ACF and PACF of model residuals showed the stable residual sequence and good fitting degrees. As a result, actual confirmed cases, and death number in Italy within the predicted 95% confidence interval. Confirmed cases number had stable and decrease for next ten days. Overall, ARIMA model fits well [9]. The limitation of the ARIMA model is that it doesn't support any fluctuations or intermediate changes during forecast period. Fig. 1 shows that ARIMA modelling requirements had been reached because there is dramatically increased in confirmed quantities and death second-order difference data sequence trends stable [9]. Seen from Fig. 2, ACF is complete autocorrelation function [7].

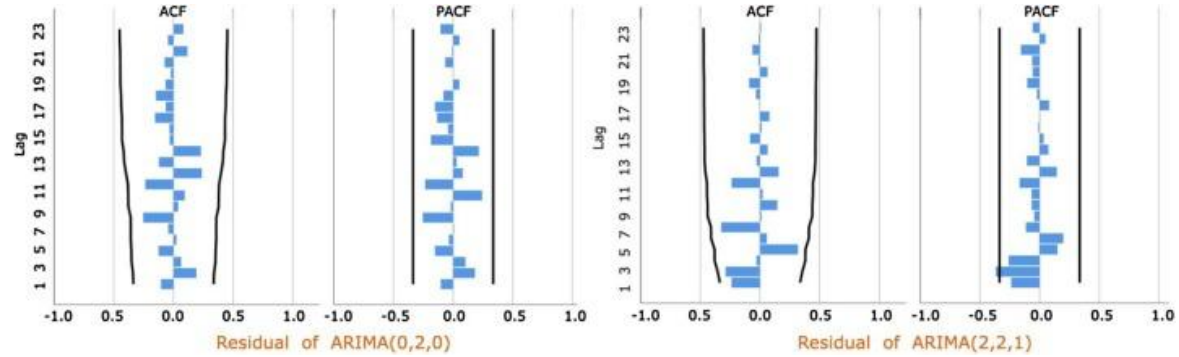


Figure 2. ACF and PACF results.

4. Machine learning

Machine learning also plays an important role in COVID-19 infection prediction. On account of the novelty and pandemic of COVID-19, the prediction task is challenging and vital [10, 11]. Ball use four basic machine learning and statistic models to forecast the short-term cumulative case, such as random

DOI: 10.54254/2753-8818/4/20220607

forest, support vector machines. Subsequently, we focus these two methods [10]. Support vector machines (SVM) is a traditional machine learning method for regression and classification. In regression task, SVM is called Support vector regression (SVR), which can be mathematically described as:

$$f(x) = \sum_{m=1}^{M} (\alpha_m^* - \alpha_m) x_m^T x + b \#(1)$$

where α_m^* and α_m are Lagrange multipliers.

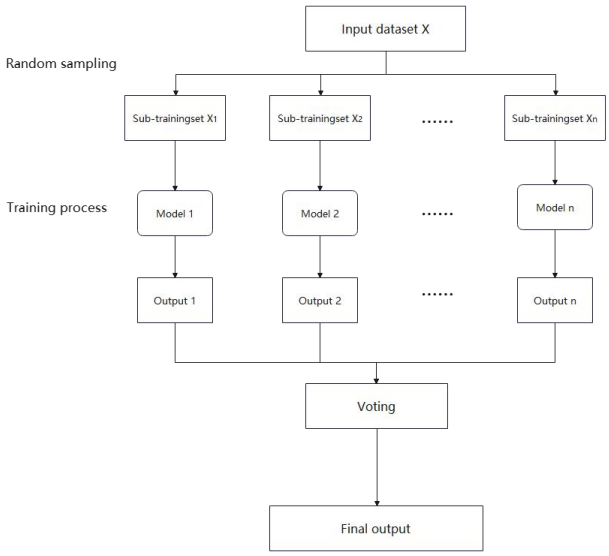


Figure 3. The structure of Random Forest.

Random Forest (RF) is a classical unsupervised machine learning method for regression and classification. The weak classifier of RF is decision tree [12]. Assuming that we have N classifier (RF), then we can get N output from N decision trees. After that, we use voting method which is a kind of ensemble learning to get the final output. The structure is shown in Fig. 3. In the study of Ball, the dataset they used is retrieved from WHO [13]. MAE, MAPE, RMSE are three performance evaluation metrics we use to evaluate the model. These three metrics values are shown in Table. 1 and the expression are shown following:

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} |y_i - \hat{y}_i| \#(2)$$

$$MAPE(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} \frac{|y_i - \hat{y}_i|}{|y_i|} \#(3)$$

$$MSE(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=0}^{n} (y_i - \hat{y}_i)^2} \#(4)$$

DOI: 10.54254/2753-8818/4/20220607

Table 1. The performance metrics of SVM and Random Forest.

Methods	Metric	Global	Germany	USA
SVM	MAE	269.274,0518	955,1736	42.496,719
	MAPE	2,0726	0,4477	1,5608
	RMSE	340.926,4251	1.387,0147	53.864,6539
Random Forest	MAE	19.771,7317	191,0731	5.852,0147
	MAPE	0,1247	0,0918	0,1406
	RMSE	25.825,8366	329,196	9.531,6776

The integration of Gradient boosting is built from a decision tree model. A tree is added to the ensemble one at a time and adjusted to correct prediction errors caused by the previous model. Olmedo et al. applied XGBoost to make COVID-19 predictions using anonymized data from a private hospital in Spain. In a preliminary evaluation of the performance of XGBoost, a comparison was made with several well-known classifiers. The results showed that the best results were achieved in terms of AUROC, AUPRC and accuracy. The model can be highly accurate in assessing mortality from laboratory values [14].

5. LSTM

As a matter of fact, LSTM is a special type of RNN, whereas they have some different.. As shown in Figure 4, there is a difference between LSTM (right) and RNN (left) [15]. Chimmula et al. used COVID-19 data to predict COVID-19 transmission in Canada based on the model mentioned above. In the studied LSTM model 1, the authors tested the network on the Canadian dataset; for the short-term predictions for Canada, the RMSE error was 34.83 with an accuracy of 93.4%. The RMSE error for the long-term prediction is about 45.70, with an accuracy of 92.67%. the prediction results of the LSTM model are shown as the red solid line in the Figure 5. Based on the results, the number of cases in Canada increases linearly until March 16, 2020. In term of the second model, the number of confirmed cases in Canada is expected to increase exponentially over a 10-day period. The LSTM network in this study is suitable for real-time data and does not make any assumptions in choosing hyperparameters. As illustrated in Figure 6, the recovery rate will start to decrease rapidly, while the mortality rate may increase [16].

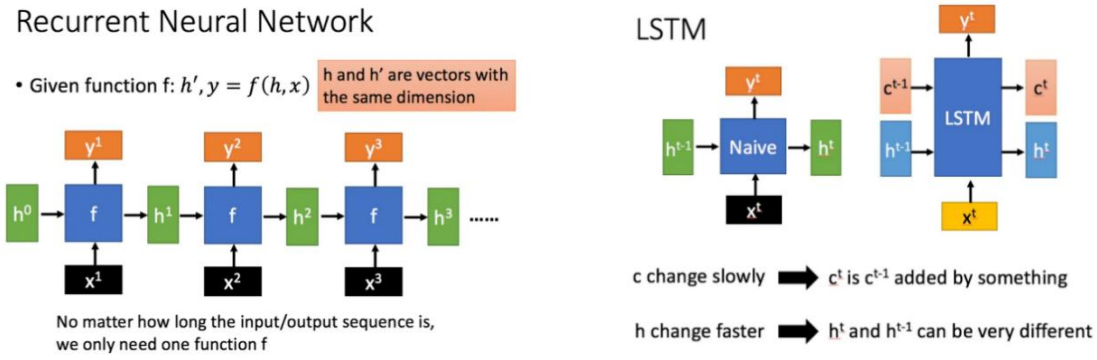


Figure 4. (a) The difference between RNN **(b.)** The difference between LSTM

DOI: 10.54254/2753-8818/4/20220607

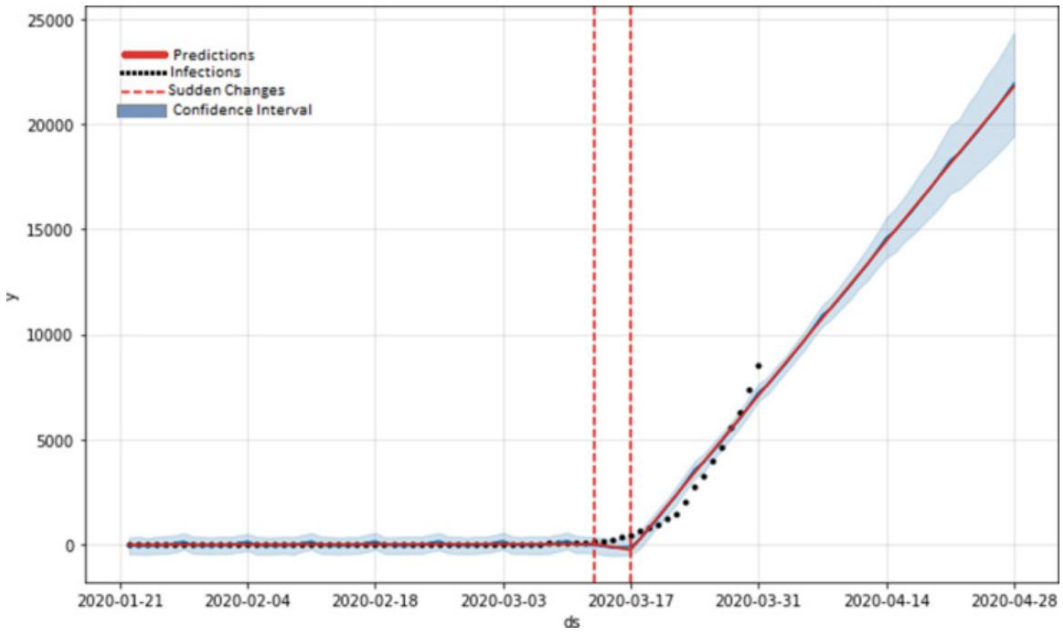


Figure 5. The predictions results of the LSTM.

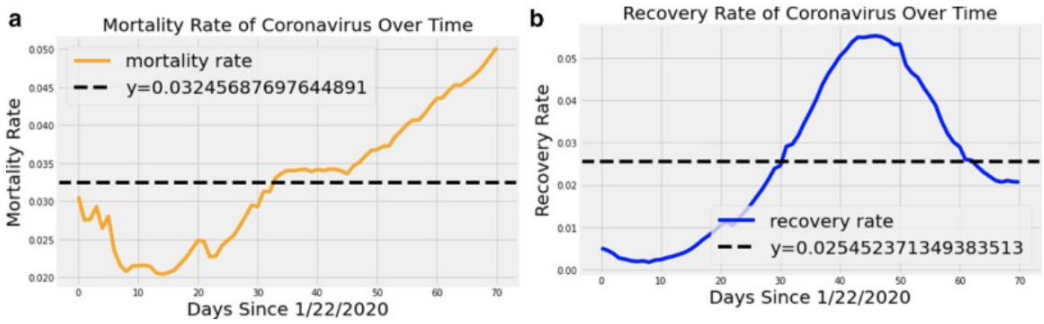


Figure 6. Mortality and Recovery rate of COVID-19.

6. Limitation & Prospects

It should be noted that the models discussed above have some defects and shortcomings. Primarily, for ARIMA, AR is autoregressive, which can derive the information of past time, and MA is moving average, which can get the current information and the residual of past time. Although ARIMA was used in time-series regression task for a long time, especially the pandemic of COVID-19, due to its unsuitability for complex and dynamic situations, it is not heavily relied upon [17]. In addition, for traditional machine learning method, like KNN and decision tree, they only focus the characteristic of the data, and ignore the properties of time sequence. But, as a kind of new method of time-series task, machine learning method can achieve the task effectively and efficiently. Moreover, although, after comparing the result of different methods, we only can find a optimal method under a specified condition including time period and area. Thus, if we want to get more strength generalization to our models, we should compare more model under more time period and countries. Last but not least, for deep learning, such as RNN or LSTM, it is easy to see that their model includes the information of time series and satisfy the property. Nevertheless, deep learning works as a black-box, the interpretability of the characteristic and result are poorer than traditional method.

For further studies, first, we can focus on the interpretability of models. This work can help us to find the more time-sires property which can be added into the model, and increase the interpretability

DOI: 10.54254/2753-8818/4/20220607

of models. Additionally, it is feasible to improve our model by more datasets including more time period and areas, to increase the generalization of model. Furthermore, one can combine the model with industry and medicine and find the value of model we proposed in the real world.

7. Conclusion

In summary, this paper discusses COVID-19 infection prediction based on statistical method, machine learning and deep learning. Specifically, ARIMA, machine learning, and LSTM models are demonstrated. According to the analysis, ARIMA can predict future trend and better explain the data set. As for machine learning (e.g., SVM and random forest), although they have good performance in many datasets, they still can focus on time properties to improve their models. Regarding to deep learning, such as RNN and LSTM, as a new method in machine learning, they can perform better in longer sequences. Nevertheless, for the existing work, they only focus the performance of model, but ignore the interpretability of model. In the future, one can improve the model to increase the performance of our model as well as find the interpretability of the proposed model. Overall, these results offer a guideline for time-series and machine leaning scenarios for COVID-19 infection prediction.

References

[1] M. Cascella, et al. Statpearls **1**, 10 (2022).
[2] A. Maher, M. Majdalawieh, N. Nizamuddin. Infectious Disease Modelling **6**, 98-111, (2021).
[3] W. B. Yu, et al. Zoological research **41**(3), 247 (2020).
[4] T. K. Lam, et al. Nature **583**(7815), 282-285 (2020).
[5] Y. Li, et al. Molecular oral microbiology **35**(4): 141-145 (2020).
[6] H. Xu, et al. Science of the Total Environment **731**, 139211 (2020).
[7] Y. Wang, et al. BMC Infectious Diseases **22**(1), 11 (2022).
[8] R. C. Sato. Einstein (São Paulo) **11**(1) 128-131 (2013).
[9] Q. Yang, et al. Journal of Infection and Public Health **13**(10) 1415-1418 (2020).
[10] S. Ballı. Chaos, Solitons & Fractals **142** 110512 (2021).
[11] G. L. Watson, et al. PLoS computational biology **17**(3) e1008837 (2021).
[12] N. K. Ahmed, et al. Econometric reviews **29**(5-6), 594-621(2010).
[13] WHO. World health organization covid cumulative dataset. September 18, 2020. Retrieved from https://covid19.who.int.
[14] D. Olmedo, et al. Journal of medical Internet research **23**(4), e26211. (2021).
[15] Tiago M. How the LSTM Improves the RNN. 1 Jan. 2021, Retrieved from: https://aditi-mittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e.
[16] V. Chimmula, K. Reddy, Z. Lei. Chaos, Solitons & Fractals **135**, 109864 (2020).
[17] H. Alabdulrazzaq, et al. Results in Physics **27**, 104509 (2021).